

# Основы машинного обучения

Лекция 5

Линейная регрессия

Евгений Соколов

[esokolov@hse.ru](mailto:esokolov@hse.ru)

НИУ ВШЭ, 2022

# Обучение линейной регрессии

$$\frac{1}{\ell} \|Xw - y\|^2 \rightarrow \min_w$$

- Вычисление MSE в NumPy:

```
np.square(X.dot(w) - y).mean()
```

# Обучение линейной регрессии

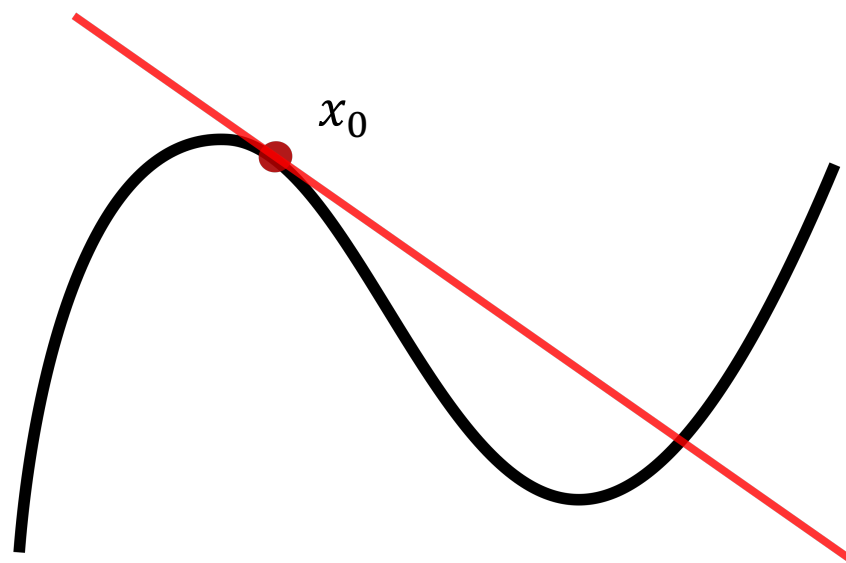
# Среднеквадратичная ошибка

- MSE для линейной регрессии:

$$Q(w_1, \dots, w_d) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\mathbf{w}_1 x_1 + \dots + \mathbf{w}_d x_d - y_i)^2$$

# Производная

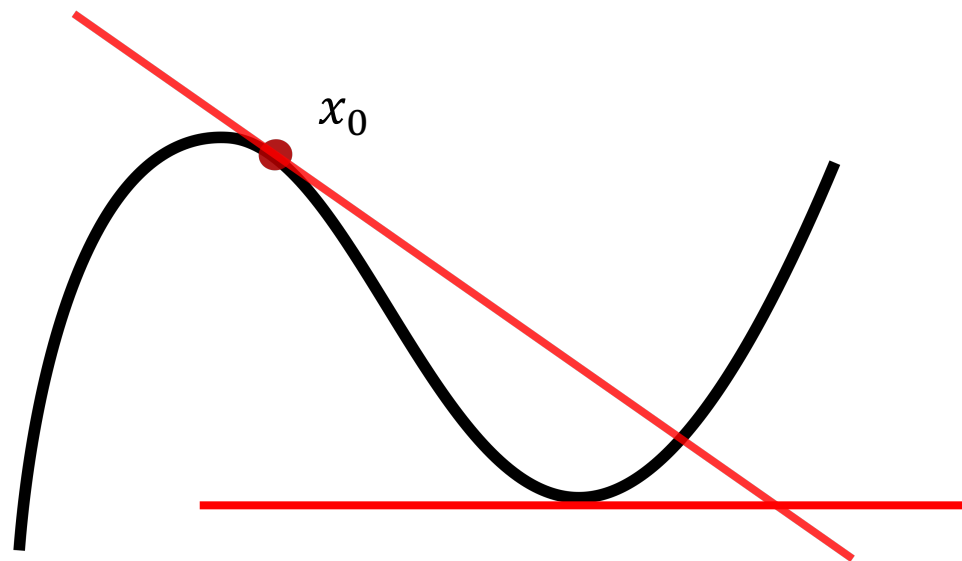
$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$



# Производная

- Если точка  $x_0$  — экстремум и в ней существует производная, то

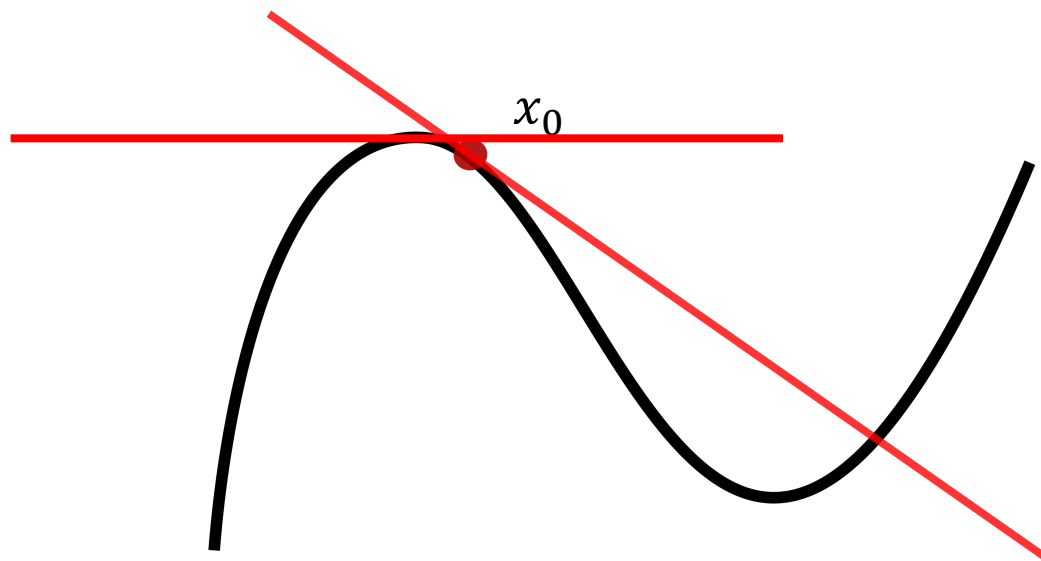
$$f'(x_0) = 0$$



# Производная

- Если точка  $x_0$  — экстремум и в ней существует производная, то

$$f'(x_0) = 0$$



# Градиент

- Градиент — вектор частных производных

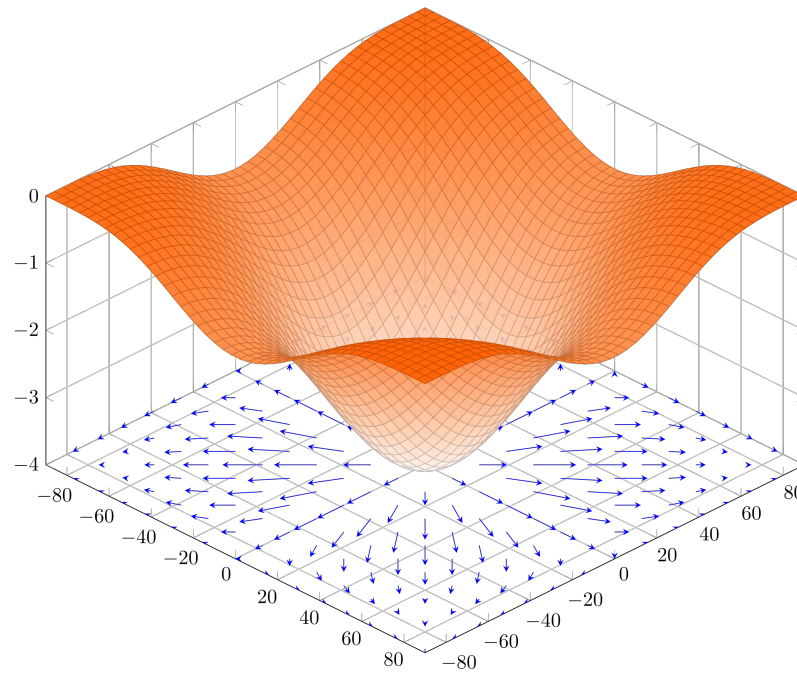
$$\nabla f(x) = \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_d} \right)$$

- У градиента есть важное свойство!



# Важное свойство

- Зафиксируем точку  $x_0$
- В какую сторону функция быстрее всего растёт?



# Важное свойство

- Зафиксируем точку  $x_0$
- В какую сторону функция быстрее всего растёт?
- В направлении градиента!
- Если градиент равен нулю, то это экстремум

# Условие экстремума

- Если точка  $x_0$  — экстремум и в ней существует производная, то

$$\nabla f(x_0) = 0$$

# Условие экстремума

- Если точка  $x_0$  — экстремум и в ней существует производная, то

$$\nabla f(x_0) = 0$$

- Если функция выпуклая, то экстремум один
- MSE для линейной регрессии — выпуклая!
  - (при некоторых условиях)

# Обучение линейной регрессии

- Можно посчитать градиент MSE:

$$\nabla \frac{1}{\ell} \|Xw - y\|^2 = \frac{2}{\ell} X^T (Xw - y)$$

- Приравниваем нулю и решаем систему линейных уравнений:

$$w = (X^T X)^{-1} X^T y$$

# Аналитическое решение

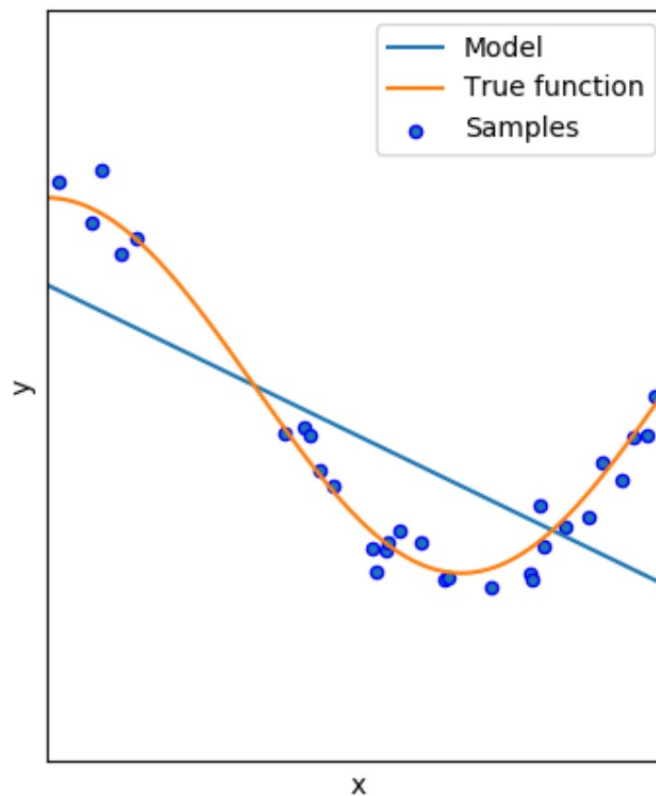
$$w = (X^T X)^{-1} X^T y$$

- Если матрица  $X^T X$  вырожденная, то будут проблемы
- Даже если она почти вырожденная, всё равно будут проблемы
- Если признаков много, то придётся долго ждать

# Переобучение и регуляризация линейных моделей

# Нелинейная задача

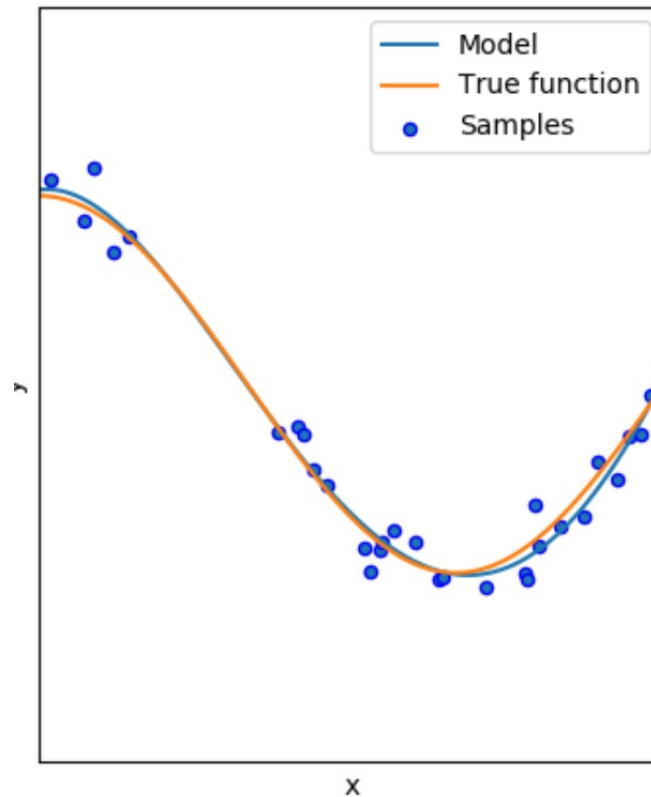
$$a(x) = w_0 + w_1 x$$





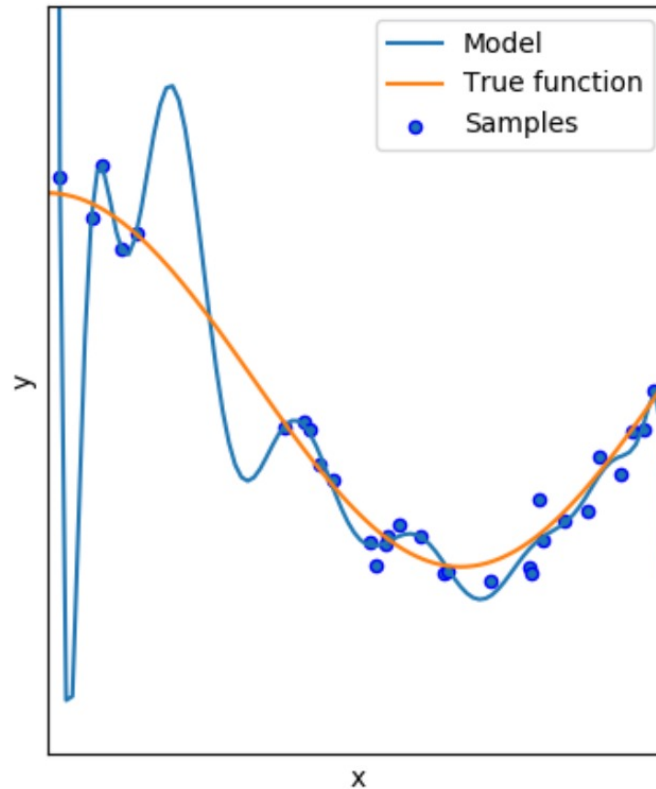
# Нелинейная задача

$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4$$



# Нелинейная задача

$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + \dots + w_{15}x^{15}$$



# Симптом переобучения

$$a(x) = 0.5 + 13458922x - 43983740x^2 + \dots$$

- Большие коэффициенты — симптом переобучения
- Эмпирическое наблюдение

# Симптом переобучения

- Большие коэффициенты в линейной модели — это плохо
- Пример: предсказание роста по весу

$$a(x) = 698x - 41714$$

- Изменение веса на 0.01 кг приведет к изменению роста на 7 см
- Не похоже на правильную зависимость

# Регуляризация

- Будем штрафовать за большие веса!
- Пример функционала:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2$$

- Регуляризатор:

$$\|w\|^2 = \sum_{j=1}^d w_j^2$$

# Регуляризация

- Регуляризованный функционал

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|^2 \rightarrow \min_w$$

- $\lambda$  — коэффициент регуляризации

# Регуляризация

- Регуляризованный функционал

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|^2 \rightarrow \min_w$$

- Аналитическое решение:

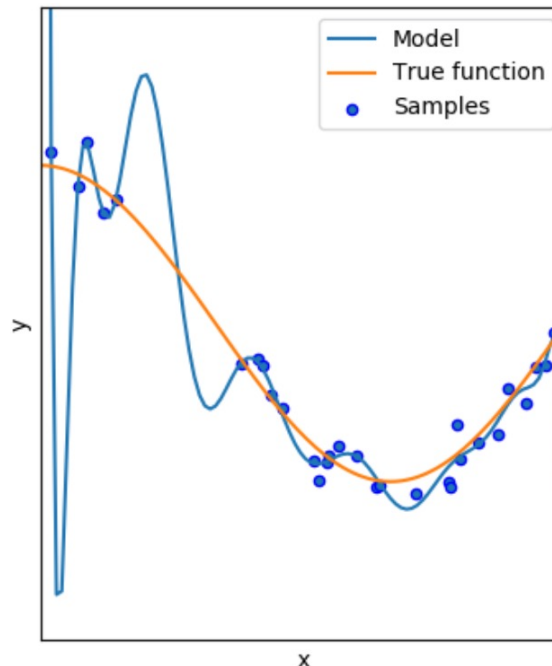
$$w = (X^T X + \lambda I)^{-1} X^T y$$

- Гребневая регрессия (Ridge regression)

# Эффект регуляризации

$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + \dots + w_{15}x^{15}$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 \rightarrow \min_w$$

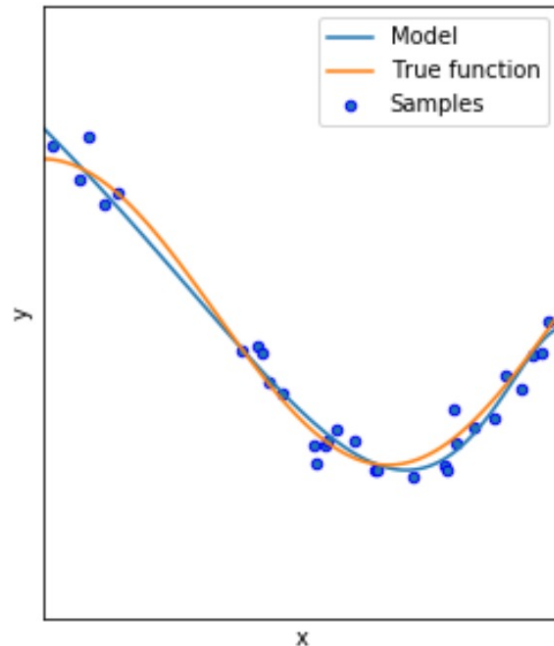




# Эффект регуляризации

$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + \dots + w_{15}x^{15}$$

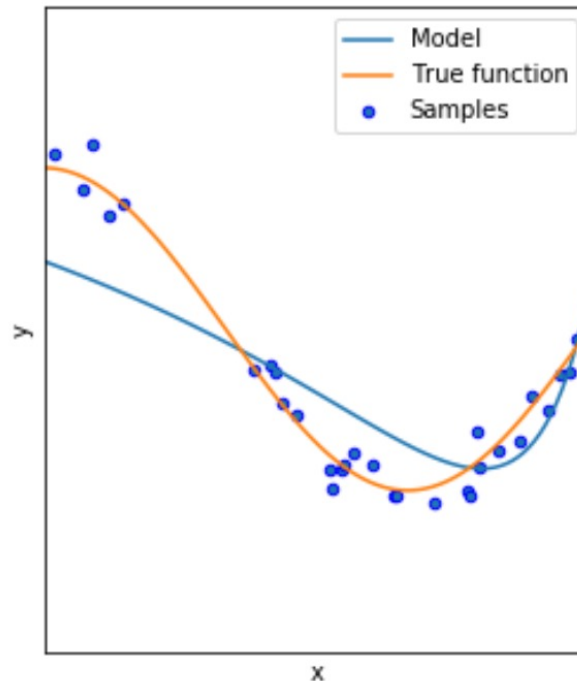
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \mathbf{0.01} \|w\|^2 \rightarrow \min_w$$



# Эффект регуляризации

$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + \dots + w_{15}x^{15}$$

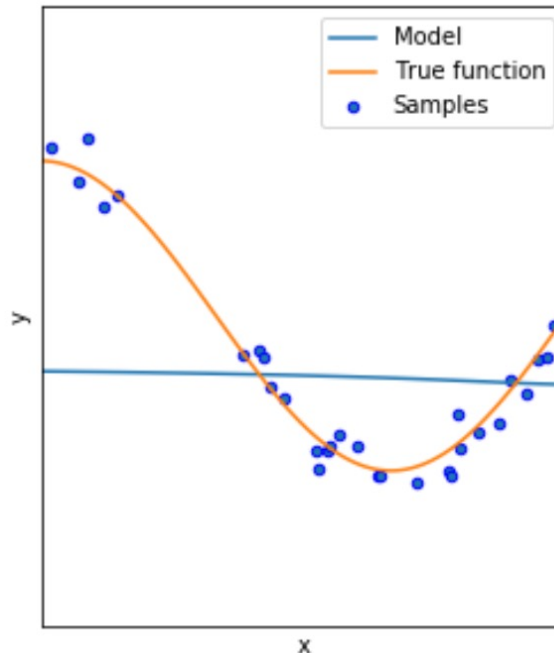
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \mathbf{1} \|w\|^2 \rightarrow \min_w$$



# Эффект регуляризации

$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + \dots + w_{15}x^{15}$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \mathbf{100} \|w\|^2 \rightarrow \min_w$$



# Лассо

- Регуляризованный функционал

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^d |w_j| \rightarrow \min_w$$

- LASSO (Least Absolute Shrinkage and Selection Operator)
- Некоторые веса зануляются
- Приводит к отбору признаков

# Регуляризаторы

- $\|z\|_2 = \sqrt{\sum_{j=1}^d z_j^2}$  —  $L_2$ -норма
- $\|z\|_1 = \sum_{j=1}^d |z_j|$  —  $L_1$ -норма

# Интерпретация линейных моделей

# Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 100.000 * (\text{площадь}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

# Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 100.000 * (\text{площадь}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

- Чем больше вес, тем важнее признак?



# Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 100.000 * (\text{площадь в кв. м.}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

- Чем больше вес, тем важнее признак?

# Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 10 * (\text{площадь в кв. см.}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

- Чем больше вес, тем важнее признак?

# Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 100.000 * (\text{площадь в кв. м.}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

- Чем больше вес, тем важнее признак?

# Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 100.000 * (\text{площадь в кв. м.}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

- Чем больше вес, тем важнее признак?
- Только если признаки масштабированы!

# Масштабирование признаков

- Отмасштабируем  $j$ -й признак
- Вычисляем среднее и стандартное отклонение признака на обучающей выборке:

$$\mu_j = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i^j$$

$$\sigma_j = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (x_i^j - \mu_j)^2}$$

# Масштабирование признаков

- Вычтем из каждого значения признака среднее и поделим на стандартное отклонение:

$$x_i^j := \frac{x_i^j - \mu_j}{\sigma_j}$$

# Регуляризация

- Если модель переобучается, то веса используются для запоминания обучающей выборки
- Правильнее масштабировать признаки и регуляризовать модель перед изучением весов

# Пример

- 1000 объектов
- Два признака
- Первый принимает значения от 0 до 1
- Второй равен единице на 10 объектах и нулю на 990 объектах
- $y = x_1 + 2x_2$



# Пример

```
[0.3175037 , 1.      ],
[0.59558502, 1.      ],
[0.48660609, 1.      ],
[0.69255463, 1.      ],
[0.81968981, 1.      ],
[0.48844247, 1.      ],
[0.13426702, 1.      ],
[0.850628   , 1.      ],
[0.57499033, 1.      ],
[0.73993748, 1.      ],
[0.70466465, 0.      ],
[0.96821177, 0.      ],
[0.29530732, 0.      ],
[0.70530677, 0.      ],
[0.36567633, 0.      ],
[0.39541072, 0.      ],
[0.23059464, 0.      ],
[0.34401018, 0.      ],
[0.94829675, 0.      ],
[0.29257085, 0.      ],
[0.24599061, 0.      ],
[0.58313798, 0.      ],
```

# Пример

$$a(x) = x_1 + 2x_2$$

- Удаляем первый признак, получаем  $MSE = 0.08$
- Удаляем второй признак, получаем  $MSE = 0.04$
- Правильнее удалить признак и посмотреть, как сильно растёт ошибка без него