# Машинное Обучение

**2-й модуль, 2-й курс**

**Евгегий Соколов**

**НИУ ВШЭ**

esokolov@hse.ru

## Описание курса

Это базовый курс по машинному обучению, представляющий наиболее популярные методы и подходы к извлечению знаний из данных. На лекциях будут рассмотрены такие темы, как линейные модели, методы, основанные на прецедентах (case-based), деревья решений и ансамбли. Мы также коснемся основ нейронных сетей. Студенты научатся предобрабатывать данные (включая категориальные и текстовые), выбирать и анализировать метрики качества для конкретной задачи, валидировать и оценивать модели. Все темы будут подкреплены практическими домашними заданиями на Python.

## Требования курса, система оценивания и посещаемость

Посещение и участие в занятиях приветствуются, но не являются обязательными. Итоговая оценка будет складываться из домашних заданий (60% оценки) и финального экзамена (40%). Экзамен пройдет в очном формате на кампусе. Домашние задания будут как теоретическими, так и практическими.

## Содержание курса

1. Типы данных, постановки задач и общий подход в машинном обучении.
2. Основные концепции машинного обучения. Метрики качества. Выбор модели, базовые функции потерь, кросс-валидация.
3. Модель линейной регрессии. Аналитическое решение и численный подход. Градиентный спуск для обучения модели. Переобучение и регуляризация. Функции потерь и работа с выбросами.
4. Модель линейной классификации. Обучение моделей классификации с использованием верхних оценок для бинарной функции потерь. Метрики качества для задач классификации. Многоклассовые и многометочные задачи и их сведение к бинарной классификации.
5. Деревья решений. Жадный подход к обучению. Функции неоднородности (impurity). Связь линейных моделей и деревьев решений.
6. Бэггинг. Разложение на смещение и дисперсию (Bias-variance decomposition). Анализ смещения и дисперсии для бэггинга. Случайный лес.
7. Градиентный бустинг на деревьях решений. Корректировка модели на основе аппроксимации остатков. Современные реализации градиентного бустинга.
8. Обучение без учителя. Кластеризация данных.
9. Визуализация данных. Метод t-SNE.

## Примеры задач для контроля знаний

Запишите модель линейной классификации. Сколько в ней параметров? Как оценить качество модели для несбалансированного набора данных? Как обучить эту модель на большом наборе данных? Как подготовить набор данных для ускорения градиентного спуска? Как применить кросс-валидацию для предотвращения переобучения? Как использовать регуляризацию и как выбрать ее силу?

## Материалы курса

Обязательного учебника нет, но полезными будут отдельные главы из книги Хейсти-Тибширани-Фридмана "The Elements of Statistical Learning".

## Политика академической честности

Списывание, плагиат и любые другие нарушения академической этики в РЭШ не допускаются.