



*Курс лекций рассматривает теоретические и прикладные аспекты порождения моделей линейной и нелинейной регрессии и алгоритмов оптимизации качества этих моделей. Рассматриваются современные алгоритмы индуктивного порождения регрессионных моделей. Приводятся примеры с использованием программ Matlab.*

## **1. Введение**

Линейные модели, метод главных компонент, сингулярное разложение, вероятностный анализ главных компонент. От непараметрических моделей к параметрическим, K-mean, RBF, MLP. Нелинейная регрессия, связь регрессии и оптимизации.

## **2. Построение моделей нелинейной регрессии**

Типы нелинейных моделей. Критерии качества, информационные критерии, методы нахождения параметров. Подготовка данных, анализ результата.

Matlab: решение задач линейной алгебры. Структуры данных, инлайн-функции. Организация системы: алгоритмы-метрики-оптимизаторы-задачи. Краткий обзор прикладных пакетов. Обзор алгоритмов оптимизации.

## **3. Двухуровневый байесовский вывод**

Сравнение моделей и бритва Оккама. Пример арифметической прогрессии и полиномиальной модели. Два уровня: байесовский метод и анализ данных. Механизм Байесовского сравнения: правдоподобие и фактор Оккама. Нахождение оптимальных параметров. Вычисление правдоподобия. Интерпретация фактора Оккама. Схема: пространство данных – пространство весов. Фактор Оккама для многих параметров. Пример целочисленной модели.

## **4. Метапараметры моделей и регуляризация**

Нейронные сети в качестве регрессионных моделей. Традиционное и Байесовское обучение сетей. Двоичные и множественные классификаторы. Установка метапараметров. Вычисление правдоподобия нейронной сети: коррекция метапараметров с использованием собственных значений матрицы Гессе функции ошибки. Сравнение сетей. Множественная регуляризация и распределение параметров в многослойных сетях.

## **5. Прогноз и ошибка в регрессионных моделях**

Прогноз для регрессии и для классификации. Вычисление значения прогностической ошибки в пространстве данных. Автоматическое определение значимости переменных в нейронных сетях. Пример регрессии для сигнала с различными уровнями шума. Использование техники ARD для построения множества скрытых переменных и их селекции. Суперпозиция нейронных сетей RBF+MLP.

## **6. Автоматический синтез моделей**

Индуктивный синтез моделей. Принципы самоорганизации. Построение критериев выбора. Полиномиальные сети. Базовый итеративный алгоритм. Алгоритм для многослойных сетей. Сходимость и регуляризация. Постановка задачи и алгоритм нахождения произвольной суперпозиции моделей.

Mathematica: функциональное программирование. Выражения, их представления и преобразования. Функции и передача аргумента. Операции с множеством функций. Работа с

шаблонами, принадлежность элемента множеству. Правила преобразований. Отложенные вычисления. Mathematica versus Matlab.

## **7. Методы однокритериальной стохастической оптимизации**

Методы моделируемого отжига, дифференциальной эволюции, генетические алгоритмы. Представление и инициализация популяции. Критерии селекции, рулетка. Кроссовер и мутация, шимы. Пополнение популяции. Множественные популяции, схемы миграции. Методы обобщающих поверхностей.

## **8. Многокритериальная оптимизация**

Постановка задачи, Парето-оптимальный фронт, пространства аргументов и значений. Агрегирующие функции: взвешенная сумма, объем фронта, целевое программирование, метод  $\epsilon$ -ограничений. DB-MOEA, SPEA. Алгоритмы без использования ПОФ: VEGA, лексикографическое упорядочивание. Визуализация ПОФ. Применение многокритериальной оптимизации. Сравнение стохастических алгоритмов.

## **9. Использование экспертной информации при построении моделей**

Пространство весов и пространство оценок. Постановка задачи согласования в линейных шкалах. Двойственная линейная регрессия. Согласование в ранговых шкалах. Поиск оптимальных весов в нелинейных моделях.

## **10. Временные ряды и регрессия**

Параметрический метод: векторная авторегрессия, связи между переменными, тест Гренджера, обращение матрицы лаговых коэффициентов, сценарный прогноз и управление в авторегрессионных моделях. Непараметрический метод: гусеница, построение матрицы развертки, интерпретация собственных значений, восстановление главных компонент.

## **11. Стохастическое интегрирование**

Задачи интегрирования, взвешенное сэмплирование, многомерное сэмплирование, отвергающее сэмплирование, метод Метрополиса-Хастингса, его сходимость, сэмплирование Гиббса, MCMC, способы построения Марковских цепей, применение MCMC.

## **12. Байесовская регрессия и гауссовские процессы**

Пространство весов и пространство функций. Гауссовские процессы. Создание моделей, нахождение гиперпараметров, обучение моделей, прогноз.

## **13. Тестирование регрессионных и оптимизационных алгоритмов**

Зачем тестировать? Теорема о бесплатных обедах. Тестовые задачи – регрессия: DELVE, ELENA, оптимизация: De Jong, DTLZ. Метрики оценки качества алгоритмов. Создание системы тестирования и отчетности.

## **14. Организация вычислительного эксперимента**

На примере ortbench: организация проекта, интерфейсов, системы настроек алгоритмов, документирование, тестопригодность кода – дилемма простота-универсальность, тестовые

задачи, отчеты, таблицы и визуализация. Некоторые источники алгоритмов, кода, данных; справочники, библиотеки, архивы.

### Основная литература

1. В. В. Стрижов, Г. О. Пташко. Алгоритмы поиска суперпозиций при выборе оптимальных регрессионных моделей. М.: ВЦ РАН. 2007. <http://strijov.com/papers>
2. Стрижов В.В., Казакова Т.В. Устойчивые интегральные индикаторы с выбором опорного множества описаний. Заводская лаборатория. Диагностика материалов. 2007 (7). С. 72-76. <http://www.ccas.ru/strijov/>
3. Стрижов В.В. Уточнение экспертных оценок с помощью измеряемых данных. Заводская лаборатория. Диагностика материалов. 2006 No 7. С.59-64. <http://www.ccas.ru/strijov/>
4. D. MacKay. Information, inference, learning algorithms. Cambridge University Press. 2003. <http://www.inference.phy.cam.ac.uk/mackay/>
5. Nabney, Yan T., Netlab: Algorithms for pattern recognition. Springer, 2004. <http://www.ccas.ru/strijov/pro>
6. Malada, H. R. and Ivakhnenko, A. G. Inductive Learning Algorithms for Complex Systems Modeling, CRC Press, 1994. <http://www.gmdh.net>
7. Mueler, J-A., and Lemke, F. Self-organising Data Mining: An Intelligent Approach To Extract Knowledge From Data. — Berlin: Dresden. 1999. <http://www.gmdh.net>
8. Брандт З. Анализ данных. М.: Мир, 2003.
9. Голуб Дж., Ван-Лоун Ч. Матричные вычисления - М.: Мир, 1999.
10. Деммель Дж. Вычислительная линейная алгебра: Теория и приложения. — М.: Мир, 2001.
11. Краснощеков П. С., Петров А. А. Принципы построения моделей. Изд. 2-е, пересмотр. и дополнен. — М.: Фазис: ВЦ РАН, 2000.

### Дополнительная литература

1. Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений. — М.: Мир, 1969.
2. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989.
3. Ивахненко А. Г., Юрачковский Ю. П. Моделирование сложных систем по экспериментальным данным. — М.: Радио и связь, 1987.
4. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Исследование зависимостей. — М.: Финансы и статистика, 1985.
5. Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений. — М.: Мир, 1969.
6. Гилязов С. Ф. Методы решения линейных некорректных задач. — М.: Изд-во МГУ, 1987.
7. Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем. — Киев: Наукова думка. 1981.
8. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. — Киев: Наукова думка. 1985.