Интеллектуальный анализ данных (БИ)

Организационная встреча

Юрий Саночкин

ysanochkin@hse.ru

НИУ ВШЭ, 2025

План курса

План курса. Лекции

- Введение в машинное обучение: примеры задач, терминология
- Постановки задач. Метод k ближайших соседей, измерение ошибки.
- Параметры и гиперпараметры, оценивание обобщающей способности. Веса в knn.
- knn для регрессии. Линейная регрессия, её применимость.
- Обучение через аналитическое решение. Регуляризация. Вычисление важности признаков.
- Градиент и градиентный спуск.

План курса. Лекции

- Стохастический градиентный спуск. Функции потерь в задачах регрессии.
- Линейная классификация. Обучение через верхнюю функцию потерь.
- Метрики качества в задачах классификации. Оценка качества ранжирования, площади под кривыми.
- Оценивание вероятности классов. Логистическая регрессия. Метод опорных векторов. Калибровка вероятностей. Многоклассовая классификация. (2 лекции)
- Решающие деревья. Структура. Критерии информативности.
- Жадное построение решающих деревьев. Счётчики для кодирования категориальных признаков.

План курса. Лекции

- Композиции моделей. Бэггинг. Смещение и разброс в бэггинге. Случайный лес.
- Градиентный бустинг, часть 1
- Градиентный бустинг, часть 2. Стекинг. +ИНТЕРПРЕТАЦИЯ (SHAP и прочее)
- Кластеризация (побольше о том, как оценивается качество кластеризации)
- Отбор признаков и метод главных компонент. Визуализация.
- Ранжирование

План курса. Семинары

- numpy и основы линейной алгебры
- ООП в питоне
- Pandas: основы работы с таблицами, вычисление статистик, устранение пропусков
- Визуализация данных
- Знакомство с scikit-learn, kNN
- Линейная регрессия
- Градиентный спуск, numpy

План курса. Семинары

- Линейная классификация и её качество
- Категориальные и текстовые данные, word2vec
- Калибровка вероятностей, логистическая регрессия и SVM
- Деревья
- Случайные леса
- Бустинг
- Контрольная работа
- Бустинг 2
- Кластеризация (побольше о том, как оценивается качество кластеризации)

План курса. Домашние задания

- numpy
- pandas
- EDA и подготовка данных
- kNN и линейные модели, работа с признаками
- Градиентный спуск
- Классификация текстов
- Деревья, леса
- Бустинг и кластеризация

Формула оценки

$$О_{\text{итоговая}} = 0.2 * ДЗ + 0.23 * МКР + 0.23 * КР + 0.34 * ЭКЗ$$

- Домашние задания
- Средняя за МКР (+ активность на семинарах)
- Контрольная работа
- Экзамен

Что нам пригодится?

Математический анализ:

- Производные и дифференцирование
- Функции многих переменных, градиенты, частные производные

Линейная алгебра:

- Векторы и матрицы
- Нормы, метрики, скалярное произведение

Что нам пригодится?

Теория вероятностей и математическая статистика:

- Основные дискретные и непрерывные распределения
- Математическое ожидание, дисперсия, моменты
- Ковариация и корреляция
- Оценки параметров
- Статистические гипотезы

Что нам пригодится?

Программирование на Python:

- Это всегда больно, нужны время и практика, чтобы привыкнуть
- Семинары и консультации помогут разобраться

О чем еще помнить?

- Мы проверяем домашние задания и другие работы на плагиат, вызываем на защиты, и в целом у нас жесткая политика по отношению к читингу
- Дедлайны к домашним заданиям бывают мягкие и жесткие
- 9 и 10 – это очень высокие оценки!
  - Но их тоже можно получить
- Делать не всё из домашних заданий – нормально
- Задавать любые вопросы – не стыдно
- Тратить много времени и немножко мучиться – нормально
  - Но это окупится!

Контакты

- Телеграм-канал курса: https://t.me/+0bkmmipS4D8yMGQy
- Телеграм-чат курса (осторожно, есть риск флуда): https://t.me/+YGWeQSCHSz84MDli
- Телеграм преподавателя: https://t.me/yury_sanochkin
- Ассистенты: @minat0nami, @iznaroda, @ll_sonya, @GroovyGrape, @russolnik, @velavokrodef, @grtfss, @bershteynm, @levisjoy, @pavelchertik, @RealRobertL, @alya_zakh, @onlyaanya, @markushalovepokushat, @annaberaya