

Семинар 4-5: Продажи и линейная регрессия

Задача 1

Предположим, Олег хочет купить автомобиль и считает сколько денег ему нужно для этого накопить¹. Он пересмотрел десяток объявлений в интернете и увидел, что новые автомобили стоят около 20000, годовалые — примерно 19000, двухлетние — 18000 и так далее.

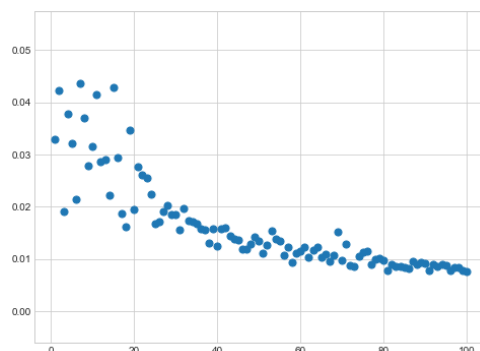
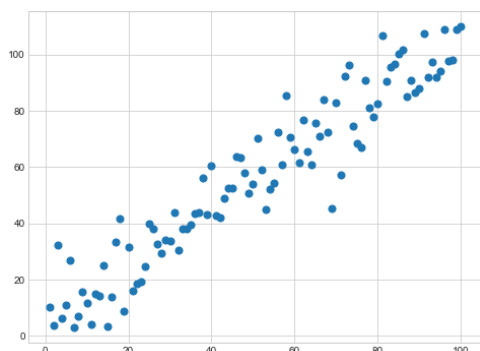
В уме Олег-аналитик выводит формулу: адекватная цена автомобиля начинается от 20000 и падает на 1000 каждый год, пока не упрётся в 10000. Олег сделал то, что в машинном обучении называют регрессией — предсказал цену по известным данным. Давайте попробуем повторить подвиг Олега.

- а) Как выглядит формула в случае Олега?
- б) За сколько продать старый айфон? Как выглядит формула?
- в) Сколько шашлыка брать на дачу? Как выглядит формула?
- г) Сколько брать шашлыка, если есть толстый друг? Как можно назвать толстого друга в терминах машинного обучения? Испортит ли толстый друг формулу?
- д) Сколько одежды брать с собой в путешествие?

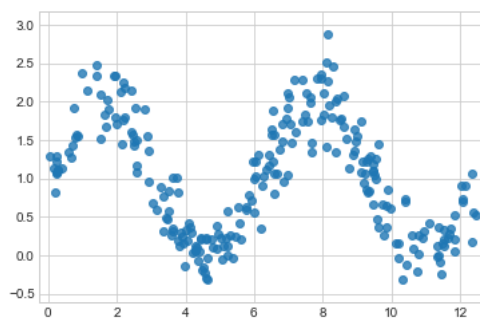
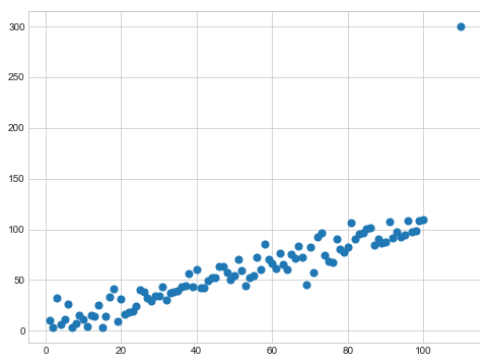
Было бы удобно иметь формулу под каждую проблему на свете. Но взять те же цены на автомобили: кроме пробега есть десятки комплектаций, разное техническое состояние, сезонность спроса и еще столько неочевидных факторов, которые Олег, даже при всём желании, не учел бы в голове. Люди тупы и ленивы — надо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

Задача 2

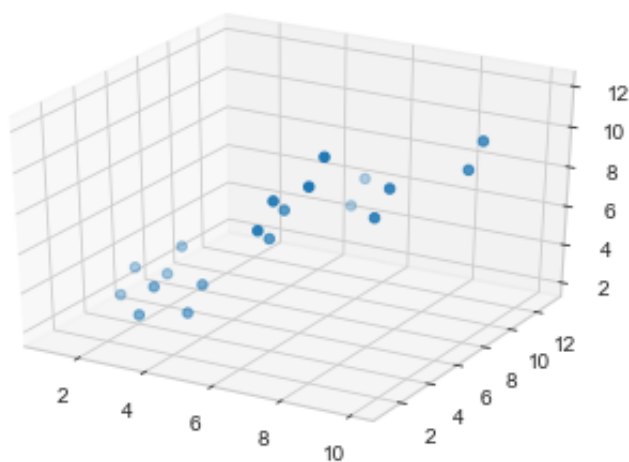
Вот несколько ситуаций, как на ваш взгляд должны пройти линии регрессии? Да, это тоже машинное обучение. Но обычно кривые рисуем не мы, а комплюхтер.



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



- Нарисуйте на каждой из картинок линию регрессии.
- Как выглядят уравнения регрессии в этих ситуациях? Какие параметры в них нам нужно обучить?
- В чём проблема на картинке слева снизу? Проинтерпретируйте её на примере шашлыков.
- В четвёртой ситуации мы выбрали для обучения полином. А почему бы не взять его в каждой ситуации и не обучить через каждую точку?
- Ещё одна, на этот раз трёхмерная картинка! Слабо дополнить её также, как мы делали это выше? Как будет выглядеть уравнение регрессии?



Задача 3

Вася измерил вес трёх упаковок с конфетками, $y_1 = 6$, $y_2 = 6$, $y_3 = 10$. Вася хочет спрогнозировать вес следующего пакетика. Модель для веса пакетиков у Васи очень простая, $y_i = \mu + u_i$, поэтому прогнозирует Вася по формуле $\hat{y}_i = \hat{\mu}$.

Для оценки параметра μ Вася использует следующую целевую функцию:

$$\sum (y_i - \hat{\mu})^2 + \lambda \cdot \hat{\mu}^2$$

- Найдите оптимальное $\hat{\mu}$ при $\lambda = 0$.
- Найдите оптимальное $\hat{\mu}$ при произвольном λ .

- в) Подберите оптимальное λ с помощью кросс-валидации «выкинь одного».
- г) Найдите оптимальное $\hat{\mu}$ при λ_{CV} .

Задача 4

Миша работает в маленькой кофейне. Харио Малабар Монсун является фирменным напитком этой кофейни. Мише интересно узнать как именно ведёт себя спрос на напиток 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

$$\sum (y_i - \hat{y}_i)^2.$$

- а) Обучите регрессионное дерево.
- б) Какой прогноз на сегодня сделает дерево Миши, если за окном 13 градусов?

Задача 5

Предположим, что в наших руках оказались исторические данные по продажам 45 магазинов Walmart, расположенных в разных регионах. Каждый магазин содержит несколько отделов. Нам хотелось бы научиться прогнозировать продажи по каждому отделу.

- а) Зачем нам может понадобиться прогнозировать продажи? Какая от этого выгода для магазина?
- б) Какую задачу машинного обучения нам предстоит решать? Какие переменные мы могли бы использовать в качестве объясняющих?
- в) Какую метрику мы могли бы использовать для оценки бизнес-эффекта от нашей модели? Отталкиваясь от каких характеристик можно было бы сконструировать её?
- г) Что такое MAE, MSE, RMSE и MAPE? Предположим, что у нас есть три магазина. Они продали товаров на 5, 10 и 100 рублей. Наша модель предсказывала, что они продадут товаров на 4, 20 и 110. Посчитайте для нашей модели все четыре метрики качества, приведённые выше.

Ещё задачи

Задача 6

Маркетологи Вова и Вася строили регрессию $y = \beta_0 + \beta_1 x$. Каждый оценивал её по своим данным. У Васи получилось, что $\hat{\beta}_1 = 2$, у Вовы получилось, что $\hat{\beta}_1 = 8$.

Пришла Алиса, отобрала у Вовы и Васи данные, соединила их вместе и построила регрессию сразу на всём. У неё получилось, что $\hat{\beta}_1 = -10$. Может ли такое быть?

Задача 7

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. Критерий остановки — три листа. Зачем нужен критерий остановки? Как дерево ведёт себя с выбросами?

Задача 8

Каждый день Маша ест конфеты и решает задачи по машинному обучению. Пусть x_i — количество решённых задач, а y_i — количество съеденных конфет.

x_i	y_i
1	1
2	2
2	8

Рассмотрим модель $y_i = \beta x_i + u_i$. Маша использует функцию потерь

$$\sum (y_i - \hat{\beta} x_i)^2$$

- Найдите МНК-оценку $\hat{\beta}$ для имеющихся трёх наблюдений.
- Нарисуйте исходные точки и полученную прямую регрессии.
- Выведите формулу для $\hat{\beta}$ в общем виде для n наблюдений.

- г) На семинаре по машинному обучению неожиданно выяснилось, что Миша тоже каждый день решает задачи по машинному обучению. Правда он более сдержан в плане конфет. Миша решил взять Машины наблюдения и с помощью функционала

$$\sum |y_i - \hat{\beta}x_i|$$

оценить β . Помогите Мише найти оценку.

- д) К поеданию конфет решает присоединиться Вадик. У него тоже есть своя функция потерь

$$\sum (y_i - \hat{\beta}x_i)^2 + 3\beta^2$$

Оцените β для его случая. Нарисуйте все три прямые на одной картинке и порассуждайте почему они получились именно такими, какими получились.