OTUS

Machine learning

Best Practice по работе с современными методами анализа данных и инструментам, необходимым для профессионального развития в качестве специалиста Data Science

Длительность курса: 160 академических часов

1 Введение

1	**Введение в машинное обучение**	обзор курса. Знакомство со стандартными задачами по машинному обучению. Понимание общего подхода: Exploratory Data Analysis and Preprocessing -> Models and experiments -> Production.
2	**Базовые инструменты анализа данных в Python**	рабочее окружение Python. Функционал базовых библиотек для работы с данными: Numpy, Pandas, scikit-learn.

3	**Exploratory Data Analysis and Preprocessing**	основные принципы и методы разведочного анализа данных. Преобразование данных в подходящий для моделей формат. Использование статистического анализа и визуализации для знакомства с данными. Домашние задания 1 Практика EDA и препроцессинга. Очистка данных, построение визуализаций и формирование признаков
4	**Задача классификации. Метод ближайших соседей (kNN)**	алгоритм kNN. Влияние нормализации данных в kNN. Структуры данных для оптимизации kNN. Метрики оценки качества классификации. Кросс- валидация.
5	**Задача регрессии. Линейная регрессия**	линейная регрессия - метод наименьших квадратов и градиентный спуск. Вероятностная трактовка линейной регрессии. Полиномиальная регрессия. Регуляризация в линейной регрессии. Метрики оценки качества регрессии. Домашние задания 1 Построение модели линейной регрессии, настройка гиперпараметров на кросс-валидации, интерпретация коэффициентов.

6	**Логистическая регрессия**	реализации логистической регрессии с помощью метода с тохас тического градиентного спуска. Регуляризация и подбор гиперпараметров. Домашние задания 1 Построение модели логистической регрессии, настройка гиперпараметров на кросс-валидации, интерпретация коэффициентов.
7	**Feature engineering & advanced preprocessing**	отбор признаков. Преобразование признаков для повышения точности модели. Устранение несбалансированности выборки. Домашние задания 1 Применение статистических и model-based методов для отбора важных признаков. Работа со SMOTE для устранения дисбаланса классов.
8	**Практическое занятие по темам, изученным в первом модуле**	повторение – мать учения.

2 Продвинутые методы машинного обучения

1 **Метод опорных векторов**

метод опорных векторов (SVM), интерпретация. Случай линейно
неразделимых данных. Kernel trick. Примеры SVM в sklearn.

Домашние задания

1 Построение SVM и выбор оптимального ядра.

---

2 **Деревья решений**

программа: Классификация и регрессия с помощью деревьев решений. Обзор
алгоритмов. Алгоритм CART. Выбор оптимального сплита, суррогатный сплит.
Обзор реализации в sklearn.

Домашние задания

1 Построение модели решающего дерева, настройка гиперпараметров, визуализация сплитов и интерпретация результатов.

---

3 **Ансамбли моделей**

ансамблирование. Случайный лес. Бэггинг, стэкинг, блэндинг.

Домашние задания

1 Построение и настройка модели случайного леса. Визуализация важности признаков.

4	**Градиентный бустинг**	теория градиентного бустинга. XGBoost, CatBoost, LightGBM. Применение библиотеки ELI5 для интерпретации моделей. Домашние задания 1 Сравнение трех разобранных алгоритмов бустинга и подбор гиперпараметров для получения лучшего качества.
5	**Обучение без учителя. K-means, EM алгоритм**	обучение без учителя. Алгоритмы кластеризации, области применения. K-means. Оценка качества обучения, ограничения и подбор алгоритма для задачи. Алгоритмы с lower-bound. EM алгоритм. Домашние задания 1 Настройка числа кластеров в алгоритме K-Means. Elbow и Silhouette метод.
6	**Обучение без учителя. Иерархическая кластеризация. DB-Scan**	иерархическая кластеризация. DB-Scan. Спектральная кластеризация. Домашние задания 1 Построение различных вариантов кластеризаций и интерпретация результатов.

7	**Методы уменьшения размерности**	метод главных компонент (Principle component analysis). Метод t-SNE. Примеры визуализации с помощью метода t-SNE.
8	**Поиск аномалий в данных**	статистические методы нахождения выбросов. Вероятностные методы. One-Class SVM, Isolation Forest.

Домашние задания

1. Практический проект по построению системы поиска аномалий.

3 Применение методов машинного обучения к разным типам данных (текст, рекомендации, графы, временные ряды)

1 **Сбор данных**

открытые источники данных. Использование API. Парсинг и
создание своих датасетов.

Домашние задания

1 Практический проект по написанию собственного парсера.

---

2 **Анализ текстовых данных. Часть 1**

задача обработки текста. Введение, обзор задач, токенизация,
лемматизация, TF-IDF. Обзор библиотек для работы с русским и английским языками.

Домашние задания

1 Практический проект по предсказанию рейтинга фильма.

---

3 **Анализ текстовых данных. Часть 2**

тематическое моделирование. Общая схема решения задач NLP.

Домашние задания

1 Тематическое моделирование на данных Вконтакте: использование модели LDA, визуализация топиков, построение тематических профилей.

4	**Анализ текстовых данных. Часть 3. Практическое занятие**	векторные представления слов, word2vec. Примеры задач NLP, создание диалоговых систем.
5	**Рекомендательные системы. Часть 1**	коллаборативная фильтрация. Проблема «холодного старта». Метрики оценки качества рекомендательной системы.
6	**Рекомендательные системы. Часть 2**	контентная фильтрация, гибридные подходы. Ассоциативные правила. Домашние задания 1 Практический проект по созданию рекомендательной системы.
7	**Анализ временных рядов. Часть 1**	постановка задачи. Простейшие методы. Экспоненциальное сглаживание. Семейство ARIMA.
8	**Анализ временных рядов. Часть 2**	извлечение признаков и применение моделей машинного обучения. Автоматическое прогнозирование. Домашние задания 1 Построение прогноза временного ряда с использованием изученных методов.

9	**Алгоритмы на графах**	анализ социальных сетей. Метрики на графах. Выделение сообществ. Домашние задания 1 Анализ графа друзей Вконтакте. Визуализация в NetworkX.
10	**АБ тестирование**	тестирование гипотез. Постановка задачи. Терминология, мощность, статистическая значимость. Параметрические и непараметрические методы. Домашние задания 1 Практика по проверке АБ-тестов.

4 Дополнительные темы. Big Data

1	**Работа с Big Data. Часть 1**	адаптация алгоритмов к batch-learning. SGD. Vowpal Wabbit. Домашние задания 1 Настройка моделей машинного обучения в Vowpal Wabbit.
2	**Работа с Big Data. Часть 2**	облачные технологии для работы с Big Data: Amazon Web Services, Google Cloud. Создание виртуальных машин, распределенные вычисления. Домашние задания 1 Запуск собственной виртуальной машины и построение моделей в облаке.
3	**Работа с Big Data. Часть 3**	spark, принципы работы и архитектуры. Построение моделей машинного обучения при помощи PySpark API.
4	**Нейронные сети и глубокое обучение**	начальные сведения о нейронных сетях. Примеры использования нейронных сетей.
5	**Бонус: поиск Data Science работы**	примеры тестовых заданий и вопросов с собеседований.

1	**Вводное занятие по проектной работе**	проект включает в себя следующие этапы: 1. Постановка задачи. Предлагается самостоятельно найти предметную область и обосновать применение в ней машинного обучения 2. Разработка данных. Одно из требований к проекту - использование данных из открытых источников. Необходимо разработать процесс сбора и очистки данных 3. Поиск алгоритма и модели для решения задачи. Необходимо выполнить подготовку данных, выбрать алгоритм и подобрать параметры для построения модели 4. Использование модели для достижения поставленной цели 5. Построение процесса. Решение задачи необходимо оформить в единый процесс по обработке данных от источника до предсказания, не требующий участия эксперта 6. Обоснование процесса. Домашние задания 1 Проектная работа