Машинное обучение на практике

Код: EAS-025

Длительность: 24 ч.

Описание:

Вводный курс - начало длинного и интересного пути по изучению машинного обучения. Вместо классического академического подхода «математика – теория ML – примеры – практика» данный курс в первую очередь ориентирован на практику. Теория и математическое обоснование, а именно: как работают алгоритмы и, главное, почему они работают в принципе, тоже очень важны. Но их изучать более разумно на следующих этапах, когда встанет вопрос об улучшении и оптимизации результатов.

Практический пример начнётся с жизненной задачи: есть набор таблиц с исходными данными (в Excel), но непонятно, что с ними можно сделать и как ко всему этому применить «волшебную» BigData. С помощью техник машинного обучения будет показано, как можно разобраться с закономерностями в этих данных и каким образом можно представить результаты в виде, понятном бизнес-заказчикам (графики, новые более простые таблицы и тому подобное).

После этого будут рассмотрены базовые классы задач, где машинное обучение эффективно, и также на примере будет показано, как данные задачи можно решить. Но, конечно, просто применение готовых формул малоэффективно – внимание будет уделено также и трактовке результатов, и в немалой степени представлению результатов конечным потребителям данных.

На дискуссионной части курса можно будет обсудить стоящие перед слушателями практические задачи, попробовать их формализовать и даже предсказать возможные трудности в реализации.

Цели:

- Понять, какие задачи можно решать машинным обучением (и узнать, что Big Data это всего лишь подраздел, а не обязательное требование).
- Научиться применять начальные методы машинного обучения и с помощью быстрого прототипирования научиться отвечать на вопрос «оценить реальную прибыль от возможного внедрения».
- Подсветить, какие данные необходимо собирать и что может потребоваться от них в ближайшем будущем. Почему «хотим хранить петабайты» это не всегда просто прихоть.

- Подготовится к более сложным темам, в частности – к полным решениям реальных сложных бизнес-задач.
- Посмотреть, как именно машинное обучение стыкуется с классической аналитикой. В частности, убедиться, что не обязательно (и даже вредно) увольнять всех существующих аналитиков для внедрения концепции.

Разбираемые темы:

1. Обзор задачи (1 час – теория).

- Какие задачи хорошо решаются машинным обучением, а какие им пытаются решать.
- Что произойдёт, если вместо Data Scientist взять неспециалиста в данной области (просто разработчика/аналитика/менеджера) с ожиданием, что в процессе научится.

2. Подготовка, очистка, исследование данных (1 час – теория, 1 час – практика).

- Как разобраться в исходных бизнес-данных (и вообще обнаружить в них какой бы то ни было порядок).
- Последовательность обработки.
- Что можно и нужно переложить на аналитиков предметной области, а что лучше сделать самому Data Scientist.
- Приоритеты решения конкретной задачи.

3. Классификаторы и Регрессоры (2 часа – теория, 2 часа – практика).

- Практический раздел - хорошо формализованные задачи с подготовленными данными.
- Разница между задачами (бинарная/небинарная/вероятностная классификация, регрессии), перераспределение задач между классами.
- Примеры классификации практических задач.

4. Кластеризация (1 час – теория, 2 часа – практика).

- Где и как проводить кластеризацию: исследование данных, проверка постановки задачи, проверки результатов.
- Какие случаи можно свести к кластеризации.

5. Что такое хорошо (1 час – теория, 1 час – практика).

- Как оценивать результаты, как привыкли это делать заказчики.
- Объяснение непривычных оценок или сведение их к привычным.
- Частные бессмысленные вопросы и что на них ответить.
- Кросс-валидация и как её делать не надо.
- Удивительные примеры оверфита и как он проникает в даже чуть небрежную

архитектуру.

6. Как улучшать модель (5 часов – теория, 3 часа – практика).

- Что делает одну модель лучше другой: параметры, признаки, ансамбли.
- Немного про параметры.
- Детально про признаки, с практикой построения и соревнованиями. Как не переборщить с признаками.
- Взгляд в бездну инструментария для поиска лучших параметров/признаков/методов.

7. Графики, отчеты, работа с живыми задачами (2 часа – теория, 2 часа – практика).

- Как доступно объяснить происходящее: себе, команде, клиенту.
- Более красивые ответы на бессмысленные вопросы.
- Как презентовать три терабайта результатов на одном слайде.
- Полуавтоматические тесты, какие точки контроля процесса действительно нужны.
- От живых задач к полному R&D процессу («НИОКР на практике») – разбор и анализ задач от аудитории.

Целевая аудитория:

Основная:

- Аналитики
- Менеджеры проектов, связанных с данными
- Технические лидеры / ведущие разработчики в любых проектах, связанных с данными
- Бизнес-аналитики

Дополнительная:

- Разработчики
- Инженеры данных (Data Engineer)
- Архитекторы, системные проектировщики

Предварительная подготовка – общее:

Умение читать простой код на Python и писать на любом скриптовом языке.

Рекомендуемые дополнительные материалы, источники:

Pandas:

- Their own doc is very nice https://pandas.pydata.org/docs/user_guide/index.html

- Book https://amzn.to/2KI5JJw
- https://mlcourse.ai topics 1, 2

Seaborn

- Their own doc https://seaborn.pydata.org/tutorial.html

Machine Learning

- Sklearn’s doc https://scikit-learn.org/stable/user_guide.html
- https://mlcourse.ai
- Kaggle, in particular https://www.kaggle.com/notebooks