

Программа курса "Введение в машинное обучение"


by Национальный исследовательский университет "Высшая школа экономики" & Yandex School of Data Analysis

НЕДЕЛЯ 1

Знакомство с анализом данных и машинным обучением

Добро пожаловать! В первом модуле курса мы расскажем о задачах, которые решает машинное обучение, определим базовый набор понятий и введем необходимые обозначения. Также мы расскажем про основные библиотеки языка Python для работы с данными (NumPy, Pandas, Scikit-Learn), которые понадобятся для выполнения практических заданий на протяжении всего курса.


 6 видео, 6 материалов для самостоятельного изучения [развернуть](#)


 **Оцениваемый:** Основные понятия машинного обучения


 **Оцениваемый:** Предобработка данных в Pandas

Логические методы классификации

Логические методы делают классификацию объектов на основе простых правил, благодаря чему являются интерпретируемыми и легкими в реализации. При объединении в композицию логические модели позволяют решать многие задачи с высоким качеством. В этом модуле мы изучим основной класс логических алгоритмов — решающие деревья. Также мы поговорим про объединение деревьев в композицию, называемую случайным лесом.

 4 видео [развернуть](#)


 **Оцениваемый:** Важность признаков


 **Оцениваемый:** Решающие деревья


НЕДЕЛЯ 2


Метрические методы классификации

Метрические методы проводят классификацию на основе сходства, благодаря чему могут работать на данных со сложной структурой — главное, чтобы между объектами можно было измерить расстояние. Мы изучим метод k ближайших соседей, а также способ его обобщения на задачи регрессии с помощью ядерного сглаживания.

 4 видео [развернуть](#)

 **Оцениваемый:** Метрические методы


 **Оцениваемый:** Выбор числа соседей


 **Оцениваемый:** Выбор метрики

Линейные методы классификации

Линейные модели — один из наиболее изученных классов алгоритмов в машинном обучении. Они легко масштабируются и широко применяются для работы с большими данными. В этом модуле мы изучим метод стохастического градиента для настройки линейных классификаторов, познакомимся с регуляризацией и обсудим некоторые тонкости работы с линейными методами.

 5 видео [развернуть](#)


 **Оцениваемый:** Линейные методы и градиентный спуск






 **Оцениваемый:** Нормализация признаков

НЕДЕЛЯ 3

Метод опорных векторов и логистическая регрессия


Линейные методы имеют несколько очень важных подвидов, о которых пойдет речь в этом модуле. Метод опорных векторов максимизирует отступы объектов, что тесно связано с минимизацией вероятности переобучения. При этом он позволяет очень легко перейти к построению нелинейной разделяющей поверхности благодаря ядровому переходу. Логистическая регрессия позволяет оценивать вероятности принадлежности классам, что оказывается полезным во многих прикладных задачах.



 5 видео [развернуть](#)

-  **Оцениваемый:** Особенности метода опорных векторов
-  **Оцениваемый:** Опорные объекты
-  **Оцениваемый:** Анализ текстов
-  **Оцениваемый:** Логистическая регрессия
-  **Оцениваемый:** Логистическая регрессия

Метрики качества классификации

В машинном обучении существует большое количество метрик качества, каждая из которых имеет свою прикладную интерпретацию и направлена на измерение конкретного свойства решения. В этом модуле мы обсудим, какие бывают метрики качества бинарной и многоклассовой классификации, а также рассмотрим способы сведения многоклассовых задач к двухклассовым.


 3 видео [развернуть](#)


-  **Оцениваемый:** Метрики качества классификации
-  **Оцениваемый:** Метрики качества классификации

НЕДЕЛЯ 4

Линейная регрессия


В этом модуле мы изучим линейные модели для регрессии и обсудим их связь с сингулярным разложением матрицы "объекты-признаки".


 3 видео [развернуть](#)

-  **Оцениваемый:** Линейная регрессия: прогноз оклада по описанию вакансии

Понижение размерности и метод главных компонент

В прикладных задачах часто возникает потребность в уменьшении количества признаков — например, для ускорения работы моделей. В этом модуле мы обсудим подходы к отбору признаков, а также изучим метод главных компонент, один из самых популярных методов понижения размерности.


 1 видео [развернуть](#)


-  **Оцениваемый:** Составление фондового индекса

НЕДЕЛЯ 5

Композиции алгоритмов

Объединение большого числа моделей в композицию может значительно улучшить итоговое качество за счет того, что отдельные модели будут исправлять ошибки друг друга. В этом модуле мы обсудим основные понятия и постановки задач, связанные с композициями, и обсудим один из наиболее распространенных способов их построения — градиентный бустинг.

 3 видео [развернуть](#)

 **Оцениваемый:** Размер случайного леса


 **Оцениваемый:** Градиентный бустинг над решающими деревьями

 **Оцениваемый:** Композиционные методы

Нейронные сети

Нейронные сети позволяют находить сложные нелинейные разделяющие поверхности, благодаря чему широко используются в таких трудных задачах, как распознавание изображений и речи. В этом модуле мы изучим многослойные нейронные сети и их настройку с помощью метода обратного распространения ошибки. Также мы поговорим о глубоких нейросетях, их архитектурах и особенностях.

 4 видео [развернуть](#)

 **Оцениваемый:** Нейронные сети

НЕДЕЛЯ 6

Кластеризация и визуализация


Этот модуль посвящен новому классу задач в машинном обучении — обучению без учителя. Под этим понимаются ситуации, в которых нужно найти структуру в данных или произвести их "разведку". В этом модуле мы обсудим две таких задачи: кластеризацию (поиск групп схожих объектов) и визуализацию (отображение объектов в двух- или трехмерное пространство).

 3 видео [развернуть](#)

Частичное обучение

Под частичным обучением понимается задача, находящаяся между обучением с учителем и кластеризацией: дана выборка, в которой значение целевой переменной известно лишь для части объектов. Такие ситуации встречаются, когда разметка объектов является дорогой операцией, но при этом достаточно дешево можно подсчитать признаки для объектов. В этом модуле мы обсудим отличия частичного обучения от рассмотренных ранее постановок, и разберем несколько подходов к

[▼Еще](#)


 3 видео [развернуть](#)

НЕДЕЛЯ 7

Машинное обучение в прикладных задачах

В этом модуле мы подведем итоги курса, вспомним основные этапы решения задачи анализа данных. Также мы разберем несколько задач из прикладных областей, чтобы подготовиться к выполнению финального проекта.

 6 видео [развернуть](#)

 **Оцениваемый:** Проект: предсказания победителя в онлайн-игре

Финальный проект

[Dota 2](#) — многопользовательская компьютерная игра жанра [MOBA](#). Игроки играют между собой матчи. В каждом матче, как правило, участвует 10 человек. Матчи формируются из живой очереди, с учётом уровня игры всех игроков. Перед началом игры игроки автоматически разделяются на две команды по пять человек. Одна команда играет за светлую сторону (The Radiant), другая — за тёмную (The Dire). Цель каждой команды — уничтожить главное здание базы противника, трон.

Вам нужно построить модель, которая по данным о первых пяти минутах матча будет предсказывать его исход — то есть определять команду-победителя.

Чтобы выполнить это задание, вам необходимо провести ряд исследований, сравнить несколько алгоритмов машинного обучения и проверить эффект от ряда манипуляций с признаками. Также, если вам понравится работать с этими данными, вы можете принять участие в [соревновании на Kaggle](#) и сравнить свои навыки с другими участниками курса!

К заданию приложены следующие файлы:

- final-statement.ipynb и final-statement.html — постановка задачи, описание данных, инструкции по выполнению
- features.zip — архив с обучающей выборкой
- features_test.zip — архив с тестовой выборкой
- data.zip — полный архив с сырыми данными и скриптом для извлечения признаков (этот архив понадобится вам только для участия в kaggle; для выполнения данного задания он не нужен)
- extract_features.py — скрипт, извлекающий признаки из сырых данных

Подход 1: градиентный бустинг "в лоб"

Один из самых универсальных алгоритмов, изученных в нашем курсе, является градиентный бустинг. Он не очень требователен к данным, восстанавливает нелинейные зависимости, и хорошо работает на многих наборах данных, что и обуславливает его популярность. В данном разделе предлагается попробовать градиентный бустинг для решения нашей задачи.

В отчете по данному этапу должны содержаться ответы на следующие вопросы:

1. Какие признаки имеют пропуски среди своих значений (приведите полный список имен этих признаков)? Что могут означать пропуски в этих признаках (ответьте на этот вопрос для двух любых признаков)?
2. Как называется столбец, содержащий целевую переменную?
3. Как долго проводилась кросс-валидация для градиентного бустинга с 30 деревьями? Инструкцию по измерению времени можно найти выше по тексту. Какое качество при этом получилось?
4. Имеет ли смысл использовать больше 30 деревьев в градиентном бустинге? Что можно сделать, чтобы ускорить его обучение при увеличении количества деревьев?

Подход 2: логистическая регрессия

Линейные методы работают гораздо быстрее композиций деревьев, поэтому кажется разумным воспользоваться именно ими для ускорения анализа данных. Одним из наиболее распространенных методов для классификации является логистическая регрессия. В данном разделе предлагается применить ее к данным, а также попробовать различные манипуляции с признаками.

В отчете по данному этапу должны содержаться ответы на следующие вопросы:

1. Какое качество получилось у логистической регрессии над всеми исходными признаками? Как оно соотносится с качеством градиентного бустинга? Чем можно объяснить эту разницу? Быстрее ли работает логистическая регрессия по сравнению с градиентным бустингом?
2. Как влияет на качество логистической регрессии удаление категориальных признаков (укажите новое значение метрики качества)? Чем можно объяснить это изменение?
3. Сколько различных идентификаторов героев существует в данной игре?
4. Какое получилось качество при добавлении "мешка слов" по героям? Улучшилось ли оно по сравнению с предыдущим вариантом? Чем можно это объяснить?
5. Какое минимальное и максимальное значение прогноза на тестовой выборке получилось у лучшего из алгоритмов?