



Онлайн-образование

Меня хорошо видно && слышно?

Ставьте + , если все хорошо
Напишите в чат, если есть проблемы

Проверить, идет ли запись!



«Задача регрессии. Линейная регрессия»



Андрей Канашов

Data Scientist
OMD OM GROUP
@Андрей Канашов

Преподаватель



Андрей Канашов

- Data Scientist в OMD OM GROUP
 - Кластерный анализ целевых аудиторий
 - Персонализация рекламы
 - Анализ социальных сетей

Правила вебинара



Активно участвуем



Задаем вопрос в чат или голосом



Off-topic обсуждаем в Slack #канал группы или #general



Вопросы вижу в чате, могу ответить не сразу

Цели вебинара | После занятия вы узнаете

1

Линейная регрессия

2

Метрики качества

3

Применение на практике

4

Переобучение и регуляризация

Задачи машинного обучения

Задачи машинного обучения

С учителем
Supervised learning



Классификация



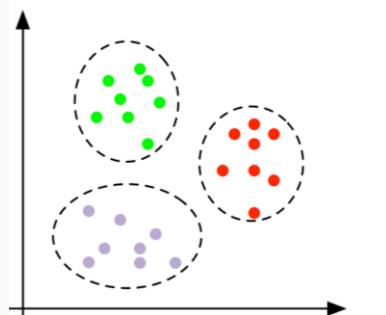
Регрессия



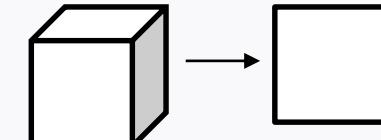
Без учителя
Unsupervised learning



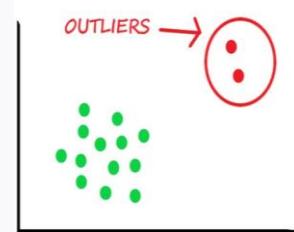
Кластеризация



Снижение размерности



Поиск аномалий

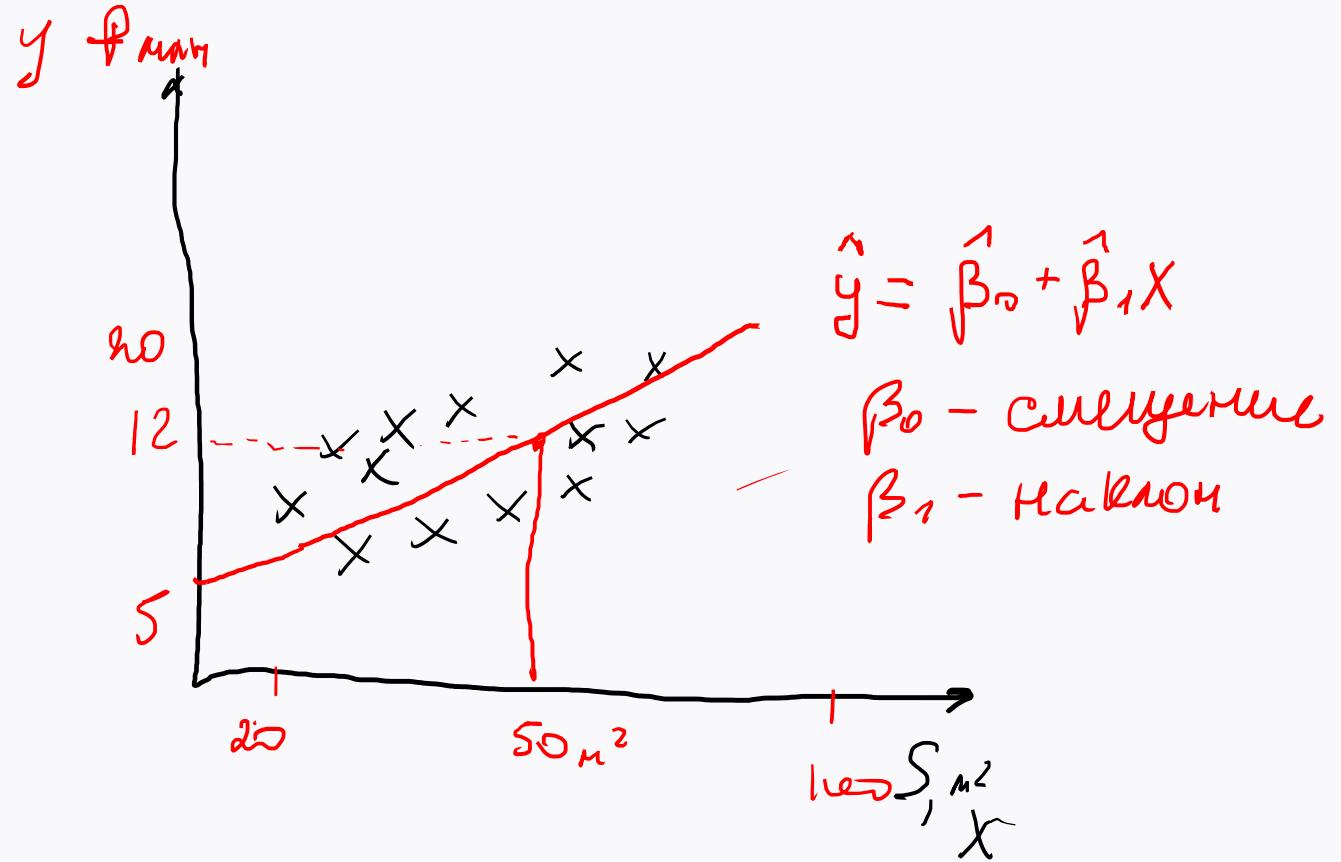
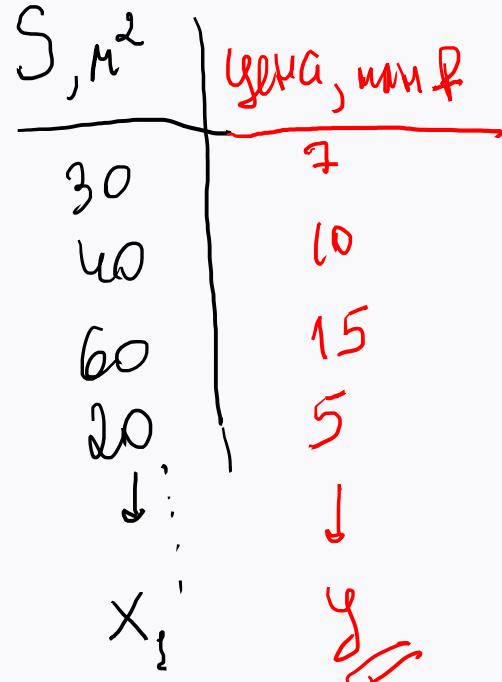




Задача регрессии Линейная регрессия

Задача регрессии

Цель – спрогнозировать непрерывную величину.

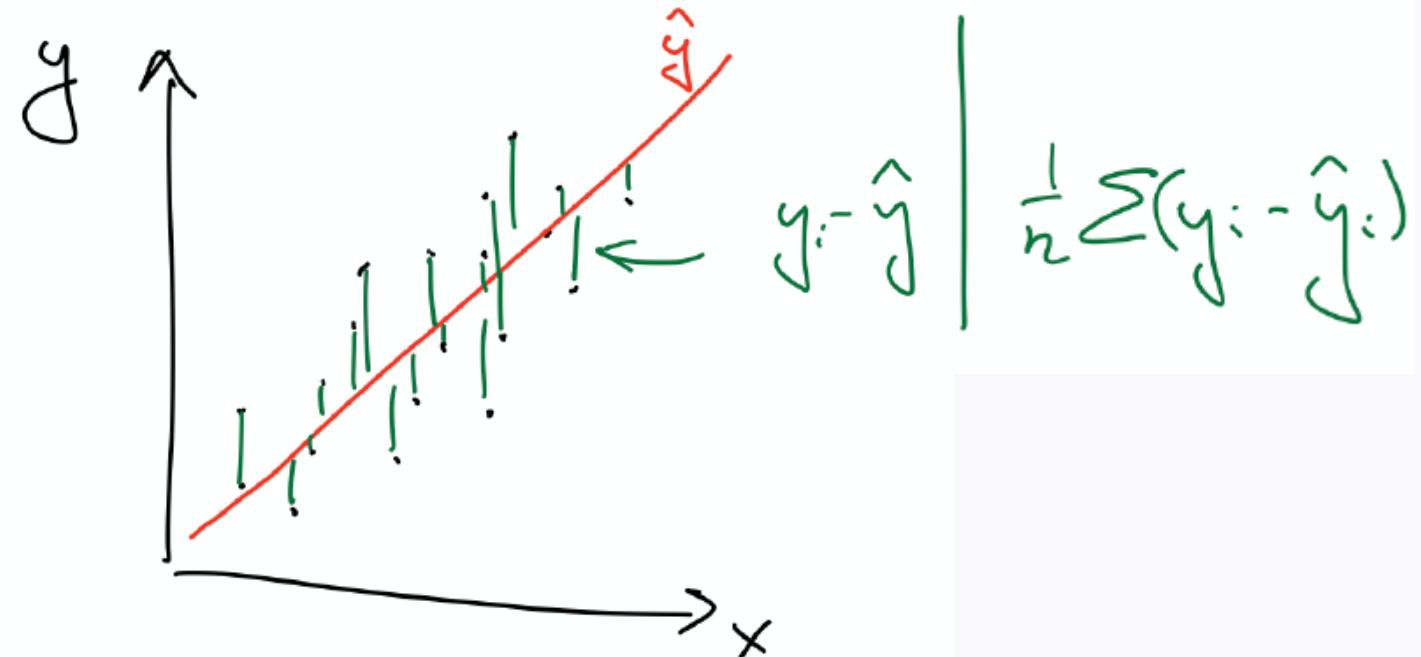
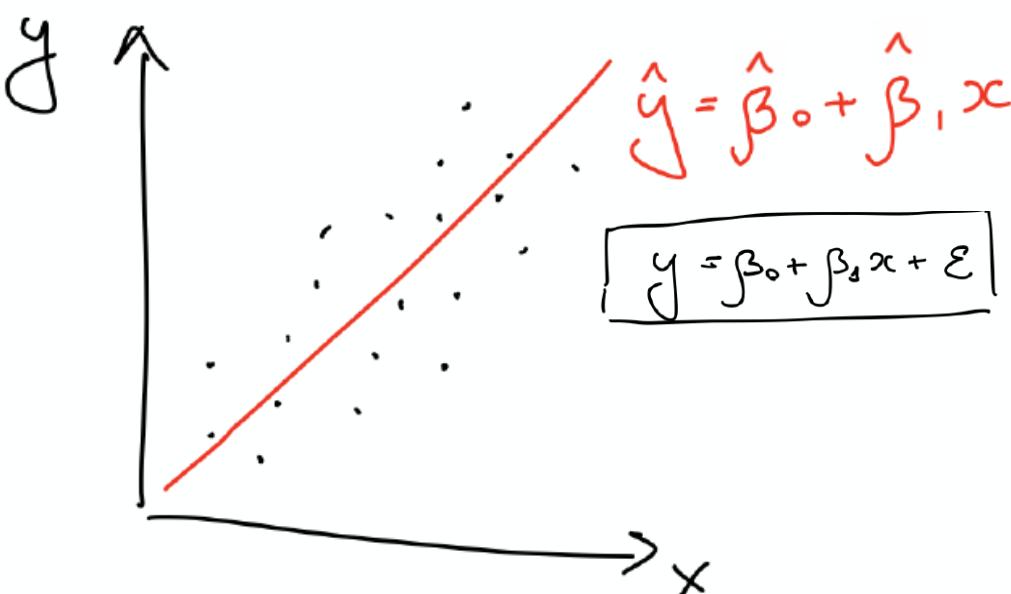


Интерактив:

https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html

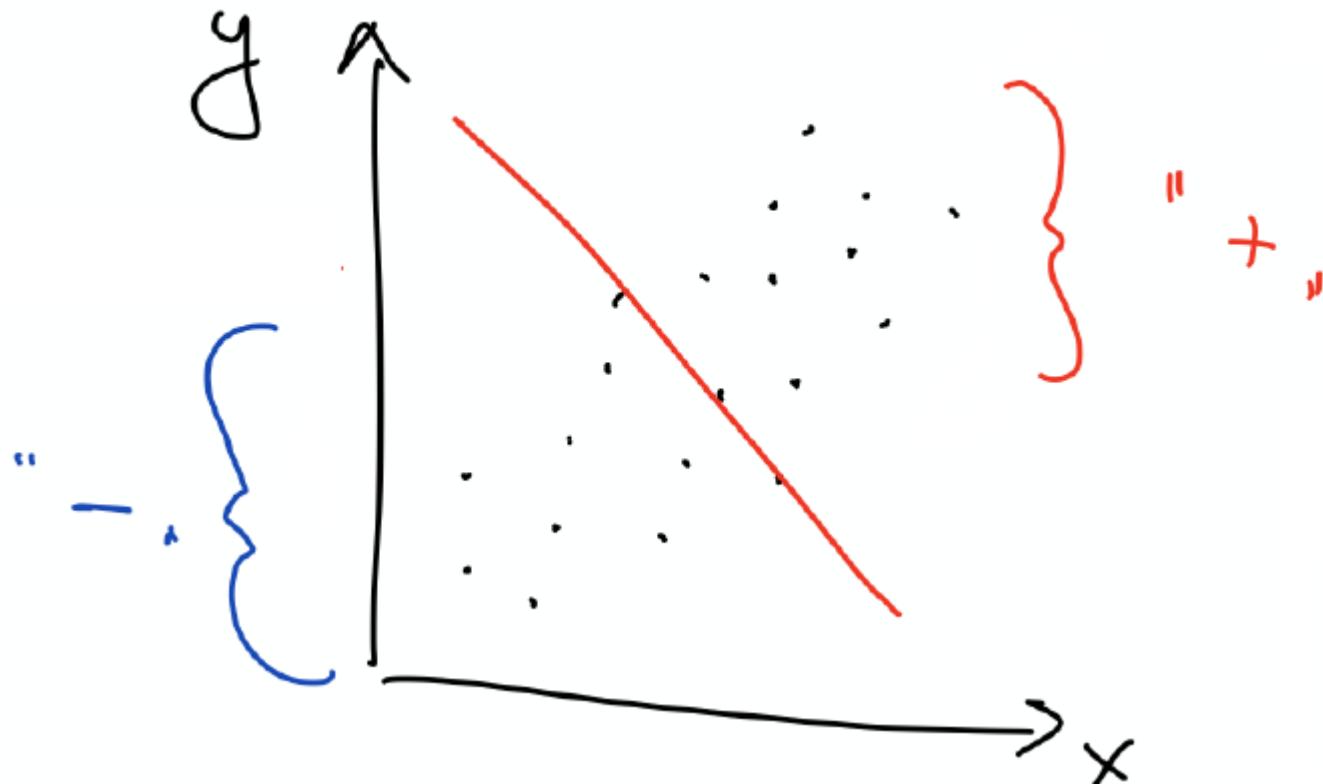
Линейная регрессия

Как обучать линейную регрессию?



В чем тут может быть проблема?

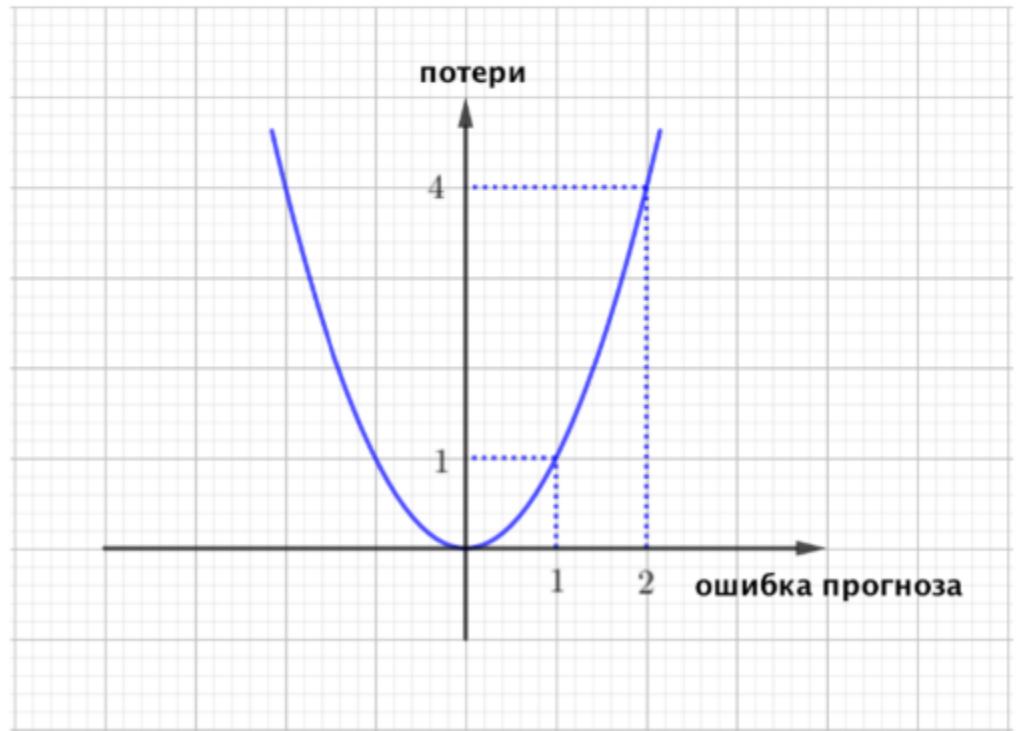
Линейная регрессия



$$\frac{1}{n} \sum (y_i - \hat{y}_i) \approx 0$$

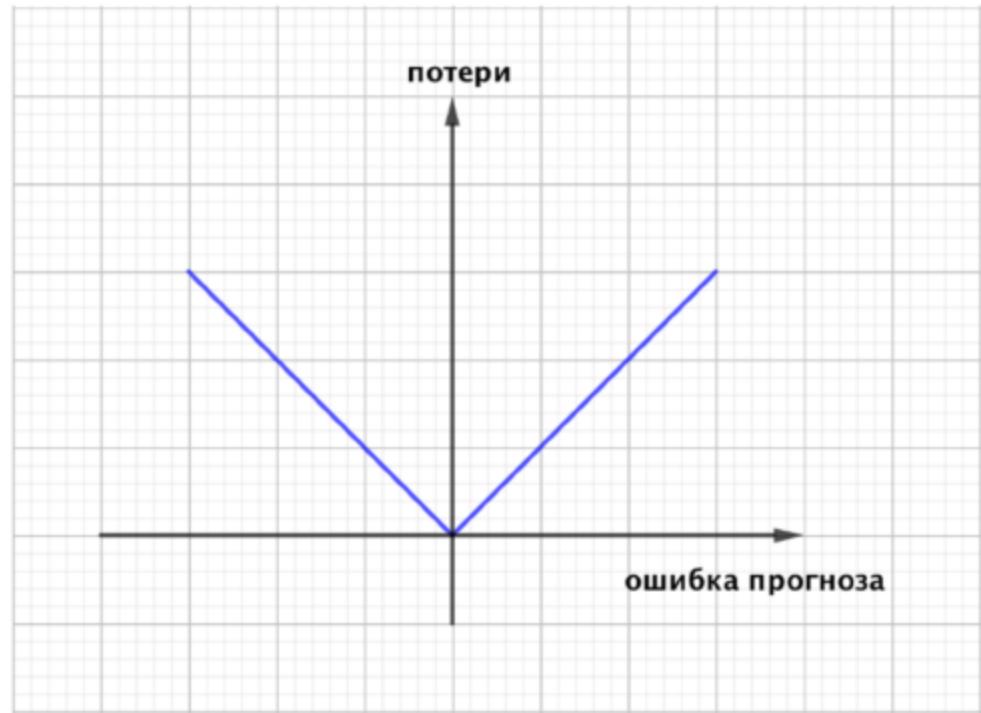
Mean Squared Error (MSE)

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

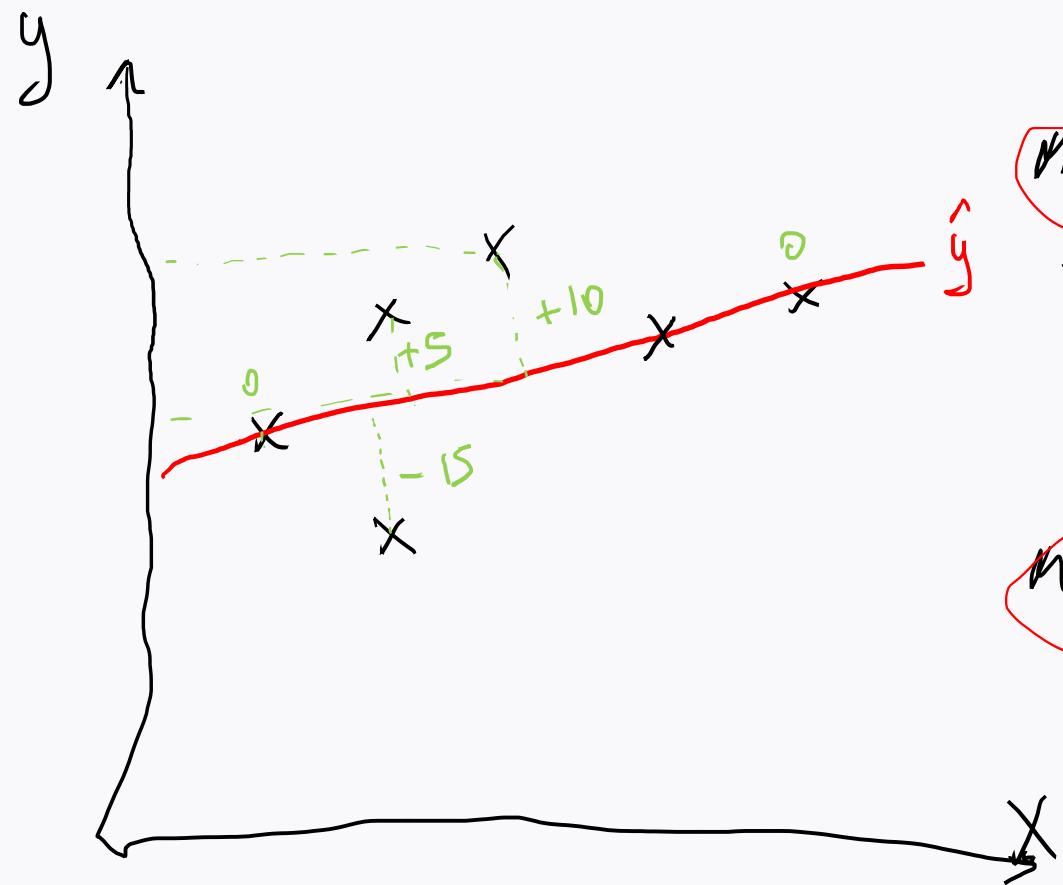


Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|.$$



Расчёт метрик



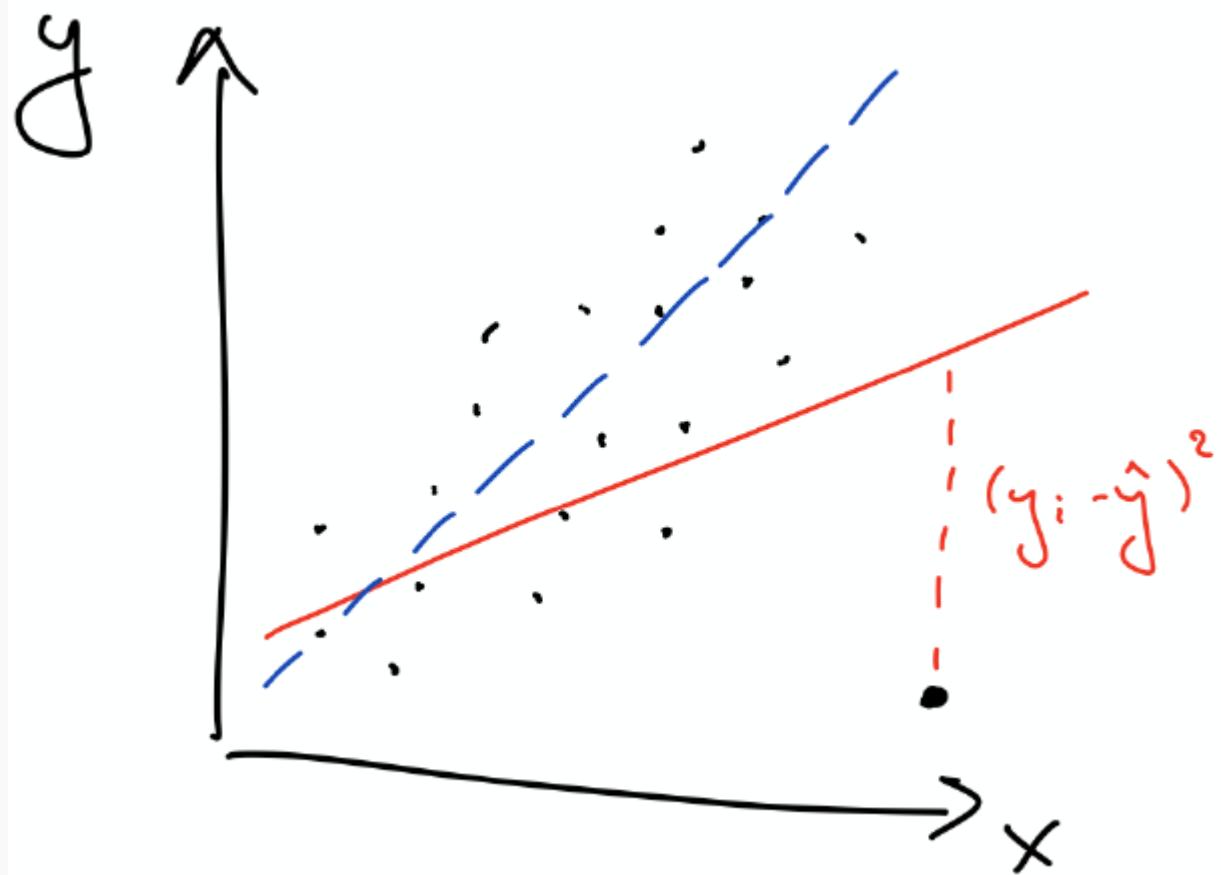
$$\frac{1}{n} \sum y_i - \hat{y}_i = \frac{1}{6} (0 - 15 + 5 + 10 + 0 + 0) = \frac{1}{6} \times 0 = 0$$

$$\text{MSE: } \frac{1}{n} \sum (y_i - \hat{y}_i)^2 = \frac{1}{6} (0 + (-15)^2 + 5^2 + 10^2 + 0^2 + 0^2) = \frac{1}{6} \cdot 225 + 25 + 100 = \frac{350}{6} \approx 60$$

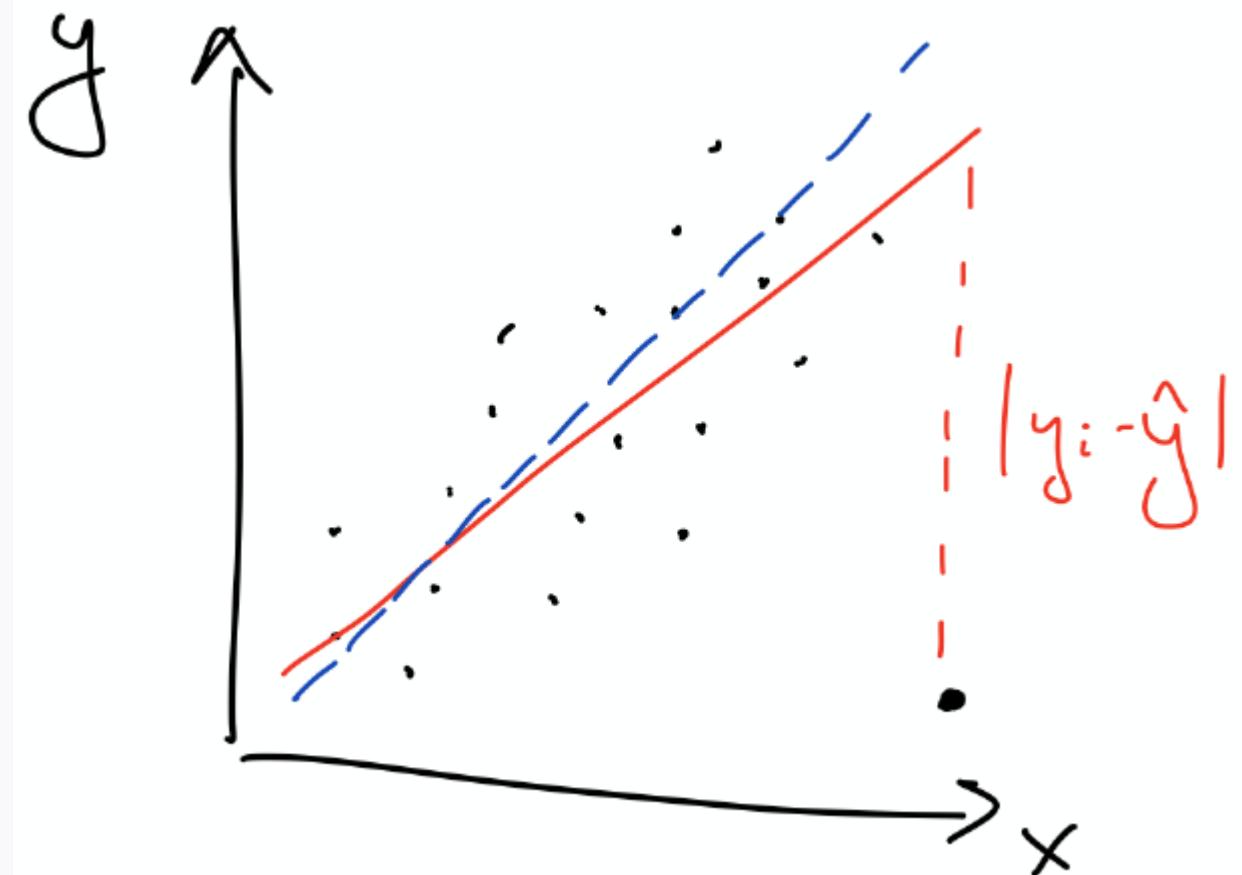
$$\text{MAE: } \frac{1}{n} \sum |y_i - \hat{y}_i| = \frac{1}{6} (0 + 15 + 5 + 10 + 0 + 0) = \frac{1}{6} \times 30 = 5$$

Влияние выбросов

MSE



MAE



Парная и множественная регрессия

- В частном случае, когда фактор единственный, говорят о **парной** или **простейшей** линейной регрессии:

$$\hat{y} = b_0 + b_1 x_1$$

Параметры:

- Свободный коэффициент (constant/intercept) b_0
- Коэффициент наклона (slope) b_1

- Когда количество факторов больше 1-го, то говорят о множественной регрессии:

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k$$

Параметры:

- Свободный коэффициент (constant/intercept) b_0
- Много коэффициентов наклона (slopes) $b_1 \dots b_n$



LIVE

Парная регрессия - минимизация β_0

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$RSS = \mathcal{L}(y, \hat{y}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min$$

Как найти минимум квадратичной функции?

$$\frac{d\mathcal{L}}{d\beta_0} = \frac{d}{d\beta_0} \left(\sum_{i=1}^n (y_i - \hat{y}_i)^2 \right) = \frac{d}{d\beta_0} \left(\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right) = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$-\sum_{i=1}^n y_i + \sum_{i=1}^n \hat{\beta}_0 + \sum_{i=1}^n \hat{\beta}_1 x_i = 0$$

$$n\hat{\beta}_0 = \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\beta}_1 x_i$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n y_i}{n} - \frac{\sum_{i=1}^n \hat{\beta}_1 x_i}{n}$$

Учитывая: $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$ $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ \Rightarrow

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Парная регрессия - минимизация β_1

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$RSS = \mathcal{L}(y, \hat{y}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{d\mathcal{L}}{d\beta_1} = \frac{d}{d\beta_1} \left(\sum_{i=1}^n (y_i - \hat{y}_i)^2 \right) = \frac{d}{d\beta_1} \left(\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right) = -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^n (x_i y_i - x_i \bar{y} + x_i \hat{\beta}_1 \bar{x} - x_i \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^n ((x_i y_i - x_i \bar{y}) + \hat{\beta}_1 (x_i \bar{x} - x_i^2)) = 0$$

$$\sum_{i=1}^n (x_i y_i - x_i \bar{y}) + \sum_{i=1}^n \hat{\beta}_1 (x_i \bar{x} - x_i^2) = 0$$

$$\hat{\beta}_1 \sum_{i=1}^n (x_i^2 - x_i \bar{x}) = \sum_{i=1}^n (x_i y_i - x_i \bar{y})$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i y_i - x_i \bar{y})}{\sum_{i=1}^n (x_i^2 - x_i \bar{x})} = \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \bar{x}}$$

Парная регрессия - минимизация β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i y_i - x_i \bar{y})}{\sum_{i=1}^n (x_i^2 - x_i \bar{x})} = \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \bar{x}} = \boxed{\frac{\sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y}))}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Учитывая: $n\bar{xy} = \sum_{i=1}^n x_i \bar{y} = \sum_{i=1}^n y_i \bar{x} = \sum_{i=1}^n \bar{yx}$

$$\sum_{i=1}^n x_i \bar{x} = \frac{n\bar{x} \sum_{i=1}^n x_i}{n} = n\bar{x}^2$$

Числитель:

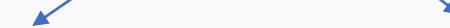
$$\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} = \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} + n\bar{xy} - n\bar{xy} = \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} + \sum_{i=1}^n \bar{xy} - \sum_{i=1}^n \bar{xy} = \sum_{i=1}^n (x_i y_i - x_i \bar{y} + \bar{xy} - \bar{xy}) = \sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y}))$$

Знаменатель:

$$\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \bar{x} = \sum_{i=1}^n (x_i^2 - x_i \bar{x}) = \sum_{i=1}^n (x_i^2 - x_i \bar{x} - x_i \bar{x} + x_i \bar{x}) = \sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + x_i \bar{x}) = \sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2) = \sum_{i=1}^n (x_i - \bar{x})^2$$

Парная регрессия - аналитическое решение

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y}))}{\sum_{i=1}^n (x_i - \bar{x})^2}$$



LIVE

Множественная регрессия – матричная форма

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_p x_p$$

Введем x_0 - вектор единиц:

$$x = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{n1} \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_p \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}$$

$$x\theta = y \quad x_{n \cdot p} \cdot \theta_{p \cdot 1} = y_{n \cdot 1}$$

$$RSS = \mathcal{L}(y, \hat{y}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - x_i^T \theta)^2 = (y_1 - x_1^T \theta)^2 + (y_2 - x_2^T \theta)^2 + \dots + (y_n - x_n^T \theta)^2 =$$

$$= [y_1 - x_1^T \theta, y_2 - x_2^T \theta, \dots, y_n - x_n^T \theta] \begin{bmatrix} y_1 - x_1^T \theta \\ y_2 - x_2^T \theta \\ \dots \\ y_n - x_n^T \theta \end{bmatrix} = (y - x\theta)^T (y - x\theta)$$

$$RSS = \mathcal{L}(y, \hat{y}) = (y - x\theta)^T (y - x\theta)$$

Множественная регрессия – аналитическое решение

Здесь под производной скалярной функции $f(x)$ по вектору x понимается градиент:

$$\frac{df(x)}{dx} = \left[\frac{df(x)}{dx_1}, \dots, \frac{df(x)}{dx_d} \right]^T$$

Свойства, которые понадобятся далее: $\frac{dx^T A}{dx} = \frac{dA^T x}{dx} = A$ и $\frac{dx^T Ax}{dx} = 2Ax$

Приравниваем производную к нулю: $\frac{d\mathcal{L}}{d\theta} = \frac{d}{d\theta}(y - x\theta)^T(y - x\theta) =$

$$= \frac{d}{d\theta}(y^T - (x\theta)^T)(y - x\theta) =$$

$$= \frac{d}{d\theta}((y^T - (x\theta)^T)y - (y^T - (x\theta)^T)(x\theta)) =$$

$$= \frac{d}{d\theta}(y^T y - (x\theta)^T y - y^T x\theta + (x\theta)^T x\theta) =$$

$$= \frac{d}{d\theta}(y^T y - \theta^T x^T y - y^T x\theta + \theta^T x^T x\theta) =$$

$$= \frac{d}{d\theta}(y^T y) - \frac{d}{d\theta}(\theta^T(x^T y)) - \frac{d}{d\theta}((y^T x)\theta) + \frac{d}{d\theta}(\theta^T(x^T x)\theta) =$$

$$= 0 - x^T y - x^T y + 2x^T x\theta = -2x^T y + 2x^T x\theta = 0$$

↓

$$x^T x\theta = x^T y$$
$$(x^T x)^{-1} x^T x\theta = (x^T x)^{-1} x^T y$$

$$\boxed{\theta = (x^T x)^{-1} x^T y}$$





LIVE

Плюсы и минусы линейной регрессии

Плюсы:

- + Очень простая и быстрая в обучении, предсказании
- + Несмотря на название, может не только в линейные зависимости
- + Легко интерпретировать

Первое, что стоит пробовать в задаче регрессии :)

Минусы:

- Не всегда предпосылка о линейной зависимости выполняется в данных
- Может быть слишком простой моделью для ваших данных
- Крайне чувствительна к предобработке данных

Проверка достижения целей

Цели вебинара | Проверка достижения целей

1

Какие параметры в линейной регрессии?

2

Какие метрики качества регрессии?

3

Как обучается линейная регрессия?

4

Что такое переобучение и как с ним бороться?

Рефлексия



Достигли ли вы цели вебинара?



С какими основными мыслями и инсайтами
уходите с вебинара?

Следующий вебинар

Тема: «Логистическая регрессия»



Понедельник 21 сентября 20:00



Ссылка на вебинар будет в ЛК за 15 минут



Материалы к занятию
в ЛК — можно изучать



Обязательный
материал обозначен
красной лентой



Заполните, пожалуйста,
опрос о занятии по ссылке в чате

Спасибо за внимание!
Приходите на следующие вебинары



Андрей Канашов
Data Scientist
OMD OM GROUP
@Андрей Канашов