Федеральное государственное бюджетное учреждение науки
Вычислительный центр им. А.А. Дородницына
Российской академии наук

На правах рукописи

Татарчук Александр Игоревич

БАЙЕСОВСКИЕ МЕТОДЫ ОПОРНЫХ ВЕКТОРОВ ДЛЯ ОБУЧЕНИЯ РАСПОЗНАВАНИЮ ОБРАЗОВ С УПРАВЛЯЕМОЙ СЕЛЕКТИВНОСТЬЮ ОТБОРА ПРИЗНАКОВ

05.13.17 – Теоретические основы информатики

диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель
д.т.н., профессор Моттль Вадим Вячеславович

Москва, 2014

Содержание

ВВЕДЕНИЕ

Работа посвящена развитию методологии селективного комбинирования признакового описания объектов при обучении двухклассовому распознаванию образов на основе байесовского обобщения метода опорных векторов.

Актуальность. В диссертации исследуется и развивается наиболее популярный в современной литературе линейный подход к обучению распознаванию двух классов объектов реального мира $\omega \in \Omega$, основанный на двух предположениях. Во-первых, предполагается, что всякий объект воспринимается компьютером через вектор его числовых признаков как точка в линейном пространстве $\mathbf{x}(\omega) \in \mathbb{R}^n$, размерность которого определяется числом признаков n. Во-вторых, предполагается, что для суждения о принадлежности объекта к одному из двух классов $y = \pm 1$ достаточно вычислить значение линейной решающей функции (decision function) $d(\mathbf{x} \mid \mathbf{a}, b) = \mathbf{a}^T \mathbf{x} + b : \mathbb{R}^n \to \mathbb{R}$, знак которой непосредственно укажет класс объекта $\mathbf{a}^T \mathbf{x} + b \gtrless 0$. Очевидно, что линейная функция $d(\mathbf{x} \mid \mathbf{a}, b)$ определяет дискриминантную гиперплоскость в $\mathbb{R}^n$, а решение о классе объекта определяется тем, по какую сторону от нее окажется вектор признаков объекта $\mathbf{x}$. Обучение линейного классификатора сводится к формированию значений параметров $(\mathbf{a}, b)$ на основе анализа конечной обучающей совокупности $\{(\mathbf{x}_j, y_j),\ j = 1, \ldots, N\}$.

Наиболее популярным методом обучения линейного классификатора по обучающей совокупности является *метод опорных векторов* (*Support Vector Machine – SVM*), разработанный Вапником В.Н. и Червоненкисом А.Я. [1], в основе которого лежит ими же ранее предложенный метод обобщенного портрета [2]. В основе метода лежит естественная идея выбирать ту гиперплоскость, которая в обучающей совокупности разделяет векторы признаков объектов разных классов с наибольшим зазором, дополнительно штрафуя возможные нарушения некоторыми объектами этого общего

«идеального» требования. В данной диссертации исследуется и развивается именно метод опорных векторов.

Одно из основных преимуществ метода опорных векторов заключается в том, что как задача обучения, так и решающее правило классификации новых объектов определяются не самими векторами признаков отдельных объектов $\mathbf{x}(\omega) \in \mathbb{R}^n$, а только их попарными скалярными произведениями $\left(\mathbf{x}(\omega')\right)^T \mathbf{x}(\omega'') : \Omega \times \Omega \to \mathbb{R}$. Из этого факта следует, что нет противоречия между традиционным признаковым способом погружения объектов распознавания в линейное пространство и беспризнаковым способом, предполагающим, что объекты $\omega \in \Omega$ могут быть представлены в компьютере только через некоторую числовую функцию парного отношения $K(\omega', \omega'')$. Правда, для того, чтобы идея поиска дискриминантной гиперплоскости в некотором линейном пространстве представления объектов реального мира имела простую математическую реализацию, функция $K(\omega', \omega'') : \Omega \times \Omega \to \mathbb{R}$ должна обладать свойствами кернела (kernel function), т.е. быть симметричной и порождать неотрицательно определенные матрицы для всех конечных комбинаций объектов. Известно, что всякий кернел погружает множество объектов реального мира $\omega \in \Omega$ в некоторое большее гильбертово пространство $\tilde{\Omega} \supseteq \Omega$, в котором он играет роль скалярного произведения [1,3].

Однако требование неотрицательной определенности для функции парного сравнения объектов является чрезвычайно ограничивающим. Альтернативный подход предложен Р. Дьюиным [4] под названием *реляционного дискриминантного анализа* (*Relational Discriminant Analysis*) [5]. Идея заключается в том, чтобы интерпретировать значения функции парного сравнения между произвольным объектом $\omega \in \Omega$ и объектами из обучающей совокупности $\{\omega_1, ..., \omega_N\}$ как вектор его вторичных признаков $\mathbf{x}(\omega) = \left(x_i(\omega) = \mathrm{S}(\omega_i, \omega), i = 1, ..., N\right)$, и применить затем обычный метод опорных векторов в $\mathbb{R}^n$. Однако за такое использование функций парного сравнения объектов приходится платить большой размерностью пространства вторичных

признаков, определяемой числом объектов в обучающей совокупности. Эта размерность многократно увеличится, если использовать несколько альтернативных функций парного сравнения [6].

Первая проблемная ситуация, определившая выбор темы данного диссертационного исследования, заключается в том, что при всей эффективности метод опорных векторов остается эвристическим по своей конструкции. С момента его создания в мировой литературе был предпринят ряд попыток снабдить его некоторой вероятностной интерпретацией [7,8]. Однако эти интерпретации оставались неполными, не позволяющими в полной мере использовать вероятностный аппарат для наделения чрезвычайно популярного метода опорных векторов принципиально новыми свойствами.

Для разрешения этой проблемной ситуации в настоящей диссертации предлагается специальная байесовская постановка обучения распознаванию двух классов объектов в линейном признаковом пространстве, приводящая к обобщению метода опорных векторов и являющаяся теоретической основой для создания новых селективных методов обучения. Основная идея байесовской постановки заключается в построении системы вероятностных предположений о паре плотностей распределения объектов двух классов $\varphi(\mathbf{x}|y=\pm 1, \mathbf{a}, b)$, определяемой объективно существующей, но неизвестной гиперплоскостью $(\mathbf{a}, b)$ в линейном пространстве признаков, при некотором априорном предположении о ее случайном выборе $\Psi(\mathbf{a}, b)$. При этом именно структура первого семейства распределений определяет характерный принцип обучения по методу опорных векторов.

Что же касается априорного распределения параметров дискриминантной гиперплоскости, а именно ее направляющего вектора, то одна из основных идей диссертации заключается в том, что характер этого распределения определяет априорную склонность к увеличению одних компонент направляющего вектора и уменьшению других, предопределяя тем самым скрытое разделение признаков объектов на информативные и

неинформативные. Выбор этого распределения играет роль регуляризации при формировании формальной байесовской постановки задачи обучения.

Вторая проблемная ситуация заключается в том, что избыточность признакового описания объектов неизбежно связана как с опасностью переобучения, выражающейся в снижении обобщающей способности результата обучения по ограниченной обучающей совокупности, так и с избыточной громоздкостью получаемых при этом решающих правил, что для некоторых практических задач являться весьма обременительным.

Стремление избежать переобучения заставляет принимать специальные меры по ограничению свободы выбора решающего правила, называемые регуляризацией процесса обучения. Наиболее популярный принцип регуляризации обучения заключается в построении селективных систем распознавания, основанных на сокращении множества признаков, характеризующих объекты распознавания. В данной диссертации мы ограничимся рассмотрением именно методов отбора признаков объектов распознавания, независимо от того, каким способом эти признаки были получены.

Для разрешения второй проблемной ситуации в диссертации предлагается использовать выбор априорного распределения направляющего вектора для наделения метода опорных векторов в его вероятностной интерпретации способностью автоматически отбирать подмножество информативных признаков с заданным уровнем селективности в результате обучения по предъявленной обучающей совокупности. Разнообразие существующих методов отбора признаков интерпретируется в диссертации как результат использования разных семейств априорных распределений направляющего вектора.

В диссертации впервые вводится понятие *параметра селективности*, являющегося параметром семейства априорных распределений направляющего вектора и выступающего в роли структурного параметра метода обучения. При нулевом значении параметра селективности оценка

направляющего вектора содержит все исходные признаки объектов, а при его бесконечном увеличении все бóльшая часть из них подавляется.

На основе анализа мировой литературы в диссертации приняты четыре критерия для оценивания «качества» селективных свойств конкретного метода обучения:

а) эффективное подавление полностью шумовых попарно независимых признаков;

б) эффективное подавление полностью шумовых признаков, имеющих значительную линейную зависимость между собой;

в) эффективное выделение группы информативных линейно независимых признаков;

г) эффективное выделение группы информативных признаков, только совместное участие которых в решении может обеспечить приемлемую точность распознавания.

По этим критериям в диссертации исследованы две наиболее популярных модификации метода опорных векторов, наделяющие его свойством отбора признаков – *Lasso SVM* (*1-norm SVM*) и *Elastic Net SVM* (*Doubly Regularized SVM*). Оба эти метода несколько разными средствами отбирают подмножество информативных признаков, число которых определяется структурными параметрами.

Третья проблемная ситуация определяется тем, что, как оказалось, эти методы далеко не удовлетворяют сочетаниям пар требований (а-б) и (в-г).

Для разрешения этой проблемной ситуации в диссертации разработаны два новых класса априорных распределений направляющего вектора дискриминантной гиперплоскости и, следовательно, две новые модификации метода опорных векторов.

Один из новых методов, названный *методом релевантных признаков* (*Relevance Feature Machine – RFM*), не выделяя строгого подмножества информативных признаков, наделяет все признаки неотрицательными весами. Чем больше значение структурного параметра селективности, тем

большее число весов приближаются к нулевым значениям, фактически исключая соответствующие признаки из принятия решения о классе объекта.

Другой предложенный метод, названный *методом опорных признаков* (*Support Feature Machine – SFM*), разбивает все множество признаков на три группы – полностью активные признаки, взвешенные признаки и полностью подавленные признаки. Можно считать, что метод *SFM* снабжает все признаки весами, но, в отличие от метода *RFM,* часть весов оказываются строгими единицами, часть принимает значения между единицей и нулем, а часть строго равны нулю.

Целью диссертации является разработка методологии селективного комбинирования признаков объектов в задаче двухклассового распознавания образов на основе байесовского обобщения метода опорных векторов.

Методы исследования. Теоретическое исследование базируется на общих принципах линейной алгебры, методе опорных векторов, методах выпуклой оптимизации и основах байесовской теории принятия решений. Экспериментальное исследование проводилось с использованием программно-алгоритмического комплекса, разработанного автором.

Научная новизна. В данной работе впервые сформулирован вероятностный подход к селективному комбинированию признаков в ходе обучения распознаванию образов в рамках метода опорных векторов. Предложены два семейства параметрических вероятностных моделей обучающей совокупности и вытекающий из него класс линейных решающих правил и критериев обучения. Разработаны соответствующие алгоритмы апостериорного оценивания разделяющей гиперплоскости, реализующие байесовский принцип обучения с заданной селективностью отбора признаков.

Положения, выносимые на защиту:

1. Общая математическая постановка задачи обучения двухклассовому распознаванию образов в линейном пространстве признаков на основе количественного измерения расстояния между вектором признаков объекта и разделяющей гиперплоскостью.

2. Вероятностная модель наблюдения объектов в пространстве векторов признаков относительно фиксированной разделяющей гиперплоскости.

3. Два семейства априорных вероятностных моделей направляющего вектора разделяющей гиперплоскости, отражающих стратегии отбора признаков на основе взвешивания всех исходно заданных признаков (feature weighting) и на основе жесткого выбора их подмножества (feature subset selection).

4. Комплекс байесовских методов и алгоритмов оценивания разделяющей гиперплоскости, реализующих принцип обучения с заданной селективностью отбора признаков объектов.

Практическая значимость. Разработанные алгоритмы позволяют строить решающие правила распознавания образов при заведомо избыточном множестве признаков представления объектов и относительно малом объеме обучающей совокупности без опасности снижения обобщающей способности результата обучения.

Связь с плановыми научными исследованиями. Работа выполнена в рамках гранта ИНТАС № 04-77-7347 «Принципы беспризнакового распознавания сигналов, символьных последовательностей и изображений» (2005-2006), грантов Российского фонда фундаментальных исследований № 05-01-00679-а «Линейные методы восстановления зависимостей в массивах данных произвольной природы», № 06-01-08042-офи «Теоретические основы, методы, инструментальные средства и новая открытая информационная технология построения систем идентификации личности по свободно пополняемому множеству биометрических характеристик», № 08-01-00695-а «Линейные методы комбинирования разнородной информации для решения задач анализа массивов данных произвольной природы», № 11-07-00409 «Методы выбора уровня сложности модели в задачах восстановления зависимостей между разнородными объектами реального мира».

Реализация и внедрение результатов работы. Результаты исследования применены для решения прикладных задач обучения распознаванию классов пространственной структуры белков по последовательности составляющих их аминокислот, распознаванию рукописных символов и подписей, вводимых в компьютер непосредственно в процессе написания, а так же в задаче распознавания рака легких.

Теоретические результаты диссертационной работы вошли в состав курса «Статистические методы анализа сигналов», читаемого профессором В.В. Моттлем студентам 5-го курса на кафедре «Интеллектуальные системы» ФУПМ МФТИ, и курсов «Теория обучения машин» и «Машинное обучение», читаемых профессором К.В. Воронцовым на ФУПМ МФТИ и в Школе анализа данных Яндекс, соответственно.

Апробация работы. Основные положения и результаты диссертации докладывались автором на конференциях:

- XIII всероссийская конференция «Математические методы распознавания образов», Зеленогорск, 2007 г.;
- The 7th International Workshop on Multiple Classifier Systems, Prague, Czech Republic, 2007 г.;
- The 6th International Conference on Machine Learning and Cybernetics, Китай, Гонконг, 2007 г.;
- Международная конференция «International Conference of Pattern Recognition», США, Тампа, 2008 г.;
- Международная конференция «Интеллектуализации обработки информации», Украина, Симферополь, 2008 г.;
- 14-ая Всероссийская конференция «Математические методы распознавания образов», Суздаль, 2009 г.;
- The 9th International Workshop, MCS 2010, Cairo, Egypt, 2010 г.;
- Международная конференция «International Conference of Pattern Recognition», Япония, Токио, 2012 г.

Кроме того, основные результаты работы докладывались на семинаре отдела Интеллектуальных систем (Москва, ВЦ РАН, 2009 г., 2014 г.).

Публикации. По тематике исследований опубликовано 18 статей, в том числе 8 статей в изданиях, входящих в список ВАК (выделены шрифтом):

1. **Середин О.С., Моттль В.В., Татарчук А.И., Разин Н.А. Выпуклые селективные критерии метода релевантных векторов в пространстве парных отношений объектов распознавания. Известия ТулГУ, Серия Естественные науки. Тула: Изд-во ТулГУ, 2013, Вып. 1, с. 165-176.**
2. **O. Seredin, V. Mottl, A. Tatarchuk, N. Razin, D. Windridge. Convex Support and Relevance Vector Machines for selective multimodal pattern recognition. Proceedings of the 21th International Conference on Pattern Recognition, Tsukuba, Japan, November 11-15, 2012. IAPR, 2012, ISSN 978-4-9906441-1-6, 2012, pp. 1647-1650.**
3. **A. Tatarchuk, E. Urlov, V. Mottl, D. Windridge. A support kernel machine for supervised selective combining of diverse pattern-recognition modalities. In: Multiple Classifier Systems. Lecture Notes In Computer Science, Vol. 5997. Springer-Verlag, Berlin \ Heidelberg, 2010, ISBN 978-3-642-12127-2_17, pp. 165-174.**
4. Татарчук А.И., Урлов Е.Н., Ляшко А.С., Моттль В.В. Экспериментальное исследование обобщающей способности методов селективного комбинирования потенциальных функций в задаче двухклассового распознавания образов. Доклады 14-й Всероссийской конференции «Математические методы распознавания образов», Суздаль, 21-26 сентября 2009 г., с. 196-199.
5. Татарчук А.И., Урлов Е.Н., Моттль В.В. Метод опорных потенциальных функций в задаче селективного комбинирования разнородной информации при обучении распознаванию образов. Доклады 14-й Всероссийской конференции «Математические методы распознавания образов», Суздаль, 21-26 сентября 2009 г., с. 192-195.
6. Татарчук А.И., Сулимова В.В., Моттль В.В., Уиндридж Д. Метод релевантных потенциальных функций для селективного комбинирования разно-

родной информации при обучении распознаванию образов на основе байесовского подхода. Доклады 14-й Всероссийской конференции «Математические методы распознавания образов», Суздаль, 21-26 сентября 2009 г., с. 188-191.

7. A. Tatarchuk, V. Sulimova, D. Windridge, V. Mottl, M. Lange. Supervised selective combining pattern recognition modalities and its application to signature verification by fusing on-line and off-line kernels. In: Multiple Classifier Systems. Lecture Notes In Computer Science, Vol. 5519. Springer-Verlag, Berlin \ Heidelberg. ISBN 978-3-642-02325-5, 2009, pp. 324-334.

8. A. Tatarchuk, V. Mottl, A. Eliseyev, D. Windridge. A Bayesian approach to multimodal pattern recognition with supervised selectivity of modalities. Proceedings of 9th International Conference on Pattern Recognition and Image Analysis: New Information Technologies, Nizhni Novgorod, September 14-20, 2008, vol. 2, p. 204.

9. В.В. Моттль, А.И. Татарчук, А.П. Елисеев Экспериментальное исследование методов многомодального распознавания образов с регулируемой селективностью. Известия Тульского государственного университета. Технические науки, вып.3. – Тула: Изд-во ТулГУ, 2008.

10. Tatarchuk A., Mottl V., Eliseyev A., Windridge D. Selectivity supervision in combining pattern-recognition modalities by feature- and kernel-selective Support Vector Machines. Proceedings of the 19th International Conference on Pattern Recognition, Vol 1-6, IEEE, ISBN 978-1-4244-2174-9, 2008, pp. 2336-2339.

11. Моттль В.В., Татарчук А. И., Елисеев А. П. Регулируемая селективность в многомодальном распознавании образов. Таврический вестник информатики и математики, 2008, том. 2, с. 89.

12. Моттль В.В., Татарчук А. И., Елисеев А. П. Многомодальное распознавание образов с регулируемой селективностью. Тезисы докладов международной научной конференции «Интеллектуализации обработки информации», Симферополь, 2008, с. 172.

13. Моттль В.В., Татарчук А. И., Елисеев А. П. Исследование методов комбинирования классификаторов и потенциальных функций в многомо-

дальном распознавании образов. Тезисы докладов международной научной конференции «Интеллектуализации обработки информации», Симферополь, 2008, с. 176.

14. Mottl V., Tatarchuk A., Krasotkina O., Seredin O., Sulimova V. Combining pattern recognition modalities at the sensor level via kernel fusion. In: Multiple Classifier Systems. Lecture Notes In Computer Science, Vol. 4472. Springer-Verlag, Berlin \ Heidelberg. ISBN 978-3-540-72481-0, 2007, pp. 1-12.

15. Моттль В.В., Татарчук А. И., Красоткина О.В., Сулимова В.В. Комбинирование потенциальных функций в многомодальном распознавании образов. Математические методы распознавания образов. 13-я Всероссийская конференция: Сборник докладов. М.: МАКС Пресс, 2007.

16. Valentina Sulimova, Vadim Mottl, Alexander Tatarchuk Multi-Kernel approach to on-line signature verification. // Proceedings of the Eighth IASTED International Conference on Signal and Image Processing, ACTA PRESS ANAHEIM, ISBN 978-0-88986-583-9, 2006, pp. 448-453.

17. В.В. Сулимова, В.В. Моттль, А.И. Татарчук. Автоматический выбор наиболее информативных фрагментов в задачах распознавания сигналов разной длительности. // Труды конференции «Интеллектуализация обработки информации»-2006, Алушта, стр. 152-154.

18. В.В. Сулимова, В.В. Моттль, А.И. Татарчук. Автоматический выбор наиболее информативных фрагментов в задачах распознавания сигналов разной длительности. // Таврический вестник математики и информатики – № 1, 2006, стр. 109-115.

Структура работы. Диссертация состоит из введения, четырех глав, заключения, приложения и списка литературы. Материал изложен на 125 страницах, содержит 2 леммы, 15 теорем, 15 рисунков, 8 таблиц и список литературы из 21 наименования.

1 ПРОБЛЕМА СЕЛЕКТИВНОГО КОМБИНИРОВАНИЯ ПРИЗНАКОВ В ЛИНЕЙНЫХ МЕТОДАХ ОБУЧЕНИЯ РАСПОЗНАВАНИЮ ОБРАЗОВ

1.1 Линейный подход к обучению распознаванию двух классов объектов

1.1.1 Разделяющая гиперплоскость в линейном пространстве признаков

В диссертации исследуется и развивается линейный подход к обучению распознаванию двух классов объектов $\omega \in \Omega$, $y(\omega) = \pm 1$, которые воспринимаются компьютером через векторы их числовых признаков $\mathbf{x}(\omega) = \left(x_1(\omega),...,x_n(\omega)\right) \in \mathbb{R}^n$. Основополагающим принципом машинного обучения является *гипотеза компактности*, которая, в самой общей формулировке для задач обучения распознаванию образов, означает, что объекты одного и того же класса имеют большее сходство между собой чем объекты, принадлежащие разным классам. Такая гипотеза, в том или ином виде, неизбежно инкорпорирована в любой метод обучения, а их многообразие определяется разными смыслами, вкладываемыми исследователями в понятие *сходство объектов*. Самым распространённым математическим инструментом измерения сходства между объектами реального мира Ω, погруженными числовыми признаками $\mathbf{x}(\omega) \in \mathbb{R}^n$ в конечномерное евклидовое пространство, является евклидова метрика

$$\rho(\omega',\omega'') = \rho\left(\mathbf{x}(\omega'),\mathbf{x}(\omega'')\right) = \rho(\mathbf{x}',\mathbf{x}'') = \| \mathbf{x}' - \mathbf{x}'' \| = \left((\mathbf{x}' - \mathbf{x}'')^T(\mathbf{x}' - \mathbf{x}'')\right)^{1/2}. \quad (1)$$

Суть линейного подхода заключается в поиске такой гиперплоскости, которая обеспечивала бы в некотором смысле оптимальное разделение объектов первого $y(\omega) = +1$ и второго $y(\omega) = -1$ классов. При этом всякий выбор направляющего вектора $\mathbf{a} \in \mathbb{R}^n$ и действительного числа $b \in \mathbb{R}$ определяет некоторую *разделяющую гиперплоскость*

$$\mathcal{H}(\mathbf{a},b)=\left\{\mathbf{z}\in\mathbb{R}^n\colon \mathbf{a}^T\mathbf{z}+b=0\right\}. \qquad (2)$$

Гиперплоскость (2) делит все признаковое пространство $\mathbb{R}^n$ на три части. Во-первых, это само множество $\mathcal{H}(\mathbf{a},b)$, и во-вторых, два множества точек, находящихся по разные от гиперплоскости стороны. Принадлежность объекта $\mathbf{x}(\omega)\in\mathbb{R}^n$ к одному из трех множеств определяет оценку его класса $\hat{y}(\omega)$, при этом множество $\mathcal{H}(\mathbf{a},b)$, как правило, для простоты объединяется с множеством точек одного из классов.

Помимо гиперплоскости (2) одним из ключевых понятий линейного подхода является *решающая функция* (*decision score function*)

$$d(\mathbf{x}\,|\,\mathbf{a},b)=\frac{\mathbf{a}^T\mathbf{x}+b}{(\mathbf{a}^T\mathbf{a})^{1/2}} \text{ или } d(\mathbf{x}\,|\,\mathbf{a},b)=\mathbf{a}^T\mathbf{x}+b \text{ при } \|\mathbf{a}\|=1. \qquad (3)$$

Важность этого понятия определяется, с одной стороны тем, что решающая функция (3) представляет собой функцию расстояния от объекта $\mathbf{x}(\omega)\in\mathbb{R}^n$ до его проекции $\mathbf{x}_{\mathcal{H}}(\omega)\in\mathbb{R}^n$ на гиперплоскость (2) с учетом знака в смысле метрики (1) (рис. 1).

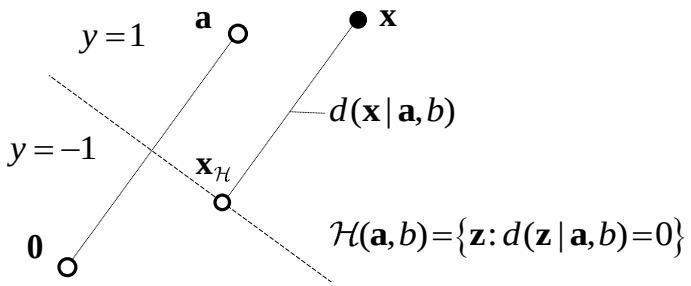


Рисунок 1. Расстояние от объекта до разделяющей гиперплоскости.

С другой стороны, итоговое суждение о классе $\hat{y}(\omega)$ объекта определяется знаком решающей функции (3)

$$\hat{y}(\omega)=\begin{cases} 1, \text{ если } d(\mathbf{x}\,|\,\mathbf{a},b)\geq 0,\\ -1, \text{ если } d(\mathbf{x}\,|\,\mathbf{a},b)<0. \end{cases} \qquad (4)$$

1.1.2 Концепция оптимальной разделяющей гиперплоскости для обучающей совокупности: Метод опорных векторов

Пусть задана обучающая совокупность объектов с указанием индексов классов, к которым они принадлежат:

$$\{(\mathbf{x}_j, y_j),\ j=1,...,N\},\ \mathbf{x}_j \in \mathbb{R}^n,\ y_j = \pm 1. \qquad (5)$$

Степень несоответствия значения решающей функции $d(\mathbf{x}\,|\,\mathbf{a},b)$ для объекта $\mathbf{x}(\omega) \in \mathbb{R}^n$ значению его целевой характеристики $y(\omega)$ выражается в виде *функции потерь*. В частности, нас далее будет интересовать функция потерь специального вида:

$$\delta(y,d\,|\,\varepsilon)=\begin{cases}0, & yd \geq \varepsilon,\\ 1-\dfrac{1}{\varepsilon}yd, & yd < \varepsilon,\end{cases} \qquad (6)$$

где $y = y(\omega)$ - истинный класс объекта, $d = d(\mathbf{x}\,|\,\mathbf{a},b)$ - значение решающей функции (3), а $\varepsilon \geq 0$ - параметр, который далее будем интерпретировать как зазор разделения двух классов объектов.

Функция потерь такого вида была впервые предложена Вапником В.Н. и Червоненкисом А.Я. и привела к методу обобщенного портрета [2], а затем к известному методу опорных векторов [1].

Такая функция потерь (рис. 2) имеет нулевой штраф $\delta(y,d\,|\,\varepsilon)=0$ для объектов, удаленных от гиперплоскости на достаточное расстояние $yd \geq \varepsilon$, и линейно растущий штраф $\delta(y,d\,|\,\varepsilon)=1-yd/\varepsilon$ для объектов слишком близко приблизившихся к границе классов или попавших в область другого класса $yd < \varepsilon$.

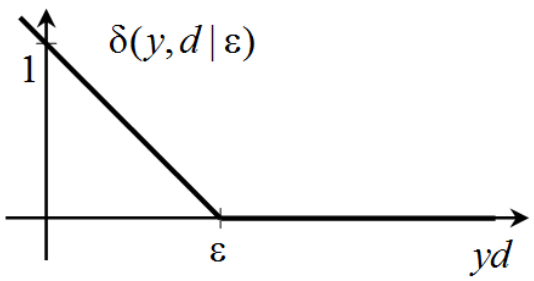


Рисунок 2. Вид функции потерь $\delta(y,d\,|\,\varepsilon)$.

В случае линейной разделимости обучающей выборки (5) естественно отдавать предпочтения такой решающей функции (3), для которой функция потерь (6) принимает нулевые значения для всех объектов обучающей выборки для как можно большего значения параметра $\varepsilon \geq 0$. Другими словами, оптимальной гиперплоскостью является такая гиперплоскость, которая без ошибок разделяет обучающую выборку с наибольшим зазором $\varepsilon \geq 0$ (рис. 3).

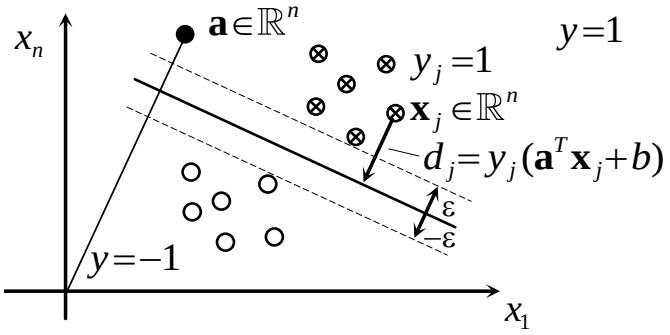


Рисунок 3. Разделение гиперплоскостью обучающей выборки с наибольшим зазором в пространстве двух вещественных признаков.

Такое требование эквивалентно невыпуклой оптимизационной задаче:

$$\begin{cases} 1/\varepsilon^2 \to \min(\mathbf{a}, b, \varepsilon), \ \mathbf{a}^T\mathbf{a}=1, \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \geq \varepsilon, \ \ j = 1,...,N, \ \varepsilon \geq 0. \end{cases} \tag{7}$$

В случае линейно неразделимой обучающей выборки (5) ненулевые значения функции потерь (6) соответствуют объектам, нарушившим границу зазора, определяемой положением гиперплоскости $(\mathbf{a}, b)$ и параметром функции потерь $\varepsilon \geq 0$ (рис. 4).

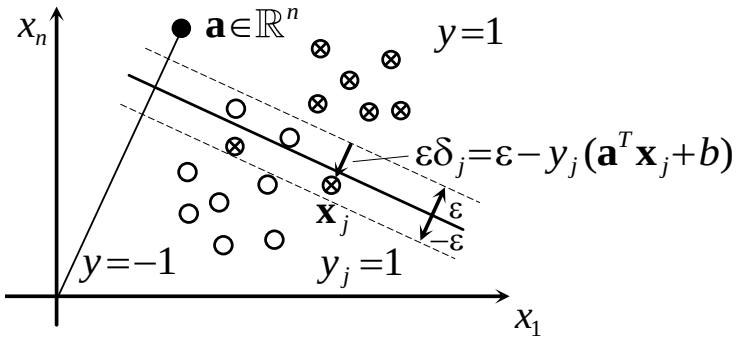


Рисунок 4. Ненулевое значение функции потерь для объекта обучающей совокупности, нарушившего границу зазора.

В качестве критерия обучения естественно искать такую гиперплоскость, которая бы разделяла обучающую выборку на два класса, с одной стороны, с как можно большей величиной зазора $\varepsilon \geq 0$, $1/\varepsilon^2 \to \min$, а с другой, с как можно меньшей величиной суммарного штрафа для ошибочно классифицированных объектов обучающей выборки (5) $\delta_j = 1-(1/\varepsilon) y_j(\mathbf{a}^T\mathbf{x}_j + b)$, $\sum_{j:(1/\varepsilon) y_j(\mathbf{a}^T\mathbf{x}_j+b)<1} \delta_j$. Нетрудно убедиться, что баланс таких весьма эвристических требований к процессу обучения выражается оптимизационным критерием

$$\begin{cases} \dfrac{1}{\varepsilon^2} + C\sum\limits_{j=1}^{N} \delta_j \to \min(\mathbf{a}, b, \boldsymbol{\delta}, \varepsilon), \ \mathbf{a}^T\mathbf{a} = 1, \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \geq \varepsilon(1-\delta_j), \delta_j \geq 0, \ j = 1,\ldots,N, \varepsilon \geq 0, \end{cases} \tag{8}$$

где C - структурный параметр, определяющий соотношение требований максимизации зазора и минимизации величины эмпирического риска. Это и есть исходный критерий обучения, получивший название метода опорных векторов.

Правда, критерий (8) не очень удобен, поскольку имеет вид задачи оптимизации на сфере $\mathbf{a}^T\mathbf{a} = 1$, представляющая собой невыпуклое множество. Классический критерий получается из (8) простой заменой переменных $\mathbf{a} = \varepsilon\tilde{\mathbf{a}}$, $b = \varepsilon\tilde{b}$, $\tilde{\mathbf{a}}^T\tilde{\mathbf{a}} = 1/\varepsilon^2$:

$$\begin{cases} J_{SVM}(\mathbf{a}, b, \boldsymbol{\delta} \,|\, C) = \mathbf{a}^T\mathbf{a} + C\sum\limits_{j=1}^{N} \delta_j \to \min(\mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R}, \boldsymbol{\delta} \in \mathbb{R}^N), \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \geq 1 - \delta_j, \delta_j \geq 0, \ j = 1,\ldots,N. \end{cases} \tag{9}$$

Этот простой и в то же время очень эффективный способ обучения распознаванию двух классов объектов путем решения задачи квадратичного программирования (минимизации квадратичной функции при линейных ограничениях) был предложен Вапником В.Н. [1] и получил за прошедшие годы огромную популярность как метод опорных векторов (*Support Vector Machine*). Его название объясняется тем, что, как показывает элементарное математическое исследование, направляющий вектор оптимальной разделяющей гиперплоскости $(\hat{\mathbf{a}}, \hat{b}, \hat{\boldsymbol{\delta}}) = \arg\min J_{SVM}(\mathbf{a}, b, \boldsymbol{\delta} \,|\, C)$ согласно (9)

есть линейная комбинация векторов признаков объектов обучающей совокупности

$$\mathbf{a} = \sum\nolimits_{j:\lambda_j>0} \lambda_j y_j \mathbf{x}_j \in \mathbb{R}^n, \tag{10}$$

где $\hat{\lambda}_j \geq 0$, $j = 1,..., N$ – решения двойственной задачи квадратичного программирования

$$\begin{cases} W_{SVM}(\lambda_1,...,\lambda_N) = \sum\limits_{j=1}^{N} \lambda_j - \frac{1}{2} \sum\limits_{j=1}^{N} \sum\limits_{l=1}^{N} (y_j y_l \mathbf{x}_j^T \mathbf{x}_l) \lambda_j \lambda_l \to \max(\lambda_1,...,\lambda_N), \\ \sum\limits_{j=1}^{N} y_j \lambda_j = 0,\ 0 \leq \lambda_j \leq C/2,\ j = 1,..., N, \end{cases} \tag{11}$$

имеющих смысл множителей Лагранжа при ограничениях-неравенствах $y_j(\mathbf{a}^T \mathbf{x}_j + b) \geq 1 - \delta_j$ в (9).

Согласно (10) направляющий вектор оптимальной разделяющей гиперплоскости и вся решающая функция

$$d(\mathbf{x}) = \sum\nolimits_{j:\hat{\lambda}_j>0} \hat{\lambda}_j y_j \mathbf{x}_j^T \mathbf{x} + b \gtrless 0, \tag{12}$$

применимая к вектору признаков нового объекта $\mathbf{x} \in \mathbb{R}^n$, полностью определяется лишь опорными объектами обучающей совокупности, для которых множители Лагранжа строго положительны $\hat{\lambda}_j > 0$. Именно это свойство метода опорных векторов определило его практическую популярность, поскольку после обучения для распознания распознавания класса нового объекта достаточно вычислить скалярное произведение $\mathbf{x}_j^T \mathbf{x}$ (12) лишь для опорных объектов обучающей совокупности (рис. 5).

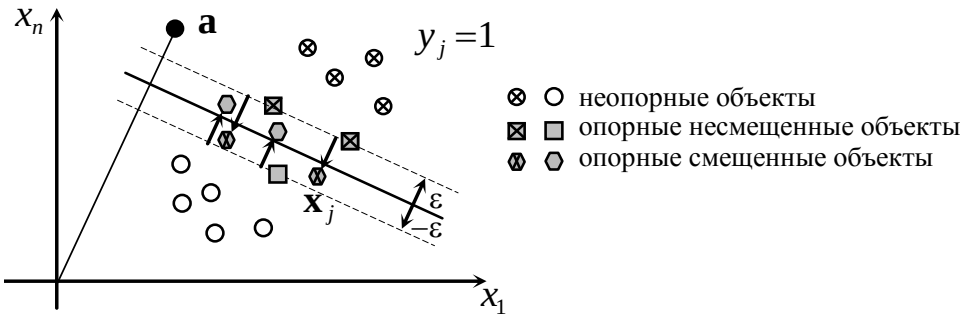


Рисунок 5. Опорные объекты обучающей выборки.

1.1.3 Вероятностная интерпретация метода опорных векторов

При всей эффективности метод опорных векторов (9)-(12) остается эвристическим по своей конструкции. С момента его создания в мировой литературе был предпринят ряд попыток снабдить его некоторой вероятностной интерпретацией [7,8]. Однако эти интерпретации оставались неполными, не позволяющими в полной мере использовать вероятностный аппарат для наделения чрезвычайно популярного метода опорных векторов принципиально новыми свойствами.

Первый подход [7] основан на калибровке по обучающей выборке монотонного преобразования значений решающей функции (12) в единичный интервал $[0,1]$. Такой способ хорошо зарекомендовал себя на практике, однако не смог удовлетворить потребности научного сообщества изложить метод опорных векторов с вероятностных позиций. Второй подход [8] заключается в переосмыслении самой задачи обучения метода опорных векторов (9) через вероятностные суждения о модели порождения данных. Ключевая идея сводится к интерпретации критерия оптимизации метода опорных векторов (9) в виде задачи максимизации логарифмированной апостериорной плотности распределения (*Maximum A Posterior estimation, MAP*) параметров неизвестной гиперплоскости

$$\ln D(\mathbf{a},b) + C\ln\prod_{j=1}^{N} Q(y\,|\,\mathbf{x}_j,\mathbf{a},b)G(\mathbf{x}_j) \to \max(\mathbf{a},b)\,. \qquad (13)$$

Первое слагаемое $D(\mathbf{a},b) \sim \exp(-0.5\mathbf{a}^T\mathbf{a} - 0.5b^2B^{-2})$ в целевой функции критерия (13) отвечает за априорное суждение об искомых параметрах, которое выражается в предположении, что параметры гиперплоскости $(\mathbf{a},b)$ попарно независимы и априори имеют нормальное распределение с единичными дисперсиями, и нулевыми математическими ожиданиями.

Второе слагаемое в (13) является логарифмированной функцией правдоподобия реализации обучающей выборки, выраженной непосредственно че-

рез функцию апостериорной вероятности $Q(y|\mathbf{x},\mathbf{a},b)$ принадлежности объекта обучающей выборки $\mathbf{x}\in\mathbb{R}^n$ к его классу $y\in\{-1,1\}$

$$Q(y=\pm1|\mathbf{x},\mathbf{a},b)=\frac{\exp\left\{-C\max\left[0,1-y(\mathbf{a}^T\mathbf{x}+b)\right]\right\}}{1+\exp(-2C)}.$$

Однако такой крайне прямолинейный способ интерпретации второго слагаемого критерия обучения (9) обладает явным недостатком, заключающийся в том, что значения функции $Q(y|\mathbf{x},\mathbf{a},b)$ строго говоря нельзя называть вероятностями, поскольку они не удовлетворяют естественному требованию $Q(y=1|\mathbf{x},\mathbf{a},b)+Q(y=-1|\mathbf{x},\mathbf{a},b)=1$. Действительно

$$Q(y=1|\mathbf{x},\mathbf{a},b)+Q(y=-1|\mathbf{x},\mathbf{a},b)=$$
$$\frac{\exp\left\{-C\max\left[0,1-(\mathbf{a}^T\mathbf{x}+b)\right]\right\}+\exp\left\{-C\max\left[0,1+(\mathbf{a}^T\mathbf{x}+b)\right]\right\}}{1+\exp(-2C)}<1,$$

исключая ситуации $|\mathbf{a}^T\mathbf{x}+b|=1$, что вынудило авторов предпринимать дополнительные меры по нормализации вероятностной модели.

В данной диссертации предлагается вероятностная модель, которая не только обобщает метод опорных векторов, но и является теоретической основой для создания методов обучения, позволяющих автоматически отбирать подмножество информативных признаков объектов, отбрасывая неинформативные. Предлагаемая вероятностная модель разработана с целью устранения необходимости «изобретения» новых эвристических алгоритмов, наделяющих метод опорных векторов свойством селективности обучения.

1.2 Проблема переобучения при обучении в признаковом пространстве большой размерности

1.2.1 Природа переобучения в линейном пространстве признаков объектов

Метод опорных векторов в своей классической постановке (9) изначально наделен простым, но очень мощным регуляризационным механизмом,

ограничивающим свободу выбора решающего правила и, который, по сути, и определил высокую эффективность, и большую популярность метода. Этот механизм заложен в само определение функции потерь (6) и принципы обучения (7), (8) и (9), который предполагает выбирать предпочтительно те гиперплоскости, которые обеспечивают разделение выборки с наибольшим зазором $\varepsilon \geq 0$. Влияние такой регуляризации на обобщающую способность обучения имеет фундаментальное теоретическое обоснование в статистической теории обучения, разработанной В.Н. Вапником и А.Я. Червоненкисом [1,9,10]. Ключевым понятием этой теории является размерность Вапика-Червоненкиса (*VC-dimension*), которая численно характеризует разнообразие выбранного семейства решающих функций, в нашем случае семейства гиперплоскостей, в котором при обучении по заданной обучающей совокупности производится поиск оптимального решения. Фактически под размерностью Вапика-Червоненкиса понимается максимальное количество объектов l, которое возможно разделить на два класса всеми возможными 2^l способами с помощью решающей функций из выбранного семейства. Нетрудно показать, что для семейства гиперплоскостей (разделяющих выборку с нулевым зазором) в $\mathbb{R}^n$ размерность Вапика-Червоненкиса равна $n+1$. В частности, в двумерном пространстве любые 3 точки еще можно разделить прямыми линиями на два класса всеми восьмью способами, а уже 4 точки - нельзя. Это означает, что при достаточно большой размерности признакового пространства для любой обучающей выборки фиксированного размера всегда можно найти подходящую гиперплоскость, которая разделить ее на два класса без единой ошибки. Именно поэтому в подобных ситуациях прямая минимизация эмпирического риска становится мало продуктивной и для решения задачи исследователь вынужден прибегать к дополнительным механизмам регуляризации обучения или, другими словами, к поиску способов ограничения свободы выбора итоговой решающей функции.

Развиваемая в диссертации вероятностная методология обеспечения селективности обучения опирается на тот факт [10], что для оптимальной

гиперплоскости, разделяющей без ошибок заданную обучающую совокупность размера N на два класса с зазором $\varepsilon > 0$, справедлива следующая оценка математического ожидания вероятности ошибки на генеральной совокупности (в среднем по всем обучающим совокупностям заданного размера):

$$EP_{error} \le E\min\left(\frac{m}{N}, \frac{[R^2\varepsilon^{-2}]}{N}, \frac{n}{N}\right), \qquad (14)$$

где m - количество опорных векторов (объектов выборки), R - радиус минимальной сферы, содержащей все данные наблюдения, и n - размерность признакового пространства, в котором производится поиск оптимальной разделяющей гиперплоскости.

Выражение (14) наглядно показывает три основных механизма, объясняющих хорошую обобщающую способность методов распознавания образов, основанных на понятии оптимальной разделяющей гиперплоскости. Во-первых, представление оптимальной гиперплоскости относительно небольшим количеством m опорных векторов (объектов выборки), во-вторых, разделение выборки с наибольшим зазором ε между объектами разных классов и, в-третьих, обучение в признаковом пространстве небольшой размерности n.

Классическая постановка метода опорных векторов (9) полностью игнорирует третий механизм повышения обобщающей способности обучения, состоящий в сокращении размерности признакового пространства, что явно указало исследователям направление для дальнейшего улучшения метода.

1.2.2 Существующие способы отбора признаков в методе опорных векторов

1.2.2.1 1-norm SVM (Lasso SVM)

Одной из первых и наиболее известных модификаций критерия метода опорных векторов (9), наделившая этот метод свойством селекции признаков

непосредственно в процессе обучения, стал метод *1-norm SVM* [11], имеющий критерий обучения:

$$\begin{cases} J_{1nSVM}(\mathbf{a},b,\boldsymbol{\delta}\,|C) = \sum_{i=1}^{n}|\,a_i\,| + C\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a},b,\boldsymbol{\delta}), \\ y_j\left(\sum_{i=1}^{n}a_i x_{ij} + b\right) \geq 1-\delta_j,\ \delta_j \geq 0,\ j=1,...,N. \end{cases} \tag{15}$$

Такой оптимизационный критерий представляет собой частный случай целевой функции общего вида $\left(\sum_{i=1}^{n}|\,x_i\,|^p\right)^{1/p} + C\sum_{j=1}^{N}\delta_j$ для $p=1$ и отличается от классического метода опорных векторов (9) исключительно видом первого слагаемого

$$\sum_{i=1}^{n}|\,a_i\,|. \tag{16}$$

Такую функцию принято называть штрафом *Lasso*, а квадратичный штраф $\sum_{i=1}^{n}a_i^2$ метода опорных векторов (9) - штрафом *Ridge*.

Критерий (15) имеет склонность в точке минимума $(\hat{\mathbf{a}},\hat{b},\hat{\boldsymbol{\delta}})$ присваивать строго нулевые значения большинству компонент направляющего вектора $\hat{a}_i = 0$, реализуя при этом на практике разумный выбор подмножества полезных для распознавания признаков. С момента своего появления, метод обучения 1-norm SVM (15) неоднократно демонстрировал свое преимущество по сравнению с классическим методом опорных векторов (9) при решении практических задач, содержащих большое количество мало полезных для распознавания признаков при относительно небольшой по размеру обучающей выборке.

1.2.2.2 Doubly Regularized SVM (Elastic Net SVM)

Несмотря на свою эффективность, метод 1-norm SVM имеет два существенных недостатка [12]. Во-первых, если исходное множество признаков содержит группу сильно линейно зависимых признаков (имеющих высокую попарную корреляцию), то такой способ обучения (15) старается в итоговом решении оставить только один из этих признаков, подавляя все остальные.

Во-вторых, количество активных признаков, т.е. количество признаков, вошедших в решение, не может быть больше чем количество объектов обучающей выборки.

Дальнейшее развитие селективных свойств метода опорных векторов получило в работе [12], в которой авторы вместо функции штрафа Lasso применили функцию штрафа *Elastic Net*, предложенную ранее для регуляризации задачи восстановления числовой зависимости [13]. Штраф Elastic Net представляет собой линейную комбинацию квадратичного штрафа Ridge $\sum_{i=1}^{n} a_i^2$ и штрафа Lasso $\sum_{i=1}^{n} |a_i|$:

$$\lambda_1 \sum_{i=1}^{n} |a_i| + \frac{\lambda_2}{2} \sum_{i=1}^{n} a_i^2 , \qquad (17)$$

где λ_1 и λ_2 - неотрицательные параметры.

Критерий обучения с функцией штрафа Elastic Net получил название *Doubly Regularized SVM* (*DrSVM*) и учитывая (17) имеет вид:

$$\begin{cases} \lambda_1 \sum_{i=1}^{n} |a_i| + \frac{\lambda_2}{2} \sum_{i=1}^{n} a_i^2 + \sum_{j=1}^{N} \delta_j \to \min(\mathbf{a}, b, \boldsymbol{\delta}), \\ y_j \left(\sum_{i=1}^{n} a_i x_{ij} + b \right) \geq 1 - \delta_j, \; \delta_j \geq 0, \; j = 1, \ldots, N. \end{cases} \qquad (18)$$

В отличие от метода 1-norm SVM (15) такой способ обучения обладает *эффектом группировки (grouping effect*), который выражается в стремлении либо удалять из решения, либо оставлять в решении сразу все признаки, имеющие высокую линейную попарную корреляцию. Кроме того, количество активных в решении признаков не ограничено размером обучающей выборки. Как следствие, такой способ обучения (18) продемонстрировал лучшую, в сравнении с 1-norm SVM, обобщающую способность на модельных и практических задачах [12].

В дальнейшем изложении результатов диссертации будет использоваться эквивалентная запись критерия обучения (18)

$$\begin{cases} J_{DrSVM}(\mathbf{a},b,\boldsymbol{\delta}\,|\,\mu,C)=\mu\sum_{i=1}^{n}|\,a_i\,|+\sum_{i=1}^{n}a_i^2+C\sum_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\boldsymbol{\delta}), \\ y_j\left(\sum_{i=1}^{n}a_i x_{ij}+b\right)\geq 1-\delta_j,\ \delta_j\geq 0,\ j=1,...,N. \end{cases} \quad (19)$$

где $C=2\lambda_2^{-1}$, $\mu=2\lambda_1\lambda_2^{-1}$.

1.2.2.3 Elastic SCAD SVM

Другим примером попытки усовершенствовать метод опорных векторов (9) наделением его селективными свойствами за счет выбора особого вида функции штрафа на компоненты направляющего вектора является работа [14]. Авторами предложена функция штрафа, которая является комбинацией квадратичного штрафа Ridge метода опорных векторов (9) и невыпуклого штрафа *SCAD* [15]:

$$\lambda_1\sum_{i=1}^{n}a_i^2+\sum_{i=1}^{n}scad(a_i\,|\,\lambda_2,\lambda_3),$$

где λ_1, λ_2 и λ_3 - неотрицательные структурные параметры, $scad(a_i\,|\,\lambda_2,\lambda_3)$ - функция штрафа SCAD

$$scad(a_i\,|\,\lambda',\lambda'')=\begin{cases} \lambda'\,|\,a_i\,|,\ \textit{если}\ |\,a_i\,|\leq\lambda', \\ -\dfrac{|\,a_i\,|^2-2a\lambda'\,|\,a_i\,|+\lambda'^2}{2(d-1)},\ \textit{если}\ \lambda'<|\,a_i\,|\leq\lambda'\lambda'', \\ \dfrac{(\lambda''+1)\lambda'^2}{2},\ \textit{если}\ |\,a_i\,|>\lambda'\lambda''. \end{cases}$$

Однако авторами метода не приводится какого-либо теоретического обоснования такого выбора штрафа и его преимуществ по сравнению с уже зарекомендовавшими себя штрафами Lasso и Elastic Net. Экспериментальное исследование, проведенное авторами, к сожалению, также не продемонстрировало какого-либо заметного превосходства предлагаемого подхода. В связи с этим, в диссертации не будет подробно рассматриваться функция штрафа Elastic SCAD и основанный на нем метод обучения Elastic SCAD SVM.

1.2.3 Свойства методов отбора признаков и недостаточность существующих подходов

1.2.3.1 Механизм селективности отбора признаков

Модификации метода опорных векторов, исследуемых в диссертации, отличаются друг от друга только видом штрафа на компоненты направляющего вектора в целевой функции соответствующих критериев (9), (15) и (19). Как будет показано далее, использование той или иной функции штрафа фактически означает привнесение в процесс обучения априорных суждений о том, какие направления, с точки зрения исследователя, являются более предпочтительными, а какие, наоборот, менее предпочтительными.

В диссертации предлагается характеризовать всякую функцию штрафа, во-первых, определяемыми ею множествами «наиболее предпочтительных» и «наименее предпочтительных» направлений направляющего вектора, и, во-вторых, количественной оценкой способности штрафа различать такие направления между собой.

Известные формулировки критериев обучения SVM (9), Lasso SVM (15) и Elastic Net SVM (19) в том виде, в котором они были предложены авторами, предполагают поиск такого направляющего вектора $\mathbf{a} \in \mathbb{R}^n$, для которого сумма эмпирического риска $\sum_{j=1}^{N} \delta_j$ и соответствующего априорного штрафа будет наименьшей. Однако здесь важно отметить, что величина эмпирического риска $\sum_{j=1}^{N} \delta_j$ зависит не только от направления вектора $\mathbf{a} \in \mathbb{R}^n$, но зависит и от его нормы $(\mathbf{a}^T\mathbf{a})^{1/2}$, непосредственно определяющей величину зазора $2\varepsilon = 2(\mathbf{a}^T\mathbf{a})^{-1/2}$, с которым разделяется обучающая выборка. Для анализа селективных свойств критерия обучения необходимо в первую очередь разделить механизмы селекции направлений и максимизации зазора. Такое разделение (8) уже было сделано в этой главе для классического метода опорных векторов (9). Действительно, в критерии (8) величина зазора ε входит в критерий как отдельная переменная оптимизации, а область поиска направ-

ляющего вектора ограничена сферой $\mathbf{a}^T\mathbf{a}=1$ единичного радиуса. При такой эквивалентной записи (8) известного критерия (9) выбор направляющего вектора определяет исключительно ориентацию разделяющей гиперплоскости и в силу фиксированной нормы $\mathbf{a}^T\mathbf{a}=1$ не несет в себе информацию о величине зазора, с которым разделяется обучающая выборка.

Выпишем критерий (8) еще раз, поделив целевую функцию на значение структурного параметра $C>0$:

$$\begin{cases} G_{SVM}(\mathbf{a},b,\boldsymbol{\delta},\varepsilon\,|\,C)=\dfrac{1}{C\varepsilon^2}+\sum\limits_{j=1}^{N}\delta_j\to\min(\varepsilon,\mathbf{a},b,\boldsymbol{\delta},\varepsilon), \\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\geq\varepsilon(1-\delta_j),\ \delta_j\geq 0,\ j=1,\ldots,N,\ \mathbf{a}^T\mathbf{a}=1,\ \varepsilon\geq 0. \end{cases} \tag{20}$$

В такой записи критерия, в отличие от классической (9), хорошо видно, что выбор оптимального направления (вектора, принадлежащего сфере с радиусом единица) определяется исключительно величиной зазора ε^{-2} и эмпирическим риском (качеством разделения выборки с зазором ε на два класса) $\sum_{j=1}^{N}\delta_j$. Такой выбор направления не зависит от каких-либо дополнительных суждений, не зависящих от выборки, т.е. об априорной предпочтительности одних направлений перед другими. Именно такое поведение метода обучения естественно характеризовать минимальной селективностью, а точнее ее полным отсутствием.

Формализуем этот подход. Для этого запишем критерий обучения (20) с функцией штрафа в самом общем виде

$$\begin{cases} G(\mathbf{a},b,\boldsymbol{\delta},\varepsilon\,|\,C)=V(\mathbf{a},\varepsilon\,|\,C)+\sum\limits_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\boldsymbol{\delta},\varepsilon), \\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\geq\varepsilon(1-\delta_j),\ \delta_j\geq 0,\ j=1,\ldots,N,\ \mathbf{a}^T\mathbf{a}=1,\ \varepsilon\geq 0, \end{cases}$$

где $V(\mathbf{a},\varepsilon\,|\,C)$ - штраф, отвечающий за максимизацию зазора ε и селекцию направлений $\mathbf{a}$, причем в (20) этот штраф не зависит от $\mathbf{a}$, C - структурный параметр обучения.

Направления из множества $\mathcal{A}=\{\mathbf{a}\in\mathbb{R}^n:\|\mathbf{a}\|=1\}$, которые в критерии обучения штрафуются сильнее всего с точки зрения выбранной функции штрафа $V(\mathbf{a},\varepsilon\,|\,C)$, будем называть *«наихудшими» направлениями*

$$\mathcal{A}_{\max}(\varepsilon \mid C) = \left\{ \mathbf{a} \in \mathbb{R}^n : V(\mathbf{a}, \varepsilon \mid C) = \max_{\mathbf{a}' \in \mathcal{A}} V(\mathbf{a}', \varepsilon \mid C) \right\}, \tag{21}$$

а направления, которые слабее всего – *«наилучшими» направлениями*

$$\mathcal{A}_{\min}(\varepsilon \mid C) = \left\{ \mathbf{a} \in \mathbb{R}^n : V(\mathbf{a}, \varepsilon \mid C) = \min_{\mathbf{a}' \in \mathcal{A}} V(\mathbf{a}', \varepsilon \mid C) \right\}. \tag{22}$$

В качестве количественной оценки способности функции штрафа различать «наихудшие» и «наилучшие» направления между собой естественно оценивать *уровень селективности* (*selectivity level – SL*), рассчитываемый как разность максимального и минимального значений штрафа для «наихудших» и «наилучших» направлений соответственно:

$$SL(\varepsilon \mid C) = \max_{\mathbf{a} \in \mathcal{A}} V(\mathbf{a}, \varepsilon \mid C) - \min_{\mathbf{a} \in \mathcal{A}} V(\mathbf{a}, \varepsilon \mid C) \geq 0. \tag{23}$$

Тогда для критерия метода опорных векторов (20) получим

$$V_{SVM}(\mathbf{a}, \varepsilon \mid C) = C^{-1}\varepsilon^{-2}, \tag{24}$$

и очевидно, что справедливы следующие утверждения:

$$\forall \varepsilon \left[\mathbf{a} \in \mathcal{A}_{\min}^{SVM} = \mathcal{A}, V_{SVM}(\mathbf{a}, \varepsilon \mid C) = C^{-1}\varepsilon^{-2} \right], \tag{25}$$

$$\forall \varepsilon \left[\mathbf{a} \in \mathcal{A}_{\max}^{SVM} = \mathcal{A}, V_{SVM}(\mathbf{a}, \varepsilon \mid C) = C^{-1}\varepsilon^{-2} \right]. \tag{26}$$

Оба множества совпадают с областью определения $\mathcal{A} = \{\mathbf{a} \in \mathbb{R}^n : \| \mathbf{a} \| = 1\}$. Уровень селективности SL (23) для штрафа (24), учитывая (25) и (26), будет наименьшим из возможных, т.е. равным нулю

$$SL_{SVM}(\varepsilon \mid C) = C^{-1}\varepsilon^{-2} - C^{-1}\varepsilon^{-2} = 0,$$

что показывает полное отсутствие селективности у классического метода опорных векторов.

Нетрудно проверить, выполнив простейшую замену переменных $\mathbf{a} = \varepsilon \tilde{\mathbf{a}}$, $b = \varepsilon \tilde{b}$ и домножив целевую функцию (15) на величину параметра $C > 0$, что оптимизационная задача

$$\begin{cases} G_{1nSVM}(\mathbf{a}, b, \boldsymbol{\delta}, \varepsilon \mid C) = \dfrac{1}{C\varepsilon} \sum\limits_{i=1}^{n} | a_i | + \sum\limits_{j=1}^{N} \delta_j \to \min(\mathbf{a}, b, \boldsymbol{\delta}, \varepsilon), \\ y_j(\mathbf{a}^T \mathbf{x}_j + b) \geq \varepsilon(1 - \delta_j), \delta_j \geq 0,\ j = 1, \ldots, N,\ \mathbf{a}^T\mathbf{a} = 1,\ \varepsilon \geq 0, \end{cases} \tag{27}$$

эквивалентна критерию обучения метода 1-norm SVM с функцией штрафа Lasso.

Однако теперь, в отличие от критерия метода опорных векторов (20), каждое направление штрафуется слагаемым $C^{-1}\varepsilon^{-1}\sum_{i=1}^{n}|a_i|$, и это слагаемое для фиксированных ε и C принимает разные значения для разных направлений $\mathcal{A}=\{\mathbf{a}\in\mathbb{R}^n:\|\mathbf{a}\|=1\}$, причем это различие не зависит от выборки, предъявленной для обучения. Кроме того, именно в такой постановке становится понятным другое свойство метода 1-norm SVM, которое неочевидно при классической записи критерия (15). Заметим, что чем больше зазор ε, тем меньший вклад в значение целевой функции делает первое слагаемое $C^{-1}\varepsilon^{-1}\sum_{i=1}^{n}|a_i|$, и, следовательно, тем меньше выражено предпочтение одних направлений перед другими.

Действительно, для метода 1-norm SVM имеем

$$V_{1nSVM}(\mathbf{a},\varepsilon\,|\,C)=C^{-1}\varepsilon^{-1}\sum_{i=1}^{n}|a_i|, \tag{28}$$

и нетрудно проверить, что справедливы следующие утверждения:

$$\forall\varepsilon\ \ \mathbf{a}\in\mathcal{A}_{\min}^{1nSVM}\leftrightarrow\underbrace{\left[|a_i|\in\{0,1\},\ \sum_{i=1}^{n}|a_i|=1,\ V_{1nSVM}(\mathbf{a},\varepsilon\,|\,C)=C^{-1}\varepsilon^{-1}\right]}_{\text{только один коэфф. }|a_i|=1\text{ отличен от нуля}}, \tag{29}$$

$$\forall\varepsilon\ \ \mathbf{a}\in\mathcal{A}_{\max}^{1nSVM}\leftrightarrow\underbrace{\left[|a_1|=\ldots=|a_n|=n^{-1/2},V_{1nSVM}(\mathbf{a},\varepsilon\,|\,C)=C^{-1}\varepsilon^{-1}n^{1/2}\right]}_{\text{все коэфф. }a_i\text{ одинаковы по модулю}}. \tag{30}$$

Содержательно эти утверждения означают, что «наилучшими» направлениями являются вектора $\mathbf{a}\in\mathcal{A}_{\min}^{1nSVM}$, совпадающие по направлению с одним из орт признакового пространства $|a_i|\in\{0,1\}$, $\sum_{i=1}^{n}|a_i|=1$, что в свою очередь означает вхождение в итоговую решающую функцию только одного признака и полное исключение из распознавания всех остальных признаков. «Наихудшие» направления $\mathbf{a}\in\mathcal{A}_{\max}^{1nSVM}$ предполагают равноправное включение в решающую функцию сразу всех признаков с одинаковыми по модулю весами $|a_1|=\ldots=|a_n|=n^{-1/2}$ (компонентами направляющего вектора). Эти

множества не зависят ни от величины зазора ε, с которым разделяется выборка, ни от значения структурного параметра обучения C, являясь фундаментальным свойством механизма селективности функции штрафа Lasso, а, следовательно, и критерия обучения метода 1-norm SVM.

Уровень селективности SL (23) для штрафа Lasso (28), учитывая (29) и (30), будет равен

$$SL_{1nSVM}(\varepsilon \mid C) = C^{-1}\varepsilon^{-1}(n^{-1/2} - 1). \qquad (31)$$

Кроме того, из (31) нетрудно увидеть, что величина «различимости направлений» (уровень селективности) уменьшается с ростом величины зазора, что характеризует способ балансировки двух разных механизмов регуляризации (отбора признаков и максимизации зазора) в методе Lasso SVM. На рисунке 6 графически проиллюстрирована зависимость уровня селективности (31) от величины зазора для $n = 2$ и трех различных значений структурного параметра C.

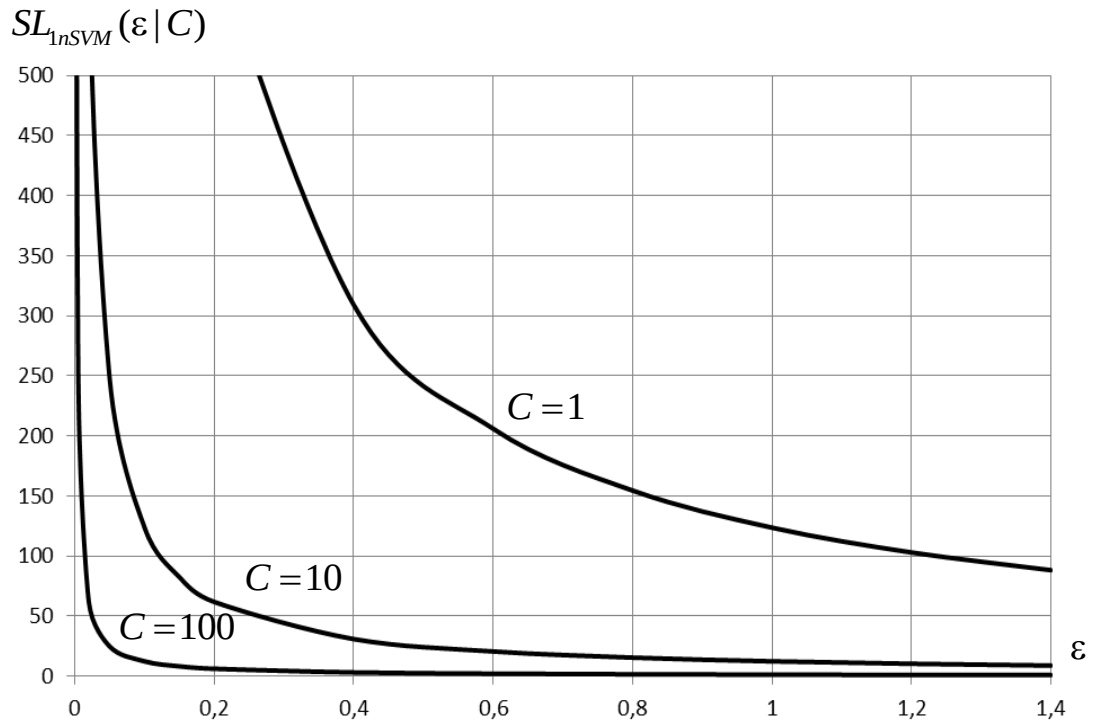


Рисунок 6. Зависимость уровня селективности штрафа Lasso от величины зазора ε и параметра разделимости выборки C.

Аналогичным образом можно показать, что критерий обучения DrSVM (18), основанный на штрафе Elastic Net, эквивалентен оптимизационной задаче

$$\begin{cases} J(\mathbf{a},b,\boldsymbol{\delta},\varepsilon\,|\,\lambda_1,\lambda_2) = V_{DrSVM}(\mathbf{a},\varepsilon\,|\,\lambda_1,\lambda_2) + \sum_{j=1}^{N}\delta_j \to \min(\mathbf{a},b,\boldsymbol{\delta},\varepsilon), \\ y_j\left(\sum_{i=1}^{n} a_i x_{ij} + b\right) \geq \varepsilon(1-\delta_j),\ \delta_j \geq 0,\ j=1,\ldots,N,\ \mathbf{a}^T\mathbf{a}=1,\ \varepsilon \geq 0, \end{cases} \tag{32}$$

где

$$V_{DrSVM}(\mathbf{a},\varepsilon\,|\,\lambda_1,\lambda_2) = \lambda_1 \frac{1}{\varepsilon}\sum_{i=1}^{n}|\,a_i\,| + \frac{\lambda_2}{2}\frac{1}{\varepsilon^2}. \tag{33}$$

Нетрудно проверить справедливость следующих утверждений:

$$\forall \varepsilon\ \mathbf{a} \in \mathcal{A}_{\min}^{DrSVM} \leftrightarrow \underbrace{\left[\,|\,a_i\,| \in \{0,1\},\ \textstyle\sum_{i=1}^{n}|\,a_i\,| = 1,\ V_{DrSVM}(\mathbf{a},\varepsilon\,|\,\lambda_1,\lambda_2) = \lambda_1\varepsilon^{-1} + 0.5\lambda_2\varepsilon^{-2}\right]}_{\text{только один коэфф. } |a_i|=1 \text{ отличен от нуля}},$$

$$\forall \varepsilon\ \mathbf{a} \in \mathcal{A}_{\max}^{DrSVM} \leftrightarrow \underbrace{\left[\,|\,a_1\,| = \ldots = |\,a_n\,| = n^{-1/2},\ V_{DrSVM}(\mathbf{a},\varepsilon\,|\,\lambda_1,\lambda_2) = \lambda_1\varepsilon^{-1}n^{1/2} + 0.5\lambda_2\varepsilon^{-2}\right]}_{\text{все коэфф. } a_i \text{ одинаковы по модулю}}.$$

Тогда уровень селективности согласно (23) определяется выражением

$$SL_{DrSVM}(\varepsilon\,|\,\lambda_1) = \lambda_1\varepsilon^{-1}(n^{-1/2}-1).$$

Таким образом, для методов 1-norm SVM, и DrSVM множества «наилучших» направлений, множества «наихудших» направлений, и уровни селективности полностью идентичны, и определяются наличием в критериях обучения штрафа Lasso.

Использование штрафов Lasso и Elastic Net фактически означает принятие специальной априорной гипотезы о модели порождения данных, которая сводится к предположению, что объекты «скорее всего» не будут представлены близкими по полезности для распознавания признаками. Следовательно, в ситуациях, когда в исходных данных есть такие признаки, обучение методами 1-norm SVM и DrSVM может приводить к снижению обобщающей способности полученных решающих функций.

В следующем параграфе будет показано принципиальное различие методов 1-norm SVM и DrSVM, и их недостаточность с иной точки зрения.

1.2.3.2 Эффект группировки признаков

В работе [12] в качестве основного недостатка метода 1-norm SVM (15) указывается тот факт, что метод полностью игнорирует при отборе признаков их взаимную линейную зависимость (попарную корреляцию), оставляя в решении только лишь небольшое (не больше количества объектов в обучающей выборке) подмножество признаков, полностью подавляя все остальные. Такое свойство метода обучения является неприемлемым для ряда прикладных задач, в частности, при анализе экспрессии генов с использованием микрочипов (*microarray analysis*), поскольку уровни экспрессии генов, которые отвечают за один и тот же биологический механизм, как правило, сильно линейно зависимы, и все такие гены вносят свой вклад в биологический процесс. Поэтому в работе [12] идеальным назван такой метод, который бы эффективно подавлял незначимые для распознавания гены и автоматически включал бы в итоговое решение все взаимосвязанные. Предлагаемый в работе [12] метод DrSVM (19), в отличие от 1-norm SVM, отчасти обладает желаемым свойством, которое выражено в виде утверждения, что чем больше корреляция ρ_{jl} между признаками с индексами j и l, тем меньше будут отличаться в итоговом решении соответствующие коэффициенты $\hat{a}_j$ и $\hat{a}_l$ направляющего вектора:

$$\left|\hat{a}_j - \hat{a}_l\right| \le \frac{\sqrt{N}}{\lambda_2}\sqrt{2(1-\rho_{jl})}, \ j,l = 1,...,n.$$

Здесь N – размер обучающей выборки, λ_2 - структурный параметр регуляризации критерия обучения (18) при квадратичном штрафе. Такое свойство метода было названо *эффектом группировки* или *grouping effect*, как в оригинальной публикации.

Заметим, что свойство группировки одинаково работает как для полезных генов или, в общем случае, признаков, так и для признаков, которые не несут никакой информации для решения задачи, но тоже участвуют в процессе обучения. Т.е. метод DrSVM будет «стараться» либо удалять все такие

«шумовые признаки», либо наоборот оставлять такую группу целиком. Последний случай, очевидно, не может способствовать увеличению обобщающей способности распознавания и интерпретации итогового решения с точки зрения содержания прикладной задачи. Другими словами, и метод DrSVM, и Lasso SVM являются представителями крайних подходов к обучению, которые либо наделяют процесс обучение эффектом группировки, либо полностью от него избавляют.

Одной из задач данной диссертации является дальнейшее развитие идеи группировки попарно зависимых признаков. Один из предлагаемых в работе методов обучения, а именно, *метод опорных признаков* (раздел 3.3), разрабатывается именно для того, чтобы обеспечить *эффект избирательной группировки признаков* (раздел 3.4.2). Избирательность предлагаемого метода выражается в стремлении сохранять в итоговом решении только попарно зависимые признаки, обладающие достаточной значимостью для распознавания, аналогично DrSVM, и в максимальном подавлении всех остальных малозначимых признаков, независимо от их попарной зависимости, аналогично Lasso SVM.

1.3 Предлагаемая концепция: Байесовский подход к одновременному построению разделяющей гиперплоскости и выбору подмножества релевантных признаков

В диссертации предлагается специальная байесовская постановка обучения распознаванию двух классов объектов в линейном признаковом пространстве, приводящая к обобщению метода опорных векторов и являющаяся теоретической основой для создания новых селективных методов обучения. Основная идея байесовской постановки заключается в построении простейшей системы вероятностных предположений о паре плотностей распределения объектов двух классов, определяемой объективно существующей, но неизвестной гиперплоскостью в линейном пространстве признаков, при некотором априорном предположении о ее случайном выборе. В диссертации предлагаются и детально рассматриваются две частные априорные модели случайного

направляющего вектора разделяющей гиперплоскости, позволяющие в процессе обучения количественно различать его компоненты, а, следовательно, и исходные признаки, по степени их полезности для распознавания двух заданных классов объектов, управляя при этом степенью селективности такого различия. Эти две модели приведут к двум разным методам обучения с отбором полезных признаков, один из которых назван *методом релевантных признаков с управляемой селективностью (метод RFM)*, а второй – *методом опорных признаков с управляемой селективностью (метод SKM)*.

1.4 Основные задачи исследования

Для реализации изложенной концепции в данной диссертации сформулированы и решены следующие задачи:

- Разработка общей математической постановки задачи обучения двухклассовому распознаванию образов в линейном пространстве признаков на основе количественного измерения расстояния между вектором признаков объекта и разделяющей гиперплоскостью.
- Разработка математической постановки общей вероятностной задачи обучения двухклассовому распознаванию объектов двух классов в линейном пространстве признаков.
- Разработка частных априорных вероятностных моделей генеральной совокупности в терминах совместного априорного распределения компонент направляющего вектора разделяющей гиперплоскости, отражающих две принятые в мировой литературе стратегии отбора признаков – взвешивания всех исходно заданных признаков (*feature weighting*) и жесткого выбора их подмножества (*feature subset selection*).
- Разработка методов и алгоритмов оценивания параметров разделяющей гиперплоскости, реализующих вероятностный принцип обучения с заданной селективностью отбора признаков при выбранных априорных предположениях о генеральной совокупности объектов двух классов.
- Количественное экспериментальное исследование разработанных методов на модельных данных и данных прикладной задачи.

2 БАЙЕСОВСКИЙ ПОДХОД К ПОИСКУ ОПТИМАЛЬНОЙ РАЗДЕЛЯЮЩЕЙ ГИПЕРПЛОСКОСТИ

Во второй главе излагается общая вероятностная байесовская постановка задачи обучения двухклассовому распознаванию образов в линейном пространстве $\mathbf{x} \in \mathbb{R}^n$ на основе метода опорных векторов, сформулированного автором в его классическом виде в сугубо детерминистских терминах [1].

Основная идея предлагаемой в диссертации байесовской постановки заключается в построении специальной системы вероятностных предположений о паре плотностей распределения объектов двух классов, определяемой объективно существующей, но неизвестной гиперплоскостью в линейном пространстве признаков, при некотором априорном предположении о ее случайном выборе.

2.1 Параметрическое семейство пары плотностей распределения, определяемое гиперплоскостью

Пусть $\mathbf{a}^T\mathbf{x} + b \gtrless 0$, т.е. $\sum_{i=1}^{n} a_i x_i + b \gtrless 0$, – некоторая гиперплоскость с направляющим вектором $\mathbf{a} = (a_1,...,a_n) \in \mathbb{R}^n$ и параметром сдвига $b \in \mathbb{R}$. Свяжем с этой гиперплоскостью пару параметрических семейств условных плотностей распределения вероятностей вектора признаков случайного объекта $\varphi(\mathbf{x} \mid \mathbf{a}, b, y; c) = \varphi(x_1,...,x_n \mid a_1,...,a_n, b, y; c)$, $\mathbf{x} \in \mathbb{R}^n$, относительно фиксированного индекса его класса $y \in \{-1, 1\}$, определяемых параметрами $\mathbf{a} \in \mathbb{R}^n, b, c \in \mathbb{R}$. Эти два семейства призваны выражать предположение, что случайные векторы признаков объектов двух классов распределены главным образом в «своих» полупространствах $\mathbf{a}^T\mathbf{x} + b > 0$ и $\mathbf{a}^T\mathbf{x} + b < 0$, однако могут попадать и в «чужие» полупространства, причем степенью возможности «ошибочных» значений управляет параметр $c > 0$. Равномерное распределение в бесконечной области является некорректным

вероятностным понятием, поэтому в диссертации предлагаются несобственные плотности, не образующие единичных интегралов по всему линейному пространству $\mathbb{R}^n$:

$$\varphi(\mathbf{x} \mid y; \mathbf{a}, b, c) = const \begin{cases} 1, \ y(\mathbf{a}^T\mathbf{x} + b) \geq 1, \\ \exp\left[-c\left(1 - y(\mathbf{a}^T\mathbf{x} + b)\right)\right], \ y(\mathbf{a}^T\mathbf{x} + b) < 1, \ c > 0. \end{cases} \quad (34)$$

Некорректность несобственной пары плотностей (34) не приводит далее к математическим противоречиям, поскольку они участвуют только в формуле Байеса [16].

Кроме того, несобственность плотностей может легко быть устранена формулированием семейства (34), например, в виде:

$$\varphi(\mathbf{x} \mid y, \mathbf{a}, b, c, d, h) = \beta \begin{cases} 1, 1 \leq y(\mathbf{a}^T\mathbf{x} + b) \leq d + 1 \text{ и } \mathbf{x}^T\mathbf{x} - (\mathbf{a}^T\mathbf{x})^2 \leq h, \\ \exp\left[-c\left(1 - y(\mathbf{a}^T\mathbf{x} + b)\right)\right], -d - 1 \leq y(\mathbf{a}^T\mathbf{x} + b) < 1 \text{ и } \mathbf{x}^T\mathbf{x} - (\mathbf{a}^T\mathbf{x})^2 \leq h, \\ 0, \ y(\mathbf{a}^T\mathbf{x} + b) > d + 1 \text{ или } -d - 1 > y(\mathbf{a}^T\mathbf{x} + b) \text{ или } \mathbf{x}^T\mathbf{x} - (\mathbf{a}^T\mathbf{x})^2 > h, \end{cases}$$

где $d, h > 0$ - параметры внешних границ распределения.

Однако такая формулировка семейства плотностей, ограниченных цилиндрической областью, только утяжелит изложение предлагаемого подхода, и поэтому детально рассматриваться в рамках диссертации не будет.

Пара плотностей распределения (34), сконцентрированных в основном «равномерно» по разные стороны заданной гиперплоскости, выражает предположение, что единственное знание о двух классах объектов заключается в их расположении преимущественно по разные стороны заданной гиперплоскости. Наглядное представление о такой паре плотностей распределения дает рисунок 7.

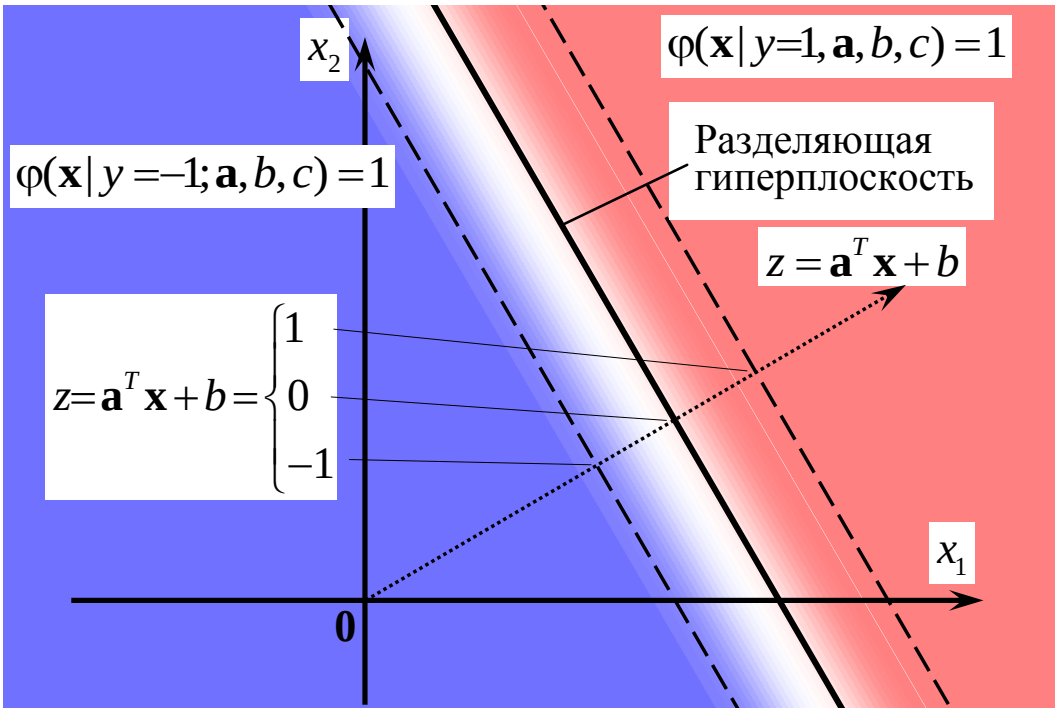


Рисунок 7. Графическое представление параметрического семейства плотностей распределения (34) двух классов объектов в двумерном пространстве.

Общий действительный параметр обеих плотностей $c>0$, определяющий степень возможности попадания вектора признаков объекта определенного класса в «чужое» полупространство (рис. 8), является структурным параметром, имеющим известное заданное значение.

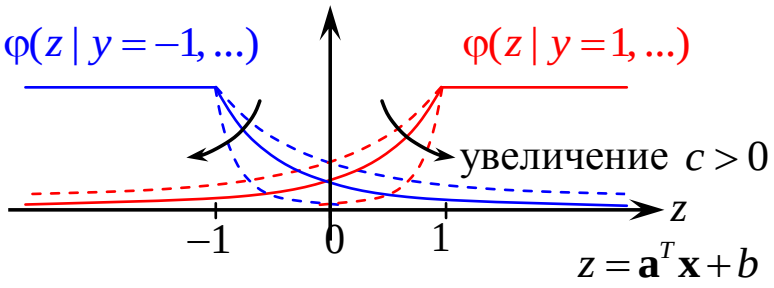


Рисунок 8. Графическое представление параметрического семейства плотностей распределения двух классов объектов в проекции на направляющий вектор разделяющей гиперплоскости.

Будем далее предполагать, что получена обучающая совокупность

$$(X,Y)=\{(\mathbf{x}_j,y_j),\ j=1,\ldots,N\}, \qquad (35)$$

образованная векторами признаков $\mathbf{x}_j = (x_{ij}, i = 1,...,n)$ с известными индексами классов y_j. Тогда совместное условное распределение обучающей совокупности $(X \mid Y)$ представимо в виде произведения плотностей (34):

$$\Phi(X \mid Y, \mathbf{a}, b; c) = \prod_{j=1}^{N} \varphi\left(\mathbf{x}_j \mid \mathbf{a}, b, y_j; c\right). \tag{36}$$

2.2 Априорное распределение параметров гиперплоскости

Другим ключевым предположением в предлагаемой вероятностной модели является суждение об априорном совместном распределении $\Psi(\mathbf{a}, b)$ параметров $(\mathbf{a}, b)$ разделяющей гиперплоскости $\mathbf{a}^T\mathbf{x} + b \gtrless 0$. Будем считать, что отсутствуют какие-либо априорные предпочтения величин порога разделяющей гиперплоскости $b \in \mathbb{R}$, тогда:

$$\Psi(\mathbf{a}, b) \propto \Psi(\mathbf{a}). \tag{37}$$

2.3 Общий байесовский критерий обучения для параметрического семейства пары плотностей распределения

Апостериорная плотность распределения $P(\mathbf{a}, b \mid X, Y, \mu)$ параметров $(\mathbf{a}, b)$ относительно обучающей совокупности (X, Y) определяется формулой Байеса:

$$P(\mathbf{a}, b \mid X, Y; c) = \frac{\Psi(\mathbf{a}, b)\Phi(X \mid Y, \mathbf{a}, b; c)}{\int\limits_{\mathbb{R}^n} \left(\int\limits_{\mathbb{R}} \Psi(\mathbf{a}', b')\Phi(X \mid Y, \mathbf{a}', b'; c) db' \right) d\mathbf{a}'} \propto \Psi(\mathbf{a}, b)\Phi(X \mid Y, \mathbf{a}, b). \tag{38}$$

Понимание обучения как задачи максимизации этой апостериорной плотности распределения в пространстве параметров модели $(\mathbf{a}, b)$ приводит к очевидному критерию:

$$\begin{aligned} (\hat{\mathbf{a}}, \hat{b} \mid X, Y; c) = &\arg\max_{\mathbf{a} \in \mathbb{R}^n,\, b \in \mathbb{R}} \Psi(\mathbf{a}, b)\Phi(X \mid Y, \mathbf{a}, b; c) = \\ &\arg\max_{\mathbf{a} \in \mathbb{R}^n,\, b \in \mathbb{R}} \left(\ln \Psi(\mathbf{a}, b) + \ln \Phi(X \mid Y, \mathbf{a}, b; c) \right). \end{aligned} \tag{39}$$

Байесовский критерий обучения (39) со структурой условного распределения наблюдений (34)-(36) является ***основным теоретическим***

предложением данной диссертации. Следующая теорема, доказанная в диссертации, показывает, что этот критерий представляет собой обобщение классического метода опорных векторов.

Теорема 1. *Критерий обучения (39) для семейства распределений (34)-(36) и априорного распределения параметров (37) эквивалентен оптимизационной задаче минимизации целевой функции* $J(\mathbf{a},b,\delta_1,...,\delta_N \mid c)$ *на выпуклом множестве, заданном набором линейных ограничений-неравенств для объектов обучающей совокупности:*

$$\begin{cases} J(\mathbf{a},b,\delta_1,...,\delta_N \mid c) = -\ln\Psi(\mathbf{a}) + c\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a},b,\delta_1,,...,\delta_N), \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \geq 1-\delta_j,\ \delta_j \geq 0,\ j=1,...,N. \end{cases} \qquad (40)$$

Доказательство теоремы приведено в приложении на с. 96.

По своей структуре оптимизационный критерий (40) отличается от классического критерия опорных векторов (9) только слагаемым $-\ln\Psi(\mathbf{a})$ вместо квадрата нормы направляющего вектора искомой разделяющей гиперплоскости $\mathbf{a}^T\mathbf{a} = \sum_{i=1}^{n} a_i^2$. Параметр $c > 0$ семейств распределений (34), отвечающий за способность объектов генеральной совокупности нарушать границу своих классов, регламентирует приоритет между качеством разделения обучающей совокупности и априорными предпочтениями по выбору направляющего вектора.

Если функция $\ln\Psi(\mathbf{a})$ строго вогнута, то целевая функция $J(\mathbf{a},b,\delta_1,...,\delta_N \mid c)$ в (40) строго выпукла. Поскольку система ограничений в (40) образует выпуклый многогранник в линейном пространстве $(\mathbf{a},b,\delta_1,...,\delta_N) \in \mathbb{R}^n \times \mathbb{R}^{N+1}$ предлагаемый в диссертации обобщенный критерий обучения по методу опорных векторов в общем виде представляет собой задачу выпуклого программирования. Это значит, что существует единственный набор оптимальных параметров $(\hat{\mathbf{a}},\hat{b})$, удовлетворяющий всем ограничениям задачи (40) и доставляющий минимум целевой функции $J(\mathbf{a},b,\delta_1,...,\delta_N \mid c)$.

Результатом обучения по обучающей совокупности (35) является решающее правило классификации $d(\mathbf{x}) = \hat{\mathbf{a}}^T\mathbf{x} + \hat{b} \gtrless 0$ (4), применимое к любому новому объекту с вектором признаков $\mathbf{x} \in \mathbb{R}^n$.

2.4 Апостериорные вероятности классов объектов

Принципиальной новизной предлагаемой в диссертации формулировки задачи обучения является тот факт, что вероятностное предположение об условном распределении векторов признаков случайных объектов (34), выражающее основную специфику метода опорных векторов, позволяет интерпретировать полученное решающее правило не только как суждение о классе нового объекта, но и наделить это суждение естественной оценкой его «уверенности».

Действительно, примем дополнительное предположение, что случайно появившийся новый объект, не содержащийся в обучающей совокупности, объективно принадлежит тому или иному классу с некоторыми априорными вероятностями

$$q_{-1} = P(y=-1),\ q_1 = P(y=1),\ q_{-1} + q_1 = 1.$$

Тогда, для параметров разделяющей гиперплоскости $(\hat{\mathbf{a}}, \hat{b})$, найденных как решение задачи обучения (40), апостериорная вероятность принадлежности объекта $\mathbf{x} \in \mathbb{R}^n$ к классу $y = 1$ определяется формулой Байеса через плотности распределения двух классов $\varphi(\mathbf{x} \mid \hat{\mathbf{a}}, \hat{b}, y; c)$:

$$p(y=1|\mathbf{x};\hat{\mathbf{a}},\hat{b},c) = 1 - p(y=-1|\mathbf{x};\hat{\mathbf{a}},\hat{b},c) = \frac{q_1\,\varphi(\mathbf{x}|y=1;\hat{\mathbf{a}},\hat{b},c)}{q_1\,\varphi(\mathbf{x}|y=1;\hat{\mathbf{a}},\hat{b},c) + q_{-1}\,\varphi(\mathbf{x}|y=-1;\hat{\mathbf{a}},\hat{b},c)}. \qquad (41)$$

Для простоты дальнейших рассуждений примем, в частности, равными априорные вероятности двух классов $q_{-1} = q_1 = 1/2$ [1]. Тогда

---

[1] Выбор различных априорных вероятностей классов $q_1 + q_{-1} = 1$ отражает другое предположение об их свойстве в сравнении с выбором различных значений параметра пересечения классов $c > 0$, характеризуя их разнонаполненность. В данной работе роль априорных вероятностей классов (q_1, q_{-1}) как параметров вероятностной модели детально не исследуется.

$$p(y=1|\mathbf{x};\hat{\mathbf{a}},\hat{b},c)=1-p(y=-1|\mathbf{x};\hat{\mathbf{a}},\hat{b},c)=\frac{\varphi(\mathbf{x}|y=1;\hat{\mathbf{a}},\hat{b},c)}{\varphi(\mathbf{x}|y=1;\hat{\mathbf{a}},\hat{b},c)+\varphi(\mathbf{x}|y=-1;\hat{\mathbf{a}},\hat{b},c)}. \quad (42)$$

Теорема 2. *Для условных плотностей распределения (34) и для случая равных априорных вероятностей* $q_{-1}=q_1=1/2$ *апостериорные вероятности классов (42) будут иметь вид сигмоидоподобной функции*

$$p(y=1|\mathbf{x},\hat{\mathbf{a}},\hat{b};c)=1-p(y=-1|\mathbf{x},\hat{\mathbf{a}},\hat{b};c)=\begin{cases}\left\{1+\exp(c)\exp\left[-c(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right]\right\}^{-1}, \hat{\mathbf{a}}^T\mathbf{x}+\hat{b}<-1,\\ \left\{1+\exp\left[-2c(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right]\right\}^{-1}, -1\le\hat{\mathbf{a}}^T\mathbf{x}+\hat{b}\le 1,\\ \left\{1+\exp(-c)\exp\left[-c(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right]\right\}^{-1}, \hat{\mathbf{a}}^T\mathbf{x}+\hat{b}>1.\end{cases} \quad (43)$$

Доказательство теоремы приведено в приложении на с. 97.

Заметим, что параметр $c>0$ условных распределений (34), определяющий степень пересечения классов, управляет крутизной сигмоидоподобного обобщения (43) ступенчатой решающей функции $\hat{\mathbf{a}}^T\mathbf{x}+\hat{b}\gtrless 0$.

Соответствующая зависимость апостериорной вероятности (43) от значения решающей функции $d(\mathbf{x}|\hat{\mathbf{a}},\hat{b})=\hat{\mathbf{a}}^T\mathbf{x}+\hat{b}$ показана на рисунке 9.

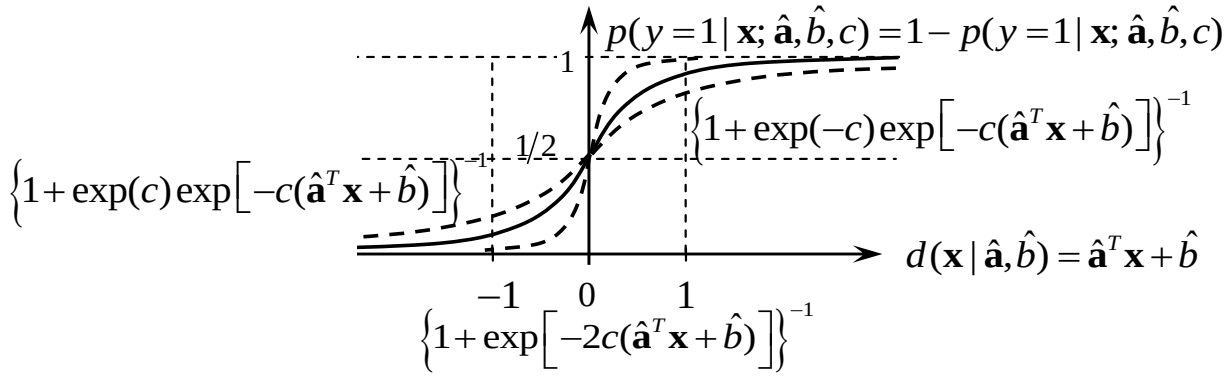


Рисунок 9. Апостериорные вероятности классов для вновь поступившего объекта в вероятностной интерпретации метода опорных векторов.

2.5 Частные априорные модели разделяющей гиперплоскости

2.5.1 Классический метод опорных векторов: Частный случай нормальных априорных распределений компонент направляющего вектора с одинаковыми дисперсиями

Рассмотрим задачу обучения (40) приняв предположение, что случайный направляющий вектор образован независимыми компонентами $\mathbf{a} = (a_1, ..., a_n)$ с нормальными плотностями распределения, имеющими нулевое математическое ожидание и одинаковые дисперсии $r_1 = ... = r_n = r$. Тогда

$$\Psi(\mathbf{a} \mid r) = \prod_{i=1}^{n} \frac{1}{r^{1/2}(2\pi)^{1/2}} \exp\left(-\frac{1}{2r} a_i^2\right) = \frac{1}{r^{n/2}(2\pi)^{n/2}} \exp\left(-\frac{1}{2r} \mathbf{a}^T \mathbf{a}\right), \qquad (44)$$

и задача обучения (40) примет вид классического критерия обучения по методу опорных векторов (9), в котором коэффициент $C = 2rc$.

Заметим, что в диссертации автором впервые предложена формальная интерпретация структурного параметра C в методе опорных векторов как относительной степени пересечения классов по сравнению с общей априорной дисперсией компонент направляющего вектора $C = 2cr$. В такой интерпретации нет смысла рассматривать иные значения априорной дисперсии, нежели $r = 1/2$, тогда $C = c$ определяется степенью «пересечения» областей концентрации распределений объектов двух классов (34).

2.5.2 Известные методы 1-norm SVM (Lasso SVM) и Doubly Regularized SVM (Elastic Net SVM)

Более общий метод Elastic Net SVM получается из задачи обучения (40) при предположении, что случайный направляющий вектор образован независимыми компонентами $\mathbf{a} = (a_1, ..., a_n)$, плотности распределения которых

$$\psi(a_i \mid \beta, \mu) = D \exp\left\{-\left(\beta a_i^2 + \mu \mid a_i \mid\right)\right\}$$

определяются одинаковыми значениями параметров (β,μ). Здесь нормирующая константа определяется обоими параметрами[2]:

$$D=\left[2\sqrt{\frac{\pi}{\beta}}\exp\left\{\frac{\mu^2}{4\beta}\right\}\Phi\left(\frac{\mu}{\sqrt{2\beta}}\right)\right]^{-1},$$

где $\Phi(u)=\frac{1}{\sqrt{2\pi}}\int_u^{\infty}\exp\left\{-\frac{z^2}{2}\right\}dz$ – функция Лапласа.

В этом случае совместная априорная плотность распределения направляющего вектора определяется выражением

$$\Psi(\mathbf{a}|\beta,\mu)=D^n\exp\left\{-\sum_{i=1}^{n}\left(\beta a_i^2+\mu|a_i|\right)\right\},\ \ln\Psi(\mathbf{a}|\beta,\mu)=const-\left(\beta\sum_{i=1}^{n}a_i^2+\mu\sum_{i=1}^{n}|a_i|\right),$$

и предлагаемая в диссертации задача обучения (40) примет вид, эквивалентный методу Doubly Regularized SVM (19):

$$\begin{cases}J_{DrSVM}(\mathbf{a},b,\boldsymbol{\delta}|c,\beta,\mu)=\beta\sum_{i=1}^{n}a_i^2+\mu\sum_{i=1}^{n}|a_i|+c\sum_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\boldsymbol{\delta}),\\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,\ldots,N.\end{cases}\qquad(45)$$

В еще более частном случае, когда $\beta=0$, получается метод обучения 1-norm SVM (Lasso SVM):

$$\begin{cases}J_{1nSVM}(\mathbf{a},b,\boldsymbol{\delta}|c,r)=\sum_{i=1}^{n}|a_i|+c(r/2)^{1/2}\sum_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\boldsymbol{\delta}),\\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,\ldots,N.\end{cases}\qquad(46)$$

Этот критерий эквивалентен (15) при интерпретации независимых компонент случайного направляющего вектора $\mathbf{a}=(a_1,\ldots,a_n)$ как распределенных по закону Лапласа $\psi(a_i|r)=(2r)^{-1/2}\exp\left(-(r/2)^{-1/2}|a_i|\right)$.

В диссертации не обсуждаются соответствующие алгоритмы оптимизации, они детально исследованы в работах [11,12,17]. Заметим лишь, что в силу специфики принятых априорных распределений критерии (45) и (46) обладают чрезвычайно важным свойством бесперебороного отбора признаков непосредственно в ходе обучения. Это свойство является

[2] Это утверждение доказано Е.О. Черноусовой.

результатом склонности этих критериев в точке минимума присваивать нулевые значения большинству компонент направляющего вектора $a_i = 0$, реализуя при этом, вообще говоря, разумный отбор подмножества «полезных» признаков.

Исследование и дальнейшее развитие методологии отбора признаков в предложенных выше вероятностных терминах является одной из задач данной диссертации. Заметим, что целевые функции критериев селективного обучения $J_{DrSVM}(\mathbf{a}, b, \delta_1, ..., \delta_N | c, \beta, \mu)$ (45) и $J_{1nSVM}(\mathbf{a}, b, \delta_1, ..., \delta_N | c, r)$ (46), как и целевая функция классического критерия опорных векторов $J_{SVM}(\mathbf{a}, b, \delta_1, ..., \delta_N | C)$ (9), представляют собой взвешенные суммы эмпирического риска $\sum_{j=1}^{N} \delta_j$, выражающего специфику именно метода опорных векторов (SVM), и той или иной функции регуляризующего штрафа, интерпретируемого в предложенной обобщенной постановке задачи SVM (40) как логарифм априорного распределения направляющего вектора дискриминантной гиперплоскости $-\ln \Psi(\mathbf{a})$.

3 БАЙЕСОВСКИЙ ПРИНЦИП УПРАВЛЕНИЯ СЕЛЕКТИВНОСТЬЮ КОМБИНИРОВАНИЯ ПРИЗНАКОВ

В третьей главе диссертации, в дополнение к существующим методам регуляризации, предложены два новых класса априорных распределений направляющего вектора, наделяющих обобщенный критерий обучения (40) свойством автоматического отбора признаков. Преимущества этих двух новых критериев селективного обучения по сравнению с известными критериями 1-norm SVM (46) и Doubly Regularized SVM (45) исследуются в настоящей диссертации с теоретической точки зрения в конце главы 3 и экспериментально в главе 4.

3.1 Обобщение метода опорных векторов: Частный случай нормальных априорных распределений компонент направляющего вектора с разными известными дисперсиями

В диссертации предлагается естественное обобщение классического метода опорных векторов за счет введения в его вероятностную постановку дополнительного предположения о том, что независимые компоненты направляющего вектора $\mathbf{a} = (a_1,...,a_n)$ распределены по нормальному закону *с разными известными дисперсиями* $r_i \geq 0,\ i = 1,...,n$ и, как и прежде, с нулевыми математическими ожиданиями.

$$\psi(a_i) = \psi(a_i \mid r_i) = \frac{1}{r_i^{1/2}(2\pi)^{1/2}} \exp\left(-(1/2r_i)a_i^2\right),\ i = 1,...,n. \qquad (47)$$

Совместная априорная плотность распределения направляющего вектора определяется выражением

$$\Psi(\mathbf{a} \mid \mathbf{r}) = \Psi(a_1,...,a_n \mid r_1,...,r_n) = \frac{1}{(2\pi)^{n/2}\prod_{i=1}^{n} r_i^{1/2}} \exp\left(-(1/2)\sum_{i=1}^{n}(1/r_i)a_i^2\right). \qquad (48)$$

В этом случае

$$\ln \Psi(\mathbf{a}\,|\,\mathbf{r}) = const - (1/2)\sum_{i=1}^{n} \ln r_i - (1/2)\sum_{i=1}^{n} (1/r_i) a_i^2 \,, \qquad (49)$$

и задача обучения (40) примет вид

$$\begin{cases} J(\mathbf{a}, b, \delta_1, ..., \delta_N | C, \mathbf{r}) = \sum_{i=1}^{n} (1/r_i) a_i^2 + C \sum_{j=1}^{N} \delta_j \to \min(\mathbf{a}, b, \delta_1, ..., \delta_N), \\ y_j (\mathbf{a}^T \mathbf{x}_j + b) \geq 1 - \delta_j, \; \delta_j \geq 0, \; j = 1, ..., N, \end{cases} \qquad (50)$$

где $C = 2c$.

Предложенный в диссертации оптимизационный критерий (50), являясь частным случаем критерия (40), представляет собой, тем не менее, обобщение метода опорных векторов (9) в случае взвешенных компонент направляющего вектора $\sum_{i=1}^{n} (1/r_i) a_i^2$ вместо $\mathbf{a}^T \mathbf{a} = \sum_{i=1}^{n} a_i^2$. В диссертации далее показано, что такая постановка задачи обучения позволяет управлять «степенью участия» отдельных признаков $(x_1, ..., x_n)$ в решающем правиле распознавания вплоть до почти полного подавления некоторых из них - чем меньше априорная дисперсия $r_i > 0$, тем меньше участие i-го признака. Этот факт выражается следующей теоремой о решении задачи (50).

Теорема 3. *В точке минимума критерия (50) значения компонент направляющего вектора* $\hat{\mathbf{a}} = (\hat{a}_1, ..., \hat{a}_n) \in \mathbb{R}^n$ *определяются равенствами*

$$\hat{a}_i = r_i \sum_{j:\lambda_j > 0} y_j \hat{\lambda}_j x_{ij} \in \mathbb{R}, \; i = 1, ..., n, \qquad (51)$$

где $\hat{\lambda}_j \geq 0$ *- множители Лагранжа при ограничениях-неравенствах, определяемые решением двойственной задачи квадратичного программирования*

$$\begin{cases} W(\lambda_1, ..., \lambda_N) = \sum_{j=1}^{n} \lambda_j - \sum_{j=1}^{n} \sum_{l=1}^{n} \lambda_j \lambda_l \left(\sum_{i=1}^{n} r_i x_{ij} x_{il} \right) y_j y_l \to \max(\lambda_1, ..., \lambda_N) \\ \sum_{j=1}^{n} y_j \lambda_j = 0, \; 0 \leq \lambda_j \leq C/2, \; j = 1, ..., N. \end{cases} \qquad (52)$$

Оптимальное значение порога определяется выражением

$$\hat{b} = -\frac{\sum_{j:0<\hat{\lambda}_j<C/2} \hat{\lambda}_j \sum_{l:\hat{\lambda}_l>0} y_l \hat{\lambda}_l \sum_{i=1}^{n} r_i x_{il} x_{ij} + (C/2) \sum_{j:\hat{\lambda}_j = C/2} y_j}{\sum_{j:0<\hat{\lambda}_j<C/2} \hat{\lambda}_j} \,. \qquad (53)$$

Доказательство теоремы приведено в приложении на с. 98.

Из утверждения (51) этой теоремы хорошо видно, что «масштаб» i-й компоненты направляющего вектора напрямую определяется значением соответствующей априорной дисперсии r_i. Соответственно, решающее правило распознавания, полученное в результате обучения и применимое к новому объекту с вектором признаков $\mathbf{x}=(x_1,...,x_n)$, будет зависеть главным образом от признаков с большими значениями r_i и почти игнорировать остальные признаки:

$$d(\mathbf{x}\,|\,\hat{\lambda}_1,...,\hat{\lambda}_N,r_1,...,r_n,\hat{b})=\sum_{j:\hat{\lambda}_j>0} y_j\hat{\lambda}_j\sum_{i=1}^{n} r_i x_{ij}x_i+\hat{b}\gtrless 0\,. \qquad (54)$$

Структура решающего правила (54), полученного по методу опорных векторов со взвешенными компонентами направляющего вектора (50), дает инструмент управления участием признаков $\mathbf{x}=(x_1,...,x_n)$ в процессе обучения. Чем больше принятое значение априорной дисперсии r_i некоторой компоненты направляющего вектора относительно дисперсий других компонент, тем большее влияние i-й признак оказывает на решение о классе объекта.

Это обстоятельство будет далее использовано для построения версии критерия обучения, *метода релевантных признаков с регулируемой селективностью,* способного автоматически оценивать дисперсии компонент направляющего вектора по обучающей совокупности при некоторых дополнительных априорных предположениях об этих дисперсиях, что позволит выделять подмножество наиболее «полезных» признаков, подавляя «лишние».

3.2 Метод релевантных признаков с регулируемой селективностью

3.2.1 Параметрическое семейство априорных нормальных-гамма распределений компонент направляющего вектора со случайными дисперсиями

Предлагаемый в диссертации метод основан на предположении, что дисперсии $(r_1,...,r_n)$ компонент направляющего вектора $\mathbf{a}=(a_1,...,a_n)$ (47) являются независимыми случайными величинами, характеризующимися некоторым априорным распределением и подлежащими оцениванию в процессе обучения по байесовскому принципу вместе с параметрами разделяющей гиперплоскости $(\mathbf{a},b)$. В литературе подобные дополнительные априорные предположения о параметрах, участвующих в некотором ранее принятом априорном предположении, принято называть *Hyper Prior Assumptions*.

Удобнее оперировать не дисперсиями r_i, а обратными к ним величинами $1/r_i$, называемыми *мерами точности* (precision measures) соответствующих случайных величин [16], предполагая, что обратные дисперсии имеют одно и то же априорное гамма-распределение:

$$\gamma(1/r\,|\,\alpha,\beta)=\frac{\beta^{\alpha}(1/r)^{\alpha-1}\exp\left(-\beta(1/r)\right)}{\Gamma(\alpha)}\propto(1/r)^{\alpha-1}\exp\left(-\beta(1/r)\right) \qquad (55)$$

здесь $\alpha>0$ и $\beta>0$ – параметры гамма-распределения, вопрос о выборе которых будет рассмотрен ниже в параграфе 3.2.3.

По-прежнему будем рассматривать параметрическое семейство нормальных совместных условных распределений параметров гиперплоскости относительно заданных дисперсий элементов направляющего вектора $\mathbf{r}=(r_1,...,r_n)^T$ (48):

$$\Psi\big((\mathbf{a},b)\,|\,\mathbf{r}\big)\propto\Psi(\mathbf{a}\,|\,\mathbf{r})=\Psi(a_1,...,a_n\,|\,r_1,...,r_n)=$$
$$\prod_{i=1}^{n}\mathcal{N}(a_i\,|\,0,r_i)=\Big(\prod_{i=1}^{n}r_i\Big)^{-1/2}(2\pi)^{-n/2}\exp\Big(-\frac{1}{2}\sum_{i=1}^{n}(1/r_i)\,a_i^2\Big). \qquad (56)$$

Тогда, вместе с априорным распределением независимых дисперсий

$$G(\mathbf{r}\,|\,\alpha,\beta)=G(1/r_1,...,1/r_n\,|\,\alpha,\beta)=\prod_{i=1}^{n}\gamma(1/r_i\,|\,\alpha,\beta)=$$
$$\Big(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\Big)^{n}\Big(\prod_{i=1}^{n}(1/r_i)\Big)^{\alpha-1}\exp\Big(-\beta\sum_{i=1}^{n}(1/r_i)\Big), \qquad (57)$$

априорное распределение совокупности параметров $(\mathbf{a},b,\mathbf{r})$ является произведением нормальных-гамма распределений [16]:

$$\bar{\Psi}\big((\mathbf{a},b,\mathbf{r})\,|\,\alpha,\beta\big)\propto\bar{\Psi}\big((\mathbf{a},\mathbf{r})\,|\,\alpha,\beta\big)=\Psi(\mathbf{a}\,|\,\mathbf{r})G(\mathbf{r}\,|\,\alpha,\beta)=$$
$$\prod_{i=1}^{n}\mathcal{N}(a_i\,|\,0,r_i)\gamma(1/r_i\,|\,\alpha,\beta)\propto\exp\Big\{-\frac{1}{2}\sum_{i=1}^{n}\Big[(1/r_i)(a_i^2+2\beta)+(2\alpha-1)\ln r_i\Big]\Big\}.$$

Как и прежде, будем исходить из условной плотности совместного распределения случайной обучающей совокупности относительно скрытых параметров $(\mathbf{a},b)$ (36)

$$\Phi(X\,|\,Y;\mathbf{a},b,c)=\prod_{j=1}^{N}\varphi(\mathbf{x}_j\,|\,y_j;\mathbf{a},b,c)\,,$$

где параметр априорной разделимости классов c в (34) полагается заданным. Тогда по формуле Байеса, аналогично (38), апостериорное совместное распределение подлежащих оцениванию параметров модели $(\mathbf{a},b,\mathbf{r})$ относительно обучающей совокупности примет вид:

$$\bar{P}\big((\mathbf{a},b,\mathbf{r})\,|\,X,Y,\alpha,\beta,c\big)=\frac{\bar{\Psi}\big((\mathbf{a},\mathbf{r})|\alpha,\beta\big)\Phi\big(X\,|Y,(\mathbf{a},b,c)\big)}{F(X\,|\,Y,c)}\propto$$
$$\propto\bar{\Psi}\big((\mathbf{a},\mathbf{r})|\alpha,\beta\big)\Phi\big(X\,|Y,(\mathbf{a},b),c\big). \qquad (58)$$

Отождествляя обучение, как и прежде, с вычислением максимума апостериорной плотности параметров, мы придем к следующему критерию

$$(\hat{\mathbf{a}},\hat{b},\hat{\mathbf{r}}\,|\,X,Y;\alpha,\beta,c)=\underset{\mathbf{a}\in\mathbb{R}^n,\,b\in\mathbb{R},\,\mathbf{r}\in\mathbb{R}}{\arg\max}\;\bar{P}\big((\mathbf{a},b,\mathbf{r})\,|\,X,Y,\alpha,\beta,c\big)=$$
$$\underset{\mathbf{a}\in\mathbb{R}^n,\,b\in\mathbb{R},\,\mathbf{r}\in\mathbb{R}}{\arg\max}\Big[\ln\Psi(\mathbf{a}\,|\,\mathbf{r})+\ln G(\mathbf{r}\,|\,\alpha,\beta)+\sum_{j=1}^{N}\ln\varphi(\mathbf{x}_j\,|\,y_j;\mathbf{a},b,c)\Big], \qquad (59)$$

являющийся обобщением критерия обучения (39).

3.2.2 Критерий обучения

Теорема 4. *Оптимизационная задача обучения (59) для (34), (49) и (57) эквивалентная критерию*

$$\begin{cases} \sum_{i=1}^{n}\left[(1/r_i)\left(a_i^2+2\beta\right)+\left(2\alpha-1\right)\ln r_i\right]+C\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}), \\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\geq 1-\delta_j,\ \delta_j\geq 0,\ j=1,\ldots,N. \end{cases} \qquad (60)$$

Доказательство теоремы приведено в приложении на с. 100.

Ключевая идея предлагаемого принципа обучения (60) заключается в том, что при подходящем выборе параметров (α,β), алгоритм демонстрирует выраженную способность подавлять неинформативные признаки выбором маленьких, но не нулевых, значений весов r_i в разделяющей гиперплоскости. Остальные признаки с бóльшими весами r_i предполагаются наиболее информативными (*relevance features*) для данной обучающей совокупности.

3.2.3 Параметры гамма распределения: Управление селективностью

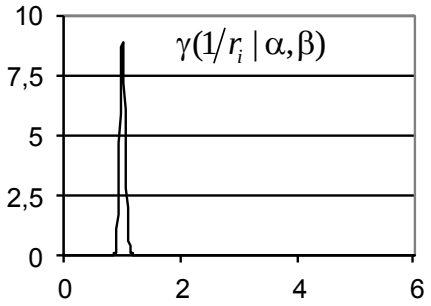
Прежде чем говорить об алгоритме решения задачи оптимизации (60), исследуем вопрос, как значения параметров α и β влияют на вид априорного гамма-распределения обратных дисперсий компонент направляющего вектора $\gamma(1/r_i\,|\,\alpha,\beta)\propto (1/r_i)^{\alpha-1}\exp\left(-\beta(1/r_i)\right)$ (55).

Известно, что математическое ожидание случайной величины, распределенной по гамма закону, равно отношению параметров $m(\alpha,\beta)=E(1/r_i\,|\,\alpha,\beta)=\alpha/\beta$, а дисперсия определяется выражением $\sigma^2(\alpha,\beta)=Var(1/r_i\,|\,\alpha,\beta)=\alpha/\beta^2$. Будем рассматривать также отношение среднеквадратичного отклонения к математическому ожиданию $\left(\sigma/m\,|\,\alpha,\beta\right)=1/\sqrt{\alpha}$.

Если $\left(\sigma/m\,|\,\alpha,\beta\right)\to 0$, априорные гамма распределения всех дисперсий r_i сконцентрированы возле общего математического ожидания α/β (рис. 10-а).

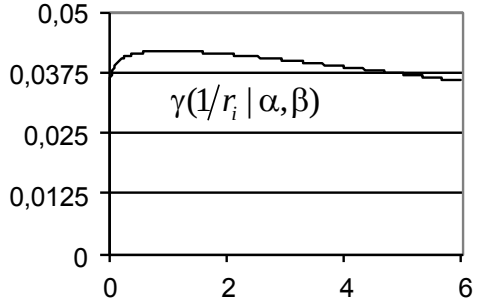
В этом случае оцененные дисперсии практически фиксированы априори и равны единице при примерно равных значениях обоих параметров $\alpha \cong \beta$. При таких значениях параметров критерий (60) эквивалентен классическому критерию опорных векторов (9), использующему все признаки объектов.

Если же $(\sigma/m \,|\, \alpha,\beta) \to 1$, то априорные распределения становятся практически равномерными (рис. 10-б). При $r_i \to 0$ соответствующее слагаемое целевой функции в (60) неограниченно уменьшается $\ln r_i \to -\infty$, и критерию выгодно уменьшать все дисперсии. Однако в этом случае невозможно выполнить ограничения, предписывающие достаточно хорошо аппроксимировать обучающую совокупность. В результате этого противоречия критерий проявляет ярко выраженную склонность к чрезмерной селективности отбора признаков, подавляя большинство из них, даже полезные.



$\alpha = 501,\ \beta = 500,\ \sigma/m = 0.045$

(а)



$\alpha = 1.05,\ \beta = 0.05,\ \sigma/m = 0.976$

(б)

Рисунок 10. Вид гамма-распределения при малом (а) и большом (б) отношении σ/m.

Управлять степенью селективности отбора признаков можно, варьируя значения параметров α и β в априорном распределении дисперсий. Будем выбирать значения двух параметров априорного гамма распределения α и β, задавая один общий параметр $0 < \mu < \infty$:

$$\alpha = (1+\mu)^2/2\mu, \quad \beta = 1/2\mu, \tag{61}$$

В этом случае параметрическое семейство гамма распределений (55) будет иметь вид

$$\gamma(1/r_i \mid \mu) \propto (1/r_i)^{(1+\mu^2)/2\mu} \exp\left[-(1/2\mu)(1/r_i)\right],\ i=1,...,n. \qquad (62)$$

Нетрудно убедиться, что

$$\begin{cases} \alpha/\beta = m(\alpha,\beta) \to 1, \\ \alpha/\beta^2 = \sigma^2(\alpha,\beta) \to 0, \\ 1/\sqrt{\alpha} = \sigma(\alpha,\beta)/m(\alpha,\beta) \to 0, \end{cases} \quad \text{при } \mu \to 0,$$

а также

$$\begin{cases} \alpha/\beta = m(\alpha,\beta) \to \infty, \\ \alpha/\beta^2 = \sigma^2(\alpha,\beta) \to \infty, \\ 1/\sqrt{\alpha} = \sigma(\alpha,\beta)/m(\alpha,\beta) \to 1, \end{cases} \quad \text{при } \mu \to \infty.$$

При увеличении μ от нуля до достаточно больших значений вид гамма-распределения плавно изменяется от сконцентрированного в окрестности $1/r_i = \alpha/\beta = 1$ до почти «равномерного» на неотрицательной полуоси (рис. 11).

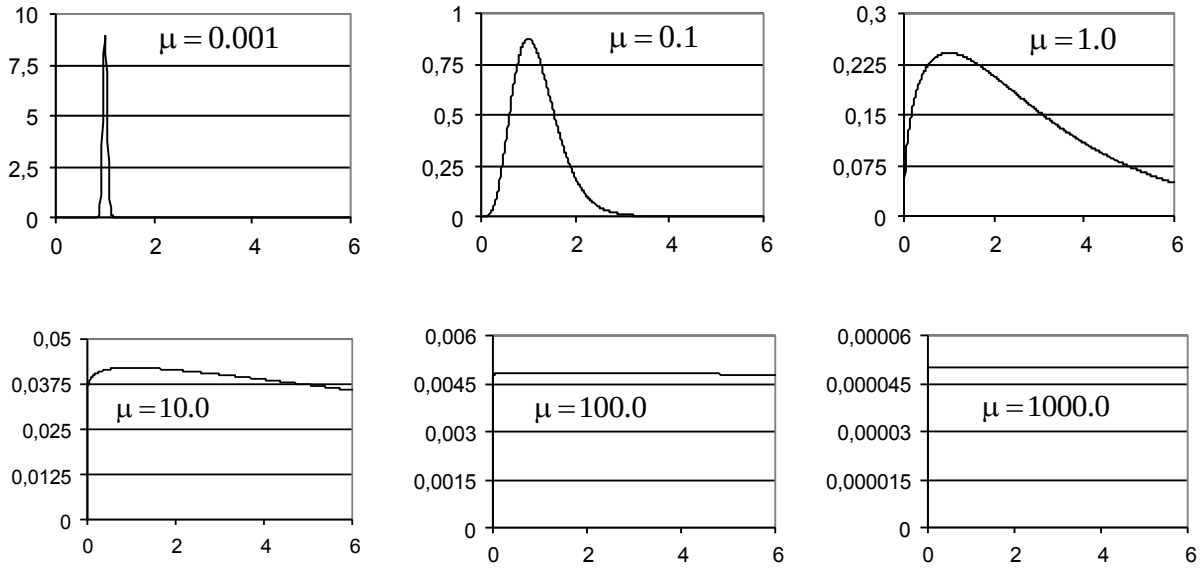


Рисунок 11. Зависимость априорного гамма-распределения дисперсий компонент направляющего вектора от параметра μ.

Обозначим через $G(\mathbf{r} \mid \mu) = G(r_1,...,r_n \mid \mu)$ априорную совместную плотность распределения обратных дисперсий:

$$G(\mathbf{r}|\mu)=G(r_1,...,r_n|\mu)=\prod_{i=1}^{n}\gamma\left((1/r_i)|\mu\right)\propto\left(\prod_{i=1}^{n}\frac{1}{r_i}\right)^{\frac{1+\mu^2}{2\mu}}\exp\left(-(1/2\mu)\sum_{i=1}^{n}(1/r_i)\right) \quad (63)$$

Тогда совместная априорная плотность распределения параметров разделяющей гиперплоскости (48) и обратных дисперсий компонент направляющего вектора (63) запишется как

$$\bar{\Psi}(\mathbf{a},b,\mathbf{r}\,|\,\mu)=\Psi(\mathbf{a}\,|\,\mathbf{r})G(\mathbf{r}\,|\,\mu).$$

Соответственно, совместное апостериорное распределение всех переменных модели относительно обучающей совокупности, аналогично (58), пропорционально произведению

$$\bar{P}(\mathbf{a},b,\mathbf{r}\,|\,X,Y;c,\mu)\propto\Psi(\mathbf{a}\,|\,\mathbf{r})G(\mathbf{r}\,|\,\mu)\Phi(X\,|\,Y,\mathbf{a},b,c).$$

Отождествляя обучение, как и прежде, с вычислением байесовской оценки неизвестных параметров, мы придем к следующему критерию

$$(\hat{\mathbf{a}},\hat{b},\hat{\mathbf{r}}|\,X,Y;\mu,c)=\underset{\mathbf{a}\in\mathbb{R}^n,\,b\in\mathbb{R},\,\mathbf{r}\in\mathbb{R}^n}{\arg\max}\;\bar{P}\left(\mathbf{a},b,\mathbf{r}\,|\,X,Y,\mu,c\right)=$$
$$\underset{\mathbf{a}\in\mathbb{R}^n,\,b\in\mathbb{R},\,\mathbf{r}\in\mathbb{R}^n}{\arg\max}\left[\ln\Psi(\mathbf{a}|\mathbf{r})+\ln G(\mathbf{r}|\mu)+\sum_{j=1}^{N}\ln\varphi(\mathbf{x}_j|y_j;\mathbf{a},b,c)\right], \quad (64)$$

представляющему собой обобщение критерия (39).

Теорема 5. *Оптимизационная задача обучения (64) для (48), (62) и (63) эквивалентна критерию*

$$\begin{cases} J_{RFM}(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}|C,\mu)= \\ \qquad\sum_{i=1}^{n}\left[(1/r_i)\left(a_i^2+(1/\mu)\right)+\left((1/\mu)+1+\mu\right)\ln r_i\right]+C\sum_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}), \\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\geq 1-\delta_j,\ \delta_j\geq 0,\ j=1,...,N. \end{cases} \quad (65)$$

где $C=2c$.

Доказательство теоремы приведено в приложении на с. 101.

Этот критерий плавно изменяет степень своей склонности к подавлению «лишних» признаков с ростом значения параметра $0<\mu<\infty$, поэтому такой

параметр уместно называть *параметром селективности* признаков в процессе обучения распознаванию образов.

При нулевом значении параметра $\mu \to 0$ и, соответственно, $1/\mu \to \infty$, штрафы $(1/\mu)\left[(1/r_i) + \ln r_i\right]$ настолько велики, что все оценки дисперсий равны единицы $r_1 = ... = r_n = 1$ в точке минимума критерия (65), который становится эквивалентным критерию классического метода опорных векторов, вообще не обладающего свойством селективности.

При неограниченном увеличении параметра $\mu \to \infty$ априорное гамма распределение обратной дисперсии каждой компоненты направляющего вектора $1/r_i$ (62) и их совместное априорное распределение (63) становятся практически «равномерными», соответственно, на положительной полуоси $0 \le 1/r_i < \infty$ (рис. 11) и в положительном квадранте пространства $\mathbb{R}^n$. Это в пределе приводит к исчезновению слагаемого $\ln G(\mathbf{r} \mid \mu)$ в критерии (64) и слагаемого $(1/\mu)\left[(1/r_i) + \ln r_i\right]$ в целевой функции критерия (65), штрафующего отклонение r_i от единицы, и критерий становится чрезмерно селективным.

Идея погружения дискретной задачи выбора подмножества в непрерывную задачу поиска в некотором смысле оптимальных неотрицательных весов, присваиваемых элементам исходного множества, лежит в основе *метода релевантных векторов* Бишопа и Типпинга [18], в которой направляющий вектор разделяющей гиперплоскости строится как взвешенная линейная комбинация векторов признаков всех объектов в обучающей совокупности. Векторы признаков объектов, получившие существенно ненулевые веса, названы Бишопом и Типпингом *релевантными векторами* в отличие от *опорных векторов* в методе опорных векторов, образующих жестко выделенное подмножество. В отличие от [18], этот прием использован в диссертации для «мягкого» отбора признаков, а не объектов обучающей совокупности, участвующих в итоговом решающем правиле

распознавания. Главным же отличием метода, предлагаемого в диссертации, является принципиально новое понятие параметра селективности μ.

В связи с этим обстоятельством данный метод обучения назван *методом релевантных признаков с регулируемой селективностью*, или, полностью характеризуя метод, *Selective Relevance Feature Support Vector Machine (RFM)*.

3.2.4 Эквивалентный критерий обучения

Критерий обучения метода релевантных признаков (65) может быть записан в эквивалентной форме, который не содержит параметров оптимизации $(r_1,...,r_n)$. Такая запись критерия будет удобна для последующего анализа свойств метода. Справедлива следующая теорема.

Теорема 6. *Критерий обучения метода релевантных признаков (65) имеет эквивалентную формулировку*

$$\begin{cases} J_{RFM}(\mathbf{a},b,\boldsymbol{\delta}\,|\,C',\mu)=\sum_{i=1}^{n}\ln(a_i^2+1/\mu)+C'\sum_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\boldsymbol{\delta}), \\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,...,N, \end{cases} \qquad (66)$$

где структурный параметр $C'\ge 0$ *связан с аналогичным параметром критерия (65) выражением*

$$C'=C/\big((1/\mu)+1+\mu\big).$$

Доказательство теоремы приведено в приложении на с. 102.

3.2.5 Алгоритм обучения

Критерий обучения (65) несложно алгоритмически реализовать, применив метод Гаусса-Зайделя с поочередной оптимизацией по двум группам переменных $(a_1,...,a_n,b,\delta_1,,...,\delta_N)$ и $(r_1,...,r_n)$. При фиксированных $(r_1,...,r_n)$, оптимизация критерия (65) только по $(a_1,...,a_n,b,\delta_1,...,\delta_N)$, что эквивалентно критерию (50), сводится к решению двойственной задачи квадратичного программирования (52). Естественно принять начальные значения $(r_i^0=1, i=1,...,n)$, что соответствует стандартному методу опорных векторов.

После того, как на очередном шаге итерационного процесса $k = 0,1,2,...$ найдено решение $(\lambda_1^k,...,\lambda_N^k)$ двойственной задачи для текущего приближения $(r_1^k,...,r_n^k)$, новые значения дисперсий $(r_1^{k+1},...,r_n^{k+1})$ на следующем шаге определяются следующей теоремой.

Теорема 7. *Новые значения дисперсий* $(r_1^{k+1},...,r_n^{k+1})$ *определяются решением* $(\lambda_1^k,...,\lambda_N^k)$ *двойственной задачи (52) для текущего приближения* $(r_1^k,...,r_n^k)$ *согласно выражению*

$$r_i^{k+1} = \frac{(r_i^k)^2 \sum_{j:\lambda_j^k>0} \sum_{l:\hat{\lambda}_l^k>0} y_j y_l x_{ij} x_{il} \lambda_j^k \lambda_l^k + 1/\mu}{\mu + 1 + 1/\mu}, \; i = 1,...,n. \qquad (67)$$

Доказательство теоремы приведено в приложении на с. 104.

Итерационная процедура обучения обычно сходится за 10-15 шагов, а сам алгоритм демонстрирует способность подавлять неинформативные признаки за счет очень маленьких весов r_i в решающем правиле (54). Оптимальное значение порога b^k определяется, как и прежде, выражением (53).

В частности, из формулы (67) хорошо видно, что при нулевой селективности $\mu \to 0$ оценки всех дисперсий тождественно равны единице $r_i^{k+1} \to 1$, и критерий (65) вырождается в обычный критерий опорных векторов, сохраняющий все признаки. При $\mu \to \infty$ алгоритм подавляет все признаки $r_i^{k+1} \to 0$.

Что же касается выбора наиболее подходящего значения параметра селективности μ, то это выбор нельзя сделать из байесовских соображений, поскольку всегда максимальное значение совместного апостериорного распределения подлежащих оцениванию параметров модели данных (58) будет наибольшим при самой «широкой» модели, т.е. при нулевой селективности. Значение параметра μ, как и любого структурного параметра модели, следует оценивать, используя внешней критерий оценивания обобщающей способности, например, тот либо иной критерий кросс-валидации на основе деления обучающей совокупности на основную и контрольную подвыборки.

3.3 Метод опорных признаков с регулируемой селективностью

3.3.1 Параметрическое семейство составных априорных распределений для компонент направляющего вектора

В разделе 3.2.1 в качестве априорного суждения о предпочитаемых значениях направляющего вектора $\mathbf{a}=(a_1,...,a_n)$ было принято, что компоненты направляющего вектора независимы и нормально распределены с некоторым дополнительным суждением о параметрах такого распределения. В этом разделе будет рассматриваться другое предположение об априорном распределении $\Psi(\mathbf{a})$ независимых компонент направляющего вектора $\mathbf{a}\in\mathbb{R}^n$.

Вид априорной плотности $\Psi(\mathbf{a})$ представляет собой комбинацию распределения Лапласа при значениях нормы компонент, не превышающих заданного порога $|a_i|\le\mu$, и нормального распределения при бóльших значениях $|a_i|>\mu$ (рис. 12):

$$\psi(a_i\,|\,\mu)\propto\exp\bigl(-q(a_i\,|\,\mu)\bigr), \tag{68}$$

$$q(a_i|\mu)=\begin{cases}2\mu|\,a_i\,|, \textit{если}\,|\,a_i\,|\le\mu,\\ \mu^2+a_i^2, \textit{если}\,|\,a_i\,|>\mu,\end{cases} \tag{69}$$

где $\mu\ge 0$ - неотрицательный параметр.

В этом случае совместное априорное распределение направляющего вектора $\mathbf{a}=(a_1,...,a_n)$ определяется выражением

$$\Psi(\mathbf{a}|\mu)=\Psi(a_1,...,a_n|\mu)=\prod_{i=1}^{n}\psi(a_i\,|\,\mu)\propto\exp\left[-\sum_{i=1}^{n}q(a_i|\mu)\right]. \tag{70}$$

Запишем совместное априорное распределение в эквивалентном виде

$$\Psi(\mathbf{a}|\mu)=\Psi(a_1,...,a_n|\mu)\propto\exp\left[-2\mu\sum_{|a_i|\le\mu}|a_i|-\sum_{|a_i|>\mu}(\mu^2+a_i^2)\right].$$

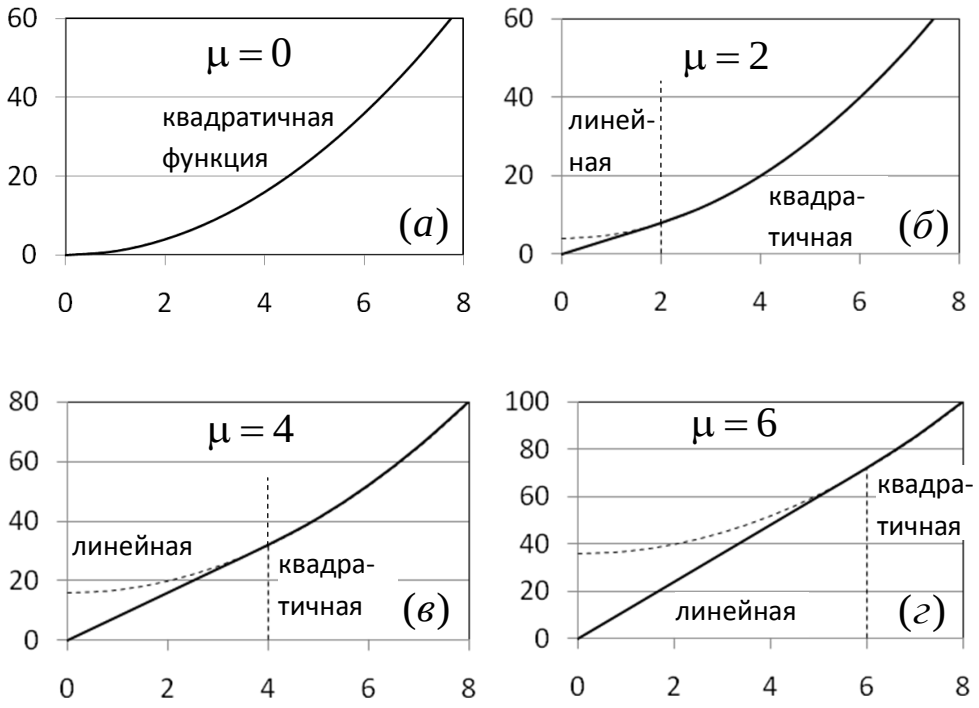


Рисунок 12. Зависимость функции $q(a_i|\mu)$ априорного распределения компонент направляющего вектора $\psi(a_i|\mu)$ от параметра μ.

Логарифм такого распределения имеет вид

$$\ln\Psi(\mathbf{a}|\mu)=\ln\Psi(a_1,...,a_n|\mu)=const-2\mu\sum_{|a_i|\le\mu}|a_i|-\sum_{|a_i|>\mu}(\mu^2+a_i^2). \qquad (71)$$

3.3.2 Критерий обучения

Вид совместного априорного распределения (70) и обобщенный критерий обучения метода опорных векторов (40) полностью определяют вид оптимизационной задачи обучения

$$\begin{cases}\sum_{i=1}^{n}q(a_i|\mu)+C\sum_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\boldsymbol{\delta}),\\ q(a_i|\mu)=\begin{cases}2\mu|a_i|, если\ |a_i|\le\mu,\\ \mu^2+a_i^2, если\ |a_i|>\mu,\end{cases}\\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge1-\delta_j,\ \delta_j\ge0,\ j=1,...,N.\end{cases} \qquad (72)$$

Для логарифма априорного распределения (71) критерий (40) примет эквивалентный вид

$$\begin{cases} J_{SFM}(\mathbf{a},b,\boldsymbol{\delta}|C,\mu)=2\mu\sum\limits_{|a_i|\le\mu}|a_i|+\sum\limits_{|a_i|>\mu}(\mu^2+a_i^2)+C\sum\limits_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\boldsymbol{\delta}), \\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,...,N. \end{cases} \quad (73)$$

Величина порога $0\le\mu<\infty$ плотности (68) выполняет роль параметра селективности обучения и при $\mu=0\Rightarrow q(a_i\,|\,\mu)=a_i^2$ оптимизационный критерий (72) эквивалентен критерию метода опорных векторов с минимальной способностью к отбору признаков, а при достаточно больших значениях селективности $\mu\gg 0\Rightarrow q(a_i\,|\,\mu)=\mu\,|\,a_i\,|$ - критерию (46) методу 1-norm SVM, с увеличивающейся селективностью обучения по мере увеличения μ относительно параметра C до полного подавления всех признаков.

3.3.3 Двойственная задача обучения

Для произвольной обучающей выборки $\{(x_{ij},i\in I),y_j,\ j=1,...,N\}$, где $I=\{1,...,n\}$ - множество индексов признаков объектов, задача (72) является задачей выпуклого программирования в пространстве $\mathbb{R}^n\times\mathbb{R}^{N+1}$, поскольку каждая из функций $q(a_i|\mu)$ выпукла в $\mathbb{R}$, а система ограничений образует выпуклый многогранник.

Обозначим через $\lambda_j\ge 0$ и $\pi_j\ge 0$ множители Лагранжа при ограничениях задачи (72), соответственно, $y_j(\mathbf{a}^T\mathbf{x}_j+b)\le 1-\delta_j$ и $\delta_j\ge 0$. Можно показать, что задача (72) является регулярной [19], и, как следствие, она эквивалентна задаче поиска седловой точки соответствующей функции Лагранжа

$$\begin{aligned} &L(a_i,i\in I,b,\delta_j,\lambda_j,\pi_j,\ j=1,...,N\,|\,\mu,C)=\frac{1}{2}\sum_{i\in I}q(a_i\,|\,\mu)+\frac{C}{2}\sum_{j=1}^{N}\delta_j- \\ &\sum_{j=1}^{N}\lambda_j\left[y_j\left(\sum_{i\in I}a_ix_{ij}+b\right)-1+\delta_j\right]-\sum_{j=1}^{N}\pi_j\delta_j\begin{cases}\to\min(a_i,i\in I,b,\delta_1,...,\delta_N),\\ \to\max(\lambda_j\ge 0,\pi_j\ge 0,\ j=1,...,N).\end{cases} \end{aligned} \quad (74)$$

Если $[(\tilde{a}_i,i\in I,\tilde{b},\tilde{\delta}_j,\ j=1,...,N),(\tilde{\lambda}_j,\tilde{\pi}_j,\ j=1,...,N)]$ есть седловая точка для (74), то ее левая часть $(\tilde{a}_i,i\in I,\tilde{b},\tilde{\delta}_j,\ j=1,...,N)$ является решением задачи (72) и наоборот.

Запишем функцию Лагранжа (74) в эквивалентном виде, как сумму функций, зависящих только от своих переменных

$$L(a_i, i \in I, b, \delta_j, \lambda_j, \pi_j, j = 1,..., N \mid \mu, C) = \sum_{i \in I} \left[\frac{1}{2} q(a_i \mid \mu) - a_i \sum_{j=1}^{N} y_j x_{ij} \lambda_j \right] + \\ + \sum_{j=1}^{N} \left(\frac{C}{2} - \lambda_j - \pi_j \right) \delta_j - b \sum_{j=1}^{N} \lambda_j y_j + \sum_{j=1}^{N} \lambda_j. \quad (75)$$

Ведем обозначение

$$L_i(a_i, \boldsymbol{\lambda} \mid \mu) = L_i(a_i, \lambda_1,..., \lambda_N \mid \mu) = \frac{1}{2} q(a_i \mid \mu) - a_i \sum_{j=1}^{N} y_j x_{ij} \lambda_j = \\ \frac{1}{2} \begin{cases} 2\mu \mid a_i \mid, \textit{если} \mid a_i \mid \le \mu \\ \mu^2 + a_i^2, \textit{если} \mid a_i \mid > \mu \end{cases} - a_i \sum_{j=1}^{N} y_j x_{ij} \lambda_j. \quad (76)$$

Функция Лагранжа примет вид

$$L(a_i, i \in I, b, \delta_j, \lambda_j, \pi_j, j = 1,..., N \mid \mu, C) = \sum_{i \in I} L_i(a_i, \lambda_1,..., \lambda_N \mid \mu) + \\ + \sum_{j=1}^{N} \left(\frac{C}{2} - \lambda_j - \pi_j \right) \delta_j - b \sum_{j=1}^{N} \lambda_j y_j + \sum_{j=1}^{N} \lambda_j. \quad (77)$$

Для того чтобы сформулировать двойственную задачу необходимо определить, как минимальное значение $L_i(a_i, \boldsymbol{\lambda} \mid \mu)$ по переменной $a_i \in \mathbb{R}$ зависит от множителей Лагранжа λ_j для каждого $i \in I$. Поскольку эти функции содержат недифференцируемые (выпуклые) функции $q(a_i \mid \mu)$, то это приводит к необходимости воспользоваться понятиями суградиент и субдифференциал вместо обычного градиента для того, чтобы определить условия минимума выпуклой функции [19].

Определение 1. *Вектор* $d \in \mathbb{R}$ *называется субградиентом выпуклой функции* $f : \mathbb{R} \to \mathbb{R}$ *в точке* $a \in \mathbb{R}$, *если соотношение* $f(a') - f(a) \ge d(a' - a)$ *выполнено для любой точки* $a' \in \mathbb{R}$.

Определение 2. *Субдифференциалом* $\partial f(a)$ *выпуклой функции* $f : \mathbb{R} \to \mathbb{R}$ *в точке* $a \in \mathbb{R}$ *называется множество всех субградиентов в данной точке.*

Лемма 1. *Субдифференциал функции* $L_i(a_i,\boldsymbol{\lambda}\,|\,\mu)$ *в точке* $a_i\in\mathbb{R}$ *состоит из:*

1. *Единственного субградиента* $\partial_{a_i}L_i(a_i,\boldsymbol{\lambda}|\mu)=a_i-\sum_{j=1}^{N}y_jx_{ij}\lambda_j$, *если* $|a_i|>\mu$.

2. *Единственного субградиента* $\partial_{a_i}L_i(a_i,\boldsymbol{\lambda}\,|\,\mu)=2\mu\left(a_i/|a_i|\right)-\sum_{j=1}^{N}y_jx_{ij}\lambda_j$, *если* $0<|a_i|\le\mu$.

3. *Множества субградиентов* $\partial_{a_i}L_i(a_i,\boldsymbol{\lambda}\,|\,\mu)=\left\{\mu d_i-\sum_{j=1}^{N}\lambda_jy_jx_{ij}:|d_i|\le1\right\}$, *если* $a_i=0$.

Доказательство леммы приведено на с. 105.

Лемма 2. *Минимум каждой функции* $L_i(a_i,\boldsymbol{\lambda}\,|\,\mu)$ *по переменной* $a_i\in\mathbb{R}$ *достигается в точках*

$$\begin{cases} a_i=\sum_{j:\lambda_j>0}y_j\lambda_jx_{ij}, & \text{если}\ \ \sum_{j=1}^{N}\sum_{l=1}^{N}y_jy_lx_{ij}x_{il}\lambda_j\lambda_l>\mu^2,\\ a_i=\eta_i\sum_{j:\lambda_j>0}y_j\lambda_jx_{ij},\ \eta_i\in[0,1], & \text{если}\ \ \sum_{j=1}^{N}\sum_{l=1}^{N}y_jy_lx_{ij}x_{il}\lambda_j\lambda_l=\mu^2,\\ a_i=0, & \text{если}\ \ \sum_{j=1}^{N}\sum_{l=1}^{N}y_jy_lx_{ij}x_{il}\lambda_j\lambda_l<\mu^2. \end{cases} \qquad (78)$$

В каждой такой точке

$$\min_{a_i\in\mathbb{R}}L_i(a_i,\boldsymbol{\lambda}\,|\,\mu)=\frac{1}{2}\min\left\{0,\ \mu^2-\sum_{j=1}^{N}\sum_{l=1}^{N}y_jy_lx_{ij}x_{il}\lambda_j\lambda_l\right\}. \qquad (79)$$

Доказательство леммы приведено на с. 105.

Справедлива следующая теорема о двойственной задаче.

Теорема 8. *Оптимизационная задача квадратичного программирования*

$$\begin{cases} W(\boldsymbol{\lambda}|\mu,C)=\sum_{i\in I}\frac{1}{2}\min\left\{0,\mu^2-\sum_{j=1}^{N}\sum_{l=1}^{N}y_jy_lx_{ij}x_{il}\lambda_j\lambda_l\right\}+\sum_{j=1}^{N}\lambda_j\to\max(\boldsymbol{\lambda}),\\ \sum_{j=1}^{N}y_j\lambda_j=0,\ 0\le\lambda_j\le C/2,\ j=1,\ldots,N. \end{cases} \qquad (80)$$

является двойственной для задачи (72).

Задача (80) может быть записана в эквивалентном виде

$$\begin{cases} L(\boldsymbol{\lambda} \mid \mu, C) = \sum_{j=1}^{N} \lambda_j + (1/2) \sum_{i \in I} \xi_i \to \max(\boldsymbol{\lambda}), \\ \xi_i \geq \sum_{j=1}^{N} \sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l - \mu^2, \ \xi_i \geq 0, \ i \in I = \{1,...,n\}, \\ \sum_{j=1}^{N} y_j \lambda_j = 0, \ 0 \leq \lambda_j \leq \dfrac{C}{2}, \ j = 1,...,N. \end{cases} \quad (81)$$

Доказательство теоремы приведено на с. 106.

Задача (81) относится к более широкому классу задач, который в литературе принято называть задачами квадратичного программирования с квадратичными ограничениями (QCQP) [20]. Для решения задач, подобных двойственной задаче (80), существует множество стандартных алгоритмов и программных решений, способных эффективно находить численное решение. Подобные алгоритмы в диссертации рассматриваться не будут.

3.3.4 Итоговое решающее правило и опорные признаки

Предположим, что оптимизационная задача (80) или (81) решена. Для дальнейших выводов потребуются только оптимальные значения $\tilde{\lambda}_1 \geq 0,...,\tilde{\lambda}_N \geq 0$, значения вспомогательных множителей $\tilde{\xi}_1 \geq 0,...,\tilde{\xi}_N \geq 0$ можно опустить. В соответствии с (78) найденное решение разделяет все множество признаков на три непересекающихся множества $I = I^- \cup I^0 \cup I^+$

$$I^- = \left\{ i \in I : \xi_i = 0, \xi_i < \mu^2 - \sum_{j=1}^{N} \sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l \right\},$$
$$I^0 = \left\{ i \in I : \xi_i = 0, \xi_i = \mu^2 - \sum_{j=1}^{N} \sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l \right\},$$
$$I^+ = \left\{ i \in I : \xi_i < 0, \xi_i = \mu^2 - \sum_{j=1}^{N} \sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l \right\},$$

или в эквивалентной форме

$$\begin{aligned} I^- &= \left\{ i \in I : \sum_{j=1}^{N} \sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l - \mu^2 < 0 \right\}, \\ I^0 &= \left\{ i \in I : \sum_{j=1}^{N} \sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l - \mu^2 = 0 \right\}, \\ I^+ &= \left\{ i \in I : \sum_{j=1}^{N} \sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l - \mu^2 > 0 \right\}. \end{aligned} \quad (82)$$

Учитывая (78) получим

$$\begin{cases} a_i = \sum_{j:\lambda_j>0} y_j \lambda_j x_{ij}, & i \in I^+, \\ a_i = \eta_i \sum_{j:\lambda_j>0} y_j \lambda_j x_{ij}, \ \eta_i \in [0,1], & i \in I^0, \\ a_i = 0, & i \in I^-. \end{cases} \qquad (83)$$

Таким образом, решение двойственной задачи полностью определяет значения компонент направляющего вектора a_i, попадающих в множества I^+ и I^-. Что же касается компонент, попадающих в множество I^0, то для них не определены пока коэффициенты η_i в (83). Кроме того, не определено значение сдвига разделяющей гиперплоскости.

Теорема 9. *Итоговое оптимальное решающее правило, определяемое решением задачи (72), имеет вид*

$$\sum_{j:\lambda_j>0} y_j \lambda_j \left[\sum_{i \in I^+} x_{ij} x_i + \sum_{i \in I^0} \eta_i x_{ij} x_i \right] + b \gtrless 0, \qquad (84)$$

где коэффициенты $(0 \le \eta_i \le 0, i = I^0)$ *и параметр сдвига* $b \in \mathbb{R}$ *определяются дополнительной задачей линейного программирования*

$$\begin{cases} 2\mu \sum_{i \in I^0} \eta_i + C \sum_{j=1}^{N} \delta_j \to \min(\eta_i, i \in I^0, b, \boldsymbol{\delta}), \\ \sum_{i \in I^+} \left(\sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_l \right) \eta_i + y_j b + \delta_j \ge 1 - \sum_{i \in I^+} \left(\sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_l \right), \\ \delta_j \ge 0, \ \ j = 1,\ldots,N, \ \ 0 \le \eta_i \le 1, \ \ i \in I^0. \end{cases} \qquad (85)$$

Доказательство. Для доказательства утверждения (84) и (85) достаточно подставить в общий вид решающего правила $\sum_{i \in I} a_i x_i + b \gtrless 0$ и в исходную задачу (72) оптимальные значения a_i (78).

Теорема доказана. ■

После того, как найдены решения двойственной задачи (81) и дополнительной задачи (85), исходная задача обучения (72) решена полностью, т.е. найдены коэффициенты $\hat{\lambda}_1 \ge 0, \ldots, \hat{\lambda}_N \ge 0$ и $0 \le \hat{\eta}_i \le 0, \ i = I^0$, определяющие

компоненты направляющего вектора оптимальной разделяющей гиперплоскости $\hat{a}_i$ согласно (83), и ее параметр сдвига $\hat{b}$.

В решающее правило (84) не вошли признаки $i = I^-$, поскольку соответствующие компоненты оптимальной разделяющей гиперплоскости оказались нулевыми. Заведомо входят в него признаки $i = I^+$, что же касается признаков из множества $i = I^0$, то в решающем правиле остаются только те из них, для которых значения коэффициентов $0 \le \eta_i \le 0$, найденные при решении задачи (85), оказались ненулевыми.

Признаки или компоненты направляющего вектора, определяющие оптимальное решающее правило

$$I_{supp} = I^+ \cup \{i \in I^0 : \eta > 0\} \subseteq I, \qquad (86)$$

естественно назвать *опорными признаками*. То обстоятельство, что изложенный метод обучения жестко выделяет некоторое подмножество признаков, полностью устраняя остальные, позволил назвать его *методом опорных признаков с управляемой селективностью* или *Selective Support Feature Support Vector Machine* (SFM) в отличие от метода релевантных признаков, предложенного в предыдущей главе, который лишь снабжал признаки неотрицательными весами в соответствии с их оцененной степенью информативности.

3.3.5 Зависимость множества опорных признаков от параметра селективности критерия обучения

Структура подмножеств признаков (82) и вид решающего правила (84) явно указывают механизм зависимости состава множества опорных признаков (компонент) от параметра $\mu \ge 0$ в критерии обучения (72).

Если $\mu = 0$, то множество заведомо опорных признаков $I^+ \subseteq I = \{1,...,n\}$ совпадает со всем множеством I. В этом частном случае функция (69) квадратична $q(a_i|\mu) = a_i^2$ при всех a_i и критерий обучения (72) не отличается от обычного критерия метода опорных векторов, не обладающего свойством се-

лективности, так что все исходные признаки входят в получаемое решающее правило распознавания, являясь опорными.

Увеличение значения параметра μ приводит к тому, что все большее число признаков попадает в множество I^- заведомо неопорных (82), соответственно уменьшается множество опорных признаков (86).

При неограниченном росте параметра селективности $\mu \to \infty$ в конце концов все признаки оказываются в множестве I^-, и опорных компонент вообще не остается $I_{supp} = \varnothing$.

3.4 Теоретическое исследование механизма селективности и эффекта группировки методов релевантных и опорных признаков

3.4.1 Механизм селективности отбора признаков

Для анализа свойств механизма селективности метода опорных признаков запишем критерий обучения (73), предложенный ранее в разделе 3.3, в эквивалентном виде, разделяющем принцип максимизации зазора $\varepsilon \geq 0$ между классами в обучающей выборке и механизм селекции направлений $\mathbf{a} \in \mathbb{R}^n$ искомой гиперплоскости:

$$\begin{cases} G(\mathbf{a}, b, \boldsymbol{\delta}, \varepsilon \,|\, C, \mu) = \dfrac{2\mu}{C\varepsilon} \sum\limits_{i:|a_i| \leq \varepsilon\mu} |a_i| + \dfrac{1}{C} \sum\limits_{i:|a_i| > \varepsilon\mu} \left(\mu^2 + \dfrac{a_i^2}{\varepsilon^2} \right) + \sum\limits_{j=1}^{N} \delta_j \to \min(\mathbf{a}, b, \boldsymbol{\delta}, \varepsilon), \\ y_j \left(\sum\limits_{i=1}^{n} a_i x_{ij} + b \right) \leq \varepsilon(1 - \delta_j),\ \delta_j \geq 0,\ j = 1, \ldots, N,\ \mathbf{a}^T\mathbf{a} = 1,\ \varepsilon \geq 0. \end{cases} \tag{87}$$

Эквивалентность оптимизационных задач нетрудно проверить, выполнив замену переменных $\tilde{\mathbf{a}} = \mathbf{a}/\varepsilon$, $\tilde{b} = b/\varepsilon$ и домножив целевую функцию на величину параметра C.

Выпишем первые два слагаемых целевой функции в (87), отвечающих за регуляризацию обучения:

$$V_{SFM}(\mathbf{a},\varepsilon\,|\,C,\mu)=\frac{2\mu}{C\varepsilon}\sum_{i:|a_i|\le\varepsilon\mu}|\,a_i\,|+\frac{1}{C}\sum_{i:|a_i|>\varepsilon\mu}\left(\mu^2+\frac{1}{\varepsilon^2}a_i^2\right). \qquad (88)$$

«Наилучшие» направления (22) критерия обучения (87), а, следовательно, и исходного критерия (73), определяются следующей теоремой.

Теорема 10. *Функция штрафа (88) для фиксированного* $\varepsilon>0$ *принимает минимальное значение на множестве* $\mathcal{A}=\{\mathbf{a}\in\mathbb{R}^n:\|\,\mathbf{a}\,\|=1\}$ *для направляющих векторов*

$$\mathcal{A}_{\min}^{SFM}(\varepsilon\,|\,C,\mu)=\left\{\mathbf{a}\in\mathbb{R}^n:V_{SFM}(\mathbf{a},\varepsilon\,|\,C,\mu)=\min_{\mathbf{a}'\in\mathcal{A}}V_{SFM}(\mathbf{a}',\varepsilon\,|\,C,\mu)\right\}, \qquad (89)$$

$$\forall\varepsilon\ \ \mathbf{a}\in\mathcal{A}_{\min}^{SFM}(\varepsilon)\leftrightarrow|\,a_i\,|\in\{0,1\},\ \sum_{i=1}^{n}|\,a_i\,|=1. \qquad (90)$$

Если $\varepsilon\mu\ge1$, *то* $\min_{\mathbf{a}'\in\mathcal{A}}V_{SFM}(\mathbf{a}',\varepsilon\,|\,C,\mu)=2\mu/C\varepsilon$.

Если $\varepsilon\mu<1$, *то* $\min_{\mathbf{a}'\in\mathcal{A}}V_{SFM}(\mathbf{a}',\varepsilon\,|\,C,\mu)=(\mu^2+\varepsilon^{-2})/C$.

Доказательство теоремы приведено на с. 107.

Как и у методов 1-norm SVM и DrSVM, селективность обучения по методу опорных признаков (73) определяется априорным предпочтением функции штрафа (88) минимально штрафовать такие направляющие вектора (89), у которых только один компонент отличен от нуля (90), что соответствует включению в решающее правило только одного признака.

«Наихудшие» направления (21) критерия обучения (87), а, следовательно, и исходного критерия (73), определяются следующей теоремой.

Теорема 11. *Функция штрафа (88) для фиксированного* $\varepsilon>0$ *принимает максимальное значение на множестве* $\mathcal{A}=\{\mathbf{a}\in\mathbb{R}^n:\|\,\mathbf{a}\,\|=1\}$ *для направляющих векторов*

$$\mathcal{A}_{\max}^{SFM}(\varepsilon\,|\,C,\mu)=\left\{\mathbf{a}\in\mathbb{R}^n:V_{SFM}(\mathbf{a},\varepsilon\,|\,C,\mu)=\max_{\mathbf{a}'\in\mathcal{A}}V_{SFM}(\mathbf{a}',\varepsilon\,|\,C,\mu)\right\},$$

Если $\varepsilon\mu\ge n^{-0.5}$, *то*

$$\forall\varepsilon\ \ \mathbf{a}\in\mathcal{A}_{\max}^{SFM}(\varepsilon)\leftrightarrow|\,a_i\,|=n^{-0.5},\ i=1,\ldots,n,$$

$$\max_{\mathbf{a}'\in\mathcal{A}}V_{SFM}(\mathbf{a}',\varepsilon\,|\,C,\mu)=2\mu n^{0.5}/C\varepsilon.$$

Если $\varepsilon\mu<n^{-0.5}$, *то*

$$\forall \varepsilon \ \ \mathbf{a} \in \mathcal{A}_{\max}^{SFM}(\varepsilon) \leftrightarrow \sum_{i=1}^{n} a_i^2 = 1, \ |a_i| \geq \varepsilon\mu,$$

$$\max_{\mathbf{a}' \in \mathcal{A}} V_{SFM}(\mathbf{a}', \varepsilon \,|\, C, \mu) = (\mu^2 n + \varepsilon^{-2})/C.$$

Доказательство теоремы приведено на с. 110.

Нетрудно теперь вычислить уровень селективности (23), который определяется разностью значений штрафа для «наихудших» и «наилучших» направлений:

$$SL_{SFM}(\varepsilon \,|\, C, \mu) = \max_{\mathbf{a} \in \mathcal{A}} V_{SFM}(\mathbf{a}, \varepsilon \,|\, C, \mu) - \min_{\mathbf{a} \in \mathcal{A}} V_{SFM}(\mathbf{a}, \varepsilon \,|\, C, \mu) =$$

$$= \frac{1}{C} \begin{cases} 2\mu(n^{0.5} - 1)\varepsilon^{-1}, \text{ если } \varepsilon\mu \geq 1, \\ 2\mu n^{0.5}\varepsilon^{-1} - \mu^2 + \varepsilon^{-2}, \text{ если } n^{-0.5} \leq \varepsilon\mu < 1, \\ \mu^2(n-1), \text{ если } \varepsilon\mu < n^{-0.5}. \end{cases}$$

Теорема 11 показывает важное отличие метода опорных признаков (73) от методов 1-norm SVM и DrSVM, которое заключается в том, что при $\varepsilon\mu < n^{-0.5}$ множество наименее предпочтительных ориентаций направляющего вектора состоит не из одного направления, оценивающего все признаки как равнозначные (методы 1-norm SVM и DrSVM), а составляет конус направлений $\{\mathbf{a} \in \mathbb{R}^n : \sum_{i=1}^{n} a_i^2 = 1, |a_i| \geq \varepsilon\mu, i = 1, \ldots, n\}$, предполагающих включение в решающую функцию $\mathbf{a}^T\mathbf{x} + b \gtrless 0$ сразу всех признаков с разными достаточно большими весами $|a_i| \geq \varepsilon\mu$, $i = 1, \ldots, n$. Это условие означает «неразличимость» для обучения таких направляющих векторов между собой с точки зрения их априорной предпочтительности. Следовательно, в ситуациях, когда в исходных данных есть несколько значимых признаков, выбор весов, с которыми они будут входить в решение, определяется исключительно качеством разделения выборки, а не априорным стремлением выделить только один из признаков, подобно методам 1-norm SVM и DrSVM.

Аналогичным образом проанализируем метод релевантных признаков с управляемой селективностью и запишем критерий обучения (66) в эквивалентном виде

$$\begin{cases} G(\mathbf{a},b,\boldsymbol{\delta},\varepsilon\,|\,C,\mu)=C^{-1}\sum\limits_{i=1}^{n}\ln(a_i^2/\varepsilon^2+1/\mu)+\sum\limits_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\boldsymbol{\delta},\varepsilon), \\ y_j\left(\sum\limits_{i=1}^{n}a_i x_{ij}+b\right)\le\varepsilon(1-\delta_j),\ \delta_j\ge 0,\ j=1,\ldots,N,\ \mathbf{a}^T\mathbf{a}=1,\ \varepsilon\ge 0. \end{cases} \tag{91}$$

Эквивалентность критериев нетрудно проверить, выполнив замену переменных $\tilde{\mathbf{a}}=\mathbf{a}/\varepsilon$, $\tilde{b}=b/\varepsilon$ и домножив целевую функцию на величину параметра $C\neq 0$.

Выпишем отдельно штраф на направление и зазор (первое слагаемое целевой функции)

$$V_{SFM}(\mathbf{a},\varepsilon\,|\,C,\mu)=C^{-1}\sum_{i=1}^{n}\ln(a_i^2/\varepsilon^2+1/\mu). \tag{92}$$

«Наилучшие» направления такого штрафа определены в следующей теореме.

Теорема 12. *Функция штрафа (92) для фиксированного* $\varepsilon>0$ *принимает минимальное значение на множестве* $\mathcal{A}=\{\mathbf{a}\in\mathbb{R}^n:\|\,\mathbf{a}\,\|=1\}$ *для направляющих векторов*

$$\mathcal{A}_{\min}^{RFM}(\varepsilon\,|\,C,\mu)=\left\{\mathbf{a}\in\mathbb{R}^n:V_{RFM}(\mathbf{a},\varepsilon\,|\,C,\mu)=\min\nolimits_{\mathbf{a}'\in\mathcal{A}}V_{RFM}(\mathbf{a}',\varepsilon\,|\,C,\mu)\right\}, \tag{93}$$

$$\forall\varepsilon\ \ \mathbf{a}\in\mathcal{A}_{\min}^{RFM}(\varepsilon)\leftrightarrow|\,a_i\,|\in\{0,1\},\ \sum_{i=1}^{n}|\,a_i\,|=1, \tag{94}$$

$$\min\nolimits_{\mathbf{a}'\in\mathcal{A}}V_{RFM}(\mathbf{a}',\varepsilon\,|\,C,\mu)=C^{-1}\ln(1/\varepsilon^2+1/\mu)+C^{-1}(n-1)\ln(1/\mu).$$

Доказательство теоремы приведено на с. 113.

«Наихудшие» направления критерия обучения (91) определены в следующей теореме.

Теорема 13. *Функция штрафа (92) для фиксированного* $\varepsilon>0$ *принимает максимальное значение на множестве* $\mathcal{A}=\{\mathbf{a}\in\mathbb{R}^n:\|\,\mathbf{a}\,\|=1\}$ *для направляющих векторов*

$$\mathcal{A}_{\max}^{RFM}(\varepsilon\,|\,C,\mu)=\left\{\mathbf{a}\in\mathbb{R}^n:V_{RFM}(\mathbf{a},\varepsilon\,|\,C,\mu)=\max\nolimits_{\mathbf{a}'\in\mathcal{A}}V_{RFM}(\mathbf{a}',\varepsilon\,|\,C,\mu)\right\}, \tag{95}$$

$$\forall \varepsilon \ \mathbf{a} \in \mathcal{A}_{\max}^{RFM}(\varepsilon) \leftrightarrow |a_i| = n^{-0.5}, \ i = 1,...,n. \qquad (96)$$

$$\max_{\mathbf{a}' \in \mathcal{A}} V_{RFM}(\mathbf{a}', \varepsilon \,|\, C, \mu) = C^{-1} n \ln(1/n\varepsilon^2 + 1/\mu).$$

Доказательство теоремы приведено на с. 115.

Уровень селективности (23) определяется разностью величин штрафов для «наихудших» и «наилучших» направлений:

$$SL_{RFM}(\varepsilon \,|\, C, \mu) = \max_{\mathbf{a}' \in \mathcal{A}} V_{RFM}(\mathbf{a}', \varepsilon \,|\, C, \mu) - \min_{\mathbf{a}' \in \mathcal{A}} V_{RFM}(\mathbf{a}', \varepsilon \,|\, C, \mu) =$$
$$C^{-1}\left[n \ln(1/n\varepsilon^2 + 1/\mu) - \ln(1/\varepsilon^2 + 1/\mu) + (n-1)\ln(1/\mu) \right] =$$
$$C^{-1}\left[n \ln(1/n\mu\varepsilon^2 + 1/\mu^2) - \ln(1/\mu\varepsilon^2 + 1/\mu^2) \right].$$

Не трудно видеть, что с точки зрения состава множеств наиболее (96) и наименее (96) предпочтительных направлений метод релевантных признаков (66) эквивалентен методам 1-norm SVM и DrSVM.

3.4.2 Эффект группировки признаков

Предложенный в диссертации метод опорных признаков (73), в отличие от методов 1-norm SVM и DrSVM, обладает *эффектом избирательной группировки признаков*. Избирательность группировки признаков определяется следующей теоремой.

Теорема 14. *Пусть обучающая выборка*

$$(X, Y) = \{(y_1, \mathbf{x}_1), ..., (y_N, \mathbf{x}_N)\} \qquad (97)$$

предварительно центрирована и нормирована таким образом, что

$$\frac{1}{N}\sum_{j=1}^{N} x_{ij} = 0, \ \sum_{j=1}^{N} x_{ij}^2 = 1, \text{ для каждого признака } i = 1,...,n. \qquad (98)$$

Пусть $\hat{\mathbf{a}} = (\hat{a}_1, ..., \hat{a}_n)^T$, $\hat{b}$, $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, ..., \hat{\delta}_N)^T$ *есть решение оптимизационной задачи метода опорных признаков с управляемой селективностью (73)*

$$\begin{cases} J_{SFM}(\mathbf{a}, b, \boldsymbol{\delta} \,|\, \mu, C) = 2\mu \sum\limits_{i:|a_i| \le \mu} |a_i| + \sum\limits_{i:|a_i| > \mu} (a_i^2 + \mu^2) + C\sum\limits_{j=1}^{N} \delta_j \to \min(\mathbf{a}, b, \boldsymbol{\delta}), \\ y_j\left(\sum\limits_{i=1}^{n} a_i x_{ij} + b \right) \ge 1 - \delta_j, \ \delta_j \ge 0, \ j = 1,...,N. \end{cases} \qquad (99)$$

для обучающей выборки (97) с учетом нормировки (98). Тогда для любой пары компонент с индексами (q,k) *направляющего вектора* $\hat{\mathbf{a}}=(\hat{a}_1,...,\hat{a}_n)^T$ *справедливы следующие неравенства.*

Если $|\hat{a}_q|>\mu$ *и* $|\hat{a}_k|>\mu$*, то имеем*

$$C\sqrt{N}\sqrt{2(1-\rho)}\geq|\hat{a}_q-\hat{a}_k|,\ \rho=\sum_{j=1}^{N}x_{qj}x_{kj},\ 0\leq\rho\leq1, \qquad (100)$$

где ρ *- есть выборочная корреляция признаков с индексами* (q,k)*, вычисленная по выборке (97) с учетом нормировки (98).*

Если $|\hat{a}_q|>\mu$, $|\hat{a}_k|\leq\mu$ *и* $0.5|\hat{a}_q+\hat{a}_k|>\mu$*, то имеем*

$$C\sqrt{N}\sqrt{2(1-\rho)}\geq|\hat{a}_q-\hat{a}_k|-2\frac{(|\hat{a}_k|-\mu)^2}{|\hat{a}_q-\hat{a}_k|}.$$

Если $|\hat{a}_q|>\mu$, $|\hat{a}_k|\leq\mu$, $0.5|\hat{a}_q+\hat{a}_k|\leq\mu$ *и компоненты* $\hat{a}_q$, $\hat{a}_k$ *одного знака, то*

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\geq\frac{(|\hat{a}_q|-\mu)^2}{|\hat{a}_q-\hat{a}_k|}.$$

Если $|\hat{a}_q|>\mu$, $|\hat{a}_k|\leq\mu$, $0.5|\hat{a}_q+\hat{a}_k|\leq\mu$ *и компоненты* $\hat{a}_q$, $\hat{a}_k$ *разных знаков, то*

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\geq\frac{(|\hat{a}_q|-\mu)^2}{|\hat{a}_q-\hat{a}_k|}-\mu\frac{|\hat{a}_q+\hat{a}_k|-|\hat{a}_k|-|\hat{a}_q|}{|\hat{a}_q-\hat{a}_k|}.$$

Если $|\hat{a}_q|\leq\mu$, $|\hat{a}_k|\leq\mu$ *и компоненты* $\hat{a}_q$, $\hat{a}_k$ *одного знака, то их значения дополнительно не ограничены.*

Если $|\hat{a}_q|\leq\mu$, $|\hat{a}_k|\leq\mu$ *и компоненты* $\hat{a}_q$, $\hat{a}_k$ *разного знака, то*

$$\frac{C\sqrt{N}}{2\mu}\sqrt{2(1-\rho)}\geq1-\frac{|\hat{a}_q+\hat{a}_k|}{|\hat{a}_q-\hat{a}_k|}.$$

Доказательство теоремы приведено на с. 116.

Для иллюстрации утверждений теоремы рассмотрим пару признаков с индексами (q,k), корреляция которых равна единице $\rho=1$. Соответствую-

щие компоненты оптимального направляющего вектора $\hat{a}_q$, $\hat{a}_k$ удовлетворяют следующим условиям:

- либо $|\hat{a}_q| > \mu$ и $|\hat{a}_k| > \mu$, либо $|\hat{a}_q| \le \mu$ и $|\hat{a}_k| \le \mu$,
- если $|\hat{a}_q| > \mu$ и $|\hat{a}_k| > \mu$, то $\hat{a}_q = \hat{a}_k$,
- если $|\hat{a}_q| \le \mu$ и $|\hat{a}_k| \le \mu$, то $\hat{a}_q \hat{a}_k \ge 0$.

Содержательно это означает, что признаки с единичной корреляцией могут входить в решение либо с одинаковыми коэффициентами $\hat{a}_q = \hat{a}_k$, когда они по модулю большие порога μ (есть эффект группировки), либо с разными, но одного знака $\hat{a}_q \hat{a}_k \ge 0$, если они по модулю меньше порога μ (нет эффекта группировки).

Таким образом, предложенный метод опорных признаков (73) характеризуется избирательным «стилем» учета линейной зависимости признаковых описаний объектов при их отборе. Избирательность предложенного метода выражается в стремлении сохранять в итоговом решении только попарно зависимые признаки, обладающие достаточной значимостью для распознавания. Для сравнения метод Elastic Net SVM сохраняет все линейно зависимые признаки. Малозначимые признаки при этом максимально подавляются в процессе обучения независимо от их попарной линейной зависимости, аналогично Lasso SVM.

Следующая теорема показывает, что метод релевантных признаков (66) также обладает эффектом группировки признаков.

Теорема 15. *Пусть обучающая выборка*

$$(X, Y) = \{(y_1, \mathbf{x}_1), ..., (y_N, \mathbf{x}_N)\} \qquad (101)$$

предварительно центрирована и нормирована таким образом, что

$$\frac{1}{N}\sum_{j=1}^{N} x_{ij} = 0,\ \sum_{j=1}^{N} x_{ij}^2 = 1, \text{ для каждого признака } i = 1, ..., n. \qquad (102)$$

Пусть $\hat{\mathbf{a}} = (\hat{a}_1, ..., \hat{a}_n)^T$, $\hat{b}$, $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, ..., \hat{\delta}_N)^T$ *есть решение оптимизационной задачи метода релевантных признаков (66)*

$$\begin{cases} J_{RFM}(\mathbf{a},b,\delta \mid \mu,C) = \sum_{i=1}^{n} \ln(a_i^2 + 1/\mu) + C\sum_{j=1}^{N} \delta_j \to \min(\mathbf{a},b,\delta), \\ y_j\left(\sum_{i=1}^{n} a_i x_{ij} + b\right) \le 1-\delta_j,\ \delta_j \ge 0,\ j=1,\ldots,N. \end{cases} \qquad (103)$$

для обучающей выборки (101) с учетом нормировки (98).

Для любой пары индексом (q,k), *для которых выполняются условия* $|\hat{a}_q| < 1/\sqrt{\mu}$, $|\hat{a}_k| < 1/\sqrt{\mu}$ *и* $\rho = \sum_{j=1}^{N} x_{qj}x_{kj} = 1$, *справедливо* $\hat{a}_q = \hat{a}_k$. *Здесь* $0 \le \rho \le 1$ - *выборочная корреляция признаков с индексами* (q,k), *вычисленная по обучающей выборке (101) с учетом нормировки (102).*

Доказательство теоремы приведено на с. 122.

3.5 Выбор оптимального уровня селективности: Метод скользящего контроля

Параметр селективности $\mu \ge 0$, как и параметр $C \ge 0$, является структурным параметром критериев обучения (65), (72), определяет последовательность вложенных классов моделей распознавания, чья размерность уменьшается с ростом μ. При этом уменьшение размерности происходит неявно для случая метода релевантных признаков (RFM) и абсолютно явно для метода опорных признаков (SFM). Однако для обоих методов нельзя указать, как количественно зависит размерность и величина параметра селективности, а также гарантировать, что эта зависимость монотонна.

Для выбора оптимальных значений параметров в экспериментальной части диссертации используется метод скользящего контроля, зарекомендовавший себя, как наиболее универсальный и один из самых эффективных способов оценивания обобщающей способности методов обучения по заданной обучающей выборке.

4 ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ МЕТОДОВ РЕЛЕВАНТНЫХ И ОПОРНЫХ ПРИЗНАКОВ С УПРАВЛЯЕМОЙ СЕЛЕКТИВНОСТЬЮ

4.1 Экспериментальное исследование на модельных данных

Основной целью экспериментального исследования является анализ предлагаемых в диссертации методов релевантных и опорных признаков по их способности сокращать признаковое описание объектов распознавания и, в конечном итоге, повышать обобщающую способность обучения по относительно малой обучающей выборке при большом числе признаков.

За основу экспериментов принята и расширена структура модельных данных, предложенная авторами метода Doubly Regularized SVM (Elastic Net SVM) [12].

4.1.1 Эксперимент А: Структура модельных данных и результаты

Модельный эксперимент организован в пространстве независимых признаков с размерностью $n=100$. Обучающая совокупность $\{\mathbf{x}_j, y_j\}_{j=1}^{N}$ представляет собой случайную выборку векторов $\mathbf{x}_j \in \mathbb{R}^n$ двух классов $y_j = \pm 1$, сгенерированных независимо по многомерному нормальному закону распределения с математическим ожиданием для первого класса $y_j = +1$

$$\mathbf{m}_{+} = (\underbrace{0.5,...,0.5}_{\substack{5 \\ \textit{информативные} \\ \textit{признаки}}}, \underbrace{0,....,0}_{\substack{95 \\ \textit{шумовые} \\ \textit{признаки}}})^T \in \mathbb{R}^n \tag{104}$$

и

$$\mathbf{m}_{-} = (\underbrace{-0.5,...,-0.5}_{\substack{5 \\ \textit{информативные} \\ \textit{признаки}}}, \underbrace{0,....,0}_{\substack{95 \\ \textit{шумовые} \\ \textit{признаки}}})^T \in \mathbb{R}^n, \tag{105}$$

для второго $y_j = -1$ класса, а также единичной ковариационной матрицей

$$\mathbf{\Sigma} = \mathbf{I}_{n \times n} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}. \tag{106}$$

При такой модели порождения данных искомая байесовская оптимальная гиперплоскость, разделяющая два класса с минимальной ошибкой распознавания, зависит только от первых 5 *информативных* признаков $\{1,2,3,4,5\}$ с равной степенью вклада в разделимость классов и не зависит от оставшихся заведомо лишних 95 *шумовых* признаков $\{5,...,100\}$ (104)-(105). Минимальная ошибка распознавания (доля неправильно распознанных объектов) для оптимальной гиперплоскости при известных заданных параметрах нормальных плотностей распределения двух классов (104), (105) и (106) составляет 0.132. Такую ошибку, недостижимую при оценивании по конечной обучающей совокупности, в литературе принято называть *Ground Truth Error*. В последней шестой строке табл. 1 она названа «оракульной» ошибкой.

Массив модельных данных включает 100 наборов обучающих совокупностей. Каждая из обучающих совокупностей содержит $N = 100$ точек по 50 в каждом классе. Такой размер обучающей совокупности вполне достаточен для оценивания разделяющей гиперплоскости в пространстве первых пяти информативных признаков, но слишком мал для ее оценивания по всем ста признакам. Контрольная совокупность образована $N_{test} = 20\,000$ точками по 10 000 точек каждого класса, что, как показали предварительные эксперименты, оказалось достаточным для ее использования в качестве модели генеральной совокупности.

Эксперимент организован следующим образом:

1. Каждая из 100 обучающих совокупностей предварительно центрировалась и нормировалась, а затем подвергалась обработке тремя алгоритмами, реализующими метод опорных векторов с двойной регуляризацией (DrSVM) (18), предлагаемые методы релевантных (RFM) и опорных признаков (SFM) согласно критериям (65) и (72), с разными значениями параметра селективности в диапазоне $10^{-6} \le \mu \le 10^{12}$ и параметра разделимости выборки $10^{-6} \le C \le 10^{12}$. Кроме того, для разных C к обучающим выборкам применялся метод опорных векторов (SVM), что эквивалентно применению любого из

сравниваемых методов при нулевом значении параметра селективности $\mu = 0$.

2. Для каждой полученной в результате обучения разделяющей гиперплоскости вычислялась доля неправильно классифицированных точек контрольной совокупности, $p_{test}^{SVM}(C)$, $p_{test}^{DrSVM}(\mu, C)$, $p_{test}^{RFM}(\mu, C)$, $p_{test}^{SFM}(\mu, C)$, которая является оценкой вероятности ошибки распознавания на генеральной совокупности. Для алгоритмов SVM и DrSVM запоминались значения компонент направляющего вектора, для RFM, кроме того, запоминалась совокупность весов признаков (128), а для алгоритма SFM - подмножество опорных признаков I_{supp} (86), активно участвующих в разделяющей гиперплоскости, вместе с их весами η_i (84).

3. К каждой обучающей совокупности применялся классический метод опорных векторов (SVM) в предположении, что известно подмножество полезных признаков, соответствующих «истинным» значениям пяти первых компонент направляющего вектора $(a_1, ..., a_5)$.

В таблице 1 приведены минимальные средние (для 100 обучающих совокупностей) ошибки распознавания на тестовой выборке («почти» генеральной совокупности) для соответствующих алгоритмов обучения по обучающим совокупностям, и соответствующие оптимальные значения параметров селективности $\hat{\mu}$ и разделимости $\hat{C}$. В скобках указана величина стандартного отклонения ошибки распознавания. Жирным шрифтом выделен метод с наименьшей ошибкой распознавания.

Таблица 1. Эксперимент А: Минимальные средние ошибки распознавания.

Метод обучения		Минимальная средняя ошибка на тестовой совокупности по 100 обучающим совокупностям	Оптимальное значение параметра селективности $\hat{\mu}$	Оптимальное значение параметра разделимости $\hat{C}$
Предложенные	**SFM**	**0.1495 (0.0146)**	**0.0516**	**0.0036**
	RFM	0.1797 (0.0249)	4 000 000	4 000 000
Существующие	SVM	0.2353 (0.0149)	-	0.0017
	DrSVM	0.1523 (0.0107)	0.0207	0.0010
	SVM на 5 заведомо информативных признаках	0.1430 (0.0055)	-	0.0164
	«Оракульная» ошибка	0.132		

При нулевом значении селективности $\mu = 0$ методы DrSVM, SFM и RFM эквивалентны и дают одну и ту же минимальную среднюю ошибку, равную ошибке метода опорных векторов, при одинаковом оптимальном значении параметра $\hat{C}_{SVM}$

$$p_{test}^{DrSVM}(0, \hat{C}_{SVM}) = p_{test}^{RFM}(0, \hat{C}_{SVM}) = p_{test}^{SFM}(0, \hat{C}_{SVM}) = p_{test}^{SVM}(\hat{C}_{SVM}) = 0.2353.$$

Такая ошибка (строка SVM в табл. 1) почти вдвое превышает минимально возможную ошибку 0.132 распознавания.

Обучение методом SVM в пространстве первых пяти информативных признаков позволяет обеспечить почти идеальное качество распознавания с ошибкой 0.1430 и с существенно меньшим среднеквадратичным отклонением. Это означает, что выборки размера 100 точек явно недостаточно для обучения методом SVM в пространстве всех признаков размерности $n = 100$.

Однако минимальные значения средней ошибки на тестовой совокупности, достигаемые селективными алгоритмами при оптимальных значениях селективности $\hat{\mu}$ и качества разделения $\hat{C}$, существенно различаются. Метод RFM обеспечивает ошибку 0.1797 и тем самым более чем наполовину уменьшает эффект переобучения SVM с ошибкой 0.2353 в сравнении с ошибкой 0.1430 при известном подмножестве информативных признаков. В то же время, ошибки 0.1523 для метода DrSVM и 0.1495 для SKM еще заметно меньше. Несмотря на отсутствие корреляции между признаками, предложенный метод SKM демонстрирует заметно лучшую обобщающую способность по сравнению с известным методом DrSVM, и практически обеспечивает ошибку распознавания SVM при обучении только по 5 информативным признакам.

Сравнивать оптимальные значения параметров селективности $\hat{\mu}$ и разделимости $\hat{C}$, обеспечивающие минимальные значения средней ошибки, нет смысла, поскольку механизм влияния этих параметров различен в разных методах.

Методы SVM и RKM не выделяют строгое подмножество нулевых признаков, поэтому для определения «фактически активных» признаков использовалось два порога значимости 10^{-3} и 10^{-6}, с которыми сравнивались по модулю соответствующие оцененные значения компонент направляющего вектора гиперплоскости. Если компонента направляющего вектора по модулю больше порога, то соответствующий признак считается активным в гиперплоскости, а если меньше, то неактивным. В таблице 2 приведено усредненное по 100 обучающим совокупностям при оптимальных значениях параметров $(\hat{\mu},\hat{C})$ количество активных признаков (для каждого порога значимости) в оптимальной гиперплоскости.

Помимо точности распознавания, селективные методы удобно сравнивать по способности сокращать размерность заданного признакового описания. В частности, для каждого метода обучения вычислялось среднее по 100 случайным обучающим совокупностям количество $0 \le n_{act} \le 100$ фактически

активных компонент направляющего вектора оптимальной гиперплоскости, количество активных информативных компонент $0 \le n_{rel} \le 5$, и среднее количество активных шумовых компонент $0 \le n_{noise} \le 95$ (104)-(105):

$$0 \le n_{act} = n_{rel} + n_{noise} \le n = 100 .$$

Таблица 2. Эксперимент A: Среднее количество активных признаков.

Метод обучения		Среднее количество n_{act} активных признаков в решении	Среднее количество n_{rel} информативных признаков в решении	Среднее количество n_{noise} шумовых признаков в решении
Предложенные	**SFM**	**5.19 / 5.19**	**4.84 / 4.84**	**0.35 / 0.35**
	RFM	8.02 / 47.40	4.43 / 4.95	3.59 / 42.45
Существующие	SVM	91.24 / 99.99	5.00 / 5.00	86.24 / 94.99
	DrSVM	7.04 / 8.99	4.97 / 4.99	2.07 / 4.00
	SVM на 5 информативных признаках	5.00 / 5.00	5.00 / 5.00	-

Нетрудно видеть, что известный метод DrSVM демонстрируют хорошую способность подавлять шумовые признаки, в отличие от SVM, который использует для распознавания практически все доступные признаки, включая 95 шумовых. Первый предложенный метод RFM не полностью справился с задачей, оставив заметное количество шумовых признаков, что и объясняет его отставание по обобщающей способности от других селективных методов (табл. 1). В то же время, второй предлагаемый в диссертации метод SKM справился с задачей практически идеально, оставляя в среднем менее одного шумового признака, в отличие от метода DrSVM, который оставляет четыре таких признака. Такое преимущество метода SFM объясняется учетом в структуре критерия обучения (функции штрафа) предположения о существовании в данных нескольких значимых признаков, в то время как, для извест-

ного метода DrSVM это предположение о модели данных является наименее предпочтительным.

На рисунке 13 приведены примеры зависимостей ошибки распознавания на тестовой совокупности от значения параметра селективности μ, полученные на одной из обучающих совокупностей при фиксированных оптимальных значениях параметра $\hat{C}$ для каждого метода. При увеличении селективности μ ошибки всех методов уменьшаются и достигают своего минимума, поскольку методы подавляют влияние шумовых признаков, компенсируя эффект переобучения. Затем ошибка снова возрастает, т.к. методы начинают удалять информативные признаки. Наконец, при слишком большом значении селективности $\mu \to \infty$ методы подавляют все признаки, и средняя ошибка увеличивается до «безразличного» значения 0.5.

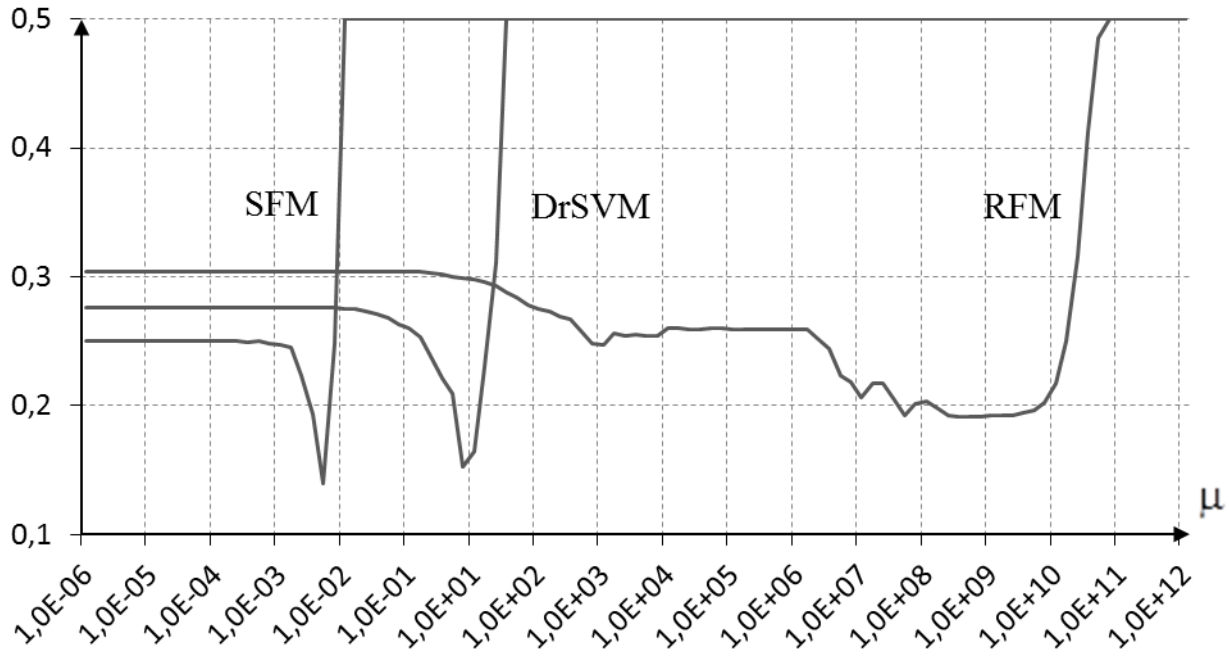


Рисунок 13. Пример эффекта регуляризации методов DrSVM, RFM и SFM на одной из модельных обучающих совокупностей для возрастающих значений параметра селективности μ при фиксированных оптимальных значениях параметра разделимости $\hat{C}$.

Для этой же обучающей совокупности на рисунке 14 показаны относительные значения 7 наибольших весов признаков в порядке убывания, полученные в результате обучения каждым из трех методов для трех значений па-

раметра селективности μ. Параметр разделимости $\hat{C}$ фиксирован для каждого метода. Заметим, что при оптимальном выборе μ по тестовой выборке оба метода правильно выделяют 5 информативных признаков.

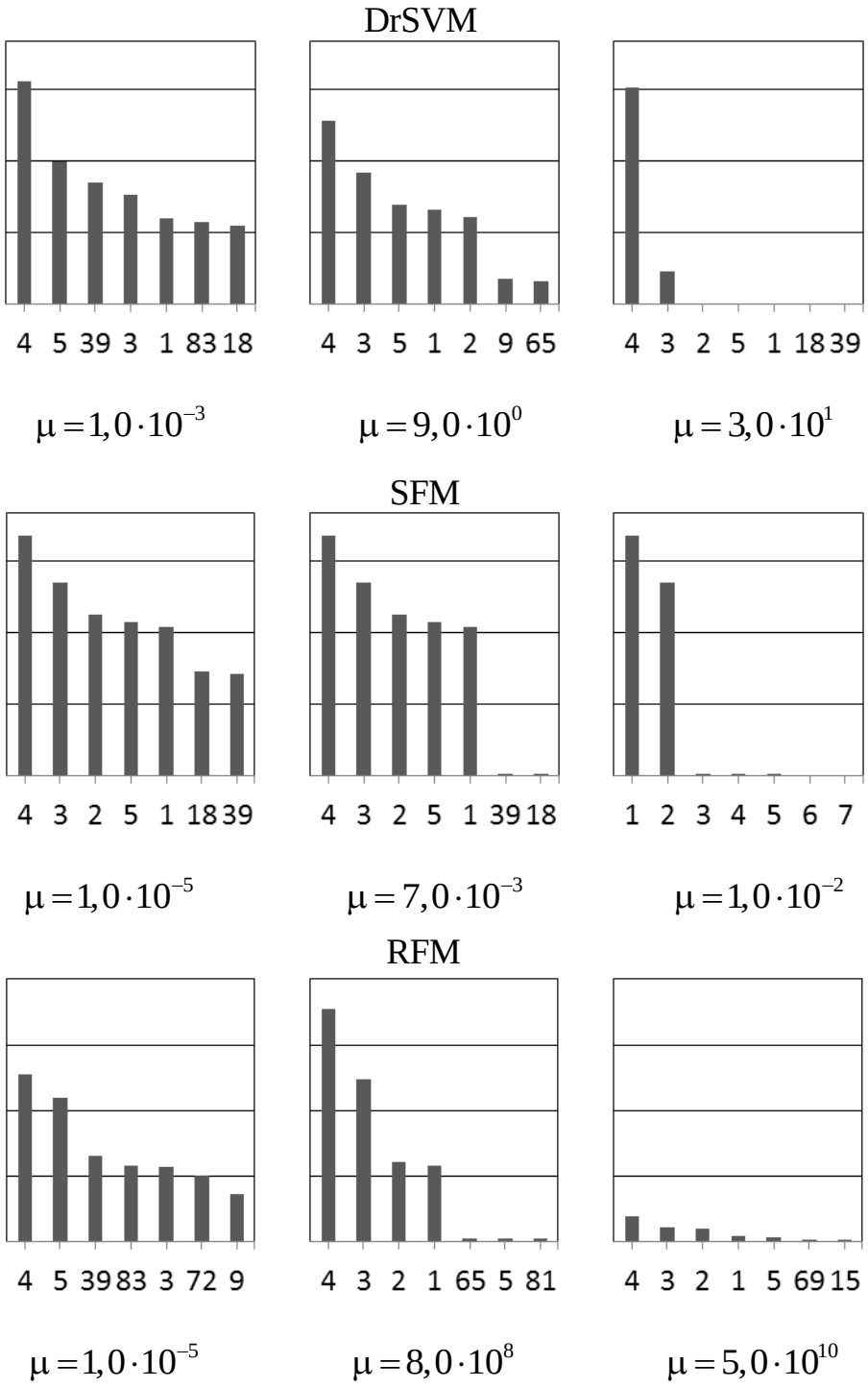


Рисунок 14. Пример выделения информативных признаков методами DrSVM, SFM и RFM для разной селективности μ и фиксированных оптимальных значений параметра разделимости $\hat{C}$. Рамки показывают оптимальные значения селективности по тестовой совокупности.

4.1.2 Эксперимент B: Структура модельных данных и результаты

Следующий эксперимент отличается от предыдущего только тем, что шумовые признаки в модели порождения данных сильно линейно зависимы. Зависимость признаков выражена через ковариационную матрицу многомерных нормальных плотностей распределения наблюдений каждого класса

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{I}_{5\times5} & \mathbf{0}_{5\times95} \\ \mathbf{0}_{5\times95} & \mathbf{\Sigma}^{*}_{95\times95} \end{pmatrix}, \tag{107}$$

где $\mathbf{I}_{5\times5}$ - единичная матрица размером 5×5, а диагональные элементы квадратной матрицы $\mathbf{\Sigma}^{*}_{95\times95}$ равны 1, а недиагональные - 0.8. Такая ковариационная матрица определяет независимые первые 5 информативных признаков и существенную попарную линейную зависимость всех 95 шумовых признаков.

Векторы математических ожиданий (104) и (105) остаются без изменений, следовательно, не изменится и байесовская ошибка распознавания при известных параметрах плотностей распределения, равная 0.132.

Как и ранее, массив модельных данных включает 100 наборов обучающих совокупностей, содержащих по $N = 100 = 50 + 50$ точек, и одну контрольную совокупность размера $N_{test} = 20\,000 = 10\,000 + 10\,000$, полученных независимо друг от друга в пространстве $n = 100$ признаков случайным генератором, описанным ранее.

В таблице 3 сведены полученные результаты, по структуре аналогичные предыдущему эксперименту.

Таблица 3. Эксперимент B: Минимальные средние ошибки распознавания.

Метод обучения		Минимальная средняя ошибка на тестовой совокупности	Оптимальное значение параметра селективности $\hat{\mu}$	Оптимальное значение параметра разделимости $\hat{C}$
Предложенные	**SFM**	**0.1429 (0.0088)**	**0.1284**	**0.0124**
	RFM	0.1611 (0.0139)	1.6384	0.1024
Существующие	SVM	0.1620 (0.0119)	-	0.0256
	DrSVM	0.1489 (0.0110)	0.4601	0.0310
	SVM на 5 информативных признаках	0.1397 (0.0052)	-	0.0128
	«Оракульная» ошибка	0.132		

Как и в предыдущем эксперименте, предложенный метод SFM хорошо справился с задачей подавления шумовых признаков, увеличив вдвое свой отрыв от существующего метода DrSVM. Такое дополнительное преимущество метода SFM над DrSVM можно объяснить особенностями эффекта группировки, который у DrSVM работает для всех коррелированных признаков независимо от их информативности, в данном случае шумовых, а у SFM только для тех признаков, которые входят в решающую функцию с достаточно большими значениями соответствующих компонент направляющего вектора, а такими признаками в данном эксперименте являются первые 5 информативных признаков (104)-(105).

У DrSVM, как и следовало ожидать, количество шумовых признаков в решении в среднем заметно выше, чем у метода SFM, который фактически всегда безошибочно выделяет правильное подмножество информативных признаков, подавляя 95 шумовых (табл. 4). RKM по-прежнему отстает от методов SFM и DrSVM, ненамного улучшая в этом эксперименте результат обычного метода опорных векторов.

Таблица 4. Эксперимент B: Среднее количество активных признаков.

Метод обучения		Среднее количество n_{act} активных признаков в решении	Среднее количество n_{rel} информативных признаков в решении	Среднее количество n_{noise} шумовых признаков в решении
Предложенные	**SFM**	**5.35 / 5.35**	**4.97 / 4.97**	**0.38 / 0.38**
	RFM	97.64 / 99.99	5.00 / 5.00	92.64 / 94.99
Существующие	SVM	98.14 / 100.00	5.00 / 5.00	93.14 / 95.00
	DrSVM	7.09 / 7.24	5.00 / 5.00	2.09 / 2.24
	SVM на 5 информативных признаках	5.00 / 5.00	5.00 / 5.00	-

4.1.3 Эксперимент C: Структура модельных данных и результаты

Следующий эксперимент принципиально отличается от предыдущих двух тем, что в модели порождения данных теперь только два информативных признака, причем таких, что только вместе они могут обеспечить приемлемую ошибку распознавания, а остальные 98 шумовых признаков попарно независимы. Количество информативных признаков было для наглядности специально сокращено до двух, однако не составит большого труда этот эксперимент обобщить на случай большего количества информативных признаков. Попарная зависимость признаков задается ковариационной матрицей следующего вида

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}^{*}_{2\times 2} & \mathbf{0}_{2\times 98} \\ \mathbf{0}_{2\times 98} & \mathbf{I}_{98\times 98} \end{pmatrix},$$

где $\mathbf{\Sigma}^{*}_{2\times 2} = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$ – матрица, определяющая большую отрицательную корреляцию двух информативных признаков, $\mathbf{I}_{98\times 98}$ - единичная квадратная матрица 98×98, определяющая независимость шумовых признаков.

Кроме того, немного изменится математическое ожидание для распределения первого класса

$$\mathbf{m}_{+} = (\underbrace{0.25, 0.25}_{\substack{2 \\ \textit{информативные} \\ \textit{признаки}}}, \underbrace{0, \ldots, 0}_{\substack{98 \\ \textit{шумовые} \\ \textit{признаки}}})^T \in \mathbb{R}^n \qquad (108)$$

и для распределения второго класса

$$\mathbf{m}_{+} = (\underbrace{-0.25, -0.25}_{\substack{2 \\ \textit{информативные} \\ \textit{признаки}}}, \underbrace{0, \ldots, 0}_{\substack{98 \\ \textit{шумовые} \\ \textit{признаки}}})^T \in \mathbb{R}^n. \qquad (109)$$

Байесовская ошибка распознавания при известных параметрах распределений классов («оракульная» ошибка) составит 0.215. Во всем остальном схема эксперимента аналогична предыдущим двум. В таблицах 5 и 6 сведены полученные результаты эксперимента.

Таблица 5. Эксперимент С: Минимальные средние ошибки распознавания.

Метод обучения		Минимальная средняя ошибка на тестовой совокупности	Оптимальное значение параметра селективности $\hat{\mu}$	Оптимальное значение параметра разделимости $\hat{C}$
Предложенные	SFM	0.3140 (0.0576)	429 496	53 687
	RFM	**0.2325 (0.0355)**	**419.43**	**209.72**
Существующие	SVM	0.4359 (0.0140)	-	0.2048
	DrSVM	0.3140 (0.0576)	858 993	53 687
	SVM для двух полезных признаков	0.2238 (0.0057)	-	429 496
	«Оракульная» ошибка	0.215		

Несмотря на достаточно хорошую разделимость классов по двум полезным признакам, обычный SVM в пространстве 100 признаков справился с задачей очень плохо, показав очень высокую долю ошибок 0.4359. Известный метод DrSVM и предложенный метод SFM справились с задачей немного лучше, заметно уменьшив ошибку распознавания до одинакового уровня в 0.3140. Однако, в отличие от предыдущих экспериментов, в этом эксперименте лучше всего справился второй предложенный в работе метод реле-

вантных признаков RKM, обеспечивший наименьшую ошибку 0.2344, практически равную ошибке метода опорных векторов 0.2238, примененного только на двух информативных признаках.

Эффективность подавления шумовых признаков методом RKM хорошо видна по среднему количеству активных признаков, оставляемых в решающей функции. Оба метода DrSVM и SFM в среднем оставляют в решении по 8 шумовых признаков, а метод RKM успешно подавил все 98 компонент направляющего вектора, соответствующих шумовым признакам, до уровня менее чем 10^{-2}, которого оказалось достаточно, чтобы не оказывать влияния на обобщающую способность обучения (табл. 6).

Таблица 6. Эксперимент C: Среднее количество активных признаков.

Метод обучения		Среднее количество n_{act} активных признаков в оптимальном решении	Среднее количество n_{rel} информативных признаков в оптимальном решении	Среднее количество n_{noise} шумовых признаков в оптимальном решении
Предложенные	SFM	9.91 / 10.19	1.97 / 1.98	7.94 / 8.21
	RFM	**5.17[3] / 82.44 / 99.99**	**1.99 / 2.00 / 2.00**	**3.18 / 80.44 / 97.99**
Существующие	SVM	99.34 / 99.99	2.00 / 2.00	97.34 / 97.99
	DrSVM	9.91 / 10.16	1.97 / 1.98	7.94 / 8.18
	SVM для двух полезных признаков	2.00 / 2.00	2.00 / 2.00	-

4.1.4 Эксперимент D: Структура модельных данных и результаты

В завершение модельных задач приведем результаты последнего эксперимента, который отличается от предыдущего только наличием корреляции между шумовыми признаками. Ковариационная матрица примет вид

[3] Первые значения количества активных признаков для метода RFM рассчитаны для порога 10^{-2}.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{*}_{2\times 2} & \mathbf{0}_{2\times 98} \\ \mathbf{0}_{2\times 98} & \boldsymbol{\Sigma}^{*}_{98\times 98} \end{pmatrix} \tag{110}$$

где матрица $\boldsymbol{\Sigma}^{*}_{2\times 2} = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$, аналогично предыдущему примеру, определяет большую отрицательную корреляцию двух информативных признаков, но шумовые признаки, в отличие от эксперимента C, теперь также сильно коррелированы в соответствии с ковариационной матрицей $\boldsymbol{\Sigma}^{*}_{98\times 98}$, в которой диагональные элементы равны 1, а недиагональные 0.8.

Векторы математических ожиданий (108) и (109) остаются неизменными. Во всем остальном схема эксперимента полностью аналогична предыдущему эксперименту. Байесовская ошибка также не изменилась и составляет 0.215. В таблицах 7 и 8 сведены полученные результаты.

Таблица 7. Эксперимент D: Минимальные средние ошибки распознавания.

	Метод обучения	Минимальная средняя ошибка на тестовой совокупности	Оптимальное значение параметра селективности $\hat{\mu}$	Оптимальное значение параметра разделимости $\hat{C}$
Предложенные	SFM	0.2364 (0.0199)	209.72	52.43
	RFM	**0.2212 (0.0122)**	**26 844**	**13 422**
Существующие	SVM	0.3406 (0.0187)	-	0.1024
	DrSVM	0.2364 (0.0199)	107 370	13 422
	SVM на двух информативных признаках	0.2177 (0.0089)	-	209.72
	«Оракульная» ошибка	0.215		

Дополнительная корреляция шумовых признаков положительно повлияла на обобщающую способность всех алгоритмов, включая метод опорных векторов, демонстрирующий на всех признаках ошибку 0.3406. Как и в

предыдущем эксперименте, методы DrSVM и SFM показали одинаковый результат, равный 0.2364, а RFM обеспечил наилучшую ошибку 0.2224, практически вплотную приблизившись к байесовской ошибке 0.215.

Корреляция шумовых признаков также положительно отразилась и на том, сколько шумовых признаков в среднем вошло в решающую функцию у каждого метода (табл. 8).

Таблица 8. Эксперимент D: Среднее количество активных признаков.

Метод обучения		Среднее количество n_{act} активных признаков в решении	Среднее количество n_{rel} информативных признаков в решении	Среднее количество n_{noise} шумовых признаков в решении
Предложенные	SFM	6.03 / 6.15	2.00 / 2.00	4.03 / 4.15
	RFM	**2.01 / 98.96**	**2.00 / 2.00**	**0.01 / 96.96**
Существующие	SVM	99.41 / 100	2.00 / 2.00	97.41 / 98.00
	DrSVM	6.03 / 6.19	2.00 / 2.00	4.03 / 4.19
	SVM на двух информативных признаках	2.0 / 2.0	2.0 / 2.0	-

Метод RFM практически безошибочно выделяет только информативные признаки, а SFM и DrSVM оставляют в среднем чуть больше 4 шумовых признаков.

4.2 Экспериментальное исследование на данных прикладной задачи

Для демонстрации работы методов использовались данные задачи распознавания рака легких из репозитория UCI [21]. Массив данных образован векторами признаков $N = 32$ пациентов, разбитых на два подмножества $N_+ = 9$ и $N_- = 23$, у которых, соответственно, диагностирован и не диагно-

стирован рак легких. Каждый вектор состоит из $n = 56$ числовых признаков, количество которых существенно превышает объем обучающей совокупности. Для сокращения признакового описания пациентов использовались предлагаемые в диссертации метод опорных и метод релевантных признаков.

Поскольку массив данных не содержит тестовой совокупности, оценка зависимости обобщающей способности обучения от параметров селективности μ и разделимости C проводилась по скользящему контролю по отдельным объектам (leave-one-out CV). Эффект регуляризации обучения для каждого метода проиллюстрирован на рисунке 15.

Рисунок 15. Результаты скользящего контроля в задаче распознавания рака легких методами SFM и RFM для возрастающих значений параметра селективности μ при фиксированном значении параметра C.

При малых значениях параметра селективности μ оба метода эквивалентны методу опорных векторов, учитывающему все признаки, что дает ошибку 0,38. С ростом значения селективности оба метода умешают ошибку примерно вдвое, однако RFM выделяет 4 признака из 56, а SFM оставляет лишь 2 признака, что, по-видимому, является причиной некоторого преимущества RFM с ошибкой 0.187 против 0.219. Наконец слишком большие значения параметра селективности подавляют все признаки, что эквивалентно

принятию решения в пользу класса, более представленного в обучающей совокупности, и ошибка стабилизируется на уровне 9/32 = 0.281.

Примеры решения других прикладных задач можно найти в публикациях автора, перечисленных во введении.

4.3 Обсуждение результатов экспериментов

Экспериментальное исследование, описанное в заключительной главе диссертации, имеет общепринятую структуру, включающую в себя серию модельных экспериментов и пример решения прикладной задачи. Целью подобных исследований является наглядная демонстрация «полезности» предлагаемых авторами новых идей на примере некоторого набора вычислительных задач по сравнению с хорошо известными и уже зарекомендовавшими себя методами. Для усиления этого эффекта модельные эксперименты настоящей диссертации основаны, по своей структуре, на экспериментальном исследовании, проведенном авторами известного метода Doubly Regularized SVM (DrSVM) [12] для иллюстрации преимущества модульно-квадратичной функции штрафа Elastic Net по сравнению с традиционным модульным Lasso.

Продемонстрированное улучшение результатов метода Doubly Regularized SVM в аналогичных модельных экспериментах является наилучшей иллюстрацией предложенных в диссертации новых методов обучения с управляемой селективностью *Relevance Feature Machine – RFM* (метод релевантных признаков) и *Support Feature Machine – SFM* (метод опорных признаков).

С одной стороны, преимущество модельных экспериментов заключается в том, что они дают возможность придать условным понятиям полезных и лишних признаков абсолютный смысл, и позволяют для каждой конкретной гиперплоскости непосредственно вычислить вероятность ошибки распознавания на генеральной совокупности. С другой стороны, это позволяет сравнивать методы в заведомо разных условиях порождения исходных данных,

скрытых для наблюдения в прикладных задачах и выраженных в знании истинного положения искомой гиперплоскости и особенностях распределений признакового описания объектов.

Именно второй аспект организации экспериментального исследования заслуживает особого внимания. Структура четырех модельных задач настоящей диссертации определяет фактически четыре простых требования к селективному обучению:

а) эффективное подавление полностью шумовых попарно независимых признаков;
б) эффективное подавление полностью шумовых признаков, имеющих значительную линейную зависимость между собой;
в) эффективное выделение группы информативных линейно независимых признаков;
г) эффективное выделение группы информативных признаков, только совместное участие которых в решении может обеспечить приемлемую точность распознавания.

Результаты экспериментов, изложенные в диссертации, наглядно иллюстрируют весьма важный факт, что ни один из существующих методов селективного обучения, включая предлагаемые, не удовлетворяет сразу всем этим весьма неизощренным требованиям к селективному обучению. Предложенный метод SFM хорошо справился с подавлением шумовых признаков (а,б) при линейно независимых информативных признаках (в), существенно улучшив и без того неплохой результат существующего метода DrSVM. Однако метод RFM в этих условиях оказался далеко не так эффективен. Требование выделять группу информативных признаков, которые только вместе обеспечивают достаточную точность распознавания (г), оказалось весьма проблематичным, как для существующего DrSVM, так и предложенного SFM, селективность которых обеспечивается за счет использования в функции штрафа модуля. Вместе с тем требование (г) не доставило никаких слож-

ностей для второго предложенного метода RFM, селективность которого имеет отличную от привычного модуля природу.

Полученные в рамках диссертационной работы результаты экспериментального исследования иллюстрируют полезность предложенных методов обучения в сравнении с известными образцами.

ЗАКЛЮЧЕНИЕ

Основные результаты диссертации:

1. Сформулирована общая математическая постановка задачи обучения двухклассовому распознаванию образов в линейном пространстве признаков на основе количественного измерения расстояния между вектором признаков объекта и разделяющей гиперплоскостью.

2. Сформулирована байесовская постановка задачи обучения двухклассовому распознаванию образов в линейном пространстве признаков, дающая вероятностное обоснование метода опорных векторов.

3. Сформулированы две новые частные априорные вероятностные модели генеральной совокупности в терминах совместного априорного распределения компонент направляющего вектора разделяющей гиперплоскости.

4. Разработаны два метода и два соответствующих алгоритма апостериорного оценивания разделяющей гиперплоскости, реализующих байесовский принцип обучения с заданной селективностью отбора признаков при выбранных априорных предположениях о генеральной совокупности объектов двух классов. Разработанные методы получили названия метод релевантных признаков и метод опорных признаков с регулируемой селективностью.

5. Теоретически исследованы преимущества предложенных двух критериев селективного обучения метода релевантных признаков и метода опорных признаков по сравнению с известными критериями Elastic Net SVM и Lasso SVM.

6. Экспериментально получены количественные оценки эффективности предложенных методов обучения на модельных данных и данных задачи распознавания рака легких.

ПРИЛОЖЕНИЕ: ДОКАЗАТЕЛЬСТВА ТЕОРЕМ

Теоремы главы 2

Доказательство теоремы 1

Подставляя (34) в (39), и учитывая (37), получим критерий обучения

$$(\hat{\mathbf{a}},\hat{b}\,|\,X,Y;c)=\arg\max_{\mathbf{a}\in\mathbb{R}^n,\,b\in\mathbb{R}}\left[\ln\Psi(\mathbf{a})+\sum\nolimits_{j=1}^{N}\ln\varphi(\mathbf{x}_j\,|\,y_j,\mathbf{a},b;c)\right]=$$

$$\arg\min_{\mathbf{a}\in\mathbb{R}^n,\,b\in\mathbb{R}}\left\{-\ln\Psi(\mathbf{a})+\underbrace{c\left[\sum\nolimits_{j:\,y_j(\mathbf{a}^T\mathbf{x}_j+b)<1}\left(1-y_j(\mathbf{a}^T\mathbf{x}_j+b)\right)\right]}_{\substack{\textit{неправильно классифицированные}\\ \textit{объекты обучающей совокупности}}}\right\}.$$

Произведем замену переменных $\delta_j=1-y_j(\mathbf{a}^T\mathbf{x}_j+b)$, сумма которых для неправильно классифицированных объектов при текущих значениях $\mathbf{a}$ и b, подлежит минимизации вместе с $-\ln\Psi(\mathbf{a})$, т.е. $-\ln\Psi(\mathbf{a})+c\sum_{j:\,y_j(\mathbf{a}^T\mathbf{x}_j+b)<1}\delta_j\to\min(\mathbf{a},b)$. Для остальных объектов, правильно классифицируемых при значениях $\mathbf{a}$ и b, сумма $\sum_{j:\,y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge1}\delta_j=\sum_{j:\,y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge1}\left(1-y_j(\mathbf{a}^T\mathbf{x}_j+b)\right)$ в критерий не входит, что позволяет записать его в виде

$$-\ln\Psi(\mathbf{a})+c\sum_{j:\,y_j(\mathbf{a}^T\mathbf{x}_j+b)<1}\delta_j+\sum_{j:\,y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge1}0\to\min\left(\mathbf{a},b,\delta_j,\ j:y_j(\mathbf{a}^T\mathbf{x}_j+b)<1\right).$$

Но для неправильно классифицированных объектов $\delta_j\ge0$, т.е. $y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge1-\delta_j$. Если рассматривать эти неравенства как дополнительные ограничения, то минимизации подлежит критерий $-\ln\Psi(\mathbf{a})+c\sum_{j=1}^{N}\delta_j$, совпадающий с (40).

Теорема доказана. ■

Доказательство теоремы 2

Выпишем плотность (34) для параметров разделяющей гиперплоскости $(\hat{\mathbf{a}},\hat{b})$ для первого класса $y=1$

$$\varphi(\mathbf{x}\mid y=1;\hat{\mathbf{a}},\hat{b},c)=const\begin{cases}1,\ \hat{\mathbf{a}}^T\mathbf{x}+\hat{b}\geq 1,\\ \exp\left[-c\left(1-(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right],\ \hat{\mathbf{a}}^T\mathbf{x}+\hat{b}<1,\ c>0,\end{cases} \tag{111}$$

для второго $y=-1$

$$\varphi(\mathbf{x}\mid y=-1;\hat{\mathbf{a}},\hat{b},c)=const\begin{cases}1,\ \hat{\mathbf{a}}^T\mathbf{x}+b\leq -1,\\ \exp\left[-c\left(1+(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right],\ \hat{\mathbf{a}}^T\mathbf{x}+\hat{b}>-1,\ c>0.\end{cases} \tag{112}$$

Пусть $\hat{\mathbf{a}}^T\mathbf{x}+\hat{b}>1$, тогда подставив (111) и (112) в выражение для апостериорной вероятности (42), получим

$$p(y=1|\mathbf{x};\hat{\mathbf{a}},\hat{b},c)=1-p(y=-1|\mathbf{x};\hat{\mathbf{a}},\hat{b},c)=\frac{1}{1+\exp\left[-c\left(1+(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right]}=$$
$$\left\{1+\exp\left[-c\left(1+(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right]\right\}^{-1}=\left\{1+\exp(-c)\exp\left[-c(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right]\right\}^{-1}. \tag{113}$$

Пусть $\hat{\mathbf{a}}^T\mathbf{x}+\hat{b}<-1$, получим

$$p(y=1|\mathbf{x};\hat{\mathbf{a}},\hat{b},c)=1-p(y=-1|\mathbf{x};\hat{\mathbf{a}},\hat{b},c)=\frac{\exp\left[-c\left(1-(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right]}{\exp\left[-c\left(1-(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right]+1}=$$
$$\frac{1}{1+\exp\left[-c\left(1-(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right]^{-1}}=\left\{1+\exp(c)\exp\left[-c(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right]\right\}^{-1}. \tag{114}$$

Пусть $-1\leq\hat{\mathbf{a}}^T\mathbf{x}+\hat{b}\leq 1$, тогда

$$p(y=1|\mathbf{x};\hat{\mathbf{a}},\hat{b},c)=1-p(y=-1|\mathbf{x};\hat{\mathbf{a}},\hat{b},c)=$$
$$\frac{\exp\left[-c\left(1-(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right]}{\exp\left[-c\left(1-(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right]+\exp\left[-c\left(1+(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right]}=$$
$$\frac{1}{1+\exp\left[-c\left(1+(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right]\exp\left[c\left(1-(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right)\right]}=$$
$$\left\{1+\exp\left[-2c(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})\right]\right\}^{-1}. \tag{115}$$

Собрав вместе выражения (113)-(115) получим (43).

Теорема доказана. ■

Теоремы главы 3

Доказательство теоремы 3

Запишем функцию Лагранжа для критерия обучения (50):

$$L(\mathbf{a},b,\boldsymbol{\delta},\boldsymbol{\lambda},\boldsymbol{\mu}\,|\,\mathbf{r},C)=\frac{1}{2}\left(\sum_{i=1}^{n}\frac{1}{r_i}a_i^2+C\sum_{j=1}^{N}\delta_j\right)-$$
$$\sum_{j=1}^{N}\lambda_j\left[y_j\left(\sum_{i=1}^{n}a_i x_{ij}+b\right)-1+\delta_j\right]-\sum_{j=1}^{N}\mu_j\delta_j\to\begin{cases}\min(\mathbf{a},b,\delta_j,\ j=1,...,N),\\ \max(\lambda_j\geq 0,\mu_j\geq 0,\ j=1,...,N).\end{cases} \tag{116}$$

Решением задачи обучения (50) является седловая точка функции Лагранжа (116).

Найдем частные производные функции Лагранжа по a_i, $i=1,...,n$, и приравняем их нулю:

$$\frac{\partial}{\partial a_i}L(\mathbf{a},b,\boldsymbol{\delta},\boldsymbol{\lambda},\boldsymbol{\mu}\,|\,\mathbf{r},C)=\frac{1}{r_i}a_i-\sum_{j=1}^{N}\lambda_j y_j x_{ij}=0\Rightarrow a_i=r_i\sum_{j:\lambda_j>0}\lambda_j y_j x_{ij}\ . \tag{117}$$

Это выражение отличается от $a_i=\sum_{j:\lambda_j>0}y_j\lambda_j x_{ij}$ традиционного метода опорных векторов тем, что каждая i-я компонента направляющего вектора, по-прежнему являясь линейной комбинацией значений i-го признака опорных объектов, вычисляется с весом, равным ее априорной дисперсии r_i.

Найдем частные производные функции Лагранжа по b

$$\frac{\partial}{\partial b}L(\mathbf{a},b,\boldsymbol{\delta},\boldsymbol{\lambda},\boldsymbol{\mu}\,|\,\mathbf{r},C)=\sum_{j=1}^{N}\lambda_j y_j=0\,. \tag{118}$$

Найдем частные производные функции Лагранжа по δ_j, $j=1,...,N$

$$\frac{\partial}{\partial\delta_j}L(\mathbf{a},b,\boldsymbol{\delta},\boldsymbol{\lambda},\boldsymbol{\mu}\,|\,\mathbf{r},C)=\frac{1}{2}C-\lambda_j-\mu_j=0\Rightarrow 0\leq\lambda_j\leq C/2\,. \tag{119}$$

Подставляя (117), (118) и (119) в (116) получим двойственную задачу обучения:

$$\begin{cases} W(\lambda_1,...,\lambda_N)=\sum\limits_{j=1}^{N}\lambda_j-\dfrac{1}{2}\sum\limits_{j=1}^{N}\sum\limits_{l=1}^{N}y_j y_l\left(\sum\limits_{i=1}^{n}r_i x_{ij}x_{il}\right)\lambda_j\lambda_l\to\max(\lambda_1,...,\lambda_N), \\ \sum\limits_{j=1}^{N}y_j\lambda_j=0,\ 0\le\lambda_j\le C/2,\ \ j=1,...,N. \end{cases} \tag{120}$$

Определим оптимальное значение порога b. Для положительных значений множителей Лагранжа $\hat{\lambda}_j>0$, являющихся решением задачи (120), выполняется равенство $y_j\left(\sum_{i=1}^{n}a_i x_{ij}+b\right)=1-\delta_j$. Кроме того, если выполняется условие $0<\hat{\lambda}_j<C/2$, то соответствующие значение сдвига в точке минимума задачи (50) равны нулю $\delta_j=0$, тогда $y_j\left(\sum_{i=1}^{n}a_i x_{ij}+b\right)=1$. Домножим это выражение на $\hat{\lambda}_j y_j$ и просуммируем по всем $0<\hat{\lambda}_j<C/2$, учитывая, что $y_j^2=1$

$$\sum_{0<\hat{\lambda}_j<C/2}\hat{\lambda}_j\left(\sum_{i=1}^{n}a_i x_{ij}+b\right)=\sum_{0<\hat{\lambda}_j<C/2}\hat{\lambda}_j y_j .$$

Преобразуем полученное выражение

$$\sum_{0<\hat{\lambda}_j<C/2}\hat{\lambda}_j\sum_{i=1}^{n}a_i x_{ij}+b\sum_{0<\hat{\lambda}_j<C/2}\hat{\lambda}_j-\sum_{0<\hat{\lambda}_j<C/2}\hat{\lambda}_j y_j=0.$$

Откуда выразим b

$$b=\frac{-\sum\limits_{0<\hat{\lambda}_j<C/2}\hat{\lambda}_j\sum\limits_{i=1}^{n}a_i x_{ij}+\sum\limits_{0<\hat{\lambda}_j<C/2}\hat{\lambda}_j y_j}{\sum\limits_{0<\hat{\lambda}_j<C/2}\hat{\lambda}_j} \tag{121}$$

Учитывая (118) получим

$$\sum_{j=1}^{N}\lambda_j y_j=\sum_{0<\hat{\lambda}_j<C/2}\hat{\lambda}_j y_j+\sum_{\hat{\lambda}_j=C/2}\hat{\lambda}_j y_j+\underbrace{\sum_{\hat{\lambda}_j=0}\hat{\lambda}_j y_j}_{=0}=0.$$

Откуда

$$\sum_{0<\hat{\lambda}_j<C/2}\hat{\lambda}_j y_j=-\sum_{\hat{\lambda}_j=C/2}\hat{\lambda}_j y_j . \tag{122}$$

Подставив (117) и (122) в (121), получим

$$\hat{b} = -\frac{\sum_{j:0<\hat{\lambda}_j<C/2}\hat{\lambda}_j \sum_{l:\hat{\lambda}_l>0} y_l\hat{\lambda}_l \sum_{i=1}^{n} r_i x_{il} x_{ij} + (C/2)\sum_{j:\hat{\lambda}_j=C/2} y_j}{\sum_{j:0<\hat{\lambda}_j<C/2}\hat{\lambda}_j}.$$

Теорема доказана. ■

Доказательство теоремы 4

Подставляя (34) в (59) получим критерий обучения

$$(\hat{\mathbf{a}},\hat{b},\hat{\mathbf{r}}\,|\,X,Y;c) = \arg\max_{\mathbf{a}\in\mathbb{R}^n,\, b\in\mathbb{R},\, \mathbf{r}\in\mathbb{R}} \left[\ln\Psi(\mathbf{a}|\mathbf{r}) + \ln G(\mathbf{r}|\alpha,\beta) + \sum_{j=1}^{N} \ln\varphi(\mathbf{x}_j\,|\,y_j,\mathbf{a},b;c) \right] =$$

$$\arg\min_{\mathbf{a}\in\mathbb{R}^n,\, b\in\mathbb{R},\, \mathbf{r}\in\mathbb{R}} \left\{ -\ln\Psi(\mathbf{a}|\mathbf{r}) - \ln G(\mathbf{r}|\alpha,\beta) + \underbrace{c\left[\sum_{j:\, y_j(\mathbf{a}^T\mathbf{x}_j+b)<1} \left(1 - y_j(\mathbf{a}^T\mathbf{x}_j+b)\right) \right]}_{\substack{\textit{неправильно классифицированные} \\ \textit{объекты обучающей совокупности}}} \right\}.$$

Произведем замену переменных $\delta_j = 1 - y_j(\mathbf{a}^T\mathbf{x}_j + b)$, сумма которых для неправильно классифицированных объектов при текущих значениях $\mathbf{a}$ и b, подлежит минимизации вместе с $-\ln\Psi(\mathbf{a}|\mathbf{r}) - \ln G(\mathbf{r}|\alpha,\beta)$, т.е. $-\ln\Psi(\mathbf{a}|\mathbf{r}) - \ln G(\mathbf{r}|\alpha,\beta) + c\sum_{j:\, y_j(\mathbf{a}^T\mathbf{x}_j+b)<1}\delta_j \to \min(\mathbf{a},b,\mathbf{r})$. Для остальных объектов, правильно классифицируемых при значениях $\mathbf{a}$ и b, сумма $\sum_{j:\, y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1}\delta_j = \sum_{j:\, y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1}\left(1 - y_j(\mathbf{a}^T\mathbf{x}_j+b)\right)$ в критерий не входит, что позволяет записать его в виде

$$-\ln\Psi(\mathbf{a}|\mathbf{r}) - \ln G(\mathbf{r}|\alpha,\beta) + c\sum_{j:\, y_j(\mathbf{a}^T\mathbf{x}_j+b)<1}\delta_j + \sum_{j:\, y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1} 0 \to \min\left(\mathbf{a},b,\delta_j,\, j: y_j(\mathbf{a}^T\mathbf{x}_j+b)<1\right).$$

Но для неправильно классифицированных объектов $\delta_j \ge 0$, т.е. $y_j(\mathbf{a}^T\mathbf{x}_j + b) \ge 1 - \delta_j$. Если рассматривать эти неравенства как дополнительные ограничения, то минимизации подлежит критерий

$$\begin{cases} -\ln\Psi(\mathbf{a}|\mathbf{r}) - \ln G(\mathbf{r}|\alpha,\beta) + c\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}), \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \ge 1 - \delta_j,\ \delta_j \ge 0,\ j = 1,\ldots,N. \end{cases}$$

Учитывая (49) получим

$$\begin{cases}(1/2)\sum_{i=1}^{n}\ln r_i+(1/2)\sum_{i=1}^{n}(1/r_i)a_i^2-\ln G(\mathbf{r}|\alpha,\beta)+c\sum_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}),\\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,...,N.\end{cases}$$

Остается подставить в критерий выражение (57). Запишем его логарифм.

$$\ln G(\mathbf{r}\,|\,\alpha,\beta)\propto(\alpha-1)\sum_{i=1}^{n}\ln(1/r_i)-\beta\sum_{i=1}^{n}1/r_i=-(\alpha-1)\sum_{i=1}^{n}\ln r_i-\beta\sum_{i=1}^{n}1/r_i\ .$$

$$G(\mathbf{r}\,|\,\alpha,\beta)=G(1/r_1,...,1/r_n\,|\,\alpha,\beta)=\prod\nolimits_{i=1}^{n}\gamma(1/r_i\,|\,\alpha,\beta)=$$
$$\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^{n}\left(\prod_{i=1}^{n}(1/r_i)\right)^{\alpha-1}\exp\left(-\beta\sum_{i=1}^{n}(1/r_i)\right),$$

Подставим его в критерий:

$$\begin{cases}(1/2)\sum_{i=1}^{n}\ln r_i+(1/2)\sum_{i=1}^{n}(1/r_i)a_i^2-(\alpha-1)\sum_{i=1}^{n}\ln(1/r_i)+\beta\sum_{i=1}^{n}1/r_i+c\sum_{j=1}^{N}\delta_j\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\to\min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}),\\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,...,N.\end{cases}$$

После элементарных преобразований получим критерий (60):

$$\begin{cases}\sum\nolimits_{i=1}^{n}\left[(1/r_i)\left(a_i^2+2\beta\right)+\left(2\alpha-1\right)\ln r_i\right]+C\sum\nolimits_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}),\\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,...,N,\end{cases}$$

где $C=2c$.

Теорема доказана. ■

Доказательство теоремы 5

Подставляя (34) в (64) и выполнив эквивалентные преобразования, аналогичные доказательству на с. 100 теоремы 4, получим критерий

$$\begin{cases}-\ln\Psi(\mathbf{a}|\mathbf{r})-\ln G(\mathbf{r}|\mu)+c\sum_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}),\\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,...,N.\end{cases}$$

Учитывая (49) получим

$$\begin{cases}(1/2)\sum_{i=1}^{n}\ln r_i+(1/2)\sum_{i=1}^{n}(1/r_i)a_i^2-\ln G(\mathbf{r}|\mu)+c\sum_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}),\\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,...,N.\end{cases}\qquad(123)$$

Остается подставить в критерий выражение (63). Запишем его логарифм

$$\ln G(\mathbf{r}|\mu)=-\frac{1+\mu^2}{2\mu}\ln r_i-(1/2\mu)\sum_{i=1}^{n}(1/r_i).$$

Подставим это выражение в критерий (123)

$$\begin{cases}\frac{1}{2}\sum_{i=1}^{n}\ln r_i+\frac{1}{2}\sum_{i=1}^{n}(1/r_i)a_i^2+\frac{1+\mu^2}{2\mu}\ln r_i+\frac{1}{2\mu}\sum_{i=1}^{n}(1/r_i)+c\sum_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}),\\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,...,N.\end{cases}$$

Умножим целевую функцию на 2 и преобразуем ее

$$\begin{cases}\sum_{i=1}^{n}(1/r_i)a_i^2+\sum_{i=1}^{n}(1/r_i)(1/\mu)+\big((1/\mu)+\mu+1\big)\ln r_i+2c\sum_{j=1}^{N}\delta_j\to\min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}),\\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,...,N.\end{cases}$$

Простейшее преобразования приводят этот критерий к виду критерия (65).

Теорема доказана. ■

Доказательство теоремы 6

Запишем функцию Лагранжа для критерия (65)

$$L(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta},\boldsymbol{\lambda},\boldsymbol{\pi},\mathbf{v}\,|\,C,\mu)=$$
$$\frac{1}{2}\sum\nolimits_{i=1}^{n}\Big[(1/r_i)\big(a_i^2+(1/\mu)\big)+\big((1/\mu)+1+\mu\big)\ln r_i\Big]+C\sum\nolimits_{j=1}^{N}\delta_j-$$
$$\sum\nolimits_{j=1}^{N}\lambda_j\Big[y_j\Big(\sum\nolimits_{i=1}^{n}a_ix_{ij}+b\Big)-1+\delta_j\Big]-\sum\nolimits_{j=1}^{N}\pi_j\delta_j-$$
$$\sum\nolimits_{i=1}^{n}v_ir_i\to\begin{cases}\min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}),\\ \max(\boldsymbol{\lambda}\ge\mathbf{0},\boldsymbol{\pi}\ge\mathbf{0},\mathbf{v}\ge\mathbf{0}).\end{cases}$$

Найдем частные производные функции Лагранжа по $r_i,\ i=1,...,n$ и приравняем их к нулю

$$\frac{\partial}{\partial r_i}L(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta},\boldsymbol{\lambda},\boldsymbol{\pi},\mathbf{v}\,|\,C,\mu)=\frac{1}{2}\Big[-\big(a_i^2+(1/\mu)\big)\frac{1}{r_i^2}+\big((1/\mu)+1+\mu\big)\frac{1}{r_i}\Big]-v_i=$$
$$\frac{1}{2}\Big[-\big(a_i^2+(1/\mu)\big)\frac{1}{r_i^2}+\big((1/\mu)+1+\mu\big)\frac{1}{r_i}\Big]-v_i=0$$

Учитывая, что для $r_i > 0 \Rightarrow v_i = 0$, получим

$$-\left(a_i^2+(1/\mu)\right)\frac{1}{r_i^2}+\left((1/\mu)+1+\mu\right)\frac{1}{r_i}=0.$$

После элементарных преобразований получим

$$r_i=\frac{a_i^2+(1/\mu)}{(1/\mu)+1+\mu}=\frac{\mu a_i^2+1}{1+\mu+\mu^2}. \qquad (124)$$

Заметим, что для любого $\mu \ge 0 \Rightarrow r_i > 0$. Подставим это выражение в исходный критерий (65)

$$\begin{cases} \sum_{i=1}^{n}\left[\frac{(1/\mu)+1+\mu}{a_i^2+(1/\mu)}\left(a_i^2+(1/\mu)\right)+\left((1/\mu)+1+\mu\right)\ln\frac{a_i^2+(1/\mu)}{(1/\mu)+1+\mu}\right]+ \\ \qquad\qquad C\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a},b,\boldsymbol{\delta}\,|\,C,\mu), \\ y_j\left(\sum_{i=1}^{n}a_i x_{ij}+b\right)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,\ldots,N. \end{cases}$$

Преобразуем критерий

$$\begin{cases} \left((1/\mu)+1+\mu\right)\sum_{i=1}^{n}\ln\frac{a_i^2+(1/\mu)}{(1/\mu)+1+\mu}+C\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a},b,\boldsymbol{\delta}\,|\,C,\mu), \\ y_j\left(\sum_{i=1}^{n}a_i x_{ij}+b\right)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,\ldots,N, \end{cases}$$

и в итоге получим критерий вида

$$\begin{cases} \sum_{i=1}^{n}\ln(a_i^2+1/\mu)+\frac{C}{1/\mu+1+\mu}\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a},b,\boldsymbol{\delta}\,|\,C,\mu), \\ y_j\left(\sum_{i=1}^{n}a_i x_{ij}+b\right)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,\ldots,N,\ \ i=1,\ldots,n. \end{cases} \qquad (125)$$

Введем новое обозначение

$$C'=\frac{C}{(1/\mu)+1+\mu}. \qquad (126)$$

С учетом (126) критерий (125) примет вид критерия (66), что и требовалось доказать.

Теорема доказана. ■

Доказательство теоремы 7

Согласно условиям теоремы на очередной итерации k получено очередной приближение значений весов $(r_1^k,...,r_n^k)$. Оптимизация (65) по $(a_1,...,a_n,b,\delta_1,,...,\delta_N)$ сводится к минимизации обобщенного критерия опорных векторов со взвешенными признаками (50):

$$\begin{cases}(a_1^k,...,a_n^k,b^k,\delta_1^k,...,\delta_N^k)=\underset{a_1,...,a_n,b,\delta_1,...,\delta_N}{\arg\min}\left(\sum_{i=1}^{n}\frac{1}{r_i^k}a_i^2+C\sum_{j=1}^{N}\delta_j\right),\\ y_j\left(\sum_{i=1}^{n}a_i x_{ij}+b\right)\geq 1-\delta_j,\ \delta_j\geq 0,\ j=1,...,N.\end{cases} \tag{127}$$

Для найденного приближения к искомому направляющему вектору $(a_1^k,...,a_n^k)$ оптимизация критерия (65) по $(r_1,...,r_n)$ сводится к решению n независимых задач:

$$r_i^{k+1}=\underset{r_i}{\arg\min}\left[(1/r_i)\left((a_i^k)^2+(1/\mu)\right)+\left((1/\mu)+1+\mu\right)\ln r_i\right],\ i=1,...,n.$$

Приравнивание производной по r_i к нулю дает условие оптимальности

$$\frac{d}{dr_i}\left[(1/r_i)\left((a_i^k)^2+(1/\mu)\right)+\left((1/\mu)+1+\mu\right)\ln r_i\right]=$$
$$-\frac{(a_i^k)^2+(1/\mu)}{r_i^2}+\frac{(1/\mu)+1+\mu}{r_i}=-\frac{1}{r_i}\left(\frac{(a_i^k)^2+(1/\mu)}{r_i}-(1/\mu)-1-\mu\right)=0.$$

Здесь существенным является только условие равенства нулю выражения в скобках, поскольку побочное условие $r_i\to\infty$ вытекает из того, что положительная производная целевой функции в пределе уменьшается до нуля при достаточно больших значениях r_i, и целевая функция в целом имеет только одну точку минимума. Как следствие, очередные значения дисперсий компонент направляющего вектора определяются отношениями

$$r_i^{k+1}=\frac{(a_i^k)^2+1/\mu}{\mu+1+1/\mu}=\frac{\mu(a_i^k)^2+1}{\mu^2+\mu+1}. \tag{128}$$

Подставив в (128) очередное приближение направляющего вектора, определяемого выражением (51) для $(\lambda_1^k,...,\lambda_N^k)$, получим (67).

Теорема доказана. ■

Доказательство леммы 1

Первые два утверждения основаны на дифференцируемости функции $L_i(a_i,\boldsymbol{\lambda}\,|\,\mu)$ (76) во всех точках $a_i \neq 0$. Третье утверждение следует из определения субдифференциала и вида функции $L_i(a_i,\boldsymbol{\lambda}\,|\,\mu)$.

Если $\mu = 0$, то субдифференциал содержит единственный субградиент, который равен $\partial_{a_i} L_i(a_i,\boldsymbol{\lambda}\,|\,\mu) = -\sum_{j=1}^{N} y_j x_{ij}\lambda_j$. Если $\mu > 0$, то субдифференциал $\partial_{a_i} L_i(a_i,\boldsymbol{\lambda}\,|\,\mu)$ это множество, причем каждый субградиент в точке $a_i = 0$ можно представить в виде суммы субградиента функции $(1/2)q(a_i\,|\,\mu)$ в этой точке и $-\sum_{j=1}^{N} y_j x_{ij}\lambda_j$, как единственного субградиента функции $-a_i\sum_{j=1}^{N} y_j x_{ij}\lambda_j$.

Учитывая, что $\mu > 0$, то в небольшой окрестности точки $a_i = 0$ функция $(1/2)q(a_i\,|\,\mu) = \mu|\,a_i\,|$, субградиент которой должен удовлетворять условию $\mu|\,a_i\,| \geq \mu d a_i$, откуда следует, что $-1 \leq d \leq 1$ или $|\,d\,| \leq 1$.

Лемма доказана. ■

Доказательство леммы 2

Необходимым и достаточным условием минимума выпуклой функции $L_i(a_i,\boldsymbol{\lambda}\,|\,\mu)$ (76) в точке $a_i \in \mathbb{R}$ является включение нуля в субдифференциал $0 \in \partial L_i(a_i\,|\,\boldsymbol{\lambda},\mu)$ этой функции в данной точке.

Учитывая лемму 1 возможны три варианта:

1. Предположим, что минимум функции $L_i(a_i,\boldsymbol{\lambda}\,|\,\mu)$ достигается в точке $|\tilde{a}_i| > \mu$, тогда имеем

$$\partial_{a_i} L_i(\tilde{a}_i,\boldsymbol{\lambda}\,|\,\mu) = \tilde{a}_i - \sum_{j=1}^{N} y_j x_{ij}\lambda_j = 0 \text{ или } \tilde{a}_i = \sum_{j=1}^{N} y_j x_{ij}\lambda_j .$$

Кроме того, учитывая $|\tilde{a}_i| > \mu$ получим

$$\left|\sum_{j=1}^{N} y_j x_{ij} \lambda_j\right| > \mu \text{ или } \sum_{j=1}^{N}\sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l > \mu^2.$$

Сама функция $L_i(a_i, \boldsymbol{\lambda} \mid \mu)$ будет иметь значение

$$L_i(\tilde{a}_i, \boldsymbol{\lambda} \mid \mu) = \frac{1}{2}\mu^2 - \frac{1}{2}\sum_{j=1}^{N}\sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l .$$

2. Предположим, что минимум функции $L_i(a_i, \boldsymbol{\lambda} \mid \mu)$ достигается в точке $0 < |a_i| \le \mu$, тогда имеем

$$\partial_{a_i} L_i(\tilde{a}_i, \lambda_1, ..., \lambda_N \mid \mu) = \mu(\tilde{a}_i / |\tilde{a}_i|) - \sum_{j=1}^{N} y_j x_{ij} \lambda_j = 0 \text{ или } \sum_{j=1}^{N}\sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l = \mu^2$$

При этом достаточное условие выполняется в точке $\tilde{a}_i = \eta_i \sum_{j=1}^{N} y_j x_{ij} \lambda_j$ для любого $\eta \in [0,1]$. Сама функция $L_i(\tilde{a}_i, \boldsymbol{\lambda} \mid \mu) = 0$ будет иметь значение нуль.

3. Предположим, что минимум достигается в точке $\tilde{a}_i = 0$, тогда получим

$$\mu d_i = \sum_{j=1}^{N} \lambda_j y_j x_{ij} \text{ для } |d_i| \le 1 \text{ или } \sum_{j=1}^{N}\sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l \le \mu^2.$$

Функция $L_i(\tilde{a}_i, \boldsymbol{\lambda} \mid \mu) = 0$ в этой точке ровна нулю.

В общем виде имеем

$$\min_{a_i \in \mathbb{R}} L_i(a_i, \boldsymbol{\lambda} \mid \mu) = \frac{1}{2}\min\left\{0,\ \mu^2 - \sum_{j=1}^{N}\sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l\right\}.$$

Лемма доказана. ■

Доказательство теоремы 8

Из (74) двойственная задача в самом общем виде примет вид

$$\begin{cases} W(\boldsymbol{\lambda}, \boldsymbol{\pi} \mid \mu, C) = \min\limits_{\substack{a_1,...,a_n,b,\\ \delta_1,...,\delta_N}} L(a_i, i \in I, b, \boldsymbol{\delta}, \boldsymbol{\lambda}, \boldsymbol{\pi} \mid \mu, C) \to \max(\boldsymbol{\lambda}, \boldsymbol{\pi}), \\ \lambda_j \ge 0,\ \pi_j \ge 0,\ j = 1, ..., N. \end{cases} \tag{129}$$

Условие минимума по b при фиксированных множителях Лагранжа:

$$\frac{\partial}{\partial b} L(a_i, i \in I, b, \boldsymbol{\delta}, \boldsymbol{\lambda}, \boldsymbol{\pi} \mid \mu) = -\sum_{j=1}^{N} y_j \lambda_j = 0,\ \sum_{j=1}^{N} y_j \lambda_j = 0. \tag{130}$$

Условия минимума по δ_j при фиксированных множителях Лагранжа:

$$\frac{\partial}{\partial \delta_j} L(a_i, i \in I, b, \boldsymbol{\delta}, \boldsymbol{\lambda}, \boldsymbol{\pi} \mid \mu) = \frac{\partial}{\partial \delta_j} \sum_{l=1}^{N} \left(\frac{C}{2} - \lambda_l - \pi_l \right) \delta_l = 0 \Rightarrow$$
$$\Rightarrow \frac{C}{2} - \lambda_j - \pi_j = 0 \Rightarrow 0 \le \lambda_j \le \frac{C}{2}, \;\; j = 1, ..., N \tag{131}$$

Подставив (130), (131) в функцию Лагранжа (77) и затем в (129) получим

$$\begin{cases} W(\boldsymbol{\lambda} \mid \mu, C) = \sum_{j=1}^{N} \lambda_j + \min\limits_{a_1, ..., a_n} \sum\limits_{i \in I} L_i(a_i, \boldsymbol{\lambda} \mid \mu) \to \max(\boldsymbol{\lambda}) \\ \sum_{j=1}^{N} y_j \lambda_j = 0, \;\; 0 \le \lambda_j \le \frac{C}{2}, \;\; j = 1, ..., N. \end{cases}$$

Или в эквивалентной форме

$$\begin{cases} W(\boldsymbol{\lambda} \mid \mu, C) = \sum_{j=1}^{N} \lambda_j + \sum\limits_{i \in I} \min\limits_{a_i} L_i(a_i, \boldsymbol{\lambda} \mid \mu) \to \max(\boldsymbol{\lambda}) \\ \sum_{j=1}^{N} y_j \lambda_j = 0, \;\; 0 \le \lambda_j \le \frac{C}{2}, \;\; j = 1, ..., N. \end{cases} \tag{132}$$

Учитывая лемму 2 получим задачу (80).

Введя вспомогательные переменные $\xi_i \ge 0, \; i \in I$, задачу (80) несложно записать в виде задачи максимизации линейной функции при квадратичных ограничениях (81).

Теорема доказана. ■

Доказательство теоремы 10

Запишем условия теоремы в виде выпуклой оптимизационной задачи

$$\begin{cases} V_{SFM}(\mathbf{a}, \varepsilon \mid C, \mu) = \dfrac{2\mu}{C\varepsilon} \sum\limits_{i:|a_i| \le \varepsilon\mu} |a_i| + \dfrac{1}{C} \sum\limits_{i:|a_i| > \varepsilon\mu} (\mu^2 + \varepsilon^{-2} a_i^2) \to \min(\mathbf{a} \mid \varepsilon, C, \mu), \\ \mathbf{a}^T \mathbf{a} = 1. \end{cases} \tag{133}$$

Заметим, что решение задачи (133) не зависит от значения $C > 0$. Домножим целевую функцию на значение параметра C

$$\begin{cases} \dfrac{2\mu}{\varepsilon} \sum\limits_{i:|a_i| \le \varepsilon\mu} |a_i| + \sum\limits_{i:|a_i| > \varepsilon\mu} (\mu^2 + \varepsilon^{-2} a_i^2) \to \min(\mathbf{a} \mid \varepsilon, \mu), \\ \mathbf{a}^T \mathbf{a} = 1. \end{cases}$$

Запишем функцию Лагранжа

$$L(\mathbf{a},\lambda\,|\,C,\mu,\varepsilon)=\frac{\mu}{\varepsilon}\sum_{i:|a_i|\le\varepsilon\mu}|\,a_i\,|+\frac{1}{2}\sum_{i:|a_i|>\varepsilon\mu}(\mu^2+\varepsilon^{-2}a_i^2)-\frac{1}{2\lambda}\left(\sum_{i=1}^{n}a_i^2-1\right)\to\begin{cases}\min(\mathbf{a}),\\ \max(\lambda).\end{cases}$$

Градиент функции Лагранжа по $|\,a_i\,|>\varepsilon\mu$

$$\partial_{a_i}L(\mathbf{a},\lambda\,|\,C,\mu,\varepsilon)=\varepsilon^{-2}a_i-\frac{1}{\lambda}a_i=a_i\left(\varepsilon^{-2}-\frac{1}{\lambda}\right)=0\Rightarrow\lambda=\varepsilon^2>0\,.$$

Градиент функции Лагранжа по $0<|\,a_i\,|\le\varepsilon\mu$

$$\partial_{a_i}L(\mathbf{a},\lambda\,|\,C,\mu,\varepsilon)=\frac{\mu}{\varepsilon}\frac{a_i}{|\,a_i\,|}-\frac{1}{\lambda}a_i=a_i\left(\frac{\mu}{\varepsilon}\frac{1}{|\,a_i\,|}-\frac{1}{\lambda}\right)=0\Rightarrow|\,a_i\,|=\mu\lambda/\varepsilon\Rightarrow0<\lambda\le\varepsilon^2\,.$$

Все компоненты делятся на три группы

1. $|\,a_i\,|>\varepsilon\mu$, $i\in I_{>\varepsilon\mu}$, тогда $\lambda=\varepsilon^2$, количество компонент $|\,I_{>\varepsilon\mu}\,|=n_{>\varepsilon\mu}$,

2. $0<|\,a_i\,|\le\varepsilon\mu$, $i\in I_{0\varepsilon\mu}$, тогда $|\,a_i\,|=\mu\lambda/\varepsilon$, $0<\lambda\le\varepsilon^2$, $|\,I_{0\varepsilon\mu}\,|=n_{0\varepsilon\mu}$,

3. $a_i=0$, $i\in I_0$, $|\,I_0\,|=n_0$, и $n_{>\varepsilon\mu}+n_{0\varepsilon\mu}+n_0=n$.

Кроме того, должно выполняться условие дополняющей нежесткости

$$\frac{1}{\lambda}\left(\sum_{i=1}^{n}a_i^2-1\right)=0\,,$$

Учитывая, что $\lambda>0$, имеем

$$\sum_{i=1}^{n}a_i^2=1\ \text{или}\ \sum_{i\in I_{>\varepsilon\mu}}a_i^2+\sum_{i\in I_{0\varepsilon\mu}}a_i^2=\sum_{i\in I_{>\varepsilon\mu}}a_i^2+n_{0\mu}(\mu\lambda/\varepsilon)^2=1.\qquad(134)$$

Рассмотрим два возможных случая:

1. Пусть $\varepsilon\mu\ge1$, тогда учитывая, что $\sum_{i=1}^{n}a_i^2=1\ge\sum_{i:|a_i|>\varepsilon\mu}a_i^2$, компонент $|\,a_i\,|>\varepsilon\mu\ge1$ в решении не будет и условие дополняющей нежесткости примет вид

$$n_{0\mu}(\mu\lambda/\varepsilon)^2=1\Rightarrow|\,a_i\,|=\mu\lambda/\varepsilon=n_{0\varepsilon\mu}^{-0.5}\le\varepsilon\mu\ \text{ и }\ n_{0\varepsilon\mu}\ge1.$$

Целевая функция в (133) имеет значение

$$\frac{2\mu}{C\varepsilon}\sum_{|a_i|\le\mu}|\,a_i\,|=\frac{2\mu}{C\varepsilon}\sum_{i\in I_{0\mu}}|\,a_i\,|=\frac{2\mu}{C\varepsilon}\sum_{i\in I_{0\mu}}n_{0\varepsilon\mu}^{-0.5}=\frac{2\mu}{C\varepsilon}n_{0\varepsilon\mu}^{0.5}\,.$$

Ее минимальное значение $2\mu/C\varepsilon \geq 0$ соответствует одной ненулевой компоненте $n_{0\varepsilon\mu}=1$, модуль которой равен

$$|a_i| = \mu\lambda/\varepsilon = n_{0\varepsilon\mu}^{-0.5} = 1.$$

2. Пусть $\varepsilon\mu < 1$, тогда в решении могут быть компоненты всех трех групп. Рассмотрим возможные два варианта.

- В решении нет компонент $|a_i| > \varepsilon\mu$, тогда, аналогично тому, как показано выше, условие дополняющей нежесткости примет вид

$$n_{0\mu}(\mu\lambda/\varepsilon)^2 = 1 \Rightarrow |a_i| = \mu\lambda/\varepsilon = n_{0\varepsilon\mu}^{-0.5}$$

Учитывая, что $|a_i| \leq \varepsilon\mu < 1$ имеем $n_{0\varepsilon\mu}^{-0.5} \leq \varepsilon\mu \Rightarrow n_{0\varepsilon\mu} \geq \varepsilon^{-2}\mu^{-2} > 1$. Целевая функция будет иметь значение $\frac{2\mu}{C\varepsilon} < \frac{2\mu}{C\varepsilon} n_{0\varepsilon\mu}^{0.5} \leq \frac{2\mu}{C\varepsilon} n^{0.5}$. Ее наименьшее значение равно $\frac{2\mu}{C\varepsilon}\lceil \varepsilon^{-1}\mu^{-1} \rceil$ при $n_{0\varepsilon\mu} = \lceil \varepsilon^{-2}\mu^{-2} \rceil$, заметим, что $\frac{2\mu}{C\varepsilon}\lceil \varepsilon^{-1}\mu^{-1} \rceil \geq \frac{2}{C\varepsilon^2}$. Если $\varepsilon^{-2}\mu^{-2} > n$, то в решении будут компоненты $|a_i| > \varepsilon\mu$.

- В решении есть компоненты $|a_i| > \varepsilon\mu$, т.е. $n_{>\varepsilon\mu} \geq 1$. Учитывая, что $\lambda = \varepsilon^2$, условие дополняющей нежесткости примет вид

$$\sum_{i \in I_{>\varepsilon\mu}} a_i^2 + \sum_{i \in I_{0\varepsilon\mu}} a_i^2 = \sum_{i \in I_{>\varepsilon\mu}} a_i^2 + \mu^2\varepsilon^2 n_{0\varepsilon\mu} = 1 \Rightarrow$$
$$\sum_{i \in I_{>\varepsilon\mu}} a_i^2 = 1 - \mu^2\varepsilon^2 n_{0\varepsilon\mu} > 0 \Rightarrow 0 \leq n_{0\mu} < \mu^{-2}\varepsilon^{-2}.$$

Целевая функция примет вид

$$\frac{2\mu}{C\varepsilon}\sum_{i:|a_i|\leq\varepsilon\mu}|a_i| + \frac{1}{C}\sum_{i:|a_i|>\varepsilon\mu}(\mu^2 + \varepsilon^{-2}a_i^2) = \frac{2\mu}{C\varepsilon}\sum_{i:|a_i|\leq\varepsilon\mu}|a_i| + \frac{1}{C}\mu^2 n_{>\varepsilon\mu} + \frac{\varepsilon^{-2}}{C}\sum_{i:|a_i|>\varepsilon\mu} a_i^2 =$$
$$\frac{2\mu}{C\varepsilon}\mu\varepsilon n_{0\varepsilon\mu} + \frac{1}{C}\mu^2 n_{>\varepsilon\mu} + \frac{1}{C\varepsilon^2}(1 - \mu^2\varepsilon^2 n_{0\varepsilon\mu}) =$$
$$\frac{1}{C}\left(2\mu^2 n_{0\varepsilon\mu} + \mu^2 n_{>\varepsilon\mu} + \frac{1}{\varepsilon^2} - \mu^2 n_{0\varepsilon\mu}\right) = \frac{1}{C}\left(\mu^2 n_{>\varepsilon\mu} + \mu^2 n_{0\varepsilon\mu} + \frac{1}{\varepsilon^2}\right).$$

Учитывая, что $n_{>\mu} \geq 1$, минимальное значение $(\mu^2 + \varepsilon^{-2})/C$ целевая функция принимает при $n_{>\mu} = 1$ и $n_{0\mu} = 0$. Из условия дополняющей нежесткости имеем величину модуля ненулевой компоненты $|a_i| = 1 > \varepsilon\mu$.

Сравним минимальные значения целевых функций $\frac{2\mu}{C\varepsilon}\left\lceil \varepsilon^{-1}\mu^{-1}\right\rceil \geq \frac{2}{C\varepsilon^2}$ и $(\mu^2+\varepsilon^{-2})/C$ в двух вариантах при $\varepsilon\mu<1$. Покажем, что справедливо неравенство $\frac{2}{C\varepsilon^2} > \frac{\mu^2+\varepsilon^{-2}}{C}$ и, следовательно, $\frac{2\mu}{C\varepsilon}\left\lceil \varepsilon^{-1}\mu^{-1}\right\rceil > \frac{\mu^2+\varepsilon^{-2}}{C}$:

$$\frac{2}{C\varepsilon^2}-\frac{\mu^2+\varepsilon^{-2}}{C}=\frac{1}{C\varepsilon^2}(2-\varepsilon^2\mu^2-1)=\frac{1}{C\varepsilon^2}(1-\varepsilon^2\mu^2)>0$$

Учитывая, что $\varepsilon\mu<1$ имеем $\varepsilon^2\mu^2<1$ и, следовательно, неравенство выполняется.

Таким образом, решением задачи (133) при $\varepsilon\mu<1$ является направляющий вектор с одной ненулевой компонентой $|a_i|=1>\varepsilon\mu$. Минимальное значение целевой функции равно $(\mu^2+\varepsilon^{-2})/C$.

Теорема доказана. ■

Доказательство теоремы 11

Запишем условия теоремы в виде выпуклой оптимизационной задачи

$$\begin{cases} V_{SFM}(\mathbf{a},\varepsilon\,|\,C,\mu)=-\dfrac{2\mu}{C\varepsilon}\sum\limits_{i:|a_i|\le\varepsilon\mu}|a_i|-\dfrac{1}{C}\sum\limits_{i:|a_i|>\varepsilon\mu}(\mu^2+\varepsilon^{-2}a_i^2)\to\min(\mathbf{a}\,|\,\varepsilon,C,\mu), \\ \mathbf{a}^T\mathbf{a}=1. \end{cases} \quad (135)$$

Заметим, что решение задачи не зависит от значения $C>0$. Домножим целевую функцию на значение параметра C

$$\begin{cases} -\dfrac{2\mu}{\varepsilon}\sum\limits_{i:|a_i|\le\varepsilon\mu}|a_i|-\sum\limits_{i:|a_i|>\varepsilon\mu}(\mu^2+\varepsilon^{-2}a_i^2)\to\min(\mathbf{a}\,|\,\varepsilon,\mu), \\ \mathbf{a}^T\mathbf{a}=1. \end{cases}$$

Запишем функцию Лагранжа

$$L(\mathbf{a},\lambda\,|\,C,\mu,\varepsilon)=-\frac{\mu}{\varepsilon}\sum_{i:|a_i|\le\varepsilon\mu}|a_i|-\frac{1}{2}\sum_{i:|a_i|>\varepsilon\mu}(\mu^2+\varepsilon^{-2}a_i^2)-\frac{1}{2\lambda}\left(\sum_{i=1}^{n}a_i^2-1\right)\to\begin{cases}\min(\mathbf{a}),\\ \max(\lambda).\end{cases}$$

Градиент функции Лагранжа по $|a_i|>\varepsilon\mu$

$$\partial_{a_i}L(\mathbf{a},\lambda\,|\,C,\mu,\varepsilon)=-\varepsilon^{-2}a_i-\frac{1}{\lambda}a_i=a_i\left(-\varepsilon^{-2}-\frac{1}{\lambda}\right)=0\Rightarrow\lambda=-\varepsilon^2<0. \quad (136)$$

Градиент функции Лагранжа по $0<|a_i|\le\varepsilon\mu$:

$$\partial_{a_i}L(\mathbf{a},\lambda\,|\,C,\mu,\varepsilon)=-\frac{\mu}{\varepsilon}\frac{a_i}{|a_i|}-\frac{1}{\lambda}a_i=a_i\left(-\frac{\mu}{\varepsilon}\frac{1}{|a_i|}-\frac{1}{\lambda}\right)=0\Rightarrow|a_i|=-\mu\lambda/\varepsilon\Rightarrow-\lambda\le\varepsilon^2.$$

Кроме того, $|a_i|=-\mu\lambda/\varepsilon\le\mu\Rightarrow-\lambda/\varepsilon\le1\Rightarrow-\lambda\le\varepsilon$ или $-\varepsilon\le\lambda<0$.

Все компоненты делятся на три группы

1. $|a_i|>\varepsilon\mu$, $i\in I_{>\varepsilon\mu}$, тогда $\lambda=-\varepsilon^2$, количество компонент $|I_{>\varepsilon\mu}|=n_{>\varepsilon\mu}$,

2. $0<|a_i|\le\varepsilon\mu$, $i\in I_{0\varepsilon\mu}$, тогда $|a_i|=-\mu\lambda/\varepsilon$, $-\varepsilon^2\le\lambda<0$, $|I_{0\varepsilon\mu}|=n_{0\varepsilon\mu}$,

3. $a_i=0$, $i\in I_0$, $|I_0|=n_0$, и $n_{>\varepsilon\mu}+n_{0\varepsilon\mu}+n_0=n$.

Кроме того, должно выполняться условие дополняющей нежесткости

$$\frac{1}{\lambda}\left(\sum_{i=1}^{n}a_i^2-1\right)=0,$$

Учитывая, что $\lambda<0$, то

$$\sum_{i=1}^{n}a_i^2=1 \text{ или } \sum_{i\in I_{>\varepsilon\mu}}a_i^2+\sum_{i\in I_{0\varepsilon\mu}}a_i^2=\sum_{i\in I_{>\varepsilon\mu}}a_i^2+n_{0\varepsilon\mu}(-\mu\lambda/\varepsilon)^2=1.$$

Рассмотрим три возможных случая

1. Пусть $\varepsilon\mu\ge1$, тогда учитывая, что $\sum_{i=1}^{n}a_i^2=1\ge\sum_{i:|a_i|>\varepsilon\mu}a_i^2$, компонент $|a_i|>\varepsilon\mu$ в решении не будет и условие дополняющей нежесткости примет вид

$$n_{0\varepsilon\mu}(-\mu\lambda/\varepsilon)^2=1\Rightarrow|a_i|=|\mu\lambda/\varepsilon|=n_{0\varepsilon\mu}^{-0.5}\le\varepsilon\mu \text{ и } \varepsilon^{-2}\mu^{-2}\le1\le n_{0\varepsilon\mu}\le n.$$

Целевая функция в (135) имеет значение

$$-\frac{2\mu}{C\varepsilon}\sum_{i\in I_{0\varepsilon\mu}}|a_i|=-\frac{2\mu}{C\varepsilon}n_{0\varepsilon\mu}|\mu\lambda/\varepsilon|=-\frac{2\mu}{C\varepsilon}n_{0\varepsilon\mu}n_{0\varepsilon\mu}^{-0.5}=-\frac{2\mu}{C\varepsilon}n_{0\varepsilon\mu}^{0.5},$$

Ее минимальное значение $-2\mu n^{0.5}/C\varepsilon\le0$, или максимальное $2\mu n^{0.5}/C\varepsilon\ge0$ для исходного критерия, соответствует $n_{0\varepsilon\mu}=n$. При этом модули всех компонент будут равны между собой

$$|a_i|=|\mu\lambda/\varepsilon|=n^{-0.5}\le\varepsilon\mu,\ i=1,...,n.$$

2. Пусть $n^{-0.5} \le \varepsilon\mu < 1$. Если в решении есть компоненты $|a_i| > \varepsilon\mu$, то $\lambda = -\varepsilon^2$ и $|a_i| = -\mu\lambda/\varepsilon = \varepsilon\mu$. Тогда не выполняется условие дополняющей нежескости

$$\sum_{i \in I_{>\varepsilon\mu}} a_i^2 + \sum_{i \in I_{0\varepsilon\mu}} a_i^2 = \sum_{i \in I_{>\varepsilon\mu}} a_i^2 + n_{0\varepsilon\mu}(-\mu\varepsilon)^2 > \mu^2\varepsilon^2 n \ge 1.$$

Следовательно, в решении есть только нулевые компоненты и $|a_i| \le \varepsilon\mu$. Тогда, как и в случае, рассмотренном выше, имеем

$$n_{0\varepsilon\mu}(-\mu\lambda/\varepsilon)^2 = 1 \Rightarrow |a_i| = |\mu\lambda/\varepsilon| = n_{0\varepsilon\mu}^{-0.5} \le \varepsilon\mu,$$

что справедливо, для $\varepsilon^{-2}\mu^{-2} \le 1 \le n_{0\varepsilon\mu} \le n$, где по условию $\varepsilon^{-2}\mu^{-2} < n$.

Целевая функция в (135) имеет значение

$$-\frac{2\mu}{C\varepsilon}\sum_{i \in I_{0\varepsilon\mu}} |a_i| = -\frac{2\mu}{C\varepsilon} n_{0\varepsilon\mu} |\mu\lambda/\varepsilon| = -\frac{2\mu}{C\varepsilon} n_{0\varepsilon\mu} n_{0\varepsilon\mu}^{-0.5} = -\frac{2\mu}{C\varepsilon} n_{0\varepsilon\mu}^{0.5},$$

Ее минимальное значение $-2\mu n^{0.5}/C\varepsilon \le 0$ соответствует $n_{0\varepsilon\mu} = n$.

3. Пусть $\varepsilon\mu < n^{-0.5}$, тогда в решении могут быть компоненты всех трех типов. Рассмотрим два возможных варианта.

- В решении нет компоенент $|a_i| > \varepsilon\mu$. Учитывая условие дополняющей нежесткости имеем

$$n_{0\varepsilon\mu}(-\mu\lambda/\varepsilon)^2 = 1 \Rightarrow |a_i| = |\mu\lambda/\varepsilon| = n_{0\varepsilon\mu}^{-0.5}.$$

Т.к. должно выполняться условие $|a_i| \le \varepsilon\mu$, а, следовательно, $|a_i| \le \varepsilon\mu < n^{-0.5}$, что неверно, т.к. $n_{0\varepsilon\mu}^{-0.5} \ge n^{-0.5}$. Т.е. в решении будут компоненты $|a_i| > \varepsilon\mu$.

- В решении есть компоненты $|a_i| > \varepsilon\mu$, т.е. $n_{>\varepsilon\mu} \ge 1$. Учитывая, что $\lambda = -\varepsilon^2$, компоненты $i : |a_i| \le \varepsilon\mu$ равны $|a_i| = \mu\varepsilon$. Учитывая по условию $\varepsilon\mu < n^{-0.5} \Rightarrow n < \varepsilon^{-2}\mu^{-2}$ условие дополняющей нежесткости примет вид

$$\sum_{i \in I_{>\varepsilon\mu}} a_i^2 + \sum_{i \in I_{0\varepsilon\mu}} a_i^2 = \sum_{i \in I_{>\varepsilon\mu}} a_i^2 + n_{0\varepsilon\mu}(-\mu\varepsilon)^2 = 1 \Rightarrow \sum_{i \in I_{>\varepsilon\mu}} a_i^2 = 1 - \mu^2\varepsilon^2 n_{0\varepsilon\mu} > 0,$$

$$\mu^2\varepsilon^2 n_{0\varepsilon\mu} < 1 \Rightarrow 0 \le n_{0\varepsilon\mu} < \mu^{-2}\varepsilon^{-2} \text{ и } n_{0\varepsilon\mu} \le n - 1.$$

Целевая функция примет вид

$$-\frac{2\mu}{C\varepsilon}\sum_{i:|a_i|\le\varepsilon\mu}|a_i|-\frac{1}{C}\sum_{i:|a_i|>\varepsilon\mu}(\mu^2+\varepsilon^{-2}a_i^2)=-\frac{2\mu}{C\varepsilon}\sum_{i:|a_i|\le\varepsilon\mu}|a_i|-\frac{1}{C}\mu^2 n_{>\varepsilon\mu}-\frac{\varepsilon^{-2}}{C}\sum_{i:|a_i|>\varepsilon\mu}a_i^2=$$

$$-\frac{2\mu}{C\varepsilon}\mu\varepsilon n_{0\varepsilon\mu}-\frac{1}{C}\mu^2 n_{>\varepsilon\mu}-\frac{\varepsilon^{-2}}{C}(1-\mu^2\varepsilon^2 n_{0\varepsilon\mu})=$$

$$-\frac{2\mu^2}{C}n_{0\varepsilon\mu}-\frac{1}{C}\mu^2 n_{>\varepsilon\mu}-\frac{\varepsilon^{-2}}{C}+\frac{1}{C}\mu^2 n_{0\varepsilon\mu}=$$

$$\frac{1}{C}\left(-2\mu^2 n_{0\varepsilon\mu}-\mu^2 n_{>\varepsilon\mu}-\frac{1}{\varepsilon^2}+\mu^2 n_{0\varepsilon\mu}\right)=\frac{1}{C}\left(-\mu^2 n_{0\varepsilon\mu}-\mu^2 n_{>\varepsilon\mu}-\frac{1}{\varepsilon^2}\right)$$

Ее минимум соответствует любому сочетанию $n_{>\varepsilon\mu}+n_{0\varepsilon\mu}=n$, учитывая $n_{>\varepsilon\mu}\ge 1$, при этом все компоненты $|a_i|\ge\varepsilon\mu$ и $\sum_{i=1}^{n}a_i^2=1$. Целевая функция принимает минимальное значение $-(\mu^2 n+\varepsilon^{-2})/C$.

Теорема доказана. ■

Доказательство теоремы 12

Запишем условия теоремы в виде выпуклой оптимизационной задачи

$$\begin{cases} V_{RFM}(\mathbf{a},\varepsilon\,|\,C,\mu)=C^{-1}\sum_{i=1}^{n}\ln(a_i^2/\varepsilon^2+1/\mu)\to\min(\mathbf{a}\,|\,\varepsilon,C,\mu), \\ \mathbf{a}^T\mathbf{a}=1. \end{cases} \tag{137}$$

Запишем функцию Лагранжа

$$L(\mathbf{a},\lambda\,|\,C,\varepsilon)=\frac{1}{2}C^{-1}\sum_{i=1}^{n}\ln(a_i^2/\varepsilon^2+1/\mu)-\frac{1}{2\lambda}\left(\sum_{i=1}^{n}a_i^2-1\right)\to\begin{cases}\min(\mathbf{a}),\\ \max(\lambda).\end{cases}$$

Градиент функции Лагранжа по $a_i\ne 0$

$$\partial_{a_i}L(\mathbf{a},\lambda)=C^{-1}\frac{a_i/\varepsilon^2}{a_i^2/\varepsilon^2+1/\mu}-\frac{1}{\lambda}a_i=0\Rightarrow a_i^2=\lambda/C-\varepsilon^2/\mu\Rightarrow\lambda>C\varepsilon^2/\mu.$$

Заметим, что все ненулевые компоненты равны между собой. Кроме того, должно выполняться условие дополняющей нежесткости

$$\frac{1}{\lambda}\left(\sum_{i=1}^{n}a_i^2-1\right)=\frac{1}{\lambda}\left(\sum_{i:\,a_i\ne 0}a_i^2-1\right)=0.$$

Учитывая, что $\lambda>0$ и $a_i^2=\lambda/C-\varepsilon^2/\mu$, если $a_i\ne 0$, получим

$$a_i^2=1/n_+ \text{ или } |a_i|=n_+^{-0.5}, \tag{138}$$

где n_+ - количество ненулевых компонент.

Подставим (138) в целевую функцию в (137)

$$C^{-1}\sum_{i=1}^{n}\ln(a_i^2/\varepsilon^2+1/\mu)=C^{-1}\sum_{i:\,a_i\neq 0}\ln(a_i^2/\varepsilon^2+1/\mu)+C^{-1}\sum_{i:\,a_i=0}\ln(1/\mu)=$$
$$C^{-1}n_+\ln(1/\varepsilon^2 n_++1/\mu)+C^{-1}(n-n_+)\ln(1/\mu).$$

Сама задача (137) примет эквивалентный вид

$$C^{-1}n_+\ln(1/\varepsilon^2 n_++1/\mu)+C^{-1}(n-n_+)\ln(1/\mu)\to\min(n_+).$$

Упростим целевую функцию

$$C^{-1}n_+\ln(1/\varepsilon^2 n_++1/\mu)-C^{-1}n_+\ln(1/\mu)\to\min(n_+).$$

Задача (137) примет эквивалентный вид

$$C^{-1}n_+\ln(\mu/\varepsilon^2 n_++1)\to\min(n_+). \qquad (139)$$

Покажем, что функция $n_+\ln(\mu/\varepsilon^2 n_++1)$ строго монотонно возрастает для $n_+\geq 1$. Рассмотрим поведение функции на непрерывном интервале $\tilde{n}_+\in[1,n]$ и найдем ее производную

$$\left[\tilde{n}_+\ln(\mu/\varepsilon^2\tilde{n}_++1)\right]'=[\tilde{n}_+\ln(\mu+\varepsilon^2\tilde{n}_+)-\tilde{n}_+\ln(\varepsilon^2\tilde{n}_+)]'=$$
$$\ln(\mu+\varepsilon^2\tilde{n}_+)+\tilde{n}_+\frac{\varepsilon^2}{\mu+\varepsilon^2\tilde{n}_+}-\ln(\varepsilon^2\tilde{n}_+)-\tilde{n}_+\frac{\varepsilon^2}{\varepsilon^2\tilde{n}_+}=$$
$$\underbrace{\ln(\mu+\varepsilon^2\tilde{n}_+)-\ln(\varepsilon^2\tilde{n}_+)}_{>0}+\underbrace{\frac{\varepsilon^2\tilde{n}_+}{\mu+\varepsilon^2\tilde{n}_+}}_{>0}-1>0.$$

Нетрудно видно, что производная строго больше нуля для $\mu>0$, а, следовательно, целевая функция (139) строго возрастает с ростом n_+. Решением задачи (139) будет $n_+=1$.

Учитывая (138) решением оптимизационной задачи (137) будут вектора

$$\forall\varepsilon\ \ \mathbf{a}\in\mathcal{A}_{\min}^{RFM}(\varepsilon)\leftrightarrow|a_i|\in\{0,1\},\ \sum_{i=1}^{n}|a_i|=1.$$

Значение целевой функции в точке минимума равно

$$\min_{\mathbf{a}'\in\mathcal{A}}V_{RFM}(\mathbf{a}',\varepsilon\,|\,C,\mu)=C^{-1}\ln(1/\varepsilon^2+1/\mu)+C^{-1}(n-1)\ln(1/\mu).$$

Теорема доказана. ■

Доказательство теоремы 13

Запишем условия теоремы в виде выпуклой оптимизационной задачи

$$\begin{cases} -V_{RFM}(\mathbf{a},\varepsilon\,|\,C,\mu) = -C^{-1}\sum_{i=1}^{n}\ln(a_i^2/\varepsilon^2+1/\mu) \to \min(\mathbf{a}\,|\,\varepsilon,C,\mu), \\ \mathbf{a}^T\mathbf{a}=1. \end{cases} \tag{140}$$

Запишем функцию Лагранжа

$$L(\mathbf{a},\lambda\,|\,C,\varepsilon) = -\frac{1}{2}C^{-1}\sum_{i=1}^{n}\ln(a_i^2/\varepsilon^2+1/\mu) - \frac{1}{2\lambda}\left(\sum_{i=1}^{n}a_i^2-1\right) \to \begin{cases}\min(\mathbf{a}),\\ \max(\lambda).\end{cases}$$

Градиент функции Лагранжа по $a_i \neq 0$

$$\partial_{a_i}L(\mathbf{a},\lambda) = -C^{-1}\frac{a_i/\varepsilon^2}{a_i^2/\varepsilon^2+1/\mu} - \frac{1}{\lambda}a_i = 0 \Rightarrow a_i^2 = -\lambda/C - \varepsilon^2/\mu \Rightarrow \lambda < -C\varepsilon^2/\mu$$

Кроме того, должно выполняться условие дополняющей нежесткости

$$\frac{1}{\lambda}\left(\sum_{i=1}^{n}a_i^2-1\right) = \frac{1}{\lambda}\left(\sum_{i:\,a_i\neq 0}a_i^2-1\right) = 0,$$

Учитывая, что $|a_i|>0$ получим

$$\frac{1}{\lambda}\left[n_+\left(-\lambda/C-\varepsilon^2/\mu\right)-1\right]=0,$$

где n_+ - количество ненулевых компонент. Условие дополняющей нежестко-сти эквивалентно определяет значение ненулевых компонент

$$a_i^2 = 1/n_+ \text{ или } |a_i| = n_+^{-0.5}. \tag{141}$$

Запишем значение целевой функции

$$-C^{-1}\sum_{i=1}^{n}\ln(a_i^2/\varepsilon^2+1/\mu) = -C^{-1}\sum_{i:\,a_i\neq 0}\ln(a_i^2/\varepsilon^2+1/\mu) - C^{-1}\sum_{i:\,a_i=0}\ln(1/\mu) =$$
$$-C^{-1}n_+\ln(1/\varepsilon^2 n_+ + 1/\mu) - C^{-1}(n-n_+)\ln(1/\mu).$$

Сама задача (140) примет эквивалентный вид

$$C^{-1}n_+\ln(1/\varepsilon^2 n_+ + 1/\mu) + C^{-1}(n-n_+)\ln(1/\mu) \to \max(n_+).$$

Преобразуем целевую функцию

$$C^{-1}n_+\ln(1/\varepsilon^2 n_+ + 1/\mu) + C^{-1}n_+\ln(1/\mu) \to \max(n_+).$$

Задача (140) примет эквивалентный вид

$$C^{-1}n_+ \ln(\mu/\varepsilon^2 n_+ + 1) \to \max(n_+). \qquad (142)$$

Аналогично доказательству предыдущей теоремы нетрудно показать, что максимальное значение целевая функция принимает при $n_+ = n$.

Учитывая (141) решением оптимизационной задачи (142) будут вектора

$$\forall \varepsilon \ \mathbf{a} \in \mathcal{A}^{RFM}_{\max}(\varepsilon) \leftrightarrow |a_i| = n^{-0.5}, \ i = 1,...,n.$$

Значение целевой функции (142) в точке максимума равно

$$\max_{\mathbf{a}' \in \mathcal{A}} V_{RFM}(\mathbf{a}', \varepsilon \,|\, C, \mu) = C^{-1} n \ln(1/n\varepsilon^2 + 1/\mu).$$

Теорема доказана. ■

Доказательство теоремы 14

Пусть $\hat{\mathbf{a}} = (\hat{a}_1,...,\hat{a}_n)^T$, $\hat{b}$, $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1,...,\hat{\delta}_N)^T$ – решение задачи (99). Пусть $\tilde{\mathbf{a}} = (\tilde{a}_1,...,\tilde{a}_n)^T$, $\tilde{b}$, $\tilde{\boldsymbol{\delta}} = (\tilde{\delta}_1,...,\tilde{\delta}_N)^T$ – произвольная совокупность значений аргументов целевой функции (99), удовлетворяющая ограничениям задачи, но не являющаяся ее решением. Тогда справедливо неравенство $J_{SFM}(\tilde{\mathbf{a}}, \tilde{b}, \tilde{\boldsymbol{\delta}}) - J_{SFM}(\hat{\mathbf{a}}, \hat{b}, \hat{\boldsymbol{\delta}}) \geq 0$, т.е.

$$C\sum_{j=1}^{N}\tilde{\delta}_j - C\sum_{j=1}^{N}\hat{\delta}_j + 2\mu\sum_{i:|a_i|\le\mu}|\tilde{a}_i| + \sum_{i:|a_i|>\mu}(\tilde{a}_i^2+\mu^2) - 2\mu\sum_{i:|a_i|\le\mu}|\hat{a}_i| - \sum_{i:|a_i|>\mu}(\hat{a}_i^2+\mu^2) \geq 0. \quad (143)$$

Выберем любые два признака, обозначив их индексы как (q,k). Рассмотрим совокупность значений аргументов $(\tilde{\mathbf{a}}, \tilde{b}, \tilde{\boldsymbol{\delta}})$, в которой направляющий вектор $\tilde{\mathbf{a}}$ отличается от оптимального $\hat{\mathbf{a}}$ только тем, что

$$\begin{cases} \tilde{a}_i = \hat{a}_i, \text{ если } i \neq q \text{ и } i \neq k, \\ \tilde{a}_q = \tilde{a}_k = 0.5(\hat{a}_q + \hat{a}_k), \\ \tilde{b} = \hat{b}. \end{cases} \qquad (144)$$

Обратим внимание на то, что функция потерь $\delta_j \geq 1 - y_j\left(\sum_{i=1}^{n} a_i x_{ij} + b\right)$, $\delta_j \geq 0$ критерия (99), как функция от $y_j\left(\sum_{i=1}^{n} a_i x_{ij} + b\right)$ (аналогично функции потерь на рисунке 2), является непрерывной по Липшицу с константой 1. С учетом этого факта предположение (144) приводит к неравенству

$$C\sum_{j=1}^{N}\tilde{\delta}_j - C\sum_{j=1}^{N}\hat{\delta}_j \le \frac{C}{2}\,|\,\hat{a}_k - \hat{a}_q\,|\sum_{j=1}^{N}\left|x_{qj} - x_{kj}\right| = \frac{C}{2}\,|\,\hat{a}_q - \hat{a}_k\,|\cdot\|\,\dot{\mathbf{x}}_q - \dot{\mathbf{x}}_k\,\|_1 . \qquad (145)$$

где $\dot{\mathbf{x}}_q$ и $\dot{\mathbf{x}}_k$ вектора значений признаков q и k, соответственно, для объектов обучающей выборки (97). Действительно,

$$C\sum_{j=1}^{N}\tilde{\delta}_j - C\sum_{j=1}^{N}\hat{\delta}_j = C\left(\sum_{j=1}^{N}\tilde{\delta}_j - \sum_{j=1}^{N}\hat{\delta}_j\right) = C\sum_{j=1}^{N}\left(\tilde{\delta}_j - \hat{\delta}_j\right) \le$$

$$C\sum_{j=1}^{N}\left|\tilde{\delta}_j - \hat{\delta}_j\right| \le C\sum_{j=1}^{N}\left|y_j\left(\sum_{i=1}^{n}\tilde{a}_i x_{ij} + \tilde{b}\right) - y_j\left(\sum_{i=1}^{n}\hat{a}_i x_{ij} + \hat{b}\right)\right| =$$

$$C\sum_{j=1}^{N}\left|\sum_{i=1}^{n}\tilde{a}_i x_{ij} - \sum_{i=1}^{n}\hat{a}_i x_{ij}\right| = C\sum_{j=1}^{N}\left|\sum_{i=1}^{n}(\tilde{a}_i - \hat{a}_i)x_{ij}\right| =$$

$$C\sum_{j=1}^{N}\left|\left[0.5(\hat{a}_q + \hat{a}_k) - \hat{a}_q\right]x_{qj} + \left[0.5(\hat{a}_q + \hat{a}_k) - \hat{a}_k\right]x_{kj}\right| =$$

$$C\sum_{j=1}^{N}\left|0.5(\hat{a}_k - \hat{a}_q)x_{qj} - 0.5(\hat{a}_k - \hat{a}_q)x_{kj}\right| =$$

$$C\sum_{j=1}^{N}\left|0.5(\hat{a}_k - \hat{a}_q)(x_{qj} - x_{kj})\right| = 0.5C\,|\,\hat{a}_k - \hat{a}_q\,|\sum_{j=1}^{N}\left|x_{qj} - x_{kj}\right|.$$

Для центрированных и нормированных признаков (98) справедливо неравенство

$$\|\,\dot{\mathbf{x}}_q - \dot{\mathbf{x}}_k\,\|_1 \le \sqrt{N}\sqrt{\|\,\dot{\mathbf{x}}_q - \dot{\mathbf{x}}_k\,\|_2^2} = \sqrt{N}\sqrt{2(1-\rho)}\,, \qquad (146)$$

где ρ - значение выборочной корреляции (100) признаков $\dot{\mathbf{x}}_q, \dot{\mathbf{x}}_k$. С учетом (98) корреляция ρ признаков $\dot{\mathbf{x}}_q, \dot{\mathbf{x}}_k$ выражается в виде

$$\rho(\dot{\mathbf{x}}_q, \dot{\mathbf{x}}_k) = \frac{\sum_{j=1}^{N}(x_{qj} - \overline{x}_{qj})(x_{kj} - \overline{x}_{kj})}{\sqrt{\sum_{j=1}^{N}(x_{qj} - \overline{x}_{qj})^2}\sqrt{\sum_{j=1}^{N}(x_{kj} - \overline{x}_{kj})^2}} = \sum_{j=1}^{N} x_{qj}x_{kj} \quad .$$

.

Правое равенство в (146) может быть записано в эквивалентной форме:

$$\sqrt{\|\,\dot{\mathbf{x}}_q - \dot{\mathbf{x}}_k\,\|_2^2} = \sqrt{\sum_{j=1}^{N}(x_{qj} - x_{kj})^2} = \sqrt{\sum_{j=1}^{N}(x_{qj}^2 - 2x_{qj}x_{kj} + x_{kj}^2)} =$$

$$\sqrt{\sum_{j=1}^{N}(x_{qj}^2 + x_{kj}^2) - 2\sum_{j=1}^{N}x_{qj}x_{kj}} = \sqrt{2\left(1 - \sum_{j=1}^{N}x_{qj}x_{kj}\right)} = \sqrt{2(1-\rho)}.$$

Левое неравенство в (146) доказывается через вогнутость квадратного корня. Запишем это неравенство в эквивалентной форме

$$\sum_{j=1}^{N}|x_{qj}-x_{kj}|\le\sqrt{N}\sqrt{\sum_{j=1}^{N}(x_{qj}-x_{kj})^2}\,.$$

Для доказательства рассмотрим абстрактную запись этого неравенства для произвольных чисел d_j, $j=1,...,N$

$$\sum_{j=1}^{N}|d_j|\le\sqrt{N}\sqrt{\sum_{j=1}^{N}d_j^2}\ \text{ или }\ \frac{\sum_{j=1}^{N}|d_j|}{N}\le\sqrt{\frac{\sum_{j=1}^{N}d_j^2}{N}}\,.$$

В силу вогнутости функции квадратного корня имеем для любого $0\le t\le 1$

$$\sqrt{t\min_j(d_j'^2)+(1-t)\max_j(d_j'^2)}\ge t\sqrt{\min_j(d_j'^2)}+(1-t)\sqrt{\max_j(d_j'^2)}\,.$$

В качестве t возьмем такое значение, для которого $t\min_j(d_j'^2)+(1-t)\max_j(d_j'^2)=\sum_{j=1}^{N}d_j'^2\Big/N$, тогда в силу линейности $t\sqrt{\min_j(d_j'^2)}+(1-t)\sqrt{\max_j(d_j'^2)}=\sum_{j=1}^{N}|d_j'|\Big/N$, что и требовалось показать.

Учитывая (146) неравенство (145) примет вид

$$C\sum_{j=1}^{N}\tilde{\delta}_j-C\sum_{j=1}^{N}\hat{\delta}_j\le\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|. \qquad (147)$$

Тогда неравенство (143) можно записать в виде

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+2\mu\sum_{i:|a_i|\le\mu}|\tilde{a}_i|+\sum_{i:|a_i|>\mu}(\tilde{a}_i^2+\mu^2)-2\mu\sum_{i:|a_i|\le\mu}|\hat{a}_i|-\sum_{i:|a_i|>\mu}(\hat{a}_i^2+\mu^2)\ge 0. \qquad (148)$$

Значение суммы в левой части этого неравенства зависит от значений компонент векторов $\hat{\mathbf{a}}=(\hat{a}_1,...,\hat{a}_n)$ и $\tilde{\mathbf{a}}=(\tilde{a}_1,...,\tilde{a}_n)$. Рассмотрим возможные варианты этих значений.

1) Пусть $|\hat{a}_q|>\mu$ и $|\hat{a}_k|>\mu$, тогда согласно (144) справедливо неравенство $|\tilde{a}_q|=|\tilde{a}_k|=0.5|\hat{a}_q+\hat{a}_k|>\mu$, а неравенство (148) примет эквивалентный вид

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+0.25|\hat{a}_q+\hat{a}_k|^2+0.25|\hat{a}_q+\hat{a}_k|^2-\hat{a}_q^2-\hat{a}_k^2\ge 0\,.$$

Элементарные преобразования этого неравенства приводят к следующей последовательности неравенств:

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+0.5(\hat{a}_q^2+2\hat{a}_q\hat{a}_k+\hat{a}_k^2)-\hat{a}_q^2-\hat{a}_k^2\geq 0,$$

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|-0.5(\hat{a}_q^2-2\hat{a}_q\hat{a}_k+\hat{a}_k^2)\geq 0,$$

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|-0.5|\hat{a}_q-\hat{a}_k|^2\geq 0,$$

и, наконец,

$$C\sqrt{N}\sqrt{2(1-\rho)}\geq|\hat{a}_q-\hat{a}_k|.$$

Заметим, что при $\rho=1$ имеем $|\hat{a}_q-\hat{a}_k|=0$, т.е. $\hat{a}_q=\hat{a}_k$.

2) Пусть $|\hat{a}_q|>\mu, |\hat{a}_k|\leq\mu$ и $|\tilde{a}_q|=|\tilde{a}_k|=0.5|\hat{a}_q+\hat{a}_k|>\mu$. Тогда неравенство (148) примет вид

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+0.25|\hat{a}_q+\hat{a}_k|^2+0.25|\hat{a}_q+\hat{a}_k|^2+\mu^2-2\mu|\hat{a}_k|-\hat{a}_q^2\geq 0.$$

Последовательность эквивалентных неравенств

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+0.5|\hat{a}_q+\hat{a}_k|^2+\mu^2-2\mu|\hat{a}_k|-\hat{a}_q^2\geq 0,$$

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+0.5|\hat{a}_q+\hat{a}_k|^2+\mu^2-\hat{a}_q^2-\hat{a}_k^2+\hat{a}_k^2-2\mu|\hat{a}_k|\geq 0,$$

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|-0.5|\hat{a}_q-\hat{a}_k|^2+\hat{a}_k^2+\mu^2-2\mu|\hat{a}_k|\geq 0,$$

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|\geq 0.5|\hat{a}_q-\hat{a}_k|^2-\hat{a}_k^2-\mu^2+2\mu|\hat{a}_k|,$$

приводит, наконец, к неравенству

$$C\sqrt{N}\sqrt{2(1-\rho)}\geq|\hat{a}_q-\hat{a}_k|-2\frac{\hat{a}_k^2-2\mu|\hat{a}_k|+\mu^2}{|\hat{a}_q-\hat{a}_k|}=|\hat{a}_q-\hat{a}_k|-2\frac{(|\hat{a}_k|-\mu)^2}{|\hat{a}_q-\hat{a}_k|}.$$

При $\rho\to 1$ имеем $|\hat{a}_q-\hat{a}_k|\to 0$, $|\hat{a}_q|\to\mu$ и $|\hat{a}_k|\to\mu$.

3) Пусть $|\hat{a}_q|>\mu$, $|\hat{a}_k|\le\mu$ и $|\tilde{a}_q|=|\tilde{a}_k|=0.5|\hat{a}_q+\hat{a}_k|\le\mu$, тогда неравенство (148) примет вид

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+\mu|\hat{a}_q+\hat{a}_k|+\mu|\hat{a}_q+\hat{a}_k|-2\mu|\hat{a}_k|-\hat{a}_q^2-\mu^2\ge 0,$$

приводящий к эквивалентным неравенствам

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+2\mu|\hat{a}_q+\hat{a}_k|-2\mu|\hat{a}_k|-\hat{a}_q^2-\mu^2\ge 0,$$

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+2\mu|\hat{a}_q+\hat{a}_k|-2\mu|\hat{a}_k|-\hat{a}_q^2-\mu^2\ge 0,$$

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+2\mu\underbrace{(|\hat{a}_q+\hat{a}_k|-|\hat{a}_k|-|\hat{a}_q|)}_{\le 0}+2\mu|\hat{a}_q|-\hat{a}_q^2-\mu^2\ge 0.$$

Здесь возможны два варианта:

Первый вариант. Если компоненты $|\hat{a}_q|>\mu$ и $|\hat{a}_k|\le\mu$ одного знака, то

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+2\mu\underbrace{(|\hat{a}_q+\hat{a}_k|-|\hat{a}_k|-|\hat{a}_q|)}_{=0}+2\mu|\hat{a}_q|-\hat{a}_q^2-\mu^2\ge 0,$$

т.е.

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+2\mu|\hat{a}_q|-\hat{a}_q^2-\mu^2\ge 0,$$

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|\ge\hat{a}_q^2-2\mu|\hat{a}_q|+\mu^2,$$

и, наконец,

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\ge\frac{\hat{a}_q^2-2\mu|\hat{a}_q|+\mu^2}{|\hat{a}_q-\hat{a}_k|}=\frac{(|\hat{a}_q|-\mu)^2}{|\hat{a}_q-\hat{a}_k|}.$$

Заметим, что при $\rho\to 1$ имеем $|\hat{a}_q|\to\mu$ и $|\hat{a}_k|\le\mu$.

Второй вариант. Если знаки компонент $\hat{a}_q$ и $\hat{a}_k$ разные, то имеем

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+2\mu\underbrace{(|\hat{a}_q+\hat{a}_k|-|\hat{a}_k|-|\hat{a}_q|)}_{\le 0}+2\mu|\hat{a}_q|-\hat{a}_q^2-\mu^2\ge 0,$$

$$\frac{C\sqrt{n}}{2}\sqrt{2(1-\rho)} \geq \underbrace{\frac{(|\hat{a}_q|-\mu)^2}{|\hat{a}_q-\hat{a}_k|}}_{>0} \underbrace{-2\mu\frac{|\hat{a}_q+\hat{a}_k|-|\hat{a}_k|-|\hat{a}_q|}{|\hat{a}_q-\hat{a}_k|}}_{\geq 0}.$$

Заметим, что при $\rho \to 1$ имеем $|\hat{a}_q| \to \mu$ и $|\hat{a}_k| \to 0$.

4) Пусть $|\hat{a}_q| \leq \mu$ и $|\hat{a}_k| \leq \mu$, то и $|\tilde{a}_q| = |\tilde{a}_k| = 0.5|\hat{a}_q + \hat{a}_k| \leq \mu$, тогда (148) примет вид

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+\mu|\hat{a}_q+\hat{a}_k|+\mu|\hat{a}_q+\hat{a}_k|-2\mu|\hat{a}_q|-2\mu|\hat{a}_k| \geq 0,$$

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k|+2\mu|\hat{a}_q+\hat{a}_k|-2\mu|\hat{a}_q|-2\mu|\hat{a}_k| \geq 0,$$

и, наконец,

$$\frac{C\sqrt{N}}{4\mu}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k| \geq |\hat{a}_q|+|\hat{a}_k|-|\hat{a}_q+\hat{a}_k|.$$

Здесь возможны два варианта.

Первый вариант. Если компоненты $\hat{a}_q$, $\hat{a}_k$ имеют одинаковые знаки, то их значения дополнительно не ограничены, т.к.

$$\frac{C\sqrt{N}}{4\mu}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k| \geq |\hat{a}_q|+|\hat{a}_k|-|\hat{a}_q+\hat{a}_k| = 0.$$

Это неравенство также справедливо для $|\hat{a}_q| = 0$ и $|\hat{a}_k| = 0$.

Второй вариант. Если компоненты $\hat{a}_q$, $\hat{a}_k$, неравные одновременно нулю, имеют разные знаки, то

$$\frac{C\sqrt{N}}{4\mu}\sqrt{2(1-\rho)}\cdot|\hat{a}_q-\hat{a}_k| \geq \underbrace{|\hat{a}_q|+|\hat{a}_k|-|\hat{a}_q+\hat{a}_k|}_{\geq 0\, u\, \leq |\hat{a}_q-\hat{a}_k|}$$

или

$$\frac{C\sqrt{N}}{4\mu}\sqrt{2(1-\rho)} \geq \frac{|\hat{a}_q|+|\hat{a}_k|-|\hat{a}_q+\hat{a}_k|}{|\hat{a}_q-\hat{a}_k|} = 1-\frac{|\hat{a}_q+\hat{a}_k|}{|\hat{a}_q-\hat{a}_k|}.$$

Нетрудно видеть, что если $C\sqrt{N}\sqrt{2(1-\rho)}/4\mu \geq 1$, то значения компонент с разными знаками также дополнительно не ограничены. При $\rho = 1$ имеем $|\hat{a}_q + \hat{a}_k| / |\hat{a}_q - \hat{a}_k| = 1$, следовательно, $|\hat{a}_q| = 0$ или $|\hat{a}_k| = 0$.

Таким образом, во всех случаях выполняются утверждения теоремы.

Теорема доказана. ■

Доказательство теоремы 15

Доказательство теоремы выполняется аналогично доказательству теоремы 14. Пусть $\hat{\mathbf{a}} = (\hat{a}_1,...,\hat{a}_n)^T$, $\hat{b}$, $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1,...,\hat{\delta}_N)^T$ – решение задачи (103). Пусть $\tilde{\mathbf{a}} = (\tilde{a}_1,...,\tilde{a}_n)^T$, $\tilde{b}$, $\tilde{\boldsymbol{\delta}} = (\tilde{\delta}_1,...,\tilde{\delta}_N)^T$ – произвольная совокупность значений аргументов целевой функции (103), удовлетворяющая ограничениям задачи, но не являющаяся ее решением. Тогда справедливо неравенство $J_{RFM}(\tilde{\mathbf{a}},\tilde{b},\tilde{\boldsymbol{\delta}}) - J_{RFM}(\hat{\mathbf{a}},\hat{b},\hat{\boldsymbol{\delta}}) \geq 0$, т.е.

$$C\sum_{j=1}^{N}\tilde{\delta}_j - C\sum_{j=1}^{N}\hat{\delta}_j + \sum_{i=1}^{n}\ln(\tilde{a}_i^2 + 1/\mu) - \sum_{i=1}^{n}\ln(\hat{a}_i^2 + 1/\mu) \geq 0. \qquad (149)$$

Выберем любые два признака, обозначив их индексы как (q,k). Рассмотрим совокупность значений аргументов $(\tilde{\mathbf{a}},\tilde{b},\tilde{\boldsymbol{\delta}})$, в которой направляющий вектор $\tilde{\mathbf{a}}$ отличается от оптимального $\hat{\mathbf{a}}$ только тем, что

$$\begin{cases} \tilde{a}_i = \hat{a}_i, \text{ если } i \neq q \text{ и } i \neq k, \\ \tilde{a}_q = \tilde{a}_k = 0.5(\hat{a}_q + \hat{a}_k), \\ \tilde{b} = \hat{b}. \end{cases} \qquad (150)$$

Тогда (149) примет вид

$$C\sum_{j=1}^{N}\tilde{\delta}_j - C\sum_{j=1}^{N}\hat{\delta}_j + 2\ln(0.25(\hat{a}_q + \hat{a}_k)^2 + 1/\mu) - \ln(\hat{a}_q^2 + 1/\mu) - \ln(\hat{a}_k^2 + 1/\mu) \geq 0. \qquad (151)$$

Тогда учитывая (147) неравенство (151) примет вид

$$\frac{C\sqrt{N}}{2}\sqrt{2(1-\rho)}\cdot|\hat{a}_q - \hat{a}_k| + 2\ln(0.25(\hat{a}_q + \hat{a}_k)^2 + 1/\mu) - \ln(\hat{a}_q^2 + 1/\mu) - \ln(\hat{a}_k^2 + 1/\mu) \geq 0. \qquad (152)$$

Заметим, что функция $\ln(\hat{a}^2+1/\mu)$ является строго выпуклой для $|\hat{a}|<1/\sqrt{\mu}$. Это не трудно доказать. Запишем ее вторую производную:

$$\ln(\hat{a}^2+1/\mu)''=-\frac{2(\hat{a}^2-1/\mu)}{(\hat{a}^2+1/\mu)^2}. \qquad (153)$$

Вторая производная (153) на интервале $|\hat{a}|<1/\sqrt{\mu}$ строго больше нуля, что является достаточным условием строгой выпуклости функции вниз. В силу строгой выпуклости и учитывая, что $|\hat{a}_q|<1/\sqrt{\mu}$ и $|\hat{a}_k|<1/\sqrt{\mu}$ имеем

$$\ln(0.25(\hat{a}_q+\hat{a}_k)^2+1/\mu)\le\frac{\ln(\hat{a}_q^2+1/\mu)+\ln(\hat{a}_k^2+1/\mu)}{2}.$$

Элементарные преобразования приводят к неравенству:

$$2\ln(0.25(\hat{a}_q+\hat{a}_k)^2+1/\mu)-\ln(\hat{a}_q^2+1/\mu)-\ln(\hat{a}_k^2+1/\mu)\le 0 \qquad (154)$$

Причём при $\hat{a}_q\neq\hat{a}_k$ это неравенство строгое (в силу строгой выпуклости функции), а при $\hat{a}_q=\hat{a}_k$ она вырождается в равенство (очевидно).

С другой стороны, неравенство (152) при $\rho=1$ примет вид

$$2\ln(0.25(\hat{a}_q+\hat{a}_k)^2+1/\mu)-\ln(\hat{a}_q^2+1/\mu)-\ln(\hat{a}_k^2+1/\mu)\ge 0.$$

Тогда учитывая (154) справедливо следующее равенство

$$2\ln(0.25(\hat{a}_q+\hat{a}_k)^2+1/\mu)-\ln(\hat{a}_q^2+1/\mu)-\ln(\hat{a}_k^2+1/\mu)=0,$$

которое, как показано выше, выполняется только для $\hat{a}_q=\hat{a}_k$.

Теорема доказана. ■

ЛИТЕРАТУРА

1. Vapnik V. Statistical Learning Theory. NY.: J. Wiley, 1998.
2. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения). М.: Наука, 1974.
3. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. М.: Наука, 1970.
4. R. Duin, E. Pekalska, D. Ridder. Relational discriminant analysis. Pattern Recognition Letters, Vol. 20, 1999, pp. 1175-1181.
5. M. Tomlinson, B. Love. When learning to classify by relations is easier than by features. Thinking & Reasonong, 2010, 16 (4), pp. 372–401.
6. Середин О.С., Моттль В.В., Татарчук А.И., Разин Н.А. Выпуклые селективные критерии релевантных векторов для сокращения размерности описания объектов, представленных парными отношениями. Известия ТулГУ, Серия Естественные науки. Тула: Изд-во ТулГУ, 2013, Вып. 1, с. 165-176.
7. John C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in large margine classifiers, 1999, pp. 61-74, MIT Press.
8. Peter Sollich. Probabilistic methods for Support Vector Machines. Advances in Neural Information Processing Systems 12, 2000, pp. 349-355, MIT Press.
9. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979. — 448 с.
10. Vapnik V. N. The Nature of Statistical Learning Theory. Springer, 1995.
11. Bradley P., Mangasarian O. Feature selection via concave minimization and support vector machines. In International Conference on Machine Learning. 1998.
12. Li Wang, Ji Zhu, Hui Zou. The doubly regularized support vector machine. Statistica Sinica, 01/2006; 16:589-615.

13. Zou, H., Hastie, T.. Regularization and variable selection via the elastic net. J. Roy. Statist. Soc. Ser. B 67, 301-320, 2005.
14. Natalia Becker, Grischa Toedt, Peter Lichter, Axel Benner. Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. BMC Bioinformatics, 2011.
15. Fan J, Li R: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of American Statistical Association 2001, 96:1348-1360.
16. ДеГроот М. Оптимальные статистические решения. М.:, Мир, 1974.
17. Разин Н.А. Выпуклые критерии и параллелизуемые алгоритмы селективного комбинирования разнородных представлений объектов в задачах восстановления зависимостей по эмпирическим данным. Дисс. к.ф.-м.н. ВЦ РАН, 2013.
18. C.M. Bishop, M.E. Tipping. Variational relevance vector machines. Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, 2000, pp. 46-53.
19. J.-B. Hiriart-Urruty, C. Lemarechal. Fundamentals of Convex Analysis. Springer, 2001.
20. Boyd S., Vandenberghe L. Convex Optimization. Cambridge University Press.
21. http://archive.ics.uci.edu/ml/datasets/Lung+Cancer – UCI Machine Learning Repository: Lung Cancer Data Set.