

5/7 семестр

ТЕМА 1. ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ

История вопроса. Основные понятия машинного обучения. Типы задач: классификация, регрессия, кластеризация. Обучение с учителем и без учителя. Метрики качества (точность, полнота, F1-score, ROC-AUC). Переобучение. Кросс-валидация и оценка обобщающей способности модели.

ТЕМА 2. ЛИНЕЙНЫЕ МОДЕЛИ И ГРАДИЕНТНЫЕ МЕТОДЫ ОПТИМИЗАЦИИ

Линейная регрессия. Функция потерь и метод наименьших квадратов. Градиентный спуск и его модификации (стохастический градиентный спуск, etc). Скорость обучения. Методы второго порядка. Применение линейных моделей в практических задачах.

ТЕМА 3. РЕГУЛЯРИЗАЦИЯ И КЛАССИФИКАЦИЯ

Методы регуляризации (L1, L2, elastic net). Логистическая регрессия для задачи бинарной классификации. Нелинейная классификация, многоклассовая классификация. Машина опорных векторов (SVM). Отбор признаков.

ТЕМА 4. СНИЖЕНИЕ РАЗМЕРНОСТИ

Проблемы признакового пространства большой размерности. Метод главных компонент (PCA): геометрическая интерпретация, вычисление собственных векторов и значений, объяснённая дисперсия, выбор числа компонент. Способы визуализации: t-SNE, etc.

ТЕМА 5. ВЕРОЯТНОСТНЫЕ МЕТОДЫ И КЛАСТЕРИЗАЦИЯ

Байесовский подход к оцениванию параметров. Наивный байесовский классификатор. ЕМ-алгоритм и смеси гауссовских распределений. Кластеризация: K-means, иерархическая кластеризация, DBSCAN, etc. Оценка качества кластеризации.

ТЕМА 6. РЕШАЮЩИЕ ДЕРЕВЬЯ

Решающие деревья для классификации и регрессии. Критерии ветвления. Избежание переобучения: ограничение глубины, минимальное число объектов в листе, etc.

ТЕМА 7. АНСАМБЛИ И ИХ ПРИМЕНЕНИЕ

Принципы обучения ансамблей. Stacking, boosting, bagging, etc. Градиентный бустинг (Gradient Boosting). XGBoost, LightGBM, CatBoost. Выбор и комбинирование моделей.

ТЕМА 8. РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ

Введение в рекомендательные системы. Основные типы задач: рейтинговое предсказание, ранжирование, прямые рекомендации. Метрики качества: precision@k, recall@k. Фильтрация на основе содержимого (content-based), коллаборативная фильтрация (collaborative filtering), матричная факторизация.

ТЕМА 9. ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

Предобработка текстов: токенизация, стеминг, лемматизация, удаление стоп-слов.
Мешок слов (Bag of Words), TF-IDF. Модели эмбедингов: Word2Vec, GloVe, FastText.
Применение в задачах классификации текстов, etc.

Литература

1. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning (2nd edition). Springer, 2009.
2. Bishop C. M. Pattern Recognition and Machine Learning. Springer, 2006.
3. Mohri M., Rostamizadeh A., Talwalkar A. Foundations of Machine Learning. MIT Press, 2012.
4. Murphy K. Machine Learning: A Probabilistic Perspective. MIT Press, 2012.
5. Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014.
6. Willi Richert, Luis Pedro Coelho. Building Machine Learning Systems with Python. Packt Publishing, 2013.
7. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. Питер, 2018.

Интернет-источники

1. [http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_\(курс_лекций%20–_материалы_лекционных_занятий,_2024_\(преподаватель_К.В._Воронцов\)\)](http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций%20–_материалы_лекционных_занятий,_2024_(преподаватель_К.В._Воронцов)))
2. <https://github.com/esokolov/ml-course-hse> – материалы лекционных и семинарских занятий, 2024 (преподаватель Е.А. Соколов)
3. <https://education.yandex.ru/handbook/ml/article/about> – Учебник по машинному обучению (С. Н. Федотов, Ф. Г. Сеницын)