Семинар 5: Регрессия

> «Нарисуй линию вдоль моих точек. Да, это искусственный интеллект.»
>
> Герман Греф

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:

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

Упражнение 1 (ставим 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, которая могла бы предсказать подходит ли конкретный человек для вашей рекламной компании (поставил бы ли он в паблик лайк, если бы был на него подписан).

а) 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?

б) Какие факторы из профилей вы бы использовали, чтобы спрогнозировать подходит ли человек для рекламной кампании?

в) Приведите ещё парочку примеров задачи классификации и задачи регрессии.

Упражнение 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.

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

Упражнение 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 друг-вегетарианец? Как можно назвать этого друга в терминах машинного обучения? Испортит ли вегетарианец 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 ещё столько неочевидных факторов, которые Олег, даже при всём желании, не учёл бы в голове. Люди тупы и ленивы — надо заставить вкалывать роботов.

Упражнение 4 (как обучаются модели)

Давайте попробуем совсем-совсем на пальцах почувствовать, как модели обучаются. Пусть у Хозяина мемов есть две переменные: x — возраст подписчика и 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?

Упражнение 5 (одинокий дуб)

[1] Сделано по мотивам статьи “Машинное обучение для людей”, прочтите её: https://vas3k.ru/blog/machine_learning/

Для того чтобы решать задачу регрессии и прогнозировать что-нибудь, можно пытаться искать коэффициенты в уравнениях, которые мы выписывали выше. Это один из вариантов модели. Он называется **линейной регрессией**. Линейной, потому что мы пытаемся провести через облако точек линию. Можно пробовать оценивать и какие-то другие, более сложные, нелинейные модели. Например, можно построить **регрессионное дерево**. Было бы нечестно бросать вас не обучив ручками ни одной модели. Давайте обучим!

Миша работает в маленькой кофейне. Харио Малабар Монсун — фирменный напиток этой кофейни. Мише интересно узнать, как именно ведёт себя количество заказов напитка y_i в зависимости от температуры за окном t_i. Четыре дня Миша записывал свои наблюдения:

t_i	y_i
21	1
19	2
12	8
8	8

Сегодня он решил обучить регрессионное дерево. В качестве функции потерь он использует

$$\text{MSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

а) Обучите регрессионное дерево.

б) Какой прогноз на сегодня сделает дерево Миши, если за окном 13 градусов?

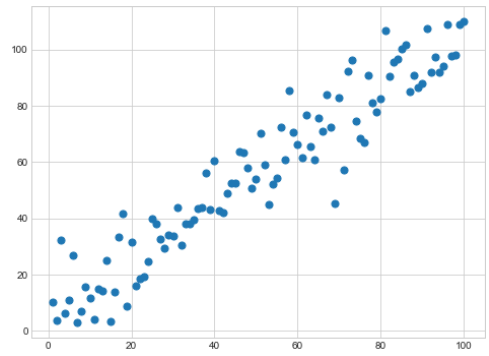
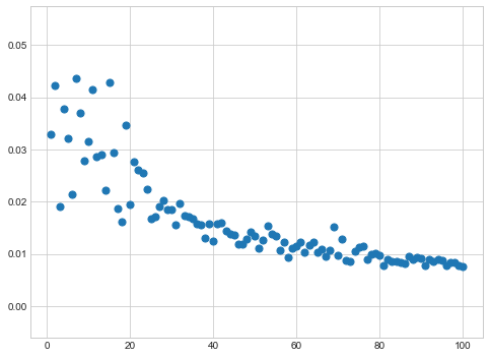
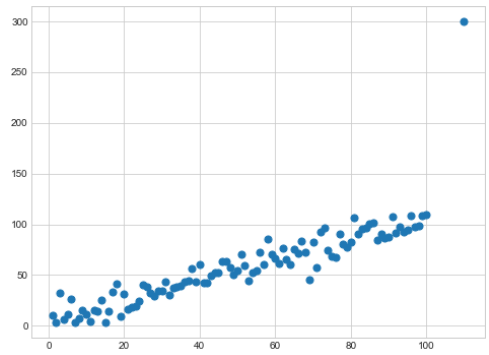
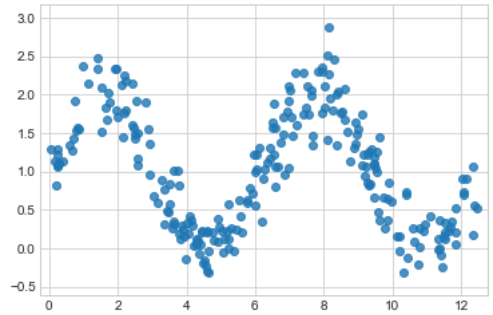
в) Можно ли для обучения дерева использовать MAE?

Ещё задачи

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

Упражнение 6

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

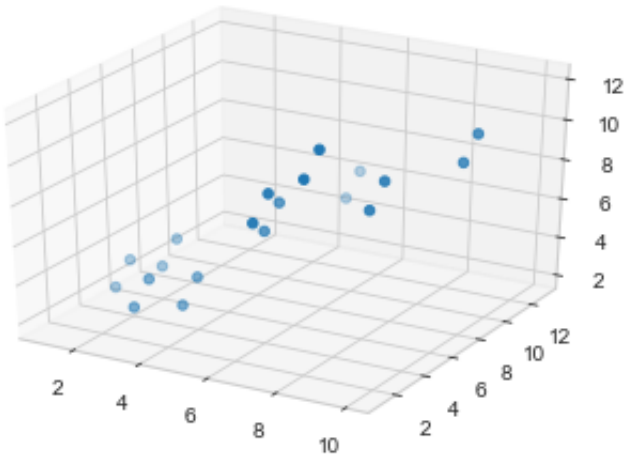
а) Нарисуйте на каждой из картинок линию регрессии.

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

в) В чём проблема на картинке слева снизу? Приведите пример ситуации, когда может наблюдаться такая картинка.

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?



Упражнение 7

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 — количество проданных экземпляров игры. 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.

Упражнение 8

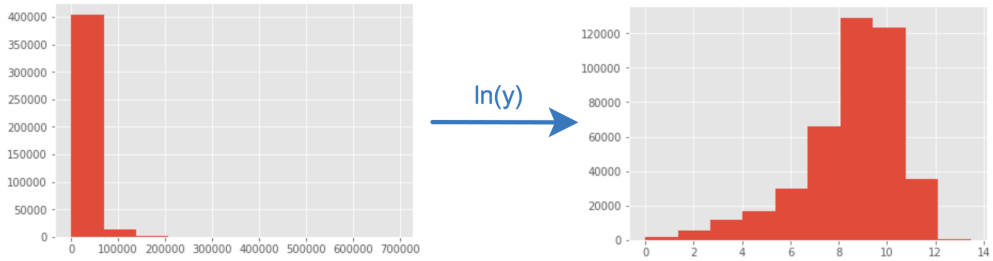
Мстислаполк, конкурент Драгомира, тоже пытается предсказать продажи видео-игр. Для моделирования он использует две переменные: x — возраст игры. Целевая переменная y — сумма продаж. Мстислаполк оценил линейную регрессию:

$$\ln y = 5 - 6 \cdot \ln x.$$

Проинтерпретируйте полученный коэффициент. Предположим, что мы отгружаем на рынок новую партию игры, выпущенной в прошлом году. Сколько экземпляров этой игры будет продано?

Упражнение 9

Логарифмирование позволяет сгладить длинные хвосты распределений, кишащие выбросами. Из-за этого на практике переменные довольно часто логарифмируют. На картинке ниже изображена гистограмма продаж в супермаркетах Wallmart. По оси x отложена сумма продаж, по оси y число продаж на такую сумму.



Понятное дело, что люди делают покупки на огромные суммы, но в маленьком количестве. Отсюда у распределения появляется огромный хвост. Если прологарифмировать продажи, распределение станет няшным.

Попробуйте посмотреть как именно происходит это сглаживание. Предположим, что в магазинах продали $y_1 = 100$, $y_2 = 200$, $y_3 = 300$ и $y_4 = 1000$ игр. Посчитайте разницу между соседними наблюдениями. Прологарифмируйте их. Что стало с этой разницей?

Упражнение 10

В один прекрасный день Маша проснулась в своей кровати и поняла, что она и есть та самая маша, которой принадлежит лёрнинг. Она решила посвятить машин лёрнингу всю свою жизнь и стала коллекционировать модели.

Вчера она пообщалась с Мишей. Он тоже коллекционер. Он спросил у неё, какое у её моделей качество. Маша не смогла ответить. Ей было очень стыдно[2]. Она решила проверить качество. У неё есть три наблюдения y_i. Она для каждого построила прогнозы. Найдите для её прогнозов MAE, MSE, RMSE и MAPE.

настоящие y_i	1	2	3
прогнозы нейросети	2	3	1
прогнозы регрессии	2	3	4
прогнозы случайного леса	1	1	1

Упражнение 11

Объясните мемас:

Упражнение 12

Выращиваем регрессионное дерево в домашних условиях! Вот вам выборка для этого:

x_i	y_i
0	5
1	6
2	4
3	100

Критерий деления вершины — минимизация квадратичной функции потерь (MSE). Критерий остановки — три листа. Зачем нужен критерий остановки? Как дерево ведёт себя с выбросами?

[2]Прям как вам после самостоятельной на следущей паре

Упражнение 13

Бернард не очень хорошо умеет в маркетинг и управление бизнесом, а ещё он — владелец книжного магазина. Он обратил внимание, что чем чаще в магазин заходят покупатели, тем чаще он пьёт. Чай. Ещё бывают постоянные покупатели, которые, в общем-то, ничего не покупают, а только делятся проблемами, выпивают чай и уходят.

Бернард хочет прикинуть свои расходы на чай в следующем месяце. Какую информацию ему надо добыть и где её достать? Какие метрики использовать при построении модели? Какие метрики использовать при оценке её качества?