**Проект месяц-бота: неделя №1 - Линейная и логистическая регрессия**

Всем привет! А вот и первая тема, на этой неделе вам предстоит разобраться в двух самых базовых линейных моделях. Линейная регрессия способна предсказать определенно числовое значение - стоимость дома, возраст человека и так далее, логистическая регрессия в свою очередь способна классифицировать объекты - определить одобрить кредит или нет, классифицировать по характеристикам котов и собак и тд. Сейчас данные алгоритмы применяют крайне редко, но все равно нельзя представить ML специалиста без их знания.

Материалы что вы можете пройти:

* Лекция о линейной регрессии от Deep Learning School - <https://www.youtube.com/watch?v=khdaLtu9i-s&t=4s&ab_channel=DeepLearningSchool>
* Основы статистики 3 пункт Karpov - <https://stepik.org/course/76/syllabus>
* Учебник от ШАД - [**https://ml-handbook.ru/chapters/linear\_models/intro**](https://ml-handbook.ru/chapters/linear_models/intro)
* Линейная регрессия в Sklearn - <https://pythobyte.com/linear-regression-in-python-with-scikit-learn-bc1b4512/>
* Основы линейной регрессии на Хабр - <https://habr.com/ru/post/514818/>
* Логистическая регрессия на Хабр - https://habr.com/ru/post/485872/

Дополнительные материалы:

* Регуляризация в линейной регрессии <https://www.youtube.com/watch?v=L_o8v5A23XA&ab_channel=DeepLearningSchool>

Небольшой дисклеймей от меня: ссылки отправленные выше содержат в себе очень много математики и в самом начале все это сложно обработать и понять. Лучше всего сначала понять как на словах работает линейная и логистическая регрессия, затем попробовать её на python в sklearn, а затем уже посмотреть как это устроено с точки зрения языка математики.

Данный курс или проект по боту прежде всего направлен на развитие у вас самостоятельности в отношении поиска информации и решении задачи. Если вы посмотрели ссылки выше и не поняли что и как, то вы можете спокойно найти в интернете как устроена линейная регрессия простыми словами. От нас же вы можете получить ответы на ваши вопросы и проверку правильности выполнения домашнего задания.

В домашнем задании вам предстоит реализовать линейную регрессию для определения цены на алмазы(предсказать колонку price).

Вам предстоит сделать несколько шагов для этого:

* Загрузить данные в python
* Обработать данные в pandas(как обработать категориальные данные - https://habr.com/ru/post/511132/), дополнительным заданием тут будет замена нескольких пропущенных знаний(вам стоит посмотреть как лучше всего заменить пропущенные значения в данных)
* После этого разделить данные на обучающую и тестовую часть(https://pythobyte.com/split-data-training-and-testing-set-34994/) и затем обучить линейную регрессию
* Проверить результаты на тестовых данных и оценить результаты, метрики для оценки качества вам нужно найти самим. Для повышения качества можно использовать множество лайфхаков, тут вам поможет гугл. вы можете как угодно изменять данные и вычленять из них новые признаки, но обязательное условие использовать Linear Regression из sklearn библиотеке.

На этом все. Всё в ваших руках, пишите по поводу любых вопросов.

ДЗ отсылается в виде ноутбуа Google Collab или Jupyter Notebook мне в личку до 3 октября.

Описание данных:

price price in US dollars (\$326--\$18,823)

carat weight of the diamond (0.2--5.01)

cut quality of the cut (Fair, Good, Very Good, Premium, Ideal)

color diamond colour, from J (worst) to D (best)

clarity a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

x length in mm (0--10.74)

y width in mm (0--58.9)

z depth in mm (0--31.8)

depth total depth percentage = z / mean(x, y) = 2 \* z / (x + y) (43--79)

table width of top of diamond relative to widest point (43--95)

**Проект месяц-бота: неделя №2 - Кластеризация и понижение размерности**

Всем привет! На этой неделе вам предстоит научиться кластерезировать данные и понижать их размерность - обучение без учителя.

Кластеризация это в какой то мере классификация, но без знания правильных ответов. Она часто применяется, так с помощью нее можно анализировать данные, создавать на ее основе новые фичи и включать в более большую архитектуру, например сначала может идти кластеризация, а к ее результатам применяться нейронные сети. Понижение размерности в свою очередь служит для удобной визуализации многомерных данных или оптимизации скорости работы модели.

Материалы что вы можете пройти:

* Лекция о кластеризации от Deep Learning School - [https://www.youtube.com/watch?v=BzbBnztHm6U&ab\_channel=DeepLearningSchool](https://www.youtube.com/watch?v=khdaLtu9i-s&t=4s&ab_channel=DeepLearningSchool)
* Основы статистики 2, 3 пункт Karpov - [https://stepik.org/course/524/promo](https://stepik.org/course/76/syllabus)
* Учебник от ШАД - [**https://ml-handbook.ru/chapters/clustering/intro**](https://ml-handbook.ru/chapters/linear_models/intro)
* Лекция о кластеризации от Deep Learning School - [https://www.youtube.com/watch?v=7wbRQF6qfRU&ab\_channel=DeepLearningSchool](https://www.youtube.com/watch?v=khdaLtu9i-s&t=4s&ab_channel=DeepLearningSchool)
* Снижение размерности PCA на Хабр - [**https://habr.com/ru/post/304214/**](https://habr.com/ru/company/ods/blog/325654/)

Дополнительные материалы:

* Что такое дендограмма <https://www.displayr.com/what-is-dendrogram/>

В домашнем задании вам предстоит реализовать кластеризацию для нескольких разных датасетов созданных мной(там нужно правильно выбрать алгоритм, для этого сильно поможет главная страничка sklearn для кластеризации) и для реальных данных. Тут вам предстоит не столько предобработать данные, сколько просто провести по ним кластеризацию и оценить результат

 Вам предстоит сделать несколько шагов для этого:

* Загрузить данные в python
* Посмотреть графики данных, применить понижение размерности до 2 при необходимости
* Отобразить на графике проведенную кластеризацию
* Как дополнительно задание на этой неделе будет оценка кластеризации. При этом где то нужно оценить не зная правильных ответов, а где то правильные ответы будут только для нескольких строк.

На этом все. Всё в ваших руках, пишите по поводу любых вопросов.

ДЗ отсылается в виде ноутбуа Google Collab или Jupyter Notebook мне в личку до 10 октября не включительно.

**Проект месяц-бота: неделя №3 - Деревья решений и случайный лес**

Всем привет! На этой неделе вам предстоит разобраться в деревьях решений и в ансамблевой модели основанной на них - случайном лесе. Деревья решений позволяют классифицировать данные и лидировали в Ml долгое время. Сейчас самый лучший классический алгоритм машинного обучения основан тоже на них, но поговорим мы про него на след неделе.

Материалы что вы можете пройти:

* Обобщенная статья на Хабр о деревьях решений - [https://habr.com/ru/company/productstar/blog/523044/](https://www.youtube.com/watch?v=khdaLtu9i-s&t=4s&ab_channel=DeepLearningSchool)
* Введение в машинное обучение 2, 3 пункт Karpov - [https://stepik.org/course/4852/syllabus](https://stepik.org/course/76/syllabus)
* Учебник от ШАД - <https://ml-handbook.ru/chapters/decision_tree/intro>
* Реализация случайного леса на Python - https://tproger.ru/translations/python-random-forest-implementation/
* Лекция о решающих деревьях от Deep Learning School - [https://www.youtube.com/watch?v=MJwAoWFTMWw&ab\_channel=DeepLearningSchool](https://www.youtube.com/watch?v=khdaLtu9i-s&t=4s&ab_channel=DeepLearningSchool)
* Лекция о случайном лесе от Deep Learning School - **<https://www.youtube.com/watch?v=vqF8wrWjR5s&ab_channel=DeepLearningSchool>**

Дополнительные материалы:

* Визуализация дерева решений [https://stackoverflow.com/questions/27817994/visualizing-decision-tree-in-scikit-learn](https://www.displayr.com/what-is-dendrogram/)

В домашнем задании вам предстоит реализовать классификацию стоимости(price\_range) мобильных телефонов по их параметрам с помощью случайного леса. В качестве дополнительных заданий будут: подбор гиперпараметров для случайного леса(GridSearchCV, RandomSearchCV), осознание вашими словами термина переобучения и как его распознать.

 Вам предстоит сделать несколько шагов для этого:

* Загрузить данные в python
* Посмотреть графики данных и обработать их
* Создать новые признаки(если это возможно)
* Обучить случайный лес на test и train
* Подобрать гиперпарметры
* Оценить результат (основная метрика тут F1-score

Ради интереса можете построить после случайного леса одно дерево решений и посмотреть результаты, удивительно как множество одинаковых алгоритмов с немного измененными данными выдают гораздо лучший результат.

На этом все. Всё в ваших руках, пишите по поводу любых вопросов.

ДЗ отсылается в виде ноутбуа Google Collab или Jupyter Notebook мне в личку до 17 октября не включительно.

Еще один совет от меня, лучшей Карпова случайный лес и деревья решений мало кто объяснит. поэтому чтобы легко все понять лучше сразу смотреть его курс, он в материалах выше.

**Проект месяц-бота: неделя №4 - Градиентный бустинг**

Всем привет! На этой неделе вам предстоит разобраться в градиентном бустинге. Это самый мощный метод из классического ML. Чаще всего именно он показывает лучший результат, при этом иногда даже без настройки. Если кратко то он состоит из множества других алгоритмов, при этом создает каждый из них учитывая ошибку предыдущего. С каждым новым алгоритмом он компенсирует недостатки предыдущих и в итоге получается хороший результат.

Материалы что вы можете пройти:

* Тяжелая статья на Хабр от ODS - [https://habr.com/ru/company/ods/blog/327250/](https://www.youtube.com/watch?v=khdaLtu9i-s&t=4s&ab_channel=DeepLearningSchool)
* Учебник от ШАД - [https://ml-handbook.ru/chapters/grad\_boost/intro](https://ml-handbook.ru/chapters/decision_tree/intro)
* Лекция о градиентном бустинге от Deep Learning School - <https://www.youtube.com/watch?v=JElfEE1OrSU&ab_channel=DeepLearningSchool>
* Статья о градиентном бустинге на Хабр - <https://habr.com/ru/company/ods/blog/645887/>

Обрадую вас, домашнее задание не меняется с прошлой недели. Вам предстоит обработать тот же самый датасет и сравнить, что себя лучше покажет randomforest и градиентный бустинг. Но на этом не все, дополнительные задания на баллы:

* Провести исследование и определить есть ли в данных мультеколенеарность
* Сравнить три главных алгоритма градиентного бустинга(Catboost, LightGBM, XGBoost)

 Вам предстоит сделать несколько шагов для этого:

* Во время обработки данных исследовать есть ли мультеколенеарность и удалить столбцы если есть
* После randomforest построить модели бустинга и возможно подобрать параметры. (GridSearchCV не работает с catboost, там есть свой поиск)
* Сравнить результаты

На этом все. Всё в ваших руках, пишите по поводу любых вопросов.

ДЗ отсылается в виде ноутбуа Google Collab или Jupyter Notebook мне в личку до 24 октября не включительно.

Материалов в этот раз не много, так как на русском языке не богатый выбор. Постарался выделить не самые плохие статьи.