Семинар 6: Регрессия

В этом семинаре мы впервые столкнёмся с настоящим машинным обучением и попробуем понять что стоит за его магией. В ручной части семинара мы пойдём по следующему плану:

- разберёмся чем классификация отличается от регрессии ;
- сформулируем задачу регрессии и поймём её специфику;
- поймём с помощью каких метрик можно оценить качество прогноза в случае регрессии;
- попробуем разобраться какой смысл стоит за этими метриками;
- разберёмся как выглядит простейшая линейная модель регрессии;
- на пальцах прикинем как она обучается.

Задача 1 (формулируем задачу)

Представьте себе, что у вас есть паблик с мемами. Вы — Хозяин мемов. Как и любой другой Хозяин мемов, вы любите лайки под мемами. Возникает желание привлечь в паблик целевую аудиторию, которая будет ставить под мемы лайки. Для этого вы хотите запустить рекламную кампанию паблика. Ясное дело, что рекламу хочется показывать не всем подряд, а только подходящим людям.

У вас есть данные по профилям всех тех людей, которые уже ставили в паблике лайки. По этим данным вам хочется построить модель, котороя могла бы предсказать подходит ли конкретный человек для вашей рекламной компании (поставил бы ли он в паблик лайк, если бы был на него подписан).

1. Сформулируйте задачу машинного обучения. Какой должна быть целевая переменная, чтобы перед вами была задача классификации. Какой должна быть целевая переменная, чтобы это была задача регрессии?
2. Какие факторы из профилей вы бы использовали, чтобы спрогнозировать подходит ли человек для рекламной кампании?
3. Приведите ещё парочку примеров задачи классификации и задачи регрессии.

Задача 2 (качество прогноза)

Добрыня, Алёша и Илья смотрят мемы и ставят на них лайки. Мы пытаемся предсказать сколько лайков они оставят под мемами на основе поведения их однокурсников. Для этого мы оценили регрессию. Ну и она нам напредсказывала, что парни поставят 4, 20 и 110 лайков. В реальности они поставили 5, 10 и 100 лайков. Возникает вопрос: насколько сильно наша модель ошиблась в прогнозировании. Для того, чтобы выяснить это используют различные метрики. Давайте посмотрим на основные.

Что такое MAE, MSE, RMSE и MAPE? Посчитайте для модели все четыре метрики качества.

Задача 3 (как выглядит модель)

Предположим, Олег хочет купить автомобиль и считает сколько денег ему нужно для этого накопить[1]. Он пересмотрел десяток объявлений в интернете и увидел, что новые автомобили стоят около 20000, годовалые — примерно 19000, двухлетние — 18000 и так далее.

В уме Олег-аналитик выводит формулу: адекватная цена автомобиля начинается от 20000 и падает на 1000 каждый год, пока не упрётся в 10000. Олег сделал то, что в машинном обучении называют регрессией — предсказал цену по известным данным. Давайте попробуем повторить подвиг Олега.

а) Как выглядит формула в случае Олега?

б) За сколько продать старый айфон? Придумайте формулу для предсказания. Проинтерпретируйте каждый коэффициент в ней.

в) Сколько одежды брать с собой в путешествие? Придумайте формулу для предсказания. Проинтерпретируйте каждый коэффициент в ней.

г) Сколько шашлыка брать на дачу? Как выглядит формула?

д) Сколько брать шашлыка, если есть толстый друг? Как можно назвать толстого друга в терминах машинного обучения? Испортит ли толстый друг формулу?

Было бы удобно иметь формулу под каждую проблему на свете. Но взять те же цены на автомобили: кроме пробега есть десятки комплектаций, разное техническое состояние, сезонность спроса и еще столько неочевидных факторов, которые Олег, даже при всём желании, не учел бы в голове. Люди тупы и ленивы — надо заставить вкалывать роботов.

Задача 4 (как обучаются модели)

Давайте попробуем совсем-совсем на пальцах почувствовать как модели обучаются. Пусть у Хозяина мемов есть две переменные: x — возраст подписчика, y — число лайков, которое он оставил. Хозяин мемов хочет оценить регрессию $y = \beta \cdot x$, то есть он хочет попытаться предсказать число лайков по возрасту подписчика. Хозяин собрал два наблюдения для оценивания модели: $x_1 = 15, y_1 = 10$ и $x_2 = 22, y_2 = 2$.

Теперь хозяину надо подобрать коэффициент β так, чтобы ошибка прогноза, измеряемая с помощью MSE оказалась поменьше.

1. Пусть $\beta = 1$. Какие значения нам спрогнозирует модель? Какая у неё будет ошибка?
2. Пусть $\beta = 0.5$. Найдите прогнозы и ошибку модели.
3. Какое значение для β нам больше подходит? Как можно найти оптимальное β?

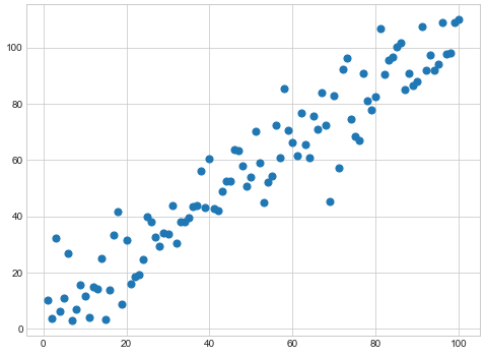
[1] сделано по мотивам https://vas3k.ru/blog/machine_learning/

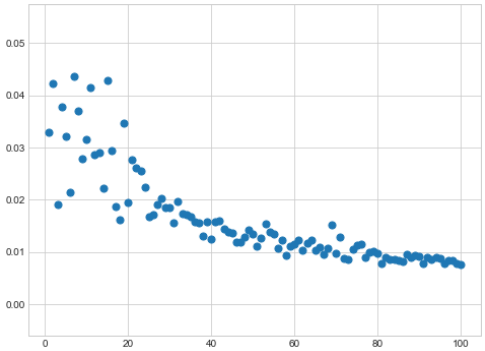
Ещё задачи

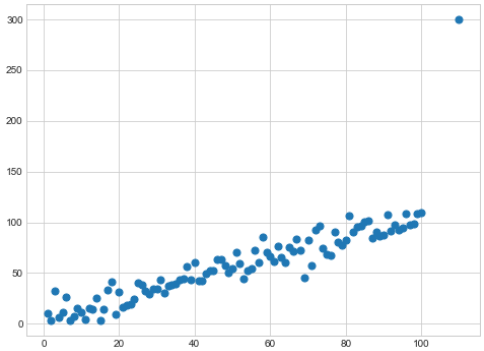
Тут находится несколько задачек, о которых вам нужно подумать самостоятельно. Возможно, что похожие задачи попадутся вам на самостоятельной работе.

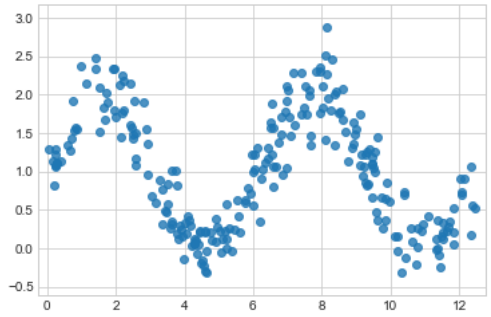
Задача 5 (картинки)

Вот несколько ситуаций, как на ваш взгляд должны пройти линии регрессии? Да, это тоже машинное обучение. Но обычно кривые рисуем не мы, а комплюхтер.









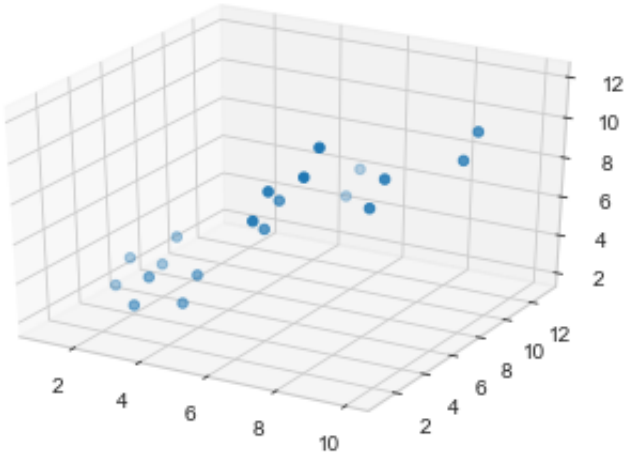
а) Нарисуйте на каждой из картинок линию регрессии.

б) Как выглядят уравнения регрессии в этих ситуациях? Какие параметры в них нам нужно обучить?

в) В чём проблема на картинке слева снизу? Проинтерпретируйте её на примере шашлыков.

г) В четвёртой ситуации мы выбрали для обучения полином. А почему бы не взять его в каждой ситуации и не обучить через каждую точку?

д) Ещё одна, на этот раз трёхмерная картинка! Слабо дополнить её также, как мы делали это выше? Как будет выглядеть уравнение регрессии?



Задача 6

Драгомир пытается предсказать продажи видео-игр. Для моделирования он использует две переменные: x_1 — возраст игры, x_2 — на кого она ориентирована. Если на мужчин, $x_2 = 1$, если на женщин, $x_2 = 0$. Целевая переменная y — сумма продаж. Драгомир оценил линейную регрессию:

$$y = 1000 - 100 \cdot x_1 + 200 \cdot x_2.$$

Проинтерпретируйте полученные коэффициенты. Предположим, что мы выпускаем на рынок свежую игру для женщин. Спрогнозируйте наши продажи.

Задача 7

Маше 13 лет. Всю свою жизнь она занималась коллекционированием моделей. Вчера она пообщалась с Мишей. Он тоже коллекционер. Он спросил у неё, какое у её моделей качество? Маша не смогла ответить и решила проверить его. У неё есть три наблюдения y_i. Она для каждого построила прогнозы. Найдите для её прогнозов MAE, MSE, RMSE и MAPE. В чём измеряются эти ошибки? Проинтерпретируйте их.

y_i	1	2	3
Нейросеть	2	3	1
Регрессия	2	3	4
Случайный лес	1	1	1

Задача 8

Объясните мемас:

Задача 9

В какой из следующих ситуаций какую метрику качества вы бы использовали? Почему? Постарайтесь обосновать свой выбор с точки зрения бизнеса. Для удобства будем в каждом пункте обозначать наблюдаемый спрос/цену/вес и тп как y_i, а наши прогнозы как $\hat{y}_i$.

- Марк и Семён решили накачаться. У них есть модель, которая прогнозирует их набор массы в зависимости от рациона. По вердиктам этой модели парни заказывают себе еду на неделю. В конце недели они взвешиваются и смотрят насколько сильно модель ошиблась.
- Коста собирается устроить вечеринку. Для неё ему понадобится шашлык (внезапно). Есть прогнозная модель, которая позволяет прикинуть сколько шашлыка ему понадобится. Он построил прогнозы для прошлой вечеринки и задумался о том какую метрику нужно выбрать для оценки качества её работы. На прошлой тусе ели только шашлык. На этой тусе, кроме шашлыка будет ещё и пицца. Предположим, что у Косты есть один толстый друг. Возможно, он будет есть шашлык, но вот вообще не факт.
- Кот Матроскин продаёт молоко. Каждую неделю он прогнозирует спрос на молоко и поставляет его в торговые точки в соответствии с прогнозами. Каждая бутылка продаётся за 40 рублей. Если он завозит слишком много, молоко портится и Матроскин теряет деньги, потраченные на производство молока и различные логистические издержки (перевозка и т.п.). Потери составляют 60 рублей с каждой просроченной бутылки.
- Мистер Белфорд торгует на бирже. Каждую неделю он отсчитывается перед инвесторами, рассказывая сколько дополнительных процентов от стартовых инвестиций принёс его алгоритм торговли. Внутрь алгоритма вшита прогнозная модель. Как лучше всего мистер Белфорд мог бы оценить её качество?