Data Scientist

Курс предназначен для программистов и аналитиков, которые интересуются машинным обучением и анализом данных.

Начало занятий

В июле

Длительность курса: 136 академических часов

1 Введение в машинное обучение

1 **Базовые инструменты анализа данных в Python.**

Цели:
Участники узнают, какие задачи они смогут решать по окончании курса, научатся настраивать рабочее окружение и узнают функционал базовых библиотек для работы с данными в python.

Программа:
1. Обзор курса. Типы решаемых задач.
2. Окружение Python. Введение в Python, Numpy, Pandas, Sklearn. API Sklearn.

Домашние задания

1 Работы с данными в библиотеках numpy и pandas и реализация библиотеки для сбора данных

см. файл Описание ДЗ.pdf

2 **Необходимые понятия из мат. анализа и линейной алгебры**

Цели:
Участники освоят весь необходмый для данного курса материал из линейной алгебры и математического анализа: научатся решать задачи на собсвенные числа и собственные вектора матриц, находить производные функций и матричных выражений и применять это для задач оптимизации функций, эффективно применять данные алгоритмы в python.

Программа:
1. Линейная алгебра: Вектор, матрица, определитель матрицы, обратная матрица, собственные числа и вектора, норма, разложения матрицы (по собственным векторам, SVD).
2. Мат. анализ: Производная, интеграл, производные матричных выражений, якобиан, дифференцирование сложной функции.
3. Оптимизация: выпуклая, поиск глобального экстремума.
4. Типы данных в Python и векторизация вычислений (примеры + время работы).

---

3 **Необходимые понятия из теории вероятности**

Цели:
Участники изучат необходимые для курса основы теории вероятности: случайная величина, основные виды распределений случайных величин, научатся считать мат. ожидание, дисперсию случайных величин.

Программа:
1. Теория вероятности: вероятность (частотная и Байесовская трактовки), случайные величины, примеры распределений, мат. ожидание и т.д.
2. Экспоненциальное семейство распределений.
3. Основы статистики.
4. Примеры на Python.

4	**Визуализация**	Цели: Участники освоят основные библиотеки для визуализации данных в python, будут правильно выбирать виды графиков для визуализации данных разных типов. Домашние задания 1 Группировка и визуализация данных в Python + EDA https://drive.google.com/file/d/1UgVyxmOumex9-gFTnX583KtCWCIdkWyQ/view?usp=sharing
5	**Feature engineering**	1. Отбор признаков. 2. Преобразование исходных данных в подходящий для модели формат. 3. Преобразование признаков для повышения точности модели. 4. Выбор части признаков.
6	**Задача классификации. Метод ближайших соседей (kNN)**	Алгоритм kNN. Влияние нормализации данных в kNN. Структуры данных для оптимизации kNN. Кросс валидация. Методы оценки качества алгоритмов классификации.

7 **Линейная регрессия**

Цели:
Участники научатся делать описательный анализ данных с помощью библиотеки pandas и визуализацию данных с помощью различных библиотек python (matplotlib, seaborn, plotly, bokeh)

Программа:
1. Линейная регрессия - метод наименьших квадратов
2. Вероятностная трактовка линейной регрессии
3. Полиномиальная регрессия
3. Регуляризация в линейной регрессии

---

8 **Обучение с учителем. Логистическая регрессия**

Реализации логистической регрессии с помощью метода стохастического градиентного спуска

Домашние задания

1 Обучение с учителем. Логистическая регрессия.

См. homework.ipynb

---

9 **Практическое занятие по темам, изученным в 1 модуле**

2 Продвинутые алгоритмы машинного обучения

1 **Метод опорных векторов**

1. Метод опорных векторов(SVM), интерпретация.
2. Случай линейно неразделимых данных.
3. kernel trick, representer theorem, примеры ядер.
4. Пример SVM в sklearn.

---

2 **Деревья решений**

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

Домашние задания

1 Реализация алгоритма дерева решений на простых данных

Необходимо реализовать алгоритм дерева решения для задачи регрессии или классификации и сравнить результат с алгоритмом из библиотеки sklearn

---

3	**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 Обучение без учителя. Кластеризация См. homework-clustering.ipynb Реализовать один из алгоритмов кластеризации. Применение готовых алгоритмов кластеризации к датасету с Kaggle
4	**Иерархическая кластеризация, DB-Scan**	1. Иерархическая кластеризация 2. DB-Scan. 3. Optics. Спектральная кластеризация.
5	**Поиск выбросов в данных**	
6	**Методы уменьшения размерности**	1. Метод главных компонент (Principle component analysis). 2. Метод t-sne. 3. Примеры визуализации с помощью метода t-sne.
7	**Ансамбли моделей.**	1. Ансамблирование. 2. Случайный лес. 3. Бустинг, бэггинг, стекинг, блендинг.

8 **Градиентный бустинг**

1. Градиентный бустинг теория
2. Примеры библиотек: xgboost, catboost, lightgbm
3. Стекинг, блендинг

Домашние задания

1 Применение бустинга для построения лучшей модели

Применение бустинга для построения лучшей модели

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

1 **Анализ текстовых данных. Часть 1**

1. Сбор данных из открытых источников.
2. Очистка данных, подготовка данных для анализа.
3. Задача обработки текста. Введение, обзор задач, токенизация, лемматизация. TF-IDF.
4. Обзор библиотек для Python для работы с русским и английским языками.

---

2 **Анализ текстовых данных. Часть 2**

1. Выделение объектов в тексте.
2. Word2vec. Fast text.
3. Анализ тональности.
4. Автоматическое реферирование и тэгирование, классификация текстов.

Домашние задания

1 Реализация процесса сбора данных через API

Реализация процесса сбора данных через API VKontakte.
Преобразование текста, подготовка текста для анализа.
Применение машинного обучения для предсказания характеристик пользователей.

---

3 **Анализ текстовых данных. Часть 3. Тематическое моделирование**

1. метод pLSA.
2. метод LDA.
3. Применеие метода LDA для тематического моделирования новостных и научных статей

---

4 **Рекомендательные системы**

1. Коллаборативная фильтрация основанная на схожести пользователей и товаров.
2. Коллаборативная фильтрация основанная на факторизации матриц.
3. Проблема "холодного старта", контентная фильтрация, гибридные подходы.
4. Ассоциативные правила.
5. Метрики оценки качества рекомендательной системы.

Домашние задания

1 Сравнение разных алгоритмов рекомендательных систем

1. На тренировочных данных с рейтингами фильмов обучить следующие алгоритмы рекомендательных систем:
- user based collaborative filtering
- item based collaborative filtering
- 3. SVD без bias
- SVD
- Факторизационные машины с дополнительной информацией по пользователям и предметом рекомендаций и эффектом времени
2. Сравнить разные алгоритмы на валидационных данных и сделать выводы

5	**Анализ временных рядов**	1. Постановка задачи. 2. Экспоненциальное сглаживание. 3. Стационарность. SARIMA. Выбор признаков во временных рядах. 4. Применение моделей машинного обучения Домашние задания 1 Предсказание временных рядов 1. Скачать датасет X. 2. Натренировать модель ARIMA, перебором подобрать наилучший набор параметров. 3. Натренировать одну из ML моделей на предсказание следующего значения временного ряда.
6	**Алгоритмы на графах**	Социальные сети, выделение сообществ
7	**АБ тестирование**	1. Тестирование гипотез. Постановка задачи. 2. Терминология, мощность, статистическая значимость. 3. Параметрические методы: t-критерий, 1p, 2p proportion. 4. Непараметрические методы bootstrap
8	**Методы оптимизации**	SGD, модификации SGD

4 Нейронные сети и глубокое обучение

1 **Простейшие нейронные сети и метод обратного распространения ошибки.**

1. Начальные сведения о нейронных сетях.
2. Теорема об универсальной аппроксимации.
3. Алгоритм обратного распространения ошибки.

---

2 **Обучение нейронных сетей**

1. Пример к предыдущей лекции: разбор word2vec.
2. Предпосылки для глубоких нейронных сетей, представления.
3. Стохастический градиентный спуск.

Домашние задания

1 Простые НС и метод обратного распространения ошибки

1. Реализовать полносвязную сеть: два скрытых слоя с функцией активации ReLU, на выходе softmax по количеству классов (задается как параметр).
2. Обучить НС на модельный датасет "make_moons" из skelarn. Визуализировать разделяющую поверхность.
3. Обучить НС на датасете MNIST.

---

3 **Сверточные нейронные сети ч.1**

1. Структура сверточных сетей.
2. Пример на MNIST (Pytorch).
3. Обзор Pytorch.
4. Примеры на Pytorch (усложняем сеть, увеличиваем точность) -- сделать inclass соревнование на Kaggle.

4	**Сверточные нейронные сети ч.2**	1. Функции активации. 2. Регуляризация (BatchNorm, Dropout) 3. Инициализация весов. 4. Модификации SGD.
5	**Сверточные нейронные сети ч.3**	1. Ансамблирование. 2. Аугментация данных, transfer learning. 3. Использование предобученных сетей. Домашние задания 1 Сверточные сети на Pytorch 1. Реализовать сверточную сеть заданной архитектуры на Pytorch. 2. Написать слой BatchNorm и добавить его в НС. 3. Написать оптимизатор RMSProp и сравнить с обычным SGD.
6	**Рекуррентные сети ч.1**	1. Простой вариант: RNN. 2. LSTM, GRU. 3. Примеры для текста (языковая модель).
7	**Рекуррентные сети ч.2**	1. Механизм внимания: пример на переводе, на картинке. Типы внимания. 2. Пример: Image captioning. Домашние задания 1 Рекуррентные сети на Pytorch 1. Реализовать language model. 2. Выполнить задачу NER на датасете ConLL2003.

8 **Примеры работы глубоких НС**

1. Изображения и видео.
2. Текст: задача POS-tagging, NER, перевод.
3. Прочее.

5 Проектная работа

1 **Вводное занятие по проектной работе**

Домашние задания

1 Проектная работа

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. Необходимо реализовать применение разработанной модели
5. Построение процесса. Решение задачи необходимо оформить в единый процесс по обработке данных от источника до предсказания, не требующий участия эксперта
6. Обоснование процесса

---

2 **Консультация по проектной работе**

---

3 **Защита проектной работы**