

ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ

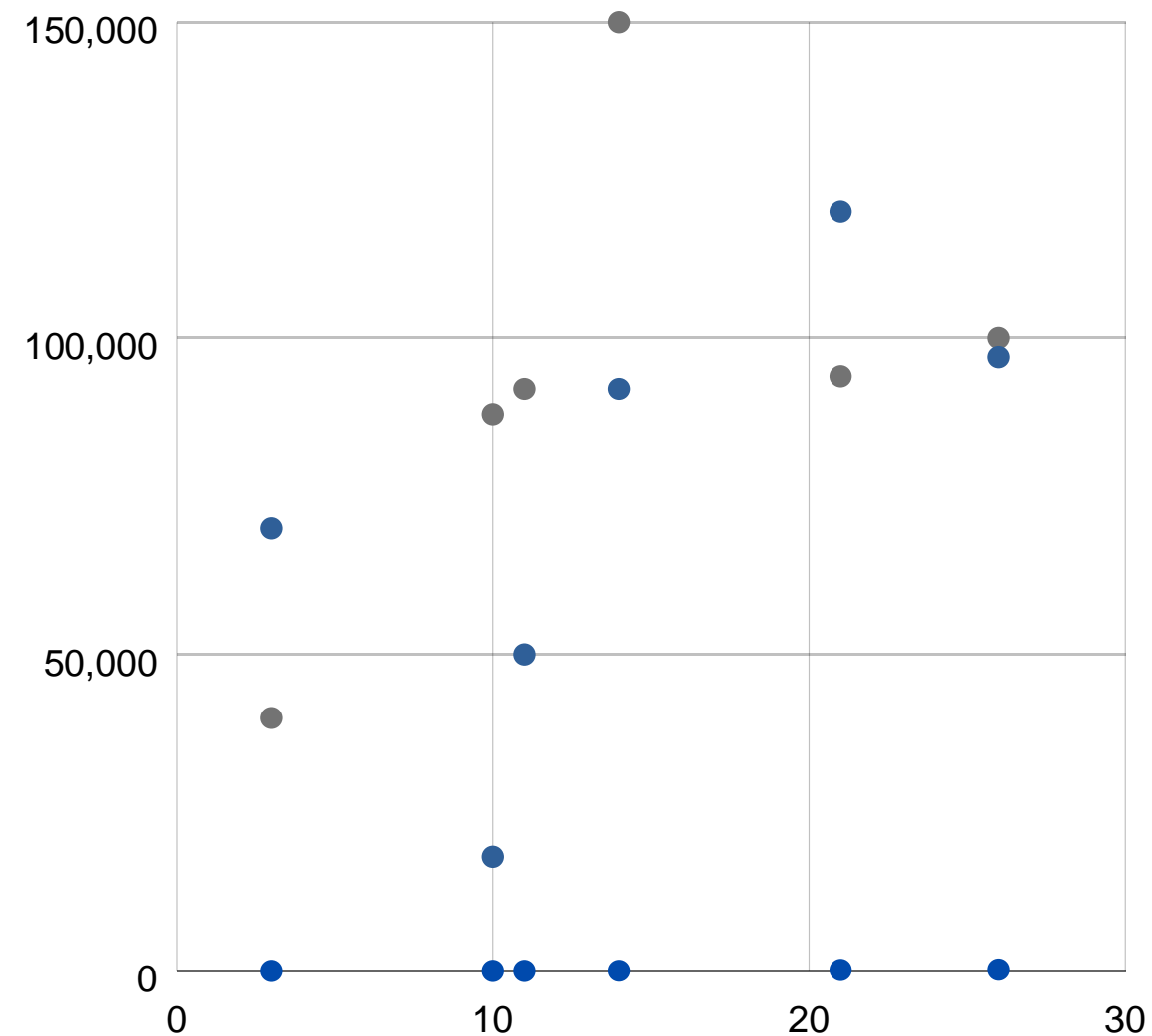
Лекция №2

01

Работа с числовыми признаками

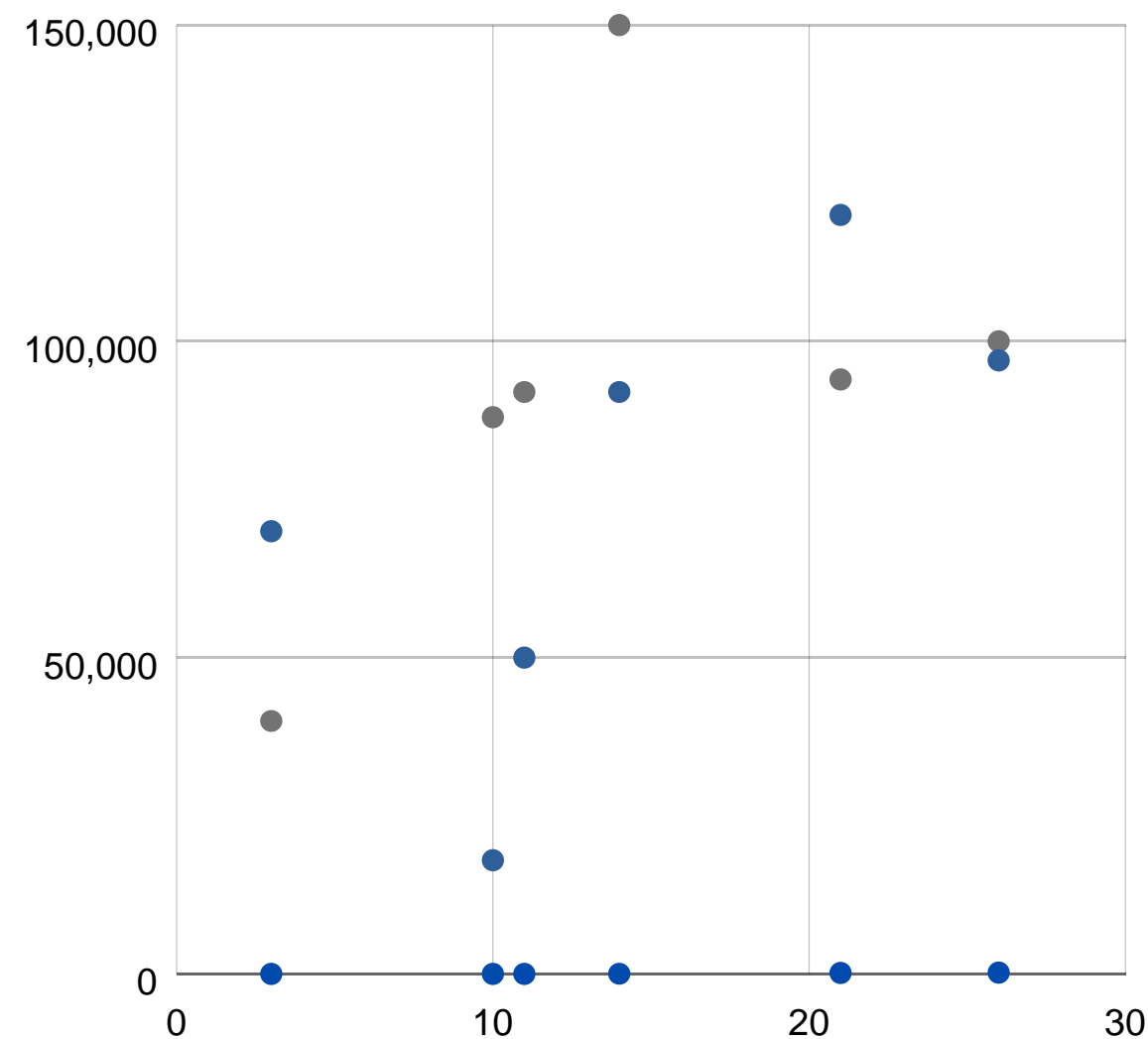
Feature Engineering

Задача: хотим применить knn вместе с Евклидовым расстоянием на набор данных, есть ли тут какая-то проблема?



Feature Engineering

Задача: хотим применить knn вместе с Евклидовым расстоянием на набор данных, есть ли тут какая-то проблема?



Мы можем заметить, что масштаб по оси Y сильно отличается от масштаба по оси X, как думаете, на что это влияет?
Как мы можем это исправить?

Feature Engineering

Масштабирование признаков (scaling)

Данный метод используется для нормализации данных и приведения их к одинаковому масштабу. Обычно применяется к числовым признакам.

MinMaxScaler:

$$x_scaled = \frac{x - x_min}{x_max - x_min}$$

- x - исходное значение признака
- x_min - минимальное значение признака в обучающей выборке
- x_max - максимальное значение признака в обучающей выборке
- x_scaled - масштабированное значение признака

Feature Engineering

Масштабирование признаков (scaling)

Standard scaler:

$$x_scaled = \frac{x - x_mean}{std \text{ по } x}$$

- x - исходное значение признака
- $mean$ - среднее значение признака в обучающей выборке
- std - стандартное отклонение признака в обучающей выборке
- x_scaled - масштабированное значение признака

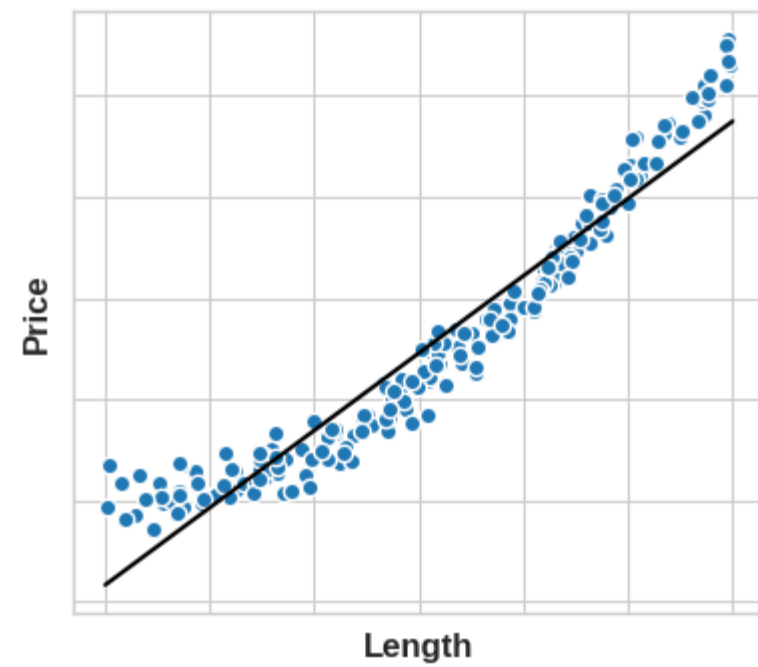
После применения среднее значение будет равно 0 и стандартное отклонение будет равно 1.

Есть ли здесь какие-то проблемы?

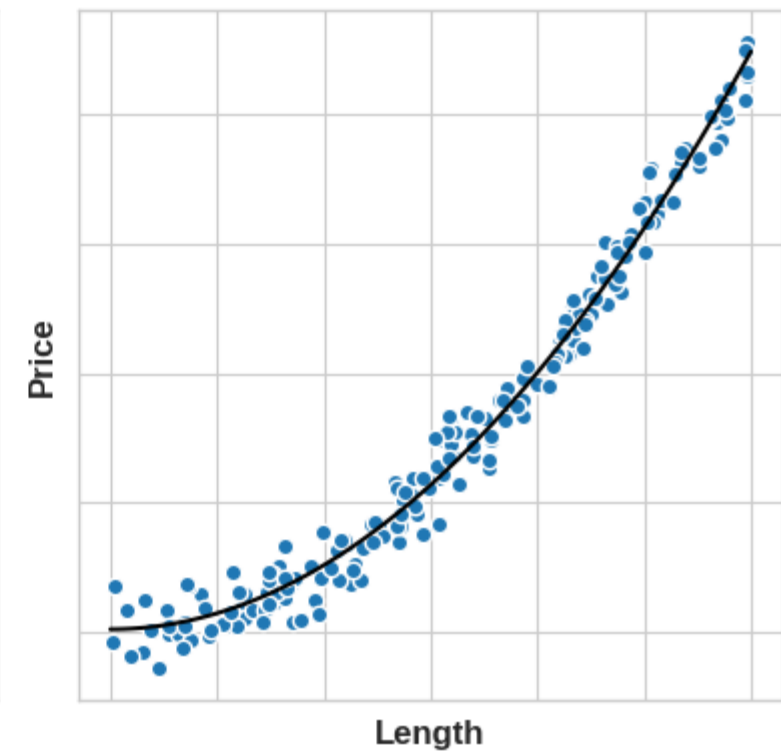
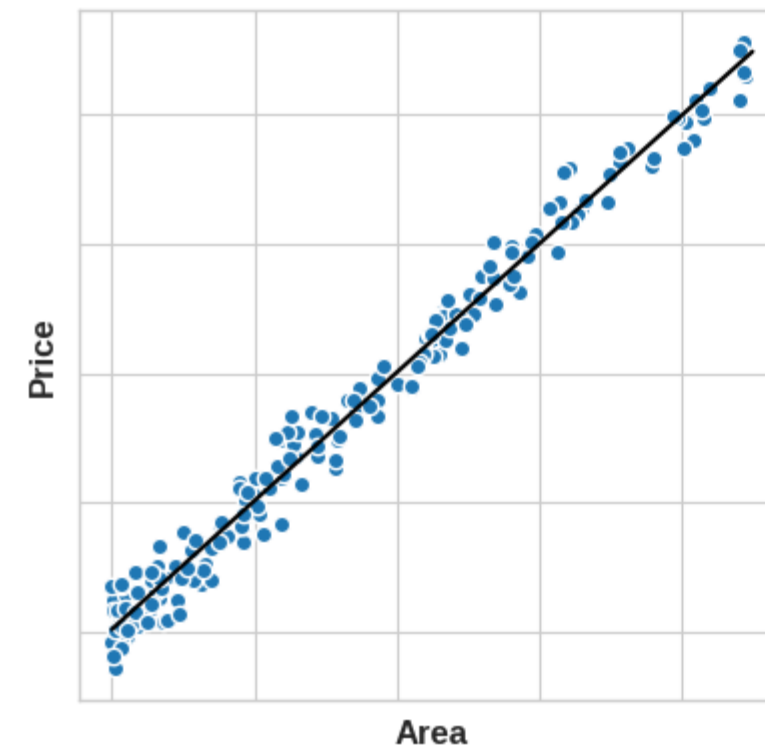
Feature Engineering

Разделимость признаков

The Base Model

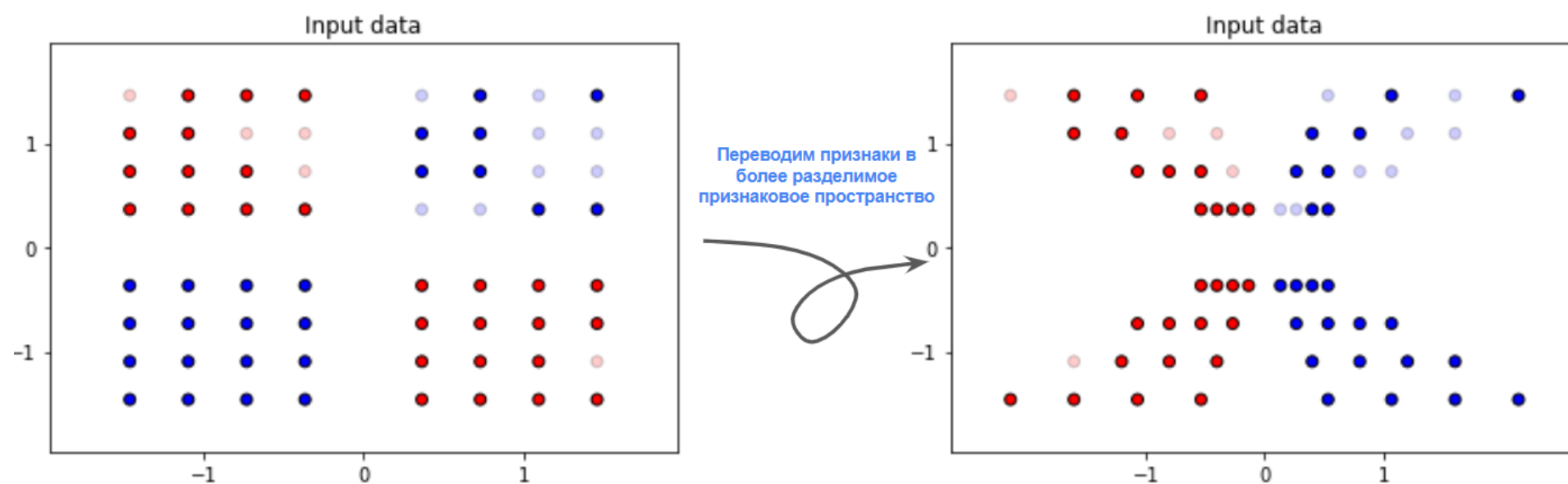


The Extended Model

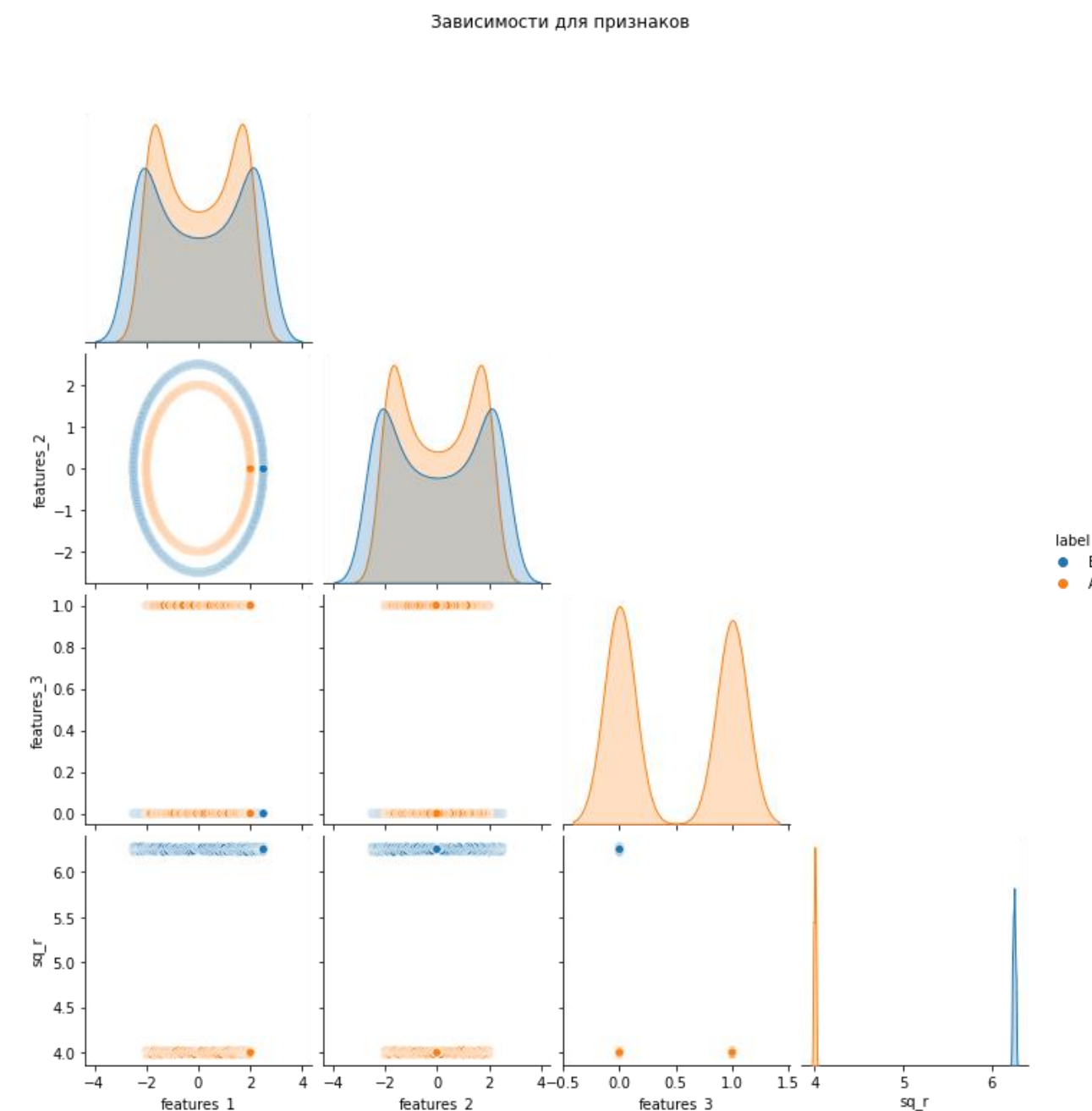


Feature Engineering

Разделимость признаков



Выборка № 1



Выборка № 2

02

Embeddings

векторное представление признаков

Embeddings

Преобразование категориальных / текстовых данных в числовые векторы фиксированной длины

- One-hot Encoding: создаем вектор фиксированной длины, состоящий из нулей и одной единицы, которая указывает на наличие или отсутствие конкретной категории. Недостатком этого метода является то, что он не сохраняет отношения между категориями. (рассмотрим сегодня позднее)
- Count Encoding: замена каждую категорию на ее количество в данных. Он сохраняет относительную частоту категории в данных, но не сохраняет семантические отношения между категориями.
- Embedding Layer: создаются числовые векторы фиксированной длины для каждой категории, используя нейронную сеть. Он позволяет сохранять семантические отношения между категориями и учитывать контекст, в котором категория встречается в данных.

Embeddings

Embedding Layer

Пример: закодировать каждое уникальное слово числовым значением

Какой результат получится? Что нужно сделать со словами?

- зеленое яблоко
- 2. красные яблоки
- 3. красный гранат



Embeddings

Embedding Layer

Пример: закодировать каждое уникальное слово числовым значением

Какой результат получится? Что нужно сделать со словами?

1. зеленое яблоко
2. красные яблоки
3. красный гранат



| | зеленый | красный | яблоко | гранат |
|---|---------|---------|--------|--------|
| 1 | 1 | 0 | 1 | 0 |
| 2 | 0 | 1 | 1 | 0 |
| 3 | 0 | 1 | 0 | 1 |

03

Feature encoding

кодирование признаков

One-hot кодирование

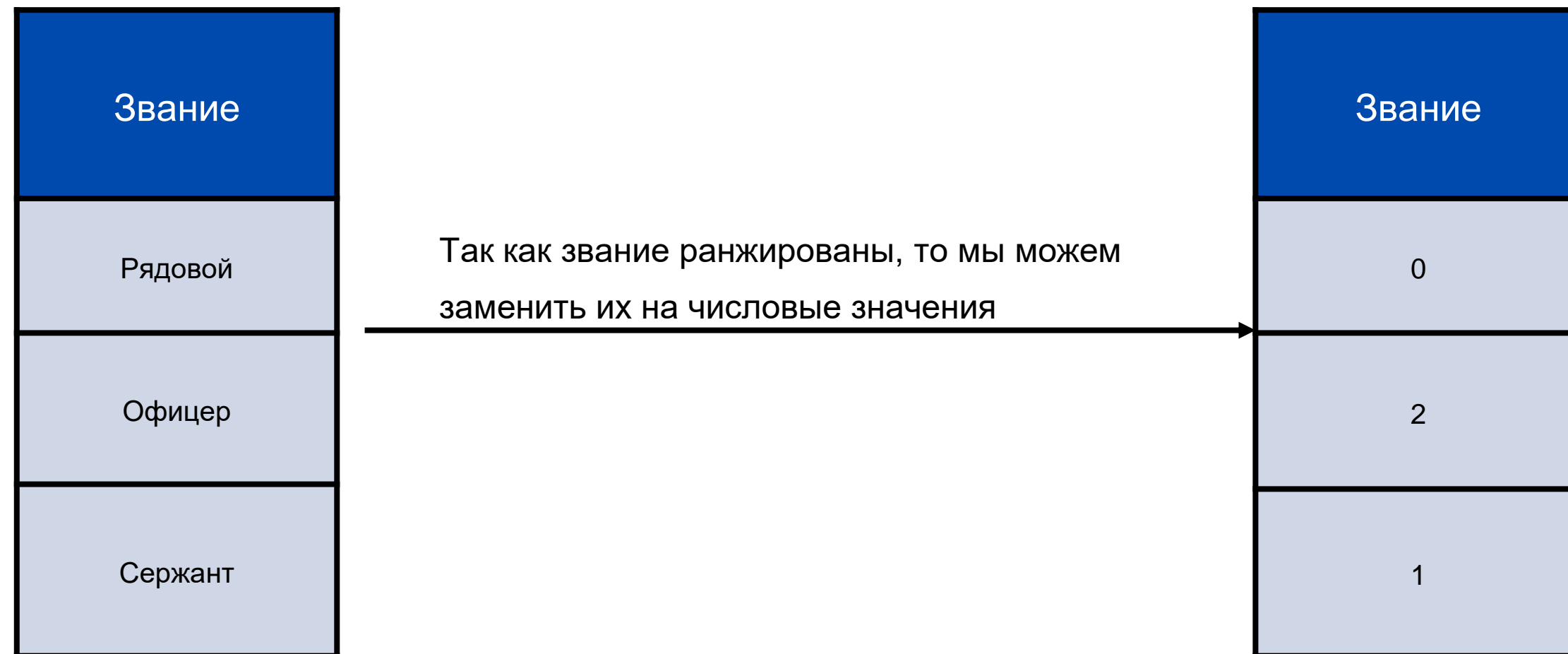
- Значения признака "район": $U = \{u_1, \dots, u_m\}$
- Новые признаки вместо x : $[x = u_1], \dots, [x = u_m]$

| Район | ЦАО | ЮАО | САО |
|-------|-----|-----|-----|
| ЦАО | 1 | 0 | 0 |
| ЮАО | 0 | 1 | 0 |
| ЦАО | 1 | 0 | 0 |
| САО | 0 | 0 | 1 |
| ЮАО | 0 | 1 | 0 |

$$\begin{aligned} a(x) = & w_0 + w_1 * (\text{площадь}) \\ & + w_2 * (\text{квартира в ЦАО?}) \\ & + w_3 * (\text{квартира в ЮАО?}) \\ & + w_4 * (\text{квартира в САО?}) \end{aligned}$$

Label Encoding

- Применимо к порядковым признакам
- Так как множество признаков является упорядоченным, то мы можем заменить значения на числа



Target Encoding

Целевая кодировка отлично подходит для:

1) **Функций с высокой мощностью:** Функция с большим количеством категорий может вызывать проблемы с кодированием: однократное кодирование привело бы к появлению слишком большого количества функций, и альтернативные варианты, такие как кодировка меток, могут не подходить для этой функции.

Целевая кодировка выводит номера для категорий, используя наиболее важное свойство объектов: их связь с целевой кодировкой.

2) **Особенности, связанные с предметной областью:** Исходя из предыдущего опыта, вы можете предположить, что категориальный признак должен быть важным, даже если он имеет низкую оценку по показателю объекта. Целевая кодировка может помочь выявить истинную информативность объекта.

$$S_i = \lambda(n_i) \frac{n_{iY}}{n_i} + (1 - \lambda(n_i)) \frac{n_Y}{n_{TR}}$$

$n(Y)$ - общее количество строк с 1 в целевой метрике,

$n(i)$ - количество строк с i -той категории,

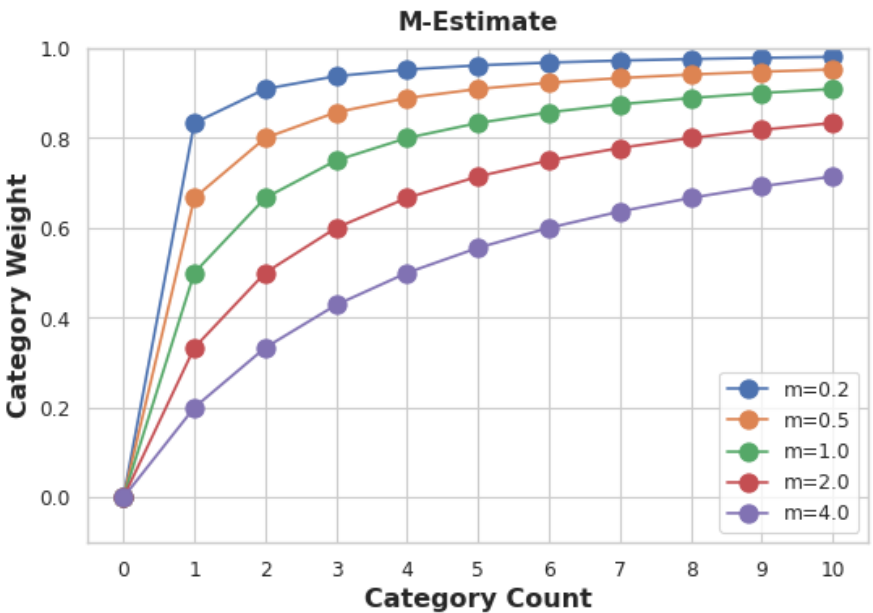
$n(iY)$ - количество строк с 1 в целевой метрике в i -той категории.

Пусть min_sample_leaf , $k = 1$ и сглаживание, $f = 1$

$\lambda(\text{'Male'}) = 1 / (1 + \exp(-(2-1)/1)) = 0.73$ # Weight Factor for 'Male'

Target Statistic = (Weight Factor * Probability of 1 for Males) + ((1-Weight Factor) * Probability of 1 Overall)
 $S(\text{'Male'}) = (0.73 * 0.5) + ((1-0.73) * 0.4) = 0.485$

$$\lambda(n) = \frac{1}{1 + e^{\frac{-(n-k)}{f}}}$$



$\lambda(\text{'Female'}) = 1 / (1 + \exp(-(4-1)/1)) = 0.95$ #Weight Factor for 'Female'


Target Statistic = (Weight Factor * Probability of 1 for Females) + ((1-Weight Factor) * Probability of 1 Overall)
 $S(\text{'Female'}) = (0.95 * 0.25) + ((1-0.95) * 0.4) = 0.259$

| | Gender | Target |
|---|--------|--------|
| 0 | Male | 1 |
| 1 | Male | 0 |
| 2 | Female | 0 |
| 3 | Female | 0 |
| 4 | Female | 0 |
| 5 | Female | 1 |
| 6 | Other | 1 |
| 7 | Other | 1 |
| 8 | Other | 0 |

Target Encoding


Для небинарного признака

| Target | |
|--------|---|
| 0 | 0 |
| 1 | 0 |
| 2 | 1 |
| 3 | 2 |
| 4 | 2 |
| 5 | 0 |
| 6 | 1 |
| 7 | 2 |




| Target_1 | Target_2 | Target_3 |
|----------|----------|----------|
| 0 | 1 | 0 |
| 1 | 1 | 0 |
| 2 | 0 | 1 |
| 3 | 0 | 0 |
| 4 | 0 | 0 |
| 5 | 1 | 0 |
| 6 | 0 | 1 |
| 7 | 0 | 0 |

| Color | Target_1 |
|---------|----------|
| 0 Red | 1 |
| 1 Red | 1 |
| 2 Red | 0 |
| 3 Red | 0 |
| 4 Red | 0 |
| 5 Green | 1 |
| 6 Green | 0 |
| 7 Green | 0 |




| Color_Target_1 | |
|----------------|----------|
| 0 | 0.400000 |
| 1 | 0.400000 |
| 2 | 0.400000 |
| 3 | 0.400000 |
| 4 | 0.400000 |
| 5 | 0.333333 |
| 6 | 0.333333 |
| 7 | 0.333333 |

| Color | Target_2 |
|---------|----------|
| 0 Red | 0 |
| 1 Red | 0 |
| 2 Red | 1 |
| 3 Red | 0 |
| 4 Red | 0 |
| 5 Green | 0 |
| 6 Green | 1 |
| 7 Green | 0 |



| Color_Target_2 | |
|----------------|----------|
| 0 | 0.200000 |
| 1 | 0.200000 |
| 2 | 0.200000 |
| 3 | 0.200000 |
| 4 | 0.200000 |
| 5 | 0.333333 |
| 6 | 0.333333 |
| 7 | 0.333333 |

| Color | Target_3 |
|---------|----------|
| 0 Red | 0 |
| 1 Red | 0 |
| 2 Red | 0 |
| 3 Red | 1 |
| 4 Red | 1 |
| 5 Green | 0 |
| 6 Green | 0 |
| 7 Green | 1 |



| Color_Target_3 | |
|----------------|----------|
| 0 | 0.400000 |
| 1 | 0.400000 |
| 2 | 0.400000 |
| 3 | 0.400000 |
| 4 | 0.400000 |
| 5 | 0.333333 |
| 6 | 0.333333 |
| 7 | 0.333333 |

Frequency Encoding

- Применимо к порядковым признакам
- Частотное кодирование - это метод кодирования, который кодирует значения категориальных признаков в соответствии с их частотами

| nom_2 |
|---------|
| Snake |
| Hamster |
| Lion |
| Snake |
| Lion |

```
:  
enc_nom_1 = (train.groupby('nom_1').size()) / len(train)  
enc_nom_1  
  
:  
nom_1  
Circle      0.124400  
Polygon     0.120477  
Square      0.165323  
Star        0.153013  
Trapezoid   0.337270  
Triangle    0.099517  
dtype: float64
```

04

Заполнение пропусков

Заполнение пропусков

Часто в мы можем найти пропуски в данных, что может помешать работе моделей.

Как можно было бы заполнить отсутствующие значения и стоит ли это делать?

Заполнение пропусков

Часто в мы можем найти пропуски в данных, что может помешать работе моделей.

Как можно было бы заполнить отсутствующие значения и стоит ли это делать?

Посмотрим на возможные варианты для числовых значений:

- мода
- медиана
- среднее

Когда какой вид значений лучше использовать? Почему?

Заполнение пропусков

Часто в мы можем найти пропуски в данных, что может помешать работе моделей.

Как можно было бы заполнить отсутствующие значения и стоит ли это делать?

Посмотрим на возможные варианты для числовых значений:

- мода - часто используется для **категориальных / порядковых** данных, так как это наиболее часто встречающееся значение в наборе данных.
- медиана - используется для **количественных** данных, если данные содержат выбросы или несимметричны. Устойчивое заполнение пропусков.
- среднее - значение используется для **количественных** данных, если данные имеют симметричное распределение и нет выбросов.

Заполнение пропусков

- Всегда стоит смотреть на логику заполняемых данных
- Нужно учитывать стоит ли вообще заполнять пропуски какими-либо не нулевыми значениями

Пример: предположим у нас есть датасет содержащий информацию о клиентах компании, занимающейся телекоммуникациями. Как стоит заполнить пропуски?

| Возраст | Самый популярный сайт по посещениям | Кол-во потраченных минут |
|---------|-------------------------------------|--------------------------|
| 20 | vk.com | |
| 45 | ok.ru | 90 |
| 75 | | 200 |

Заполнение пропусков

- Всегда стоит смотреть на логику заполняемых данных
- Нужно учитывать стоит ли вообще заполнять пропуски какими-либо не нулевыми значениями

Пример: предположим у нас есть датасет содержащий информацию о клиентах компании, занимающейся телекоммуникациями. Как стоит заполнить пропуски?

| Возраст | Самый популярный сайт по посещениям | Кол-во потраченных минут |
|---------|-------------------------------------|--------------------------|
| 20 | vk.com | |
| 45 | ok.ru | 90 |
| 75 | | 200 |

Вывод: заполнять данные нужно аккуратно и заполнение нулевыми значениями - не всегда плохо

Заполнение пропусков

Другие варианты заполнения

Интерполяция

- прогнозируем пропущенные значения на основе известных соседних наблюдений
- линейная - используем линейную функцию между двумя соседними точками
- ближайшего соседа - используется для категориальных / дискретных значений, значения заполняются на основе ближайшего известного
- сплайн-интерполяция - это метод, который используется для заполнения пропущенных значений в числовых данных, где значения могут иметь сложное нелинейное поведение. Он состоит в том, чтобы приблизить пропущенное значение кусочно-линейной функцией, которая проходит через несколько точек соседних наблюдений

Заполнение пропусков

Другие варианты заполнения

Интерполяция

Плюсы:

- достаточно эффективна (особенно, если в данных есть сезонность / тенденция)

Минусы:

- сложный процесс, особенно если множество пропусков / сложная структура данных
- требуется тщательный анализ данных перед проведением заполнения

Заполнение пропусков

Другие варианты заполнения

Конечно, существуют и другие варианты заполнения пропусков, один из них - использование моделей машинного обучения (да, так тоже можно), но сильно усложнять данный процесс не стоит, так как он может лишь помешать в дальнейшем.

**Место для ваших
вопросов**

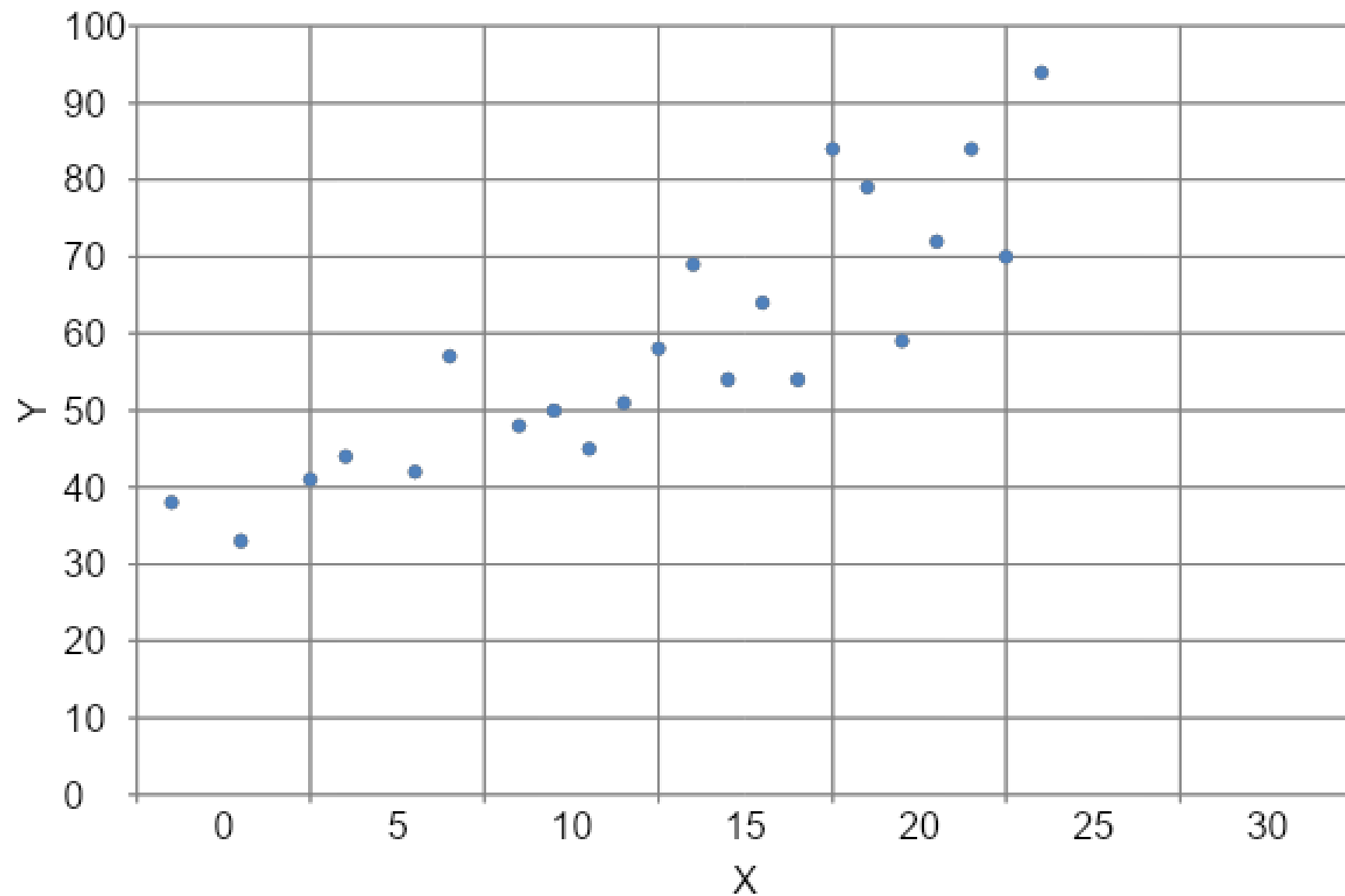


05

Линейная модель

Простейшая модель

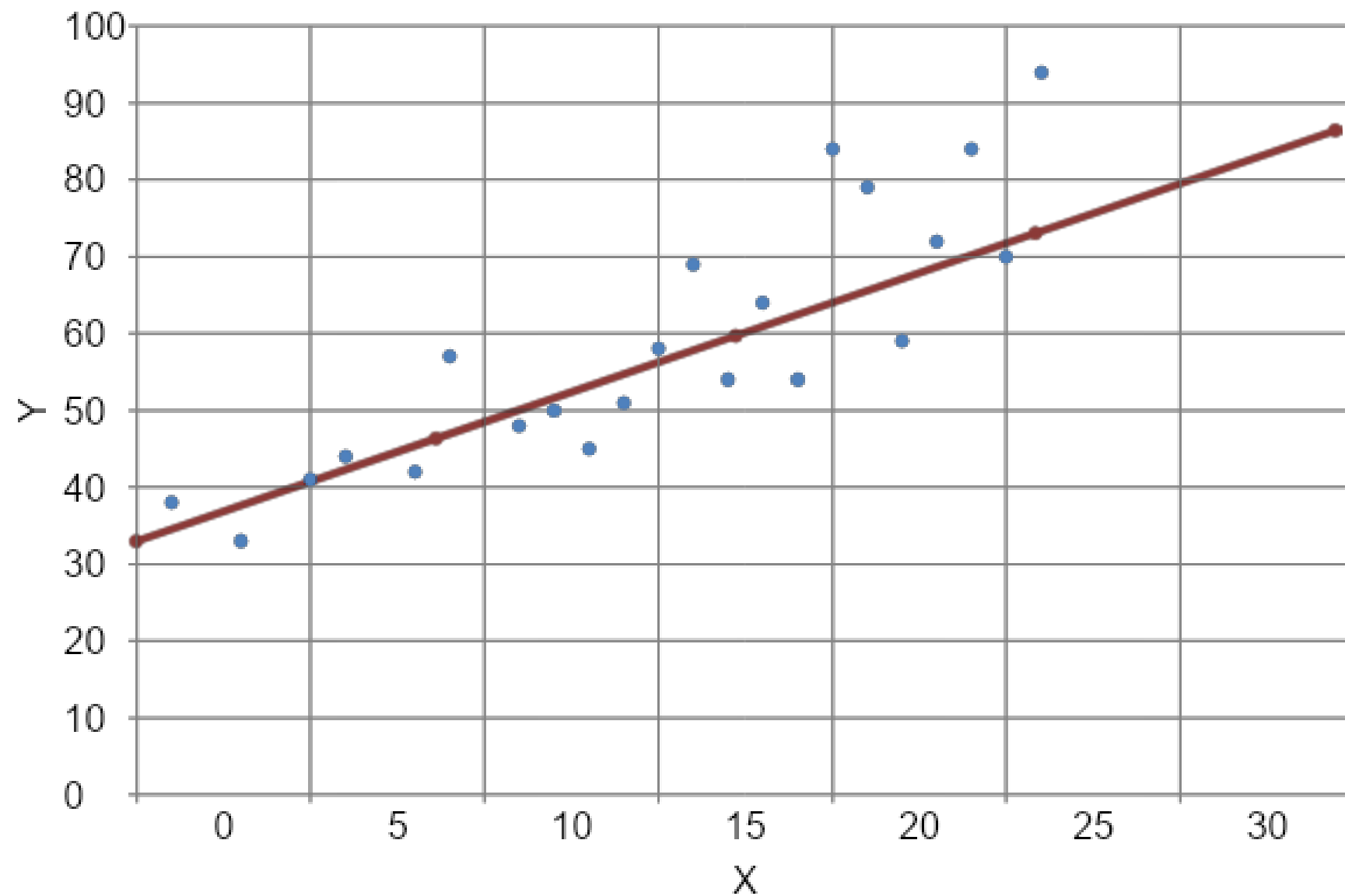
- В пространстве есть множество точек X , хотим каждому объекту сопоставить значение



Простейшая модель

- В пространстве есть множество точек X , хотим каждому объекту сопоставить значение
- Опишем одной функцией

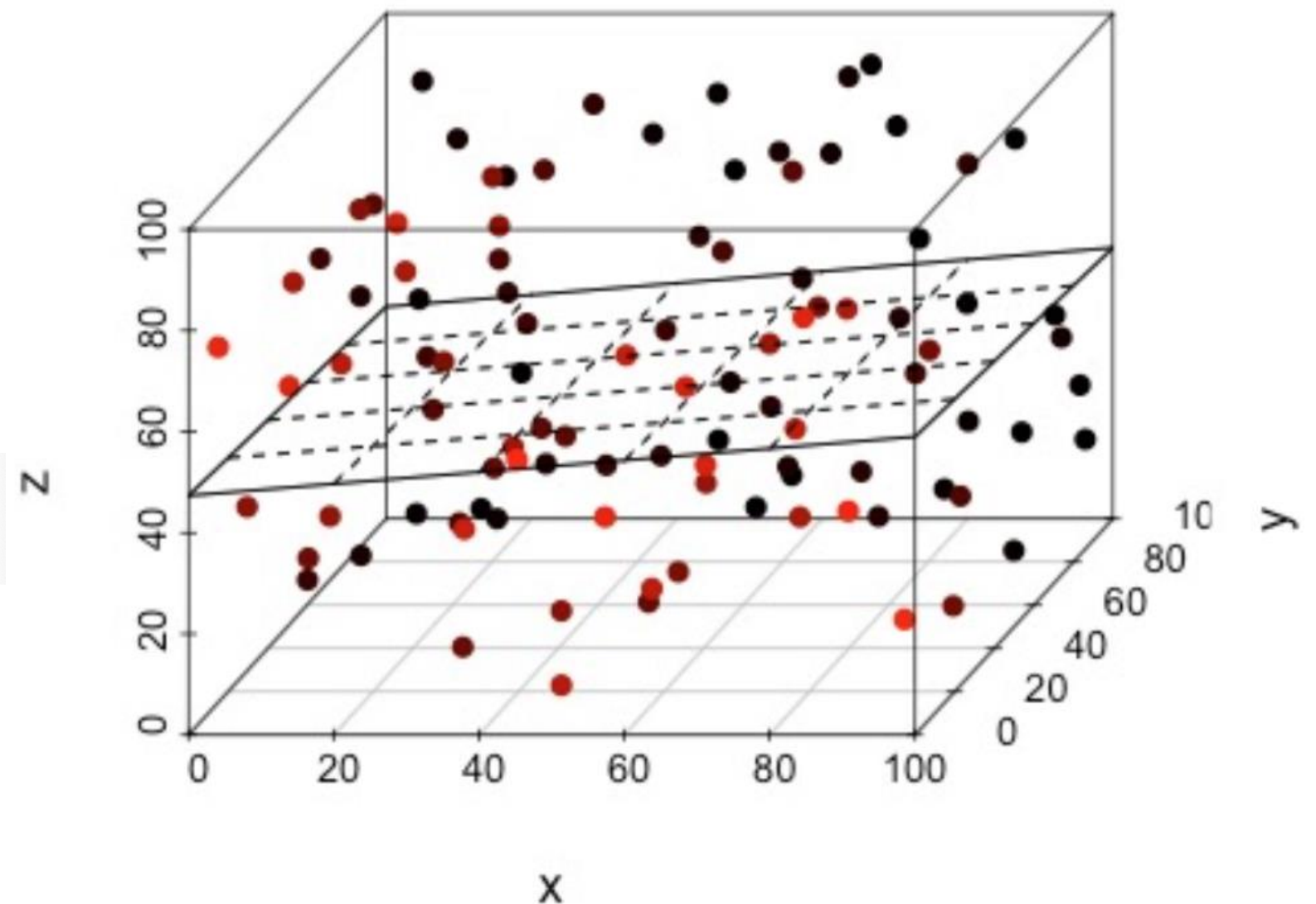
$$a(x) = w_1 x + w_0$$



Модель с двумя признаками

- В пространстве есть множество точек X , хотим каждому объекту сопоставить значение
- Опишем одной функцией

$$a(x) = w_0 + w_1 x_1 + w_2 x_2$$



Основные понятия

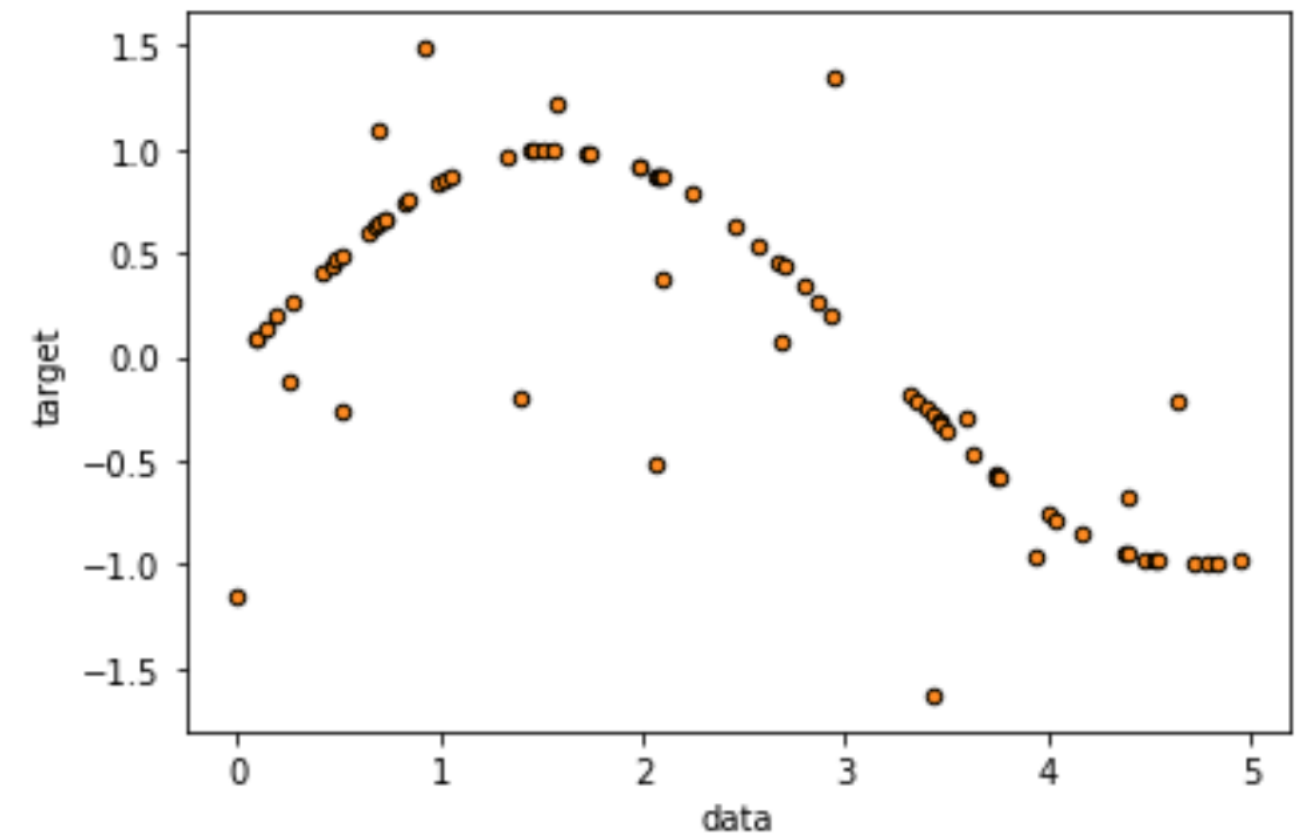
- Задача регрессии: $\mathbb{X} \rightarrow \mathbb{R}$
- В d -мерном пространстве d признаков
- Количество параметров: $d + 1$
- Вектор весов: $w = (w_0, \dots, w_d) \in \mathbb{R}^{d+1}$
- Цель найти такое w , чтобы он «лучшим образом» описывал данные
- Общий вид модели: $a(x) = w_0 + w_1x_1 + \dots + w_dx_d$

Свободный коэффициент/сдвиг/bias

Веса/коэффициенты

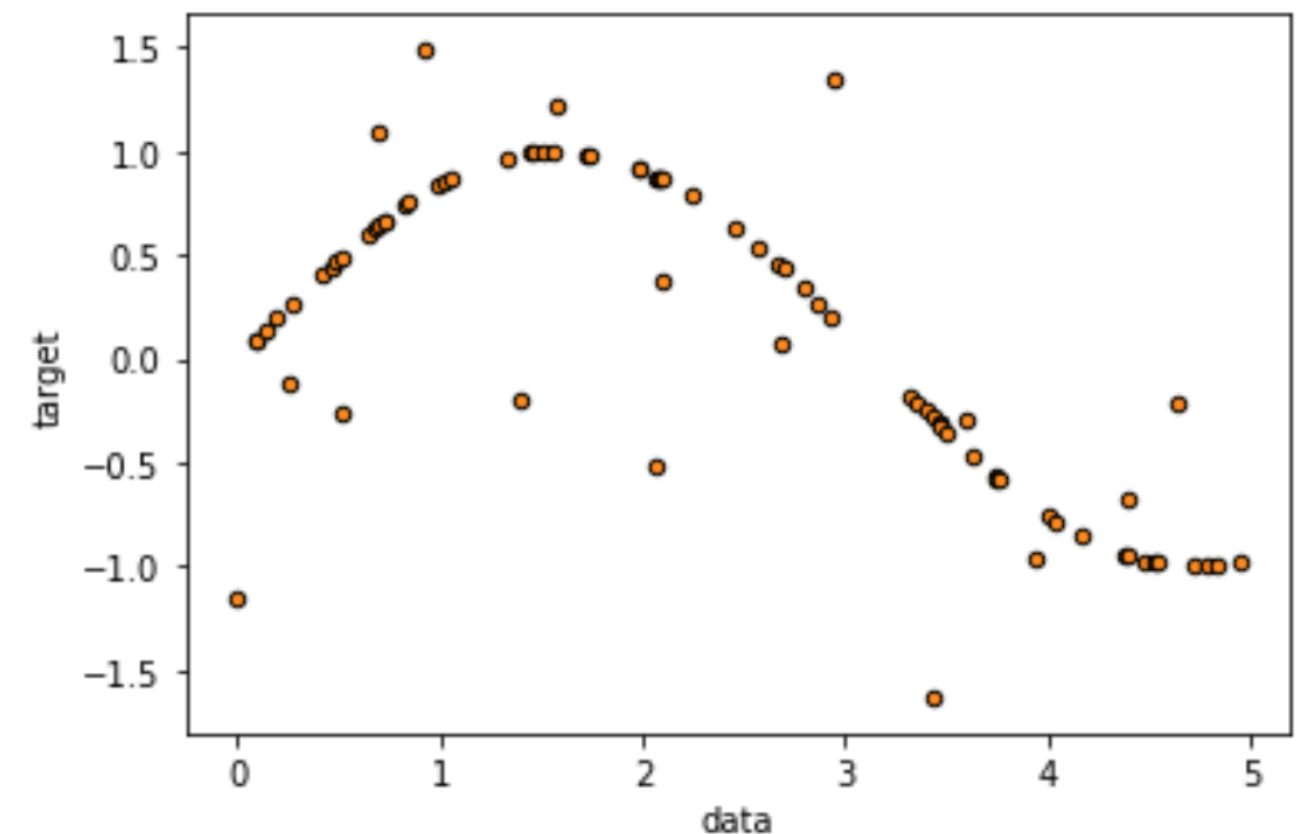
Полиномиальные признаки

- Что если целевая переменная имеет нелинейную зависимость от одного из параметров?



Полиномиальные признаки

- Что если целевая переменная имеет нелинейную зависимость от одного из параметров?
- Добавим еще один признак в модель, который будет функцией (тригонометрические, возведение в степень, произведение с другим признаком и т.д.) от данного, тогда модель сможет линейно описывать данные





06

Обучение модели

Матричный вид

- Модель: $a(x) = w_0 + \langle w, x \rangle$
- Матрица объекты-признаки:

объект и его признаки

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

значения признака на всех объектах

- Результирующий вектор ответов

$$a(x) = X\vec{w} \quad \begin{pmatrix} \sum_{i=1}^d w_i x_{1i} \\ \sum_{i=1}^d w_i x_{2i} \\ \vdots \\ \sum_{i=1}^d w_i x_{\ell i} \end{pmatrix}$$

Ошибка модели

- Будем определять насколько хорошо модель предсказывает ответы с помощью «расстояния» между целевой переменной и предсказанием модели
- Есть множество различных способов вычислить это «расстояние» (функцию потерь)

Ошибка модели

- Будем определять насколько хорошо модель предсказывает ответы с помощью «расстояния» между целевой переменной и предсказанием модели
- Есть множество различных способов вычислить это «расстояние» (функцию потерь)
- Отклонение прогнозов: $Xw - y = \begin{pmatrix} \langle w, x_1 \rangle - y_1 \\ \vdots \\ \langle w, x_\ell \rangle - y_\ell \end{pmatrix}$

Ошибка модели

- Будем определять насколько хорошо модель предсказывает ответы с помощью «расстояния» между целевой переменной и предсказанием модели
- Есть множество различных способов вычислить это «расстояние» (функцию потерь)
- Отклонение прогнозов: $Xw - y = \begin{pmatrix} \langle w, x_1 \rangle - y_1 \\ \vdots \\ \langle w, x_\ell \rangle - y_\ell \end{pmatrix}$
- Функция потерь (среднеквадратичная ошибка): $\frac{1}{\ell} \|Xw - y\|^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2$

Ошибка модели

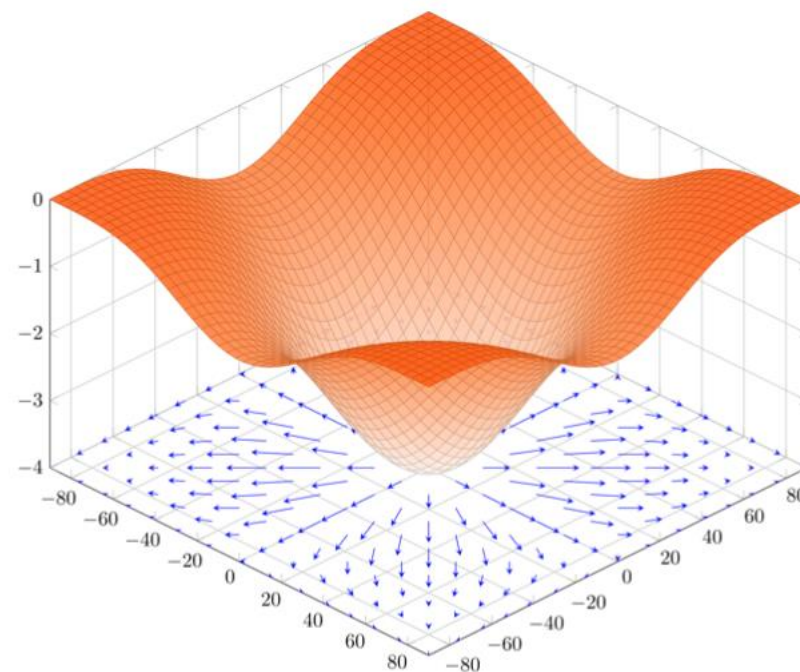
- Будем определять насколько хорошо модель предсказывает ответы с помощью «расстояния» между целевой переменной и предсказанием модели
- Есть множество различных способов вычислить это «расстояние» (функцию потерь)
- Отклонение прогнозов: $Xw - y = \begin{pmatrix} \langle w, x_1 \rangle - y_1 \\ \vdots \\ \langle w, x_\ell \rangle - y_\ell \end{pmatrix}$
- Функция потерь (среднеквадратичная ошибка): $\frac{1}{\ell} \|Xw - y\|^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2$
- Задача обучения: $\frac{1}{\ell} \|Xw - y\|^2 \rightarrow \min_w$

Градиент

- Как найти оптимальный вектор w , минимизировать функционала ошибки? $Q = \frac{1}{\ell} \|Xw - y\|^2$

Градиент

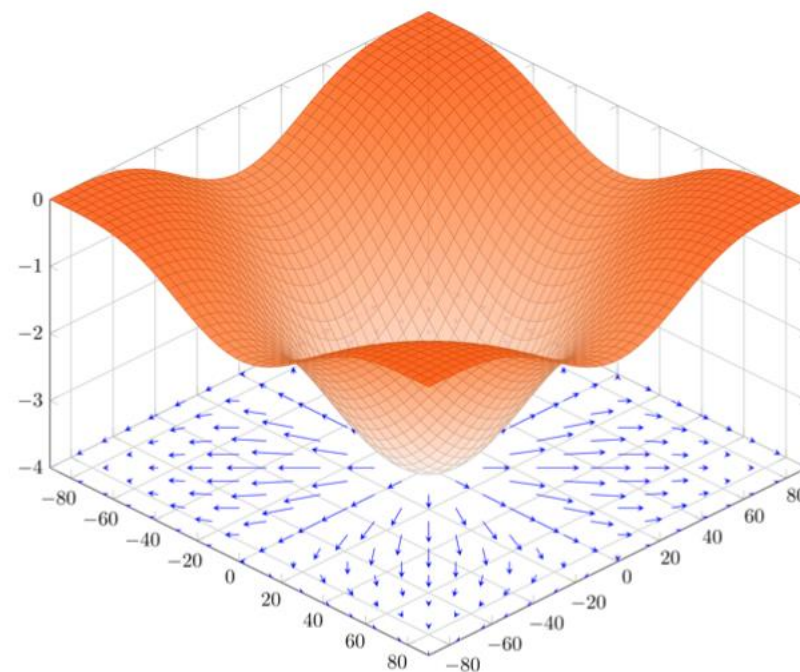
- Как найти оптимальный вектор w , минимизировать функционала ошибки? $Q = \frac{1}{\ell} \|Xw - y\|^2$
- Будем использовать градиент!



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

Градиент

- Как найти оптимальный вектор w , минимизировать функционала ошибки? $Q = \frac{1}{\ell} \|Xw - y\|^2$
- Будем использовать градиент!



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

- Градиент показывает направление, в котором функция растет быстрее всего. А антиградиент показывает обратное
- В точке оптимума градиент примет нулевое значение
- Если функция выпуклая, то экстремум один. При соблюдении $\nabla f(x_0) = 0$ условий MSE для линейной регрессии будет выпуклой

Аналитическое решение

- Вычислим градиент для MSE:

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

- Приравняем к нулю и найдем оптимальный вектор весов:

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

$$\begin{aligned} S &= \|Aw - y\|^2 = (Aw - y)^T (Aw - y) = \\ &= y^T y - y^T Aw - w^T A^T y + w^T A^T Aw = \\ &= y^T y - 2y^T Aw + w^T A^T Aw. \end{aligned}$$

$$\frac{\partial S}{\partial w} = -2A^T y + 2A^T Aw = 0.$$

$$A^T Aw = A^T y,$$

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

Аналитическое решение

- Вычислим градиент для MSE:

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

- Приравняем к нулю и найдем оптимальный вектор весов:

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

- Если матрица $X^T X$ вырожденная/почти вырожденная, возникнут проблемы

Аналитическое решение

- Вычислим градиент для MSE:

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

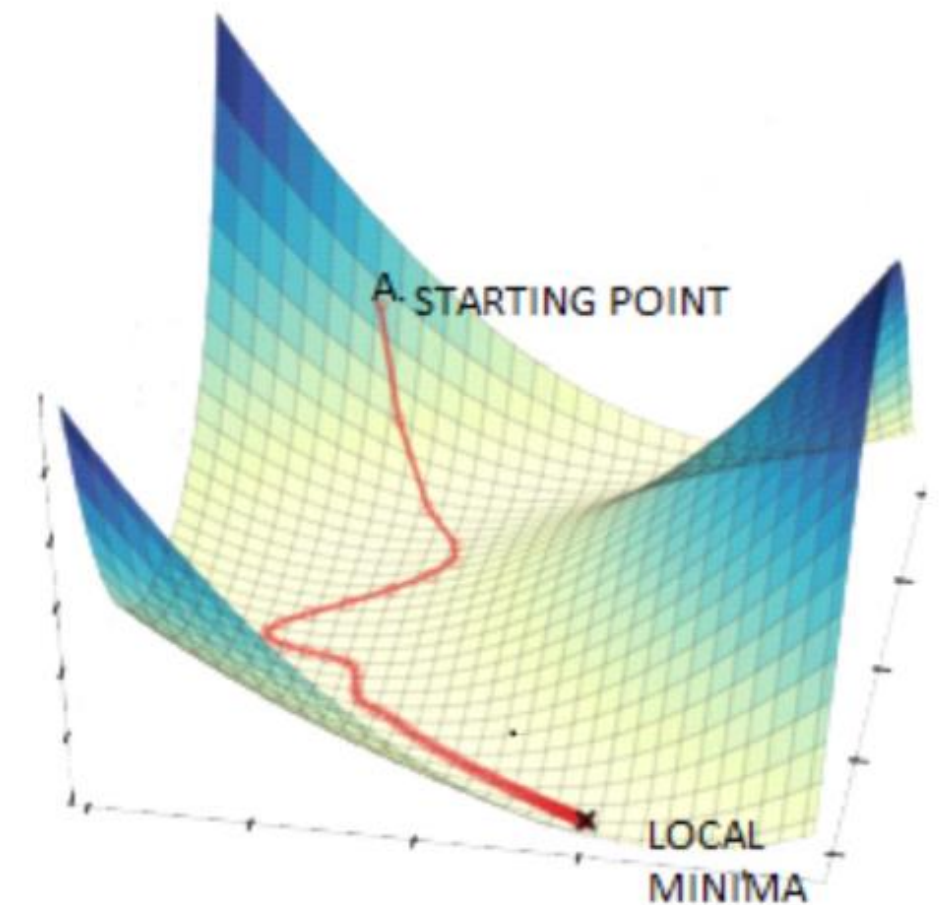
- Приравняем к нулю и найдем оптимальный вектор весов:

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

- Если матрица $X^T X$ вырожденная/почти вырожденная, возникнут проблемы
- Подумайте: как изменяется количество необходимых вычислений при увеличении матрицы X ?

Градиентный спуск

- Чтобы избежать проблем, возникающих при аналитическом решении, воспользуемся другим способом, а именно градиентным спуском



Градиентный спуск

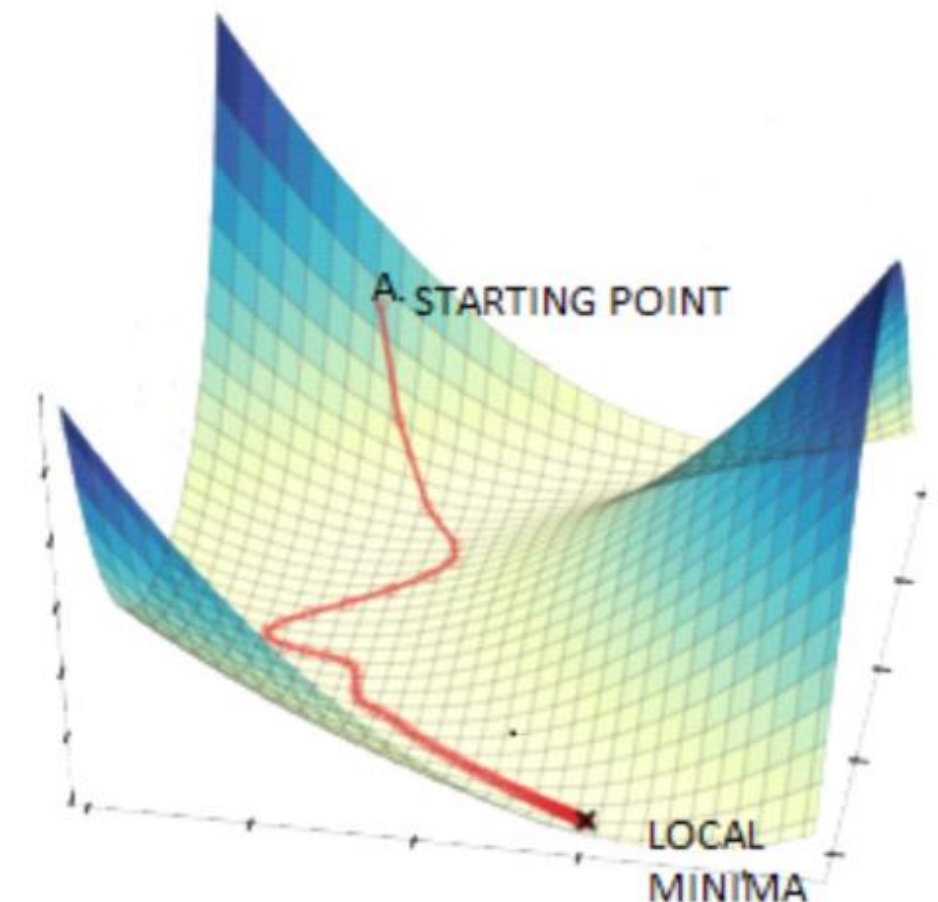
- Чтобы избежать проблем, возникающих при аналитическом решении, воспользуемся другим способом, а именно градиентным спуском
- Алгоритм:
 1. Берем случайный вектор весов
 2. Двигаемся в сторону антиградиента
 3. Повторяем, пока не выполнится критерий остановки, либо установленное количество раз

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

Новая точка

Размер шага

Градиент в предыдущей точке



Сходимость

- Останавливаем процесс, если

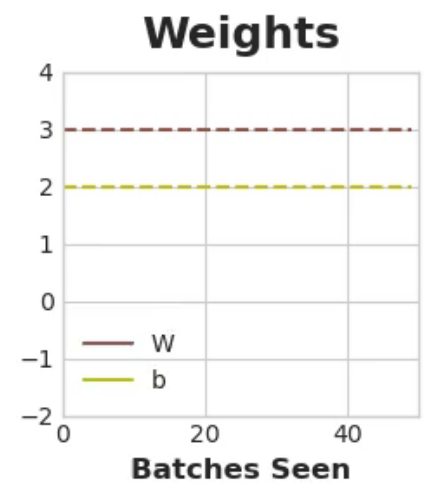
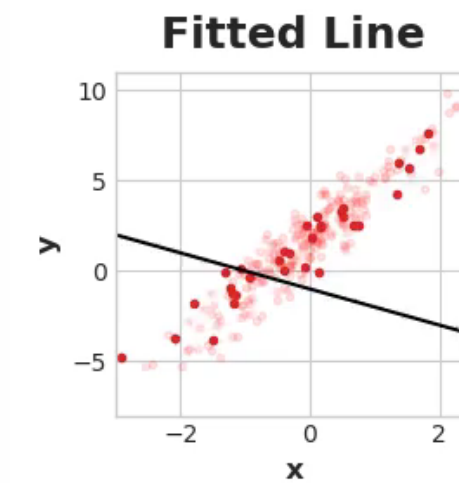
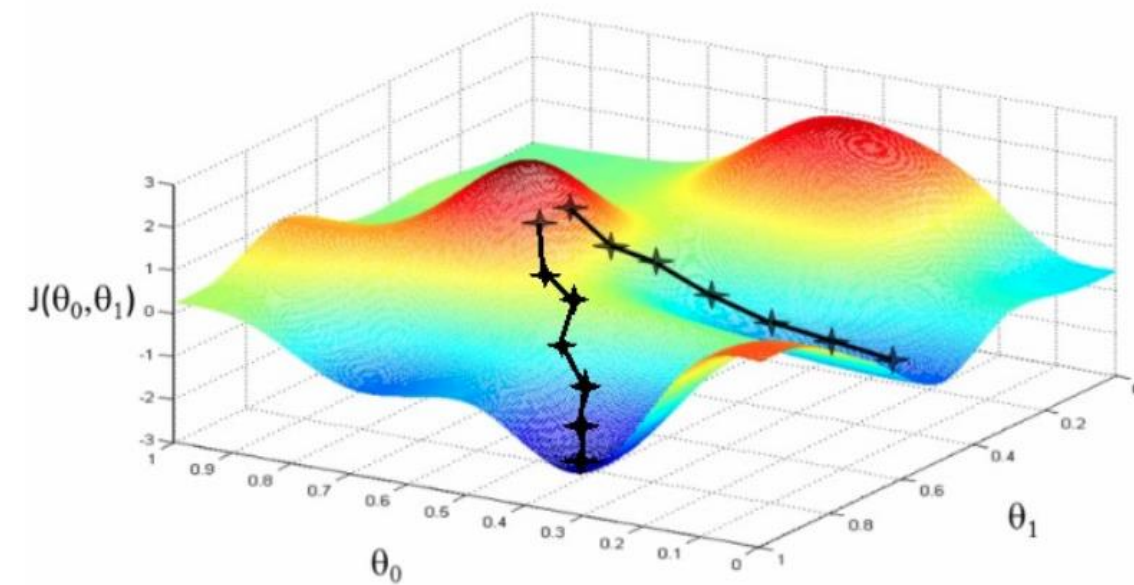
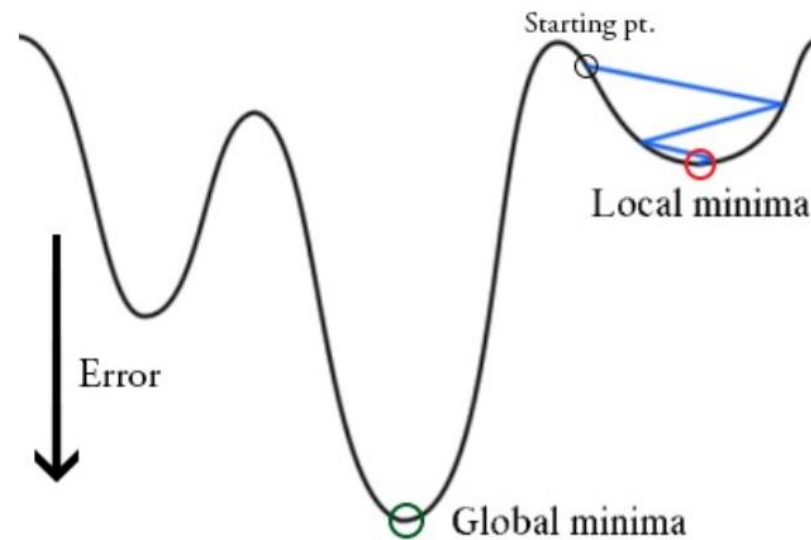
$$\|w^t - w^{t-1}\| < \varepsilon$$

- Либо:

$$\|\nabla Q(w^t)\| < \varepsilon$$

Проблема локальных МИНИМУМОВ

- Градиентный спуск находит только локальные минимумы



Подбор гиперпараметров

- Длина шага является гиперпараметром, который необходимо установить до обучения модели

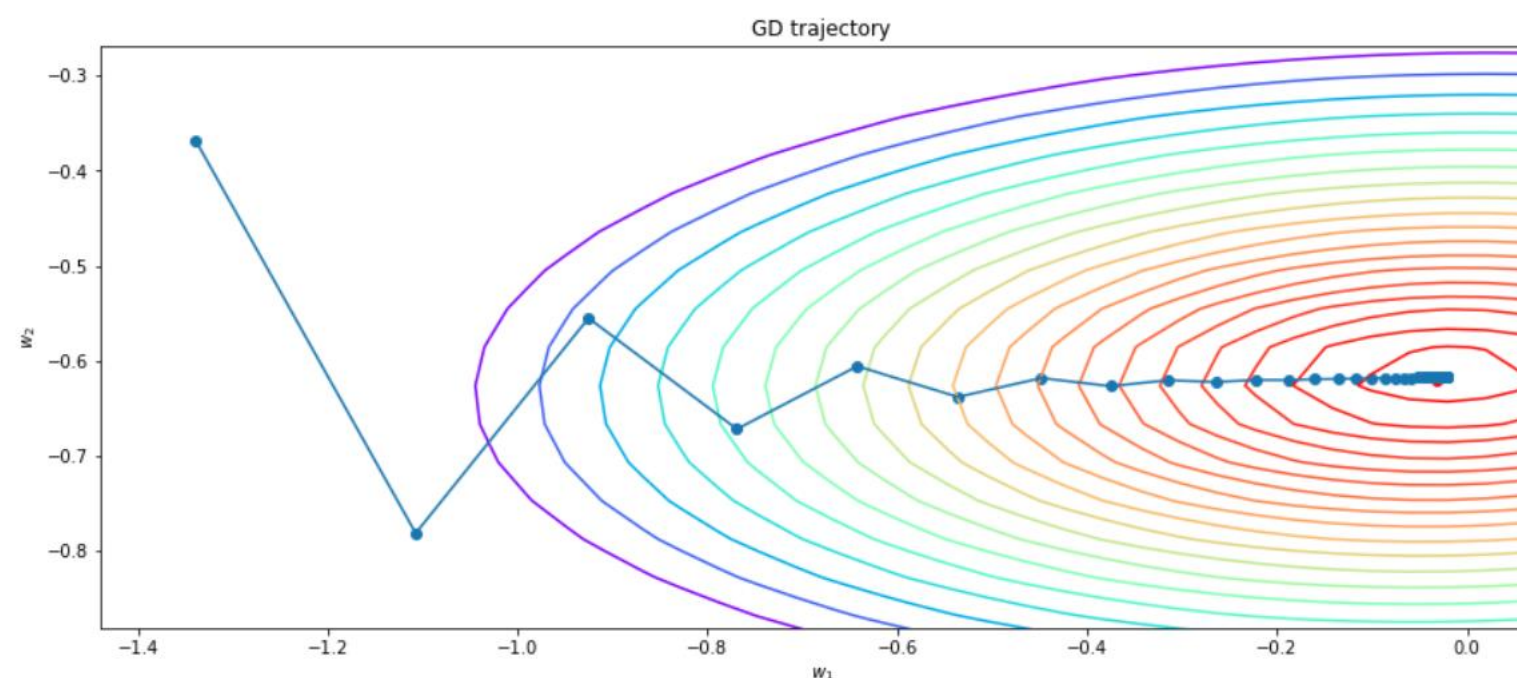
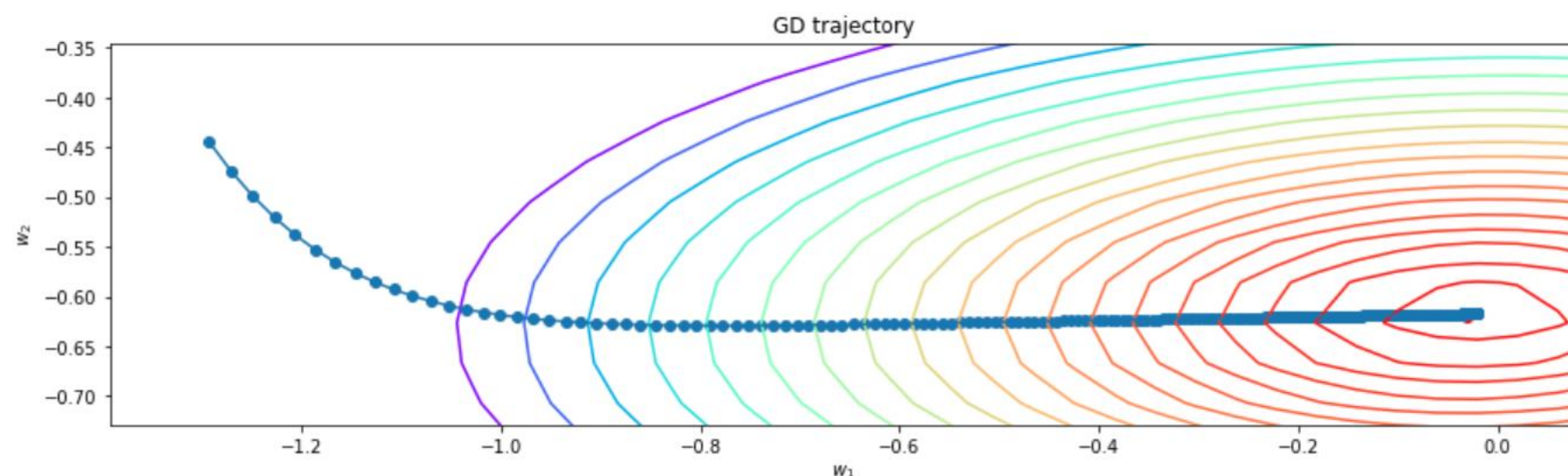
$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

Подбор гиперпараметров

- Длина шага является гиперпараметром, который необходимо установить до обучения модели

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

- Визуализация разной длины шага



- При слишком большом шаге, градиентный спуск может разойтись, а при слишком маленьком модель будет обучаться слишком долго, либо не дойдет до оптимума
- Можно менять в зависимости от номера итерации, например:

$$\eta_t = \frac{1}{t}$$



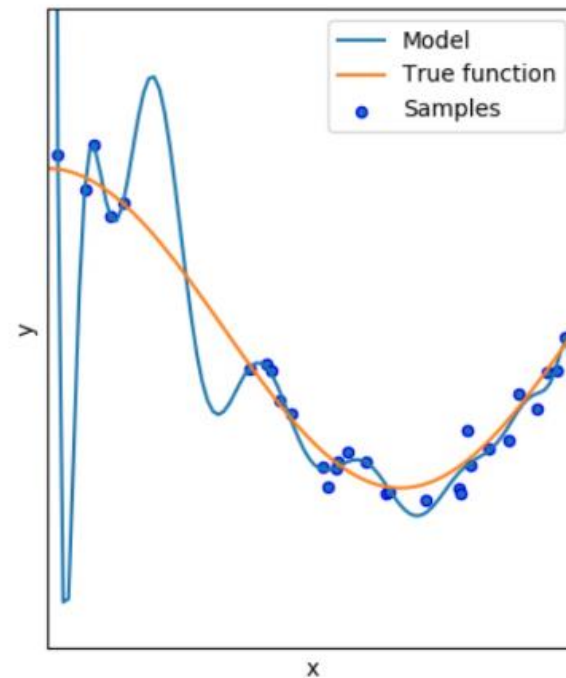
07

Регуляризация модели

Проблема переобучения

- Модель может подогнать веса под тренировочную выборку, вместо того чтобы описать истинное распределение

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



- Большие коэффициенты - симптом переобучения

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

Регуляризация

- Будем штрафовать модель за большие веса

Регуляризация

- Будем штрафовать модель за большие веса

- Регуляризатор (на примере Ridge): $\|w\|^2 = \sum_{j=1}^d w_j^2$

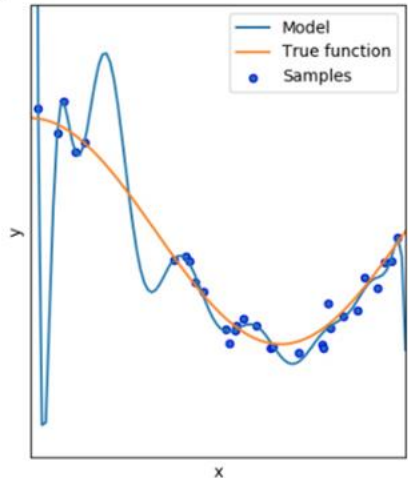
- Регуляризованный функционал: $\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|^2 \rightarrow \min_w$

- λ - коэффициент регуляризации

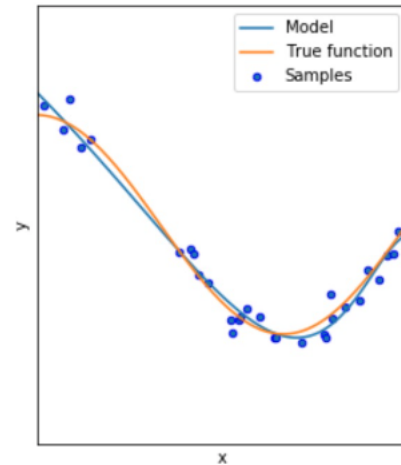
Эффект регуляризации

- Необходимо подбирать коэффициент как гиперпараметр

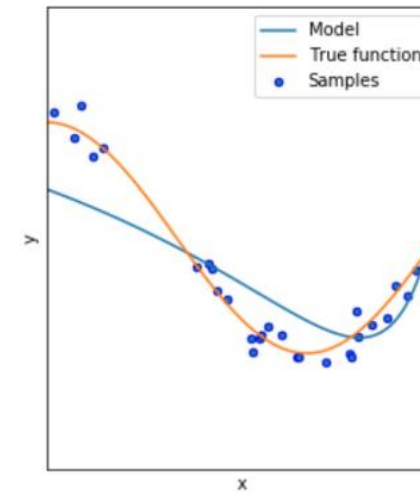
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 \rightarrow \min_w$$



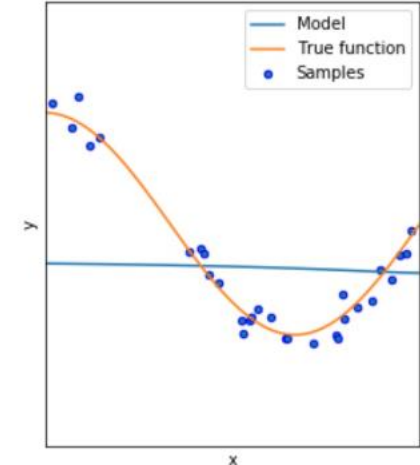
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \mathbf{0.01} \|w\|^2 \rightarrow \min_w$$



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

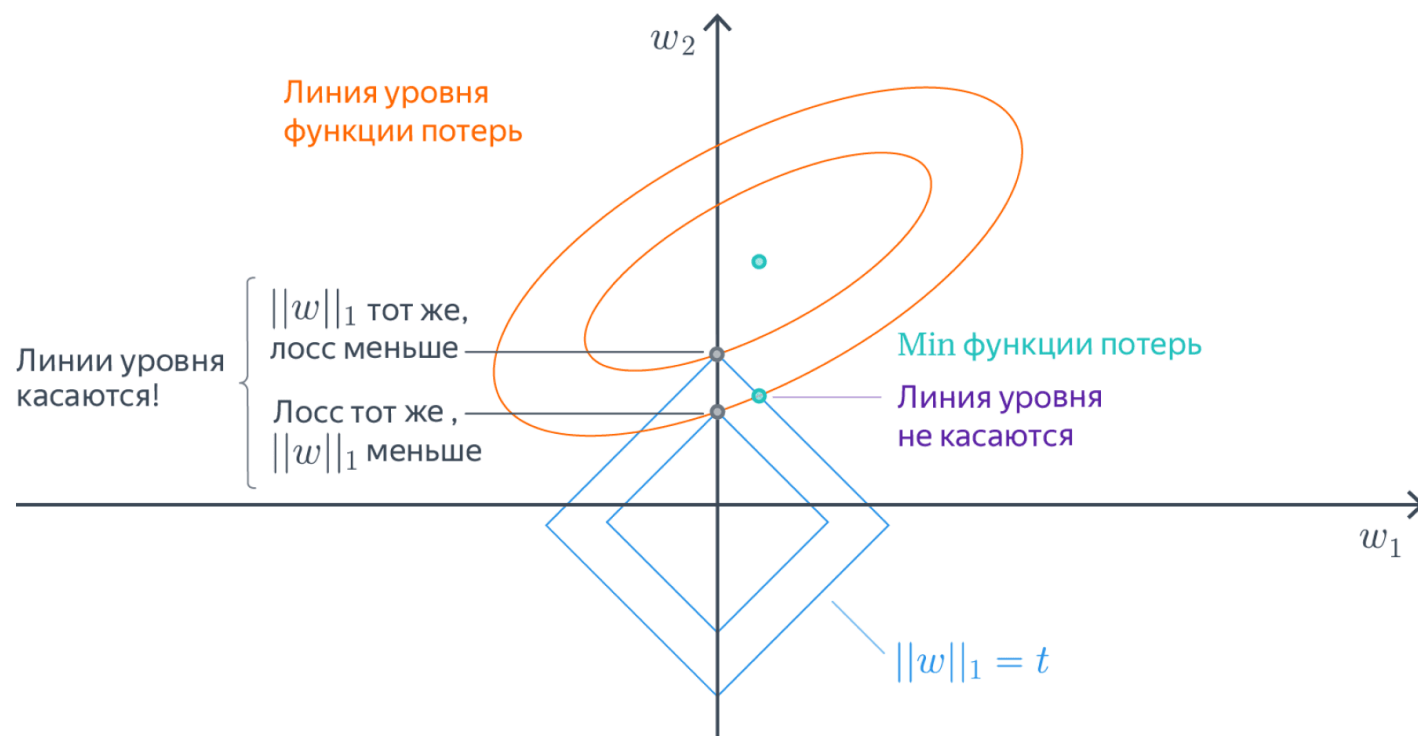


$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \mathbf{100} \|w\|^2 \rightarrow \min_w$$

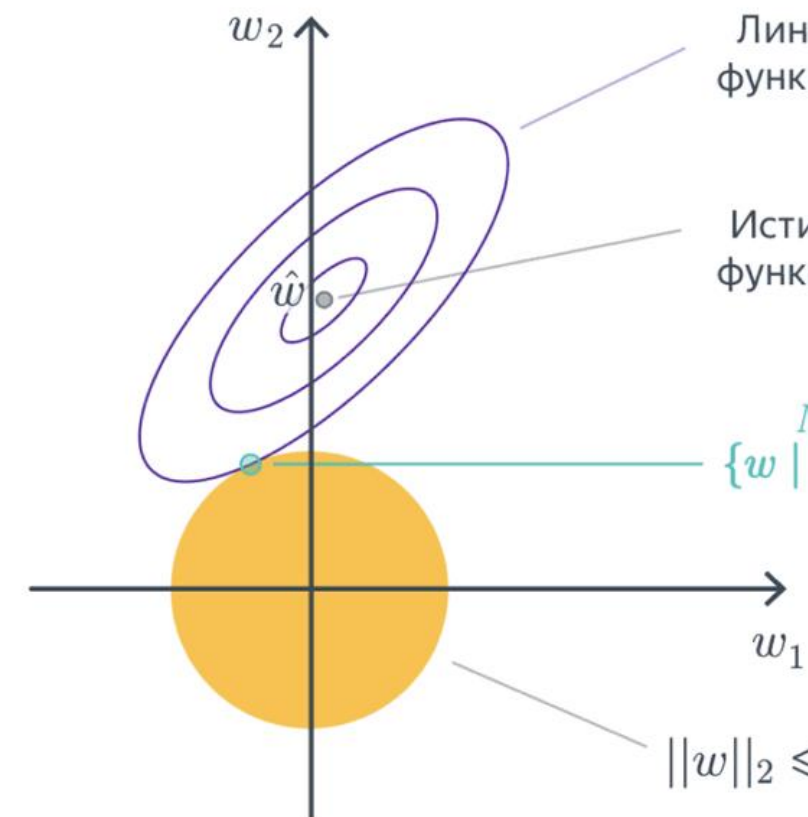


Виды регуляризаторов

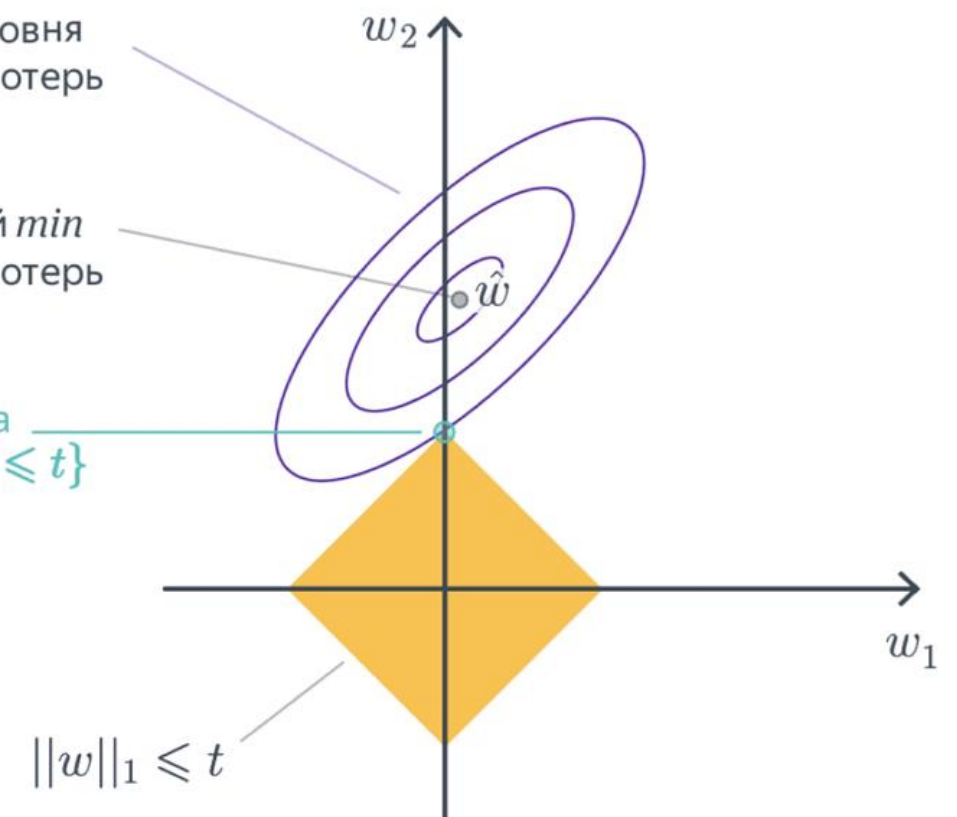
- Ridge $\|z\|_2 = \sqrt{\sum_{j=1}^d z_j^2}$ — L_2 -норма
- Lasso $\sum_{j=1}^d |z_j|$ — L_1 -норма



L_2 -регуляризация



L_1 -регуляризация



Лассо

- Регуляризованный функционал: $\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^d |w_j| \rightarrow \min_w$ $\nabla_w L(f_w, X, y) = 2X^T(Xw - y) + 2\lambda w$
- LASSO (Least Absolute Shrinkage and Selection operator) $w = (X^T X + \lambda I)^{-1} X^T y$
- Некоторые веса загибаются
- Приводит к отбору признаков



08

Функции потерь

Функции потерь

Помимо MSE существует множество других видов функций потерь:

- Функция потерь Хубера
- MAPE
- SMAPE
- ...

Функция потерь Хубера

$$L_H(y, a) = \begin{cases} \frac{1}{2}(y - a)^2, & |y - a| < \delta \\ \delta \left(|y - a| - \frac{1}{2}\delta \right), & |y - a| \geq \delta \end{cases}$$

- Функционал ошибки:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} L_H(y_i, a(x_i))$$

MAPE

- Mean Absolute Percentage Error (средний модуль относительной ошибки)

$$L(y, a) = \left| \frac{y - a}{y} \right|$$

$$Q(a, X) = \frac{100\%}{\ell} \sum_{i=1}^{\ell} \left| \frac{a(x_i) - y_i}{y_i} \right|$$

SMAPE

- Symmetric Mean Absolute Percentage Error (симметричный средний модуль относительной ошибки)

$$L(y, a) = \frac{|y - a|}{(|y| + |a|)/2}$$

$$Q(a, X) = \frac{100\%}{\ell} \sum_{i=1}^{\ell} \frac{|y_i - a(x_i)|}{(|y_i| + |a(x_i)|)/2}$$

**Место для ваших
вопросов**

Литература

1. Учебник по машинному обучению ШАД - <https://education.yandex.ru/handbook/ml>
2. Гайдбук Kaggle - <https://www.kaggle.com/learn/feature-engineering>
3. Примеры кодирования Kaggle - <https://www.kaggle.com/code/subinium/11-categorical-encoders-and-benchmark>
- 4.