

Полугодовой спецкурс кафедры «Вычислительной математики» Мехмат МГУ

Введение в методы машинного обучения (Introduction to Machine Learning Methods)

(проф., д.ф.-м.н. Кумсков М.И.)

<http://numa.math.msu.su/study/specials.html>

почта для коммуникаций 2022.ml.mkumskov@gmail.com

Аннотация

Рассматриваются основные типы моделей, используемые для поиска закономерностей в данных: регрессионный анализ, классификация, кластеризация данных, построение простых и обобщенных деревьев решений, сокращение размерности – метод главных компонент. Показывается тесная взаимосвязь методов при обработки реальных данных, а также описаны эволюционные методы отбора оптимальных подмножеств признаков – генетические алгоритмы и Метод Группового Учета Аргументов (МГУА).

Методы широко используются в различных прикладных задачах машинного обучения, включая анализ изображений, обработка биомедицинских объектов, прогнозирования свойств лекарственных соединений, а также анализ текстов естественного языка.

Кратко обсуждаются идеи и методы машины поддерживающих векторов (SVM – Support Vector Machine), а также методы bootstrap-построения оценок при недостаточном числе исходных данных. Рассматриваются основные понятия «нечеткого» (fuzzy) анализа данных и их применения для построения новых признаков и ведения «нелинейности» в модели. Предлагаемые алгоритмы могут быть использованы в базах данных для восстановления (предсказания) пропущенных значений.

Описана модель персептрона и ее ограничения, а также архитектура классической нейронной сети и сети «глубокое обучение». Рассмотрены подходы к решению проблемы переобучения. Дан обзор задач, успешно решаемых нейронными сетями. Дана характеристика вычислительным платформам реализации нейросетей – TensorFlow и Keras.

Лекция 1 «Введение в курс»

Постановки задач.
Обзор программы
Термины линейной алгебры. Примеры представления объектов.
Обучающие наборы данных - Датасеты

Лекция 2 «Основные типы моделей»

Регрессионный анализ.
Кластеризация данных.
Классификация.
Простые и обобщенные деревья решений. Случайный лес.
Сокращение размерности – метод главных компонент.
Эволюционные алгоритмы.
Коллекции моделей – семейства прогнозаторов
Нейронные сети. Глубокое обучение. Сверточные сети.
Обучение с подкреплением

Лекция 3-4 «Введение в методы машинного обучения»

Введение в анализ данных и распознавание образов.
Первичное преобразование данных, поиск выбросов.
Задача классификации и Задача регрессии.
Регрессионный анализ, скользящий контроль.
Нелинейная регрессия
Регуляризация. Гребневая регрессия
Непараметрическая регрессия
Деревья решений, простая и обобщенная формы.
Задача выбора опорных признаков в метрику
Отказ от прогнозирования.

Лекция 5-6 «Близость (похожесть) объектов. Кластеры и их поиск»

Кластер как связная компонента графа.
DBSCAN алгоритм
Построение минимального покрывающего дерева.
Метод K средних, простая и обобщенная версии.
Иерархический кластер-анализ, дендрограммы.
Алгоритм «Формальный элемент» (ФОРЭЛЬ)
EM-алгоритм
Обобщенное дерево решений

Лекция 7 «Метод главных компонент»

Факторы и их поиск, SVD разложение матрицы.
Факторная структура в данных и ее идентификация
Метод CoMFA – интерпретация факторов
Геометрический смысл факторов.
Регрессия на факторах.
Многомерное шкалирование

Лекция 8 «Продвинутые методы анализа-1»

Эволюционные алгоритмы. Задача выбора признаков.
Метод Группового Учета Аргументов (МГУА)
Линейный и нелинейный МГУА. Фашификация параметров.
Генетические Алгоритмы.
Примеры решаемых задач.

Лекция 9 «Продвинутые методы анализа-2»

Кернел функции – «беспризнаковый» анализ данных.
Алгебра кернелов.
Примеры кернелов.
SVM и поддерживающие вектора.
«Когда мало данных» – Метод Bootstrap.
Семейства прогнозирующих алгоритмов. Boosting.
«Нечеткие» признаки (Fuzzy). Лингвистическая переменная.
«Нечеткие» классификаторы

Лекция 10 «Нейронные сети-1»

Модель персептрона и ее ограничения.
Классические нейронные сети, слой нейронов, два типа нейронов.
Проблема переобучения.
Задачи, решаемые нейронными сетями, «Глубокое обучение».

Лекция 11 «Нейронные сети-2»

Реализация алгоритмов
Среда анализа данных R.
Библиотеки Python.
Реализация нейросетей. Платформа TensorFlow. Keras.
Платформа PyTorch

Лекция 12 «Нейронные сети-3»

Анализ Изображений и сверточные нейросети.
Нейросети и «Инженерия признаков».
Классический нейрон и RBF-нейроны.
Синтез архитектуры нейро сетей по результатам анализа данных
Перспективы развития нейросетей.
Графические процессоры (GPU)
Вычисления в облаке.

Лекция 13-14 «Задача «Структура-свойство»

Работа со объектами, имеющими структуру (объекты-графы)
Анализ изображений. Построение графа изображений.
Анализ молекулярных графов. Молекулярные дескрипторы.
Прогнозирование свойств лекарственных соединений.
Процесс построения модели машинного обучения. Этапы.