

Maximum likelihood (Метод максимального правдоподобия)

Пусть у нашей модели есть параметры W . Пусть у нас есть наши наблюдения Y . Хотим максимизировать $P(y|W)$

Задача

- Дана монетка. Подбрасываем ее три раза. Какая ML-оценка для вероятности выпадения орла?

Задача

- Дана монетка. Подбрасываем ее три раза. Какая ML-оценка для вероятности выпадения орла?

$$P = (\text{число выпадений орла}) / 3$$

Есть ли проблемы с этим подходом?

Задача

- Дана монетка. Подбрасываем ее три раза. Какая ML-оценка для вероятности выпадения орла?

$$P = (\text{число выпадений орла}) / 3$$

Часто будем получать, что вероятность выпадения орла - 0.

**Самые лживые слова
- "никогда" и
"навсегда". Те, кто их
говорил мне, в итоге
предавали.**

Максимум апостериорной вероятности (MAP)

Введем априорное знание о монетке. Пусть монетки с вероятностью выпадения орла p

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

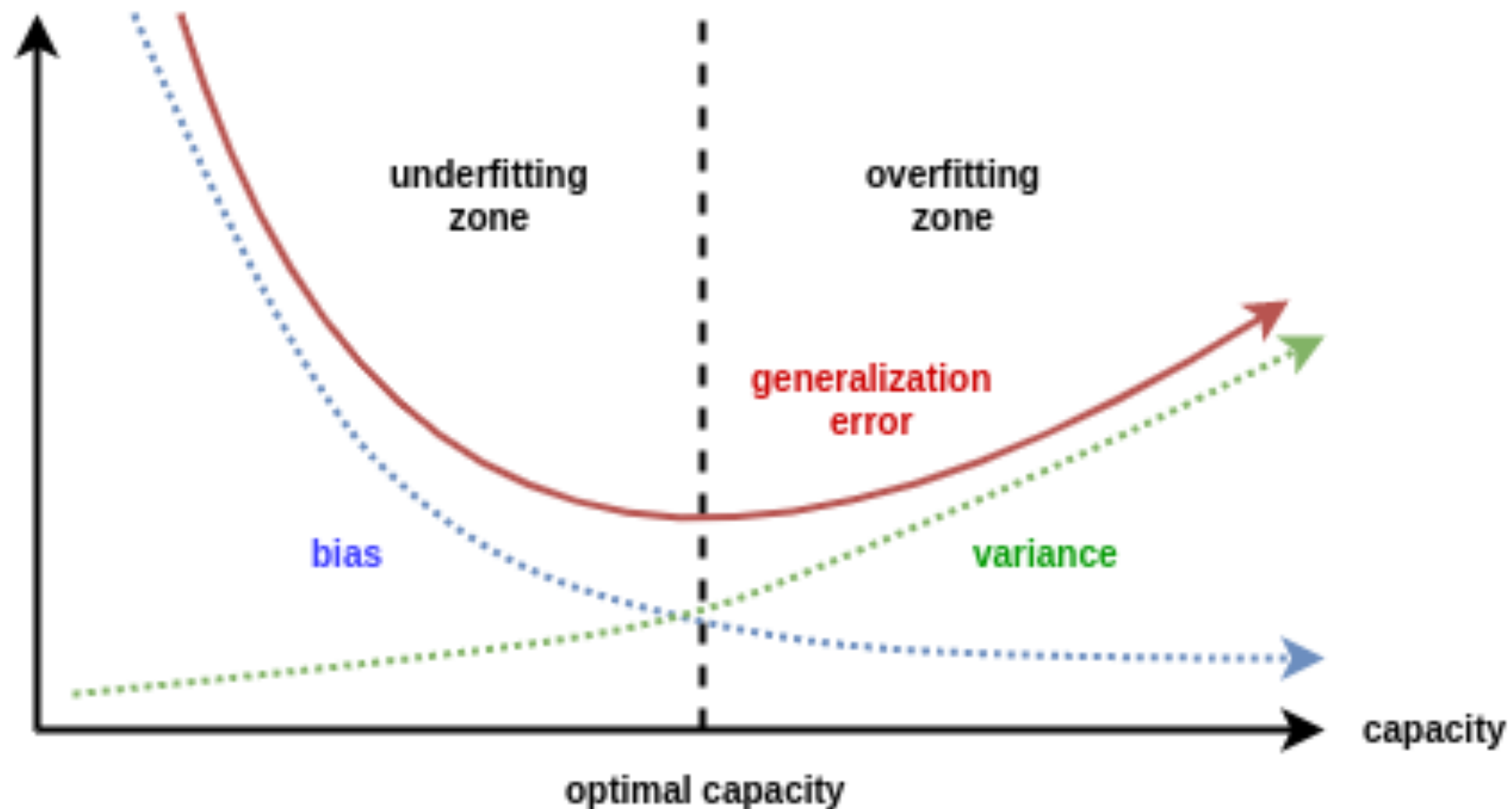


Максимум апостериорной вероятности (MAP)

- Будем максимизировать вероятность параметра

$$\hat{\theta}_{\text{MAP}}(x) = \arg \max_{\theta} f(\theta | x) = \arg \max_{\theta} \frac{f(x | \theta) g(\theta)}{\int_{\Theta} f(x | \vartheta) g(\vartheta) d\vartheta} = \arg \max_{\theta} f(x | \theta) g(\theta).$$

Bias-variance tradeoff



Линейная регрессия

Постановка в одномерном случае

x - некий признак объекта (**независимая переменная**)

y - предсказываемая величина (**зависимая переменная**)

Предположим, что $y = f(x) + \text{eps}$ (eps - шум, распределенный нормально)

Хотим найти такую функцию $h(x) = bx + a$, которая **лучше всего** аппроксимирует эту зависимость

Много переменных

x_j - некий признак объекта (**независимая переменная**)

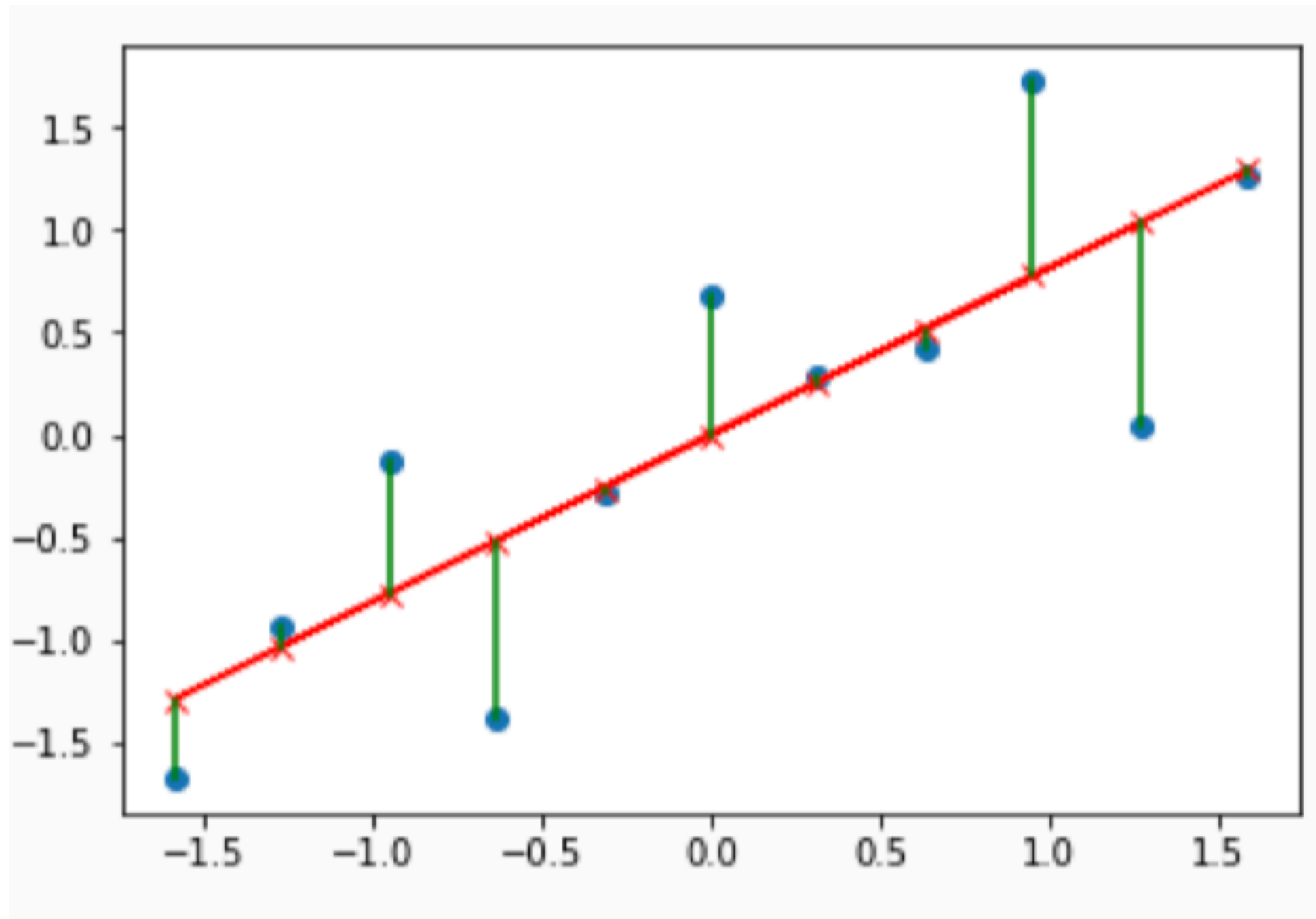
y - предсказываемая величина (**зависимая переменная**)

Предположим, что $y = f(x) + \text{eps}$ (eps - шум, распределенный нормально)

Хотим найти такую функцию $h(x) = bx_1 + \dots + bx_n + a$, которая **лучше всего** аппроксимирует эту зависимость

Решается аналогично

Residuals (остатки)



$$r_i = y - \hat{y}_i$$

MSE

$$\sum_i r_i^2 = \sum_i (y_i - \hat{y}_i)^2 = MSE$$

Хотим минимизировать эту штуку

Почему минимизируем квадраты, а не просто остатки?

Можно ли минимизировать что-то другое?

MAE

$$\sum_i |r_i| = \sum_i |y_i - \hat{y}_i| = MAE$$

Можно ли минимизировать что-то другое?

MAE

$$\sum_i |r_i| = \sum_i |y_i - \hat{y}_i| = MAE$$

Можно ли минимизировать что-то другое?

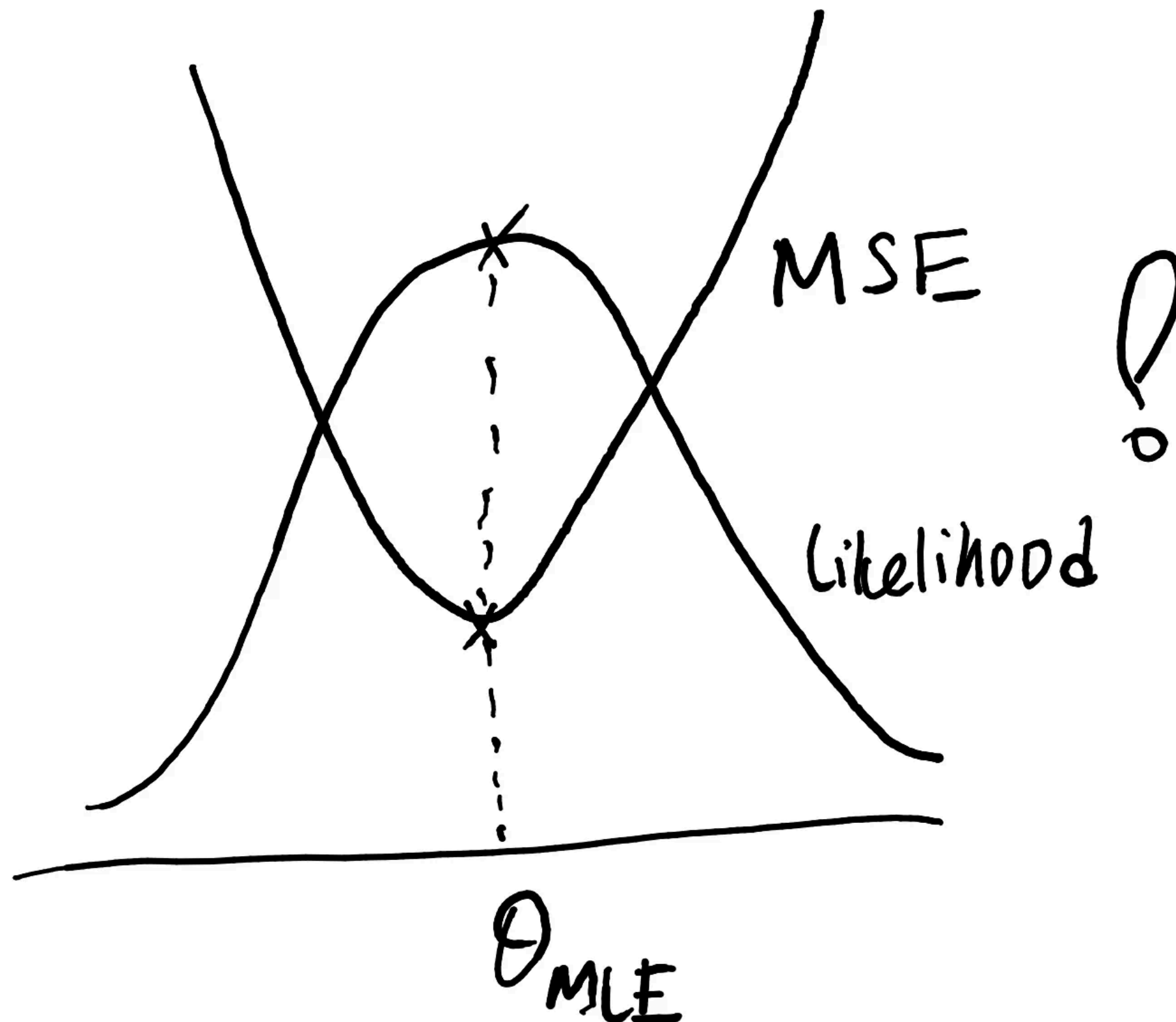
Квантильная ошибка (quantile loss)

$$L_q(\hat{y}, y_i) = \max[q \cdot (\hat{y} - y_i), (q - 1) \cdot (\hat{y} - y_i)]$$

Что выбрать?

- Зависит от задачи
- А почему вообще мы считаем, что это хорошие оценки?

MSE можно получить, используя метод максимального правдоподобия для нашей задачи



Предположения

Обычно, в линейной модели мы предполагаем, что:

$$y_i = Wx_i + \epsilon$$

$$\epsilon \sim N(0, \sigma_i^2)$$

Отсюда:

$$y_i \sim N(Wx_i, \sigma_i^2)$$

ML

$$p(Y|X, W) = \prod_i p(y|x_i, W)$$

$$\log p(Y|X, W) = \sum_i \log p(y|x_i, W)$$

Отсюда:

$$y \sim N(Wx, \sigma_e^2)$$

ML

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) &= \sum_{i=1}^N \log N(y_i; \mathbf{x}_i \mathbf{w}, \sigma^2) \\ &= \sum_{i=1}^N \log \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i \mathbf{w})^2}{2\sigma_e^2}\right) \\ &= -\frac{N}{2} \log 2\pi\sigma_e^2 - \sum_{i=1}^N \frac{(y_i - \mathbf{x}_i \mathbf{w})^2}{2\sigma_e^2}\end{aligned}$$

ML

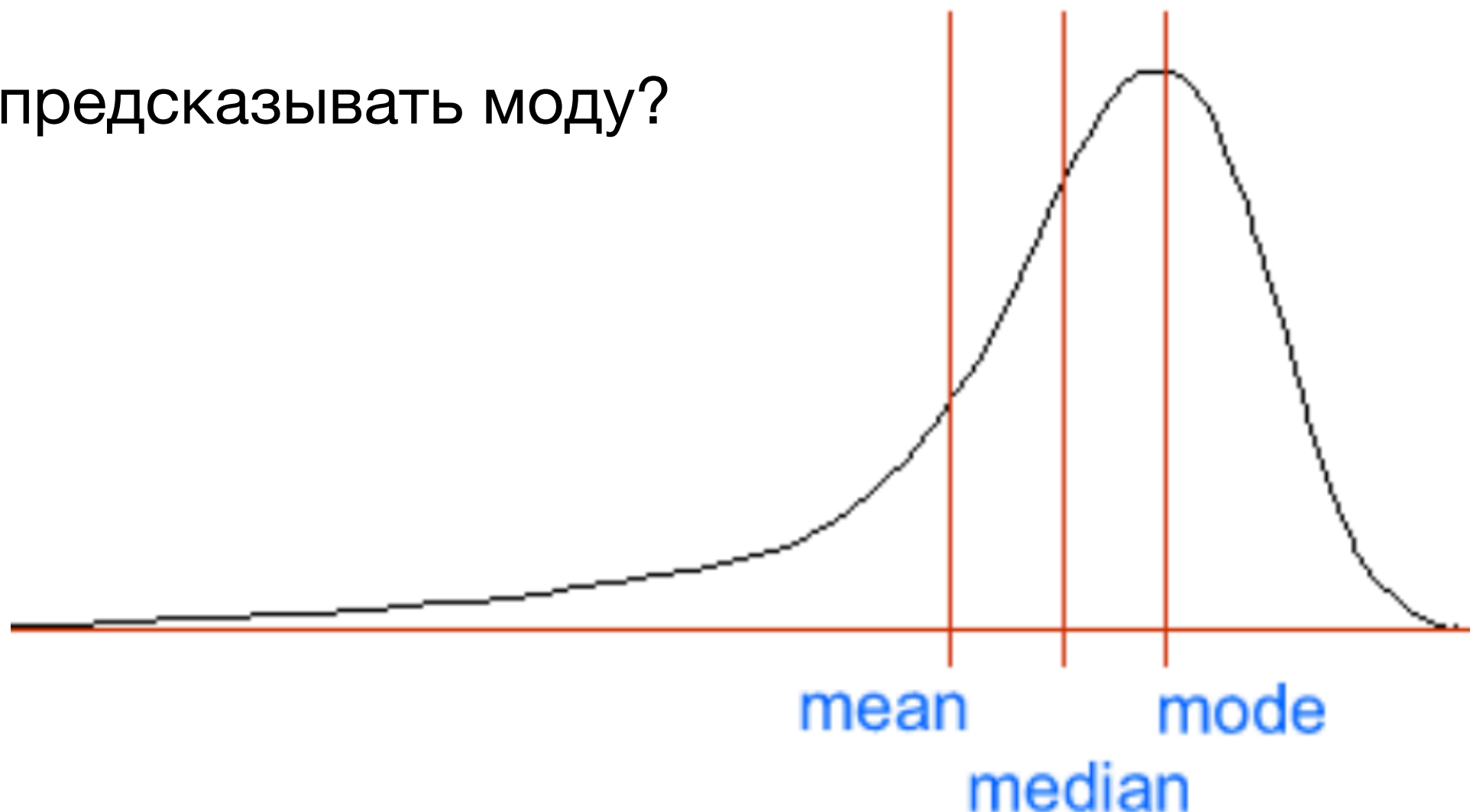
$$\mathbf{w}_{MLE} = \arg \max_{\mathbf{w}} - \sum_{i=1}^N (y_i - \mathbf{x}_i \mathbf{w})^2$$

$$= \arg \min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^N (y_i - \mathbf{x}_i \mathbf{w})^2$$

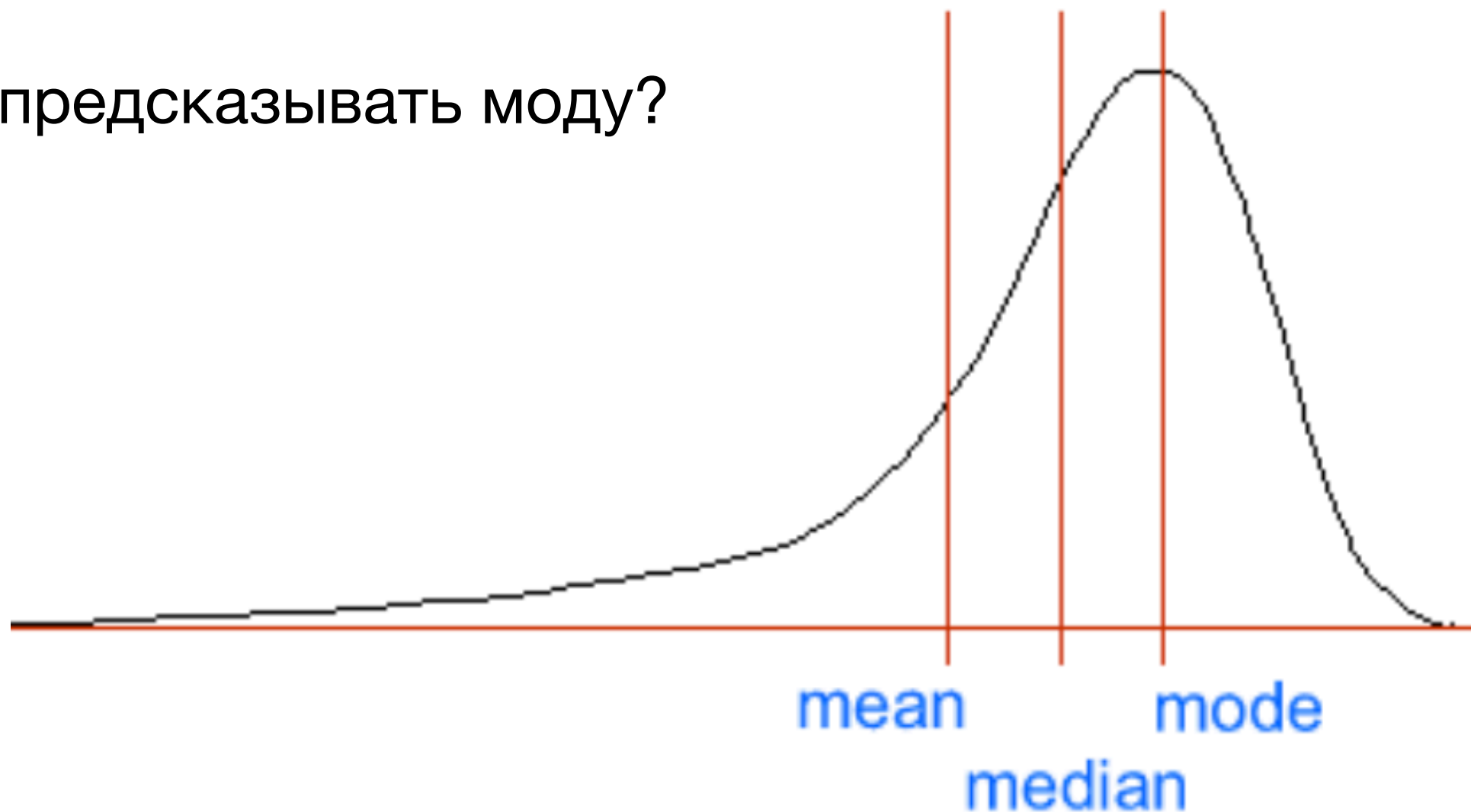
$$= \arg \min_{\mathbf{w}} \text{MSE}_{\text{train}}$$

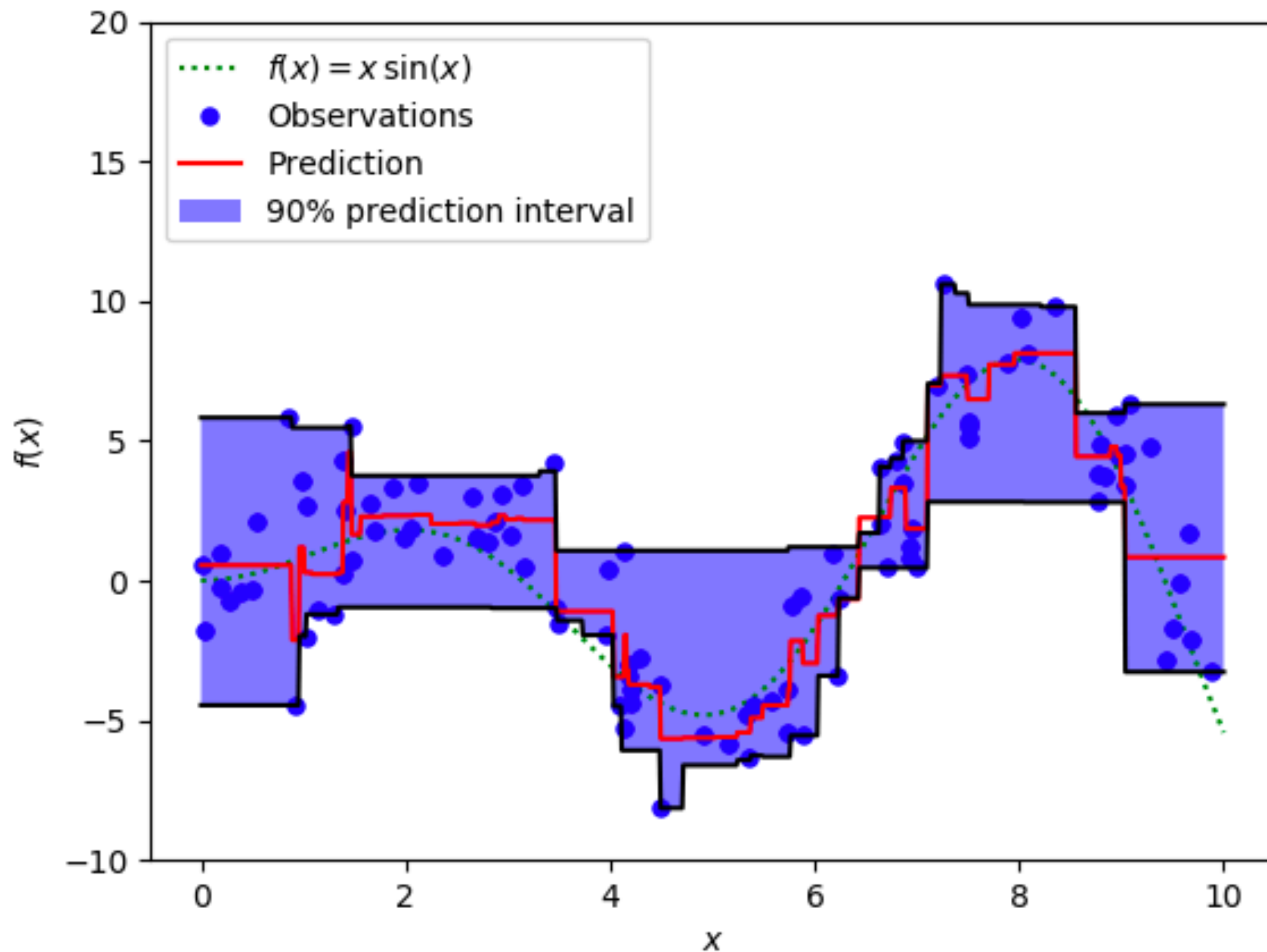
- Пусть мы хотим получить предсказание для конкретного значения x
- Что нам пытаются предсказать MSE?
- Что нам пытаются предсказать MAE?
- Что нам пытаются предсказать quantile loss

- Пусть мы хотим получить предсказание для конкретного значения x
- Что нам пытаются предсказать MSE? - среднее
- Что нам пытаются предсказать MAE? - медиану
- Что нам пытаются предсказать quantile loss - ?
- Как предсказывать моду?



- Пусть мы хотим получить предсказание для конкретного значения x
- Что нам пытаются предсказать MSE? - среднее
- Что нам пытаются предсказать MAE? - медиану
- Что нам пытаются предсказать quantile loss - квантиль
- Как предсказывать моду?





<https://heartbeat.fritz.ai/5-regression-loss-functions-all-machine-learners-should-know-4fb140e9d4b0>

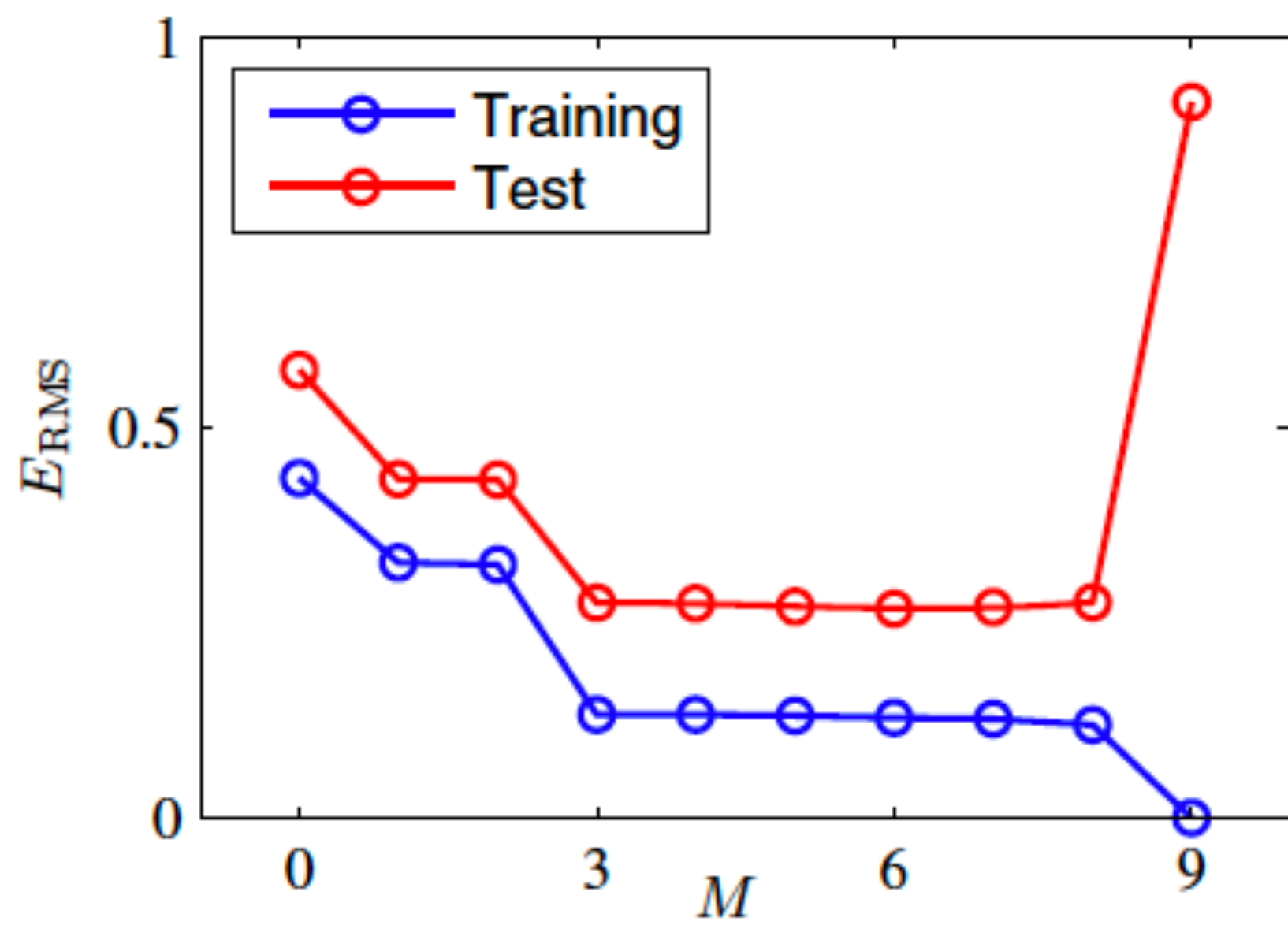
R-squared

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

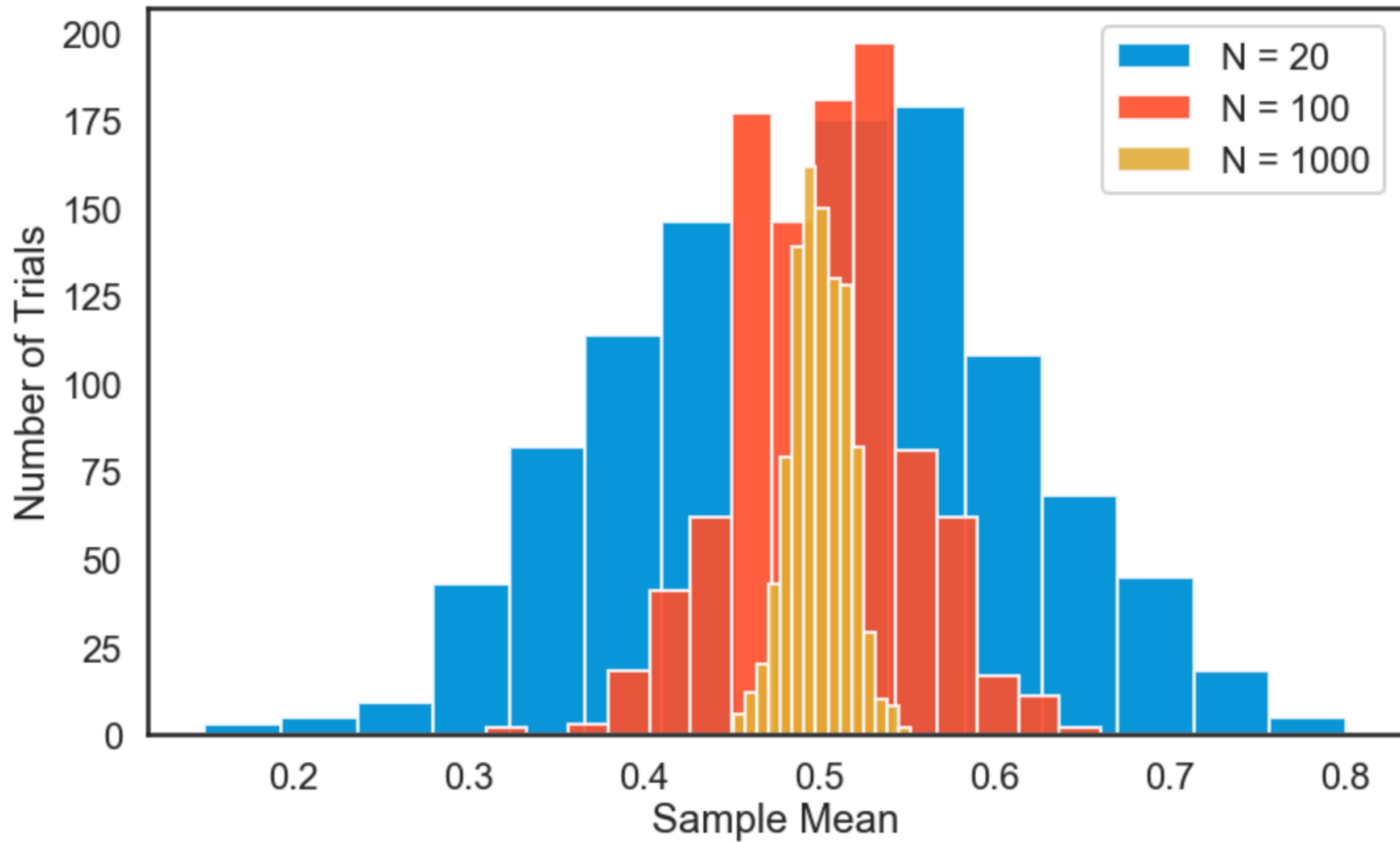
$$SS_{reg} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i r_i^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Коэффициент детерминации, в случае выполнения некоторых предположений, доля объясняемой **дисперсии**



| | $M = 0$ | $M = 1$ | $M = 6$ | $M = 9$ |
|---------|---------|---------|---------|-------------|
| w_0^* | 0.19 | 0.82 | 0.31 | 0.35 |
| w_1^* | | -1.27 | 7.99 | 232.37 |
| w_2^* | | | -25.43 | -5321.83 |
| w_3^* | | | 17.37 | 48568.31 |
| w_4^* | | | | -231639.30 |
| w_5^* | | | | 640042.26 |
| w_6^* | | | | -1061800.52 |
| w_7^* | | | | 1042400.18 |
| w_8^* | | | | -557682.99 |
| w_9^* | | | | 125201.43 |



Регуляризация

Добавляем штраф за большие веса

$$MSE + \textit{penalty}(w)$$

Виды штрафов:

$$L_1 = \alpha \sum_i |w_i|$$

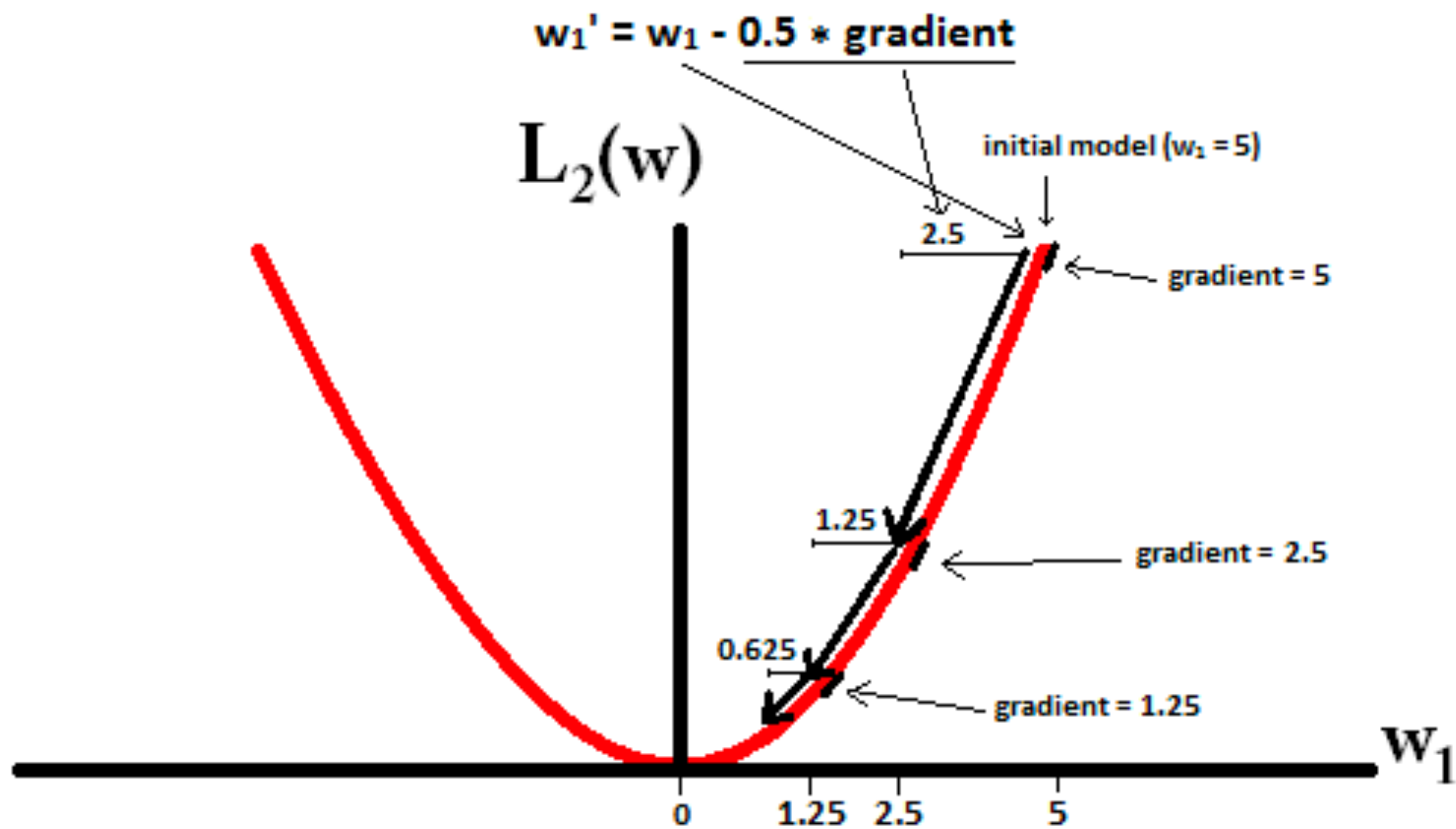
$$L_2 = \beta \sum_i w_i^2$$

$$L_{\textit{elastic}} = \alpha \sum_i |w_i| + (1 - \alpha) \sum_i w_i^2$$

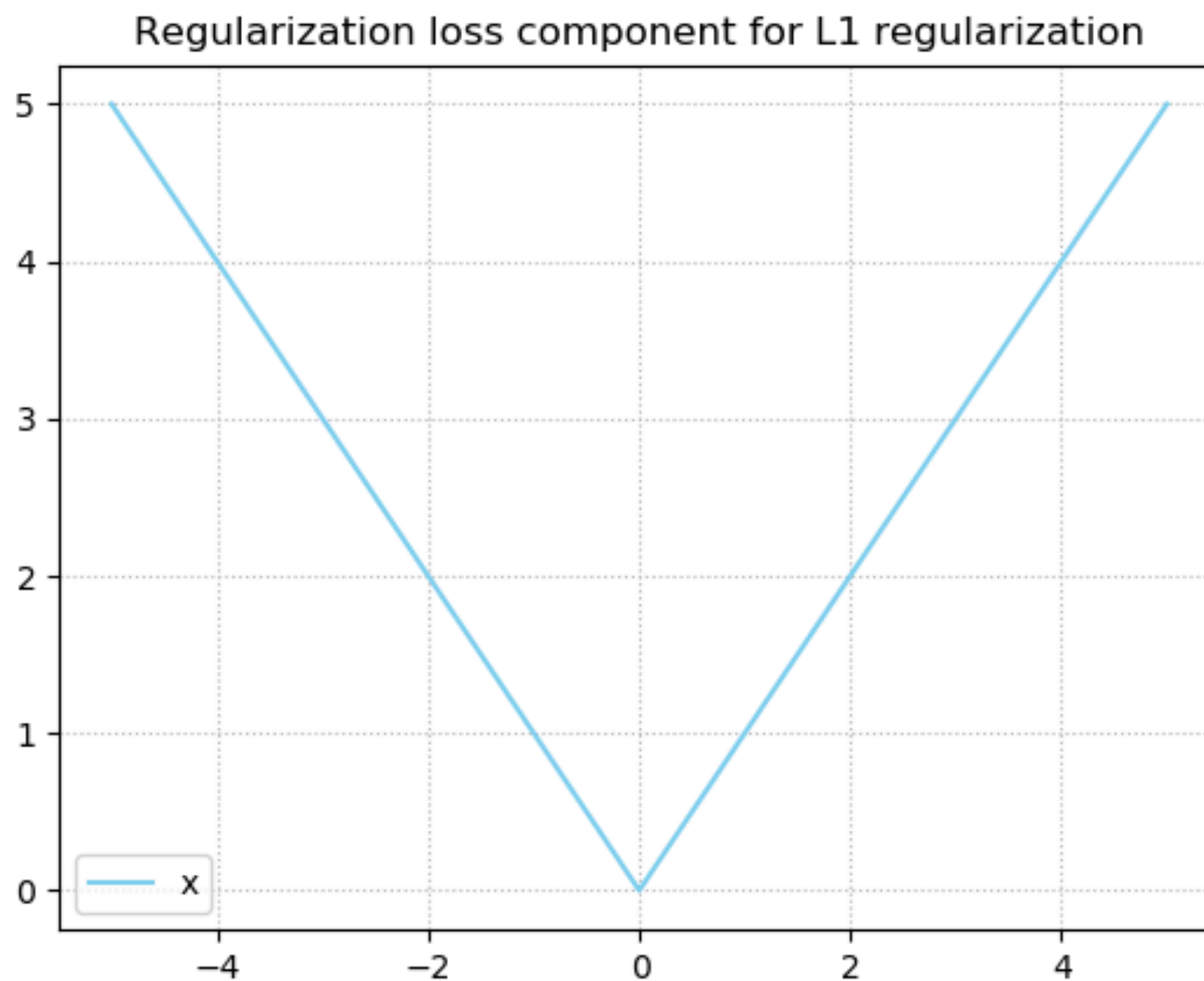
В чем отличие L1 от L2

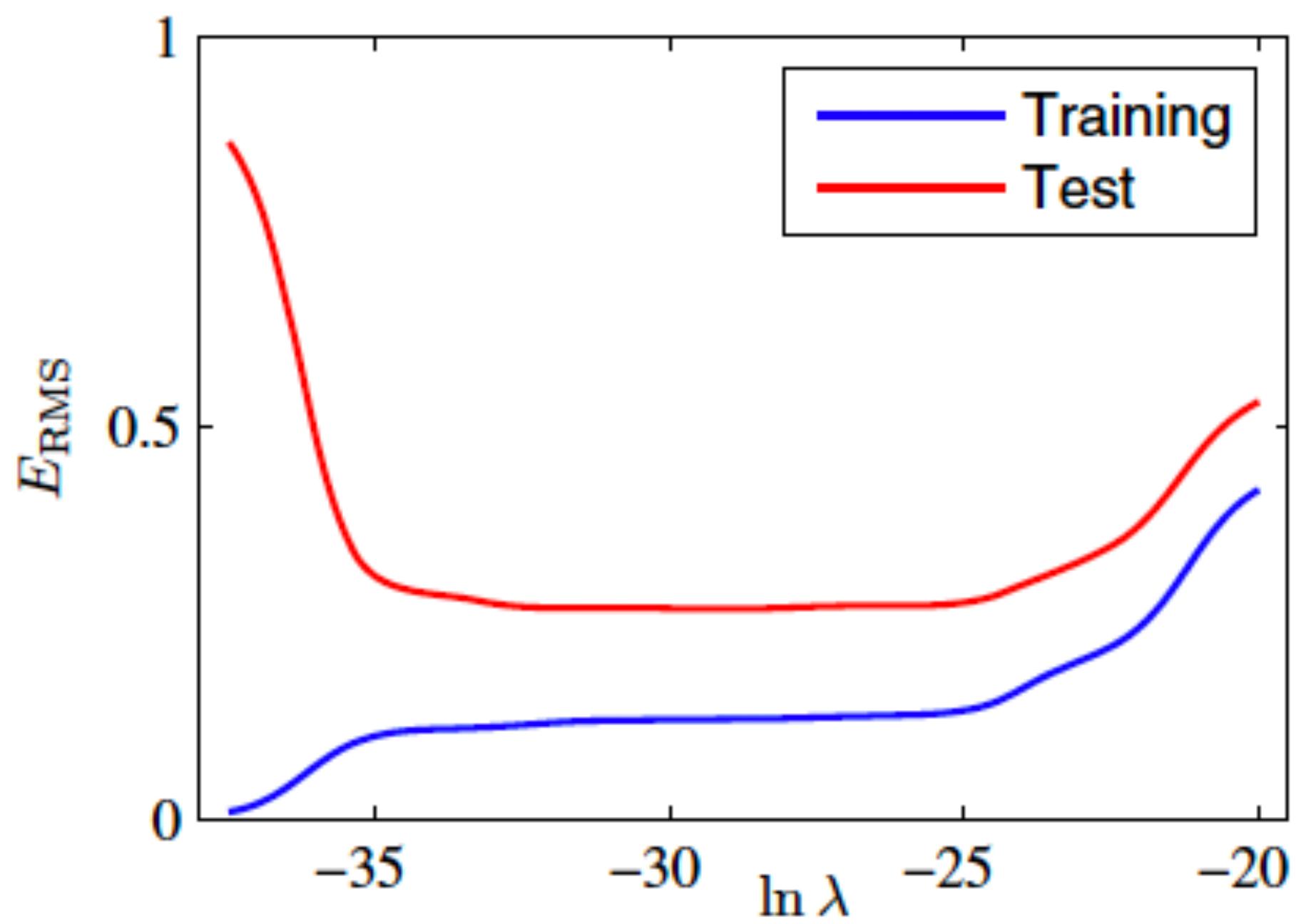
Как штрафуются модель за наличие больших весов?

Регуляризация



Регуляризация





Регуляризация

Регуляризация - это сообщение некоторой информации о весах, которую мы знаем без данных. Регуляризация - введение априорной вероятности.



Введение априора

Регуляризация - способ задания априора для нашей модели.
Априор особо полезен при малом количестве данных

А есть еще есть способы введения априора?

Введение априора

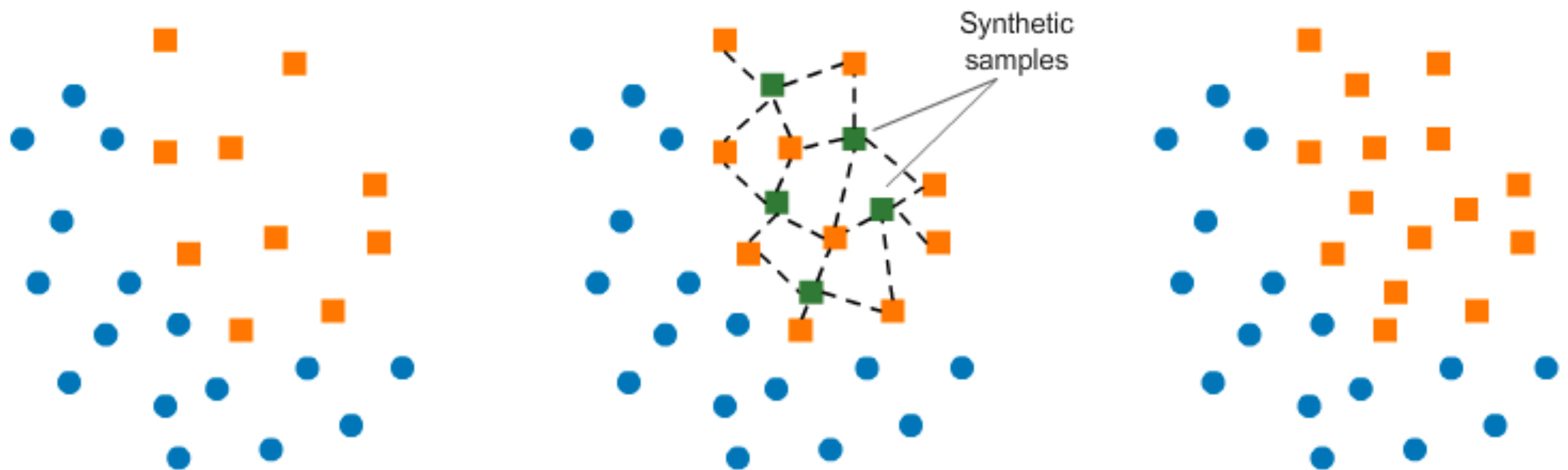
Регуляризация - способ задания априора для нашей модели.
Априор особо полезен при малом количестве данных

А есть еще есть способы введения априора?

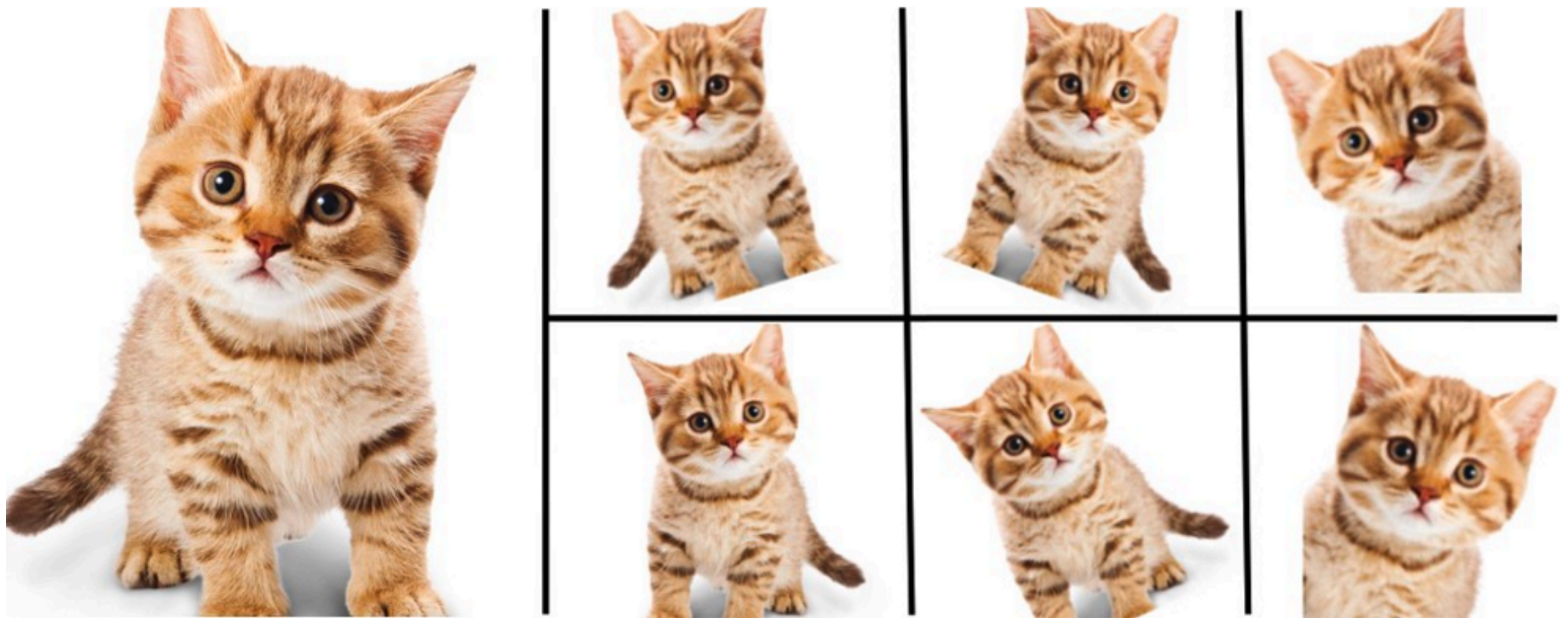
- 1) структура модели
- 2) аугментация данных
- 3)

Примеры аугментации

SMOTE - локально аппроксимирует пространство наших объектов и создает “новые” объекты



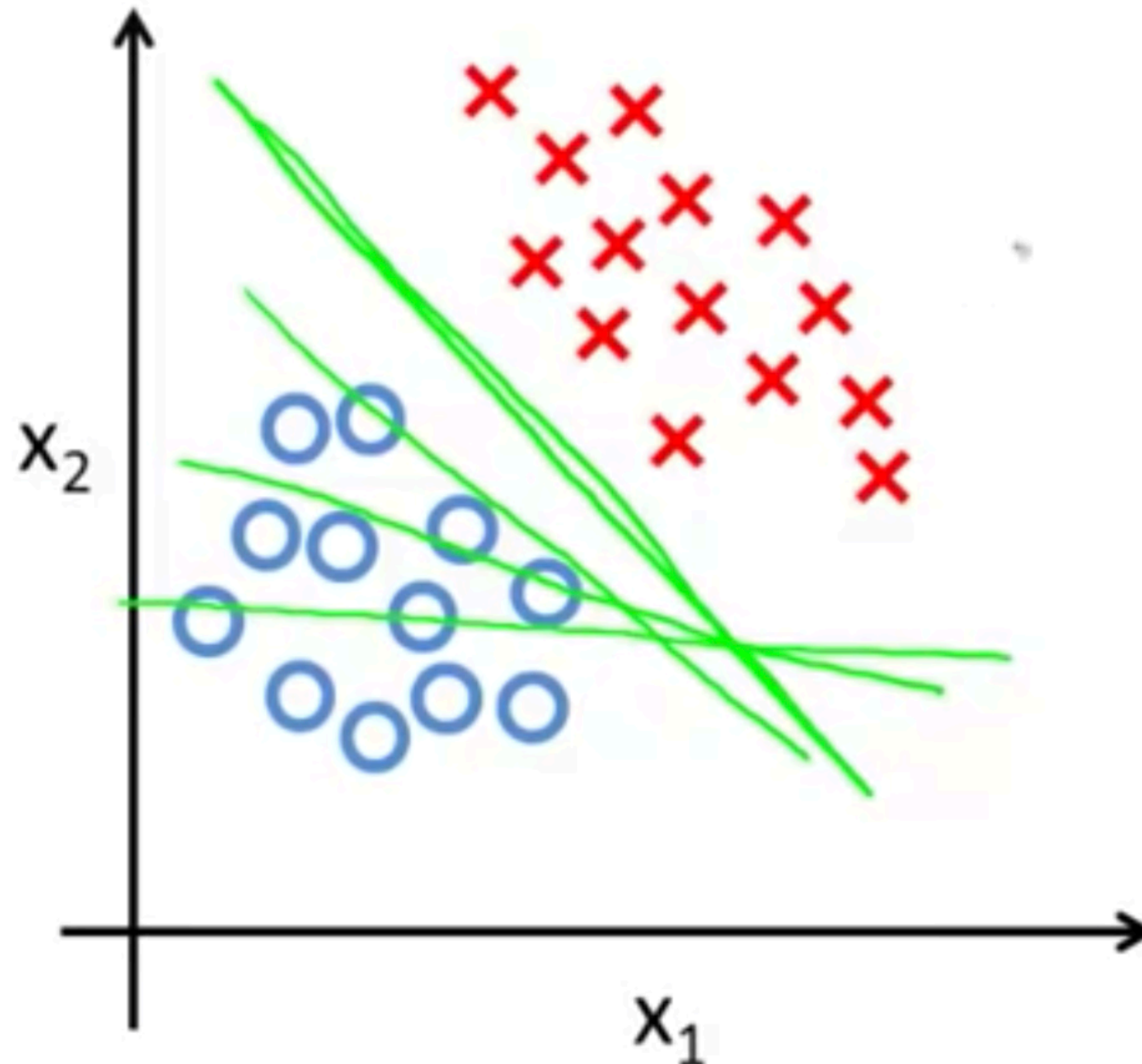
Аугментация в нейронных сетях - повороты картинок, отрезание частей и тд



Enlarge your Dataset

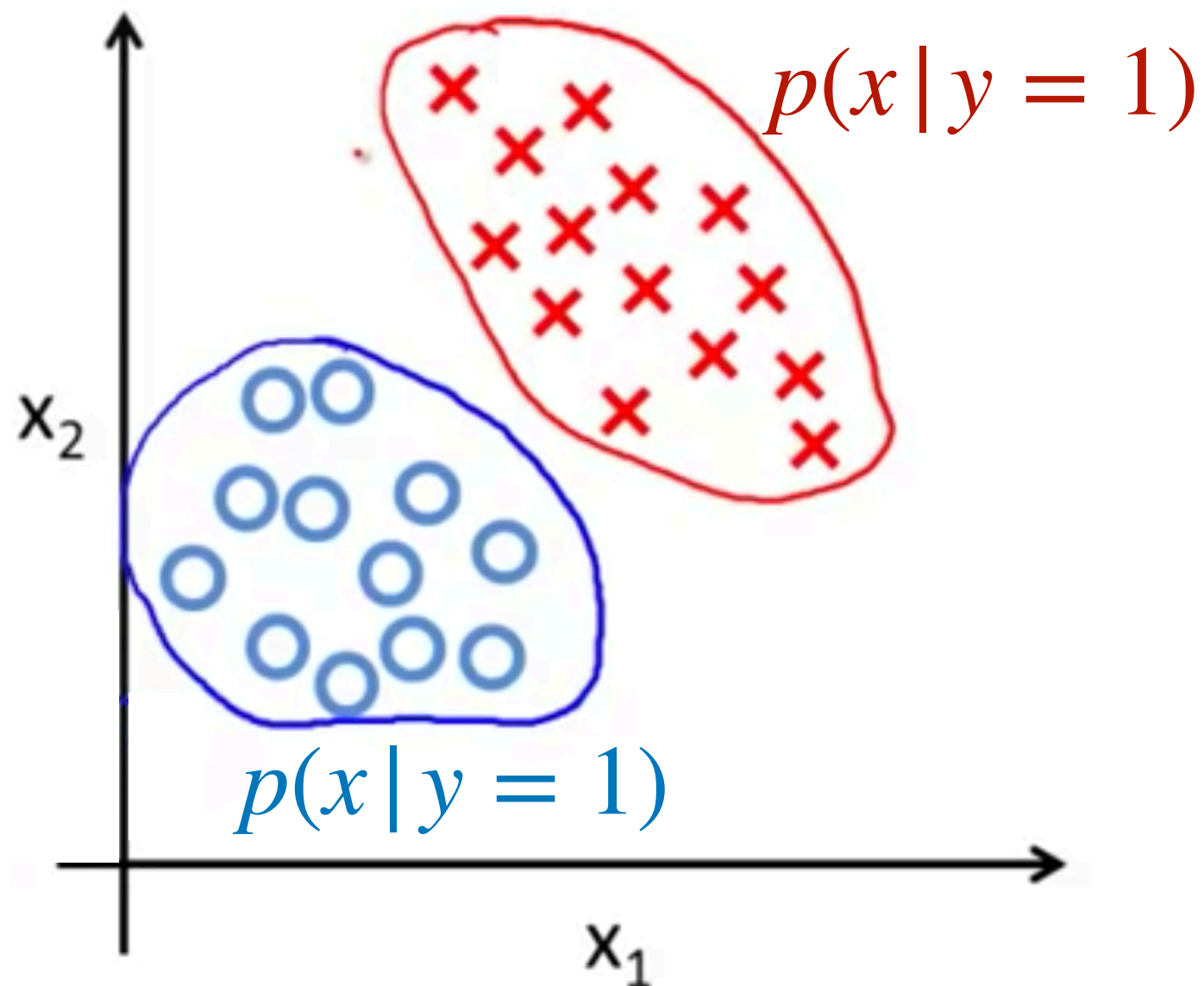
**Почему нельзя
аугментировать от балды?**

Генеративные vs дискриминативные модели



Дискриминативная модель ищет разделяющую плоскость. Задача дискриминативной модели найти $p(\text{class}|\mathbf{x})$. Если она может это делать, то дальше просто для \mathbf{x} выбираем класс с наибольшей вероятностью

Генеративная модель



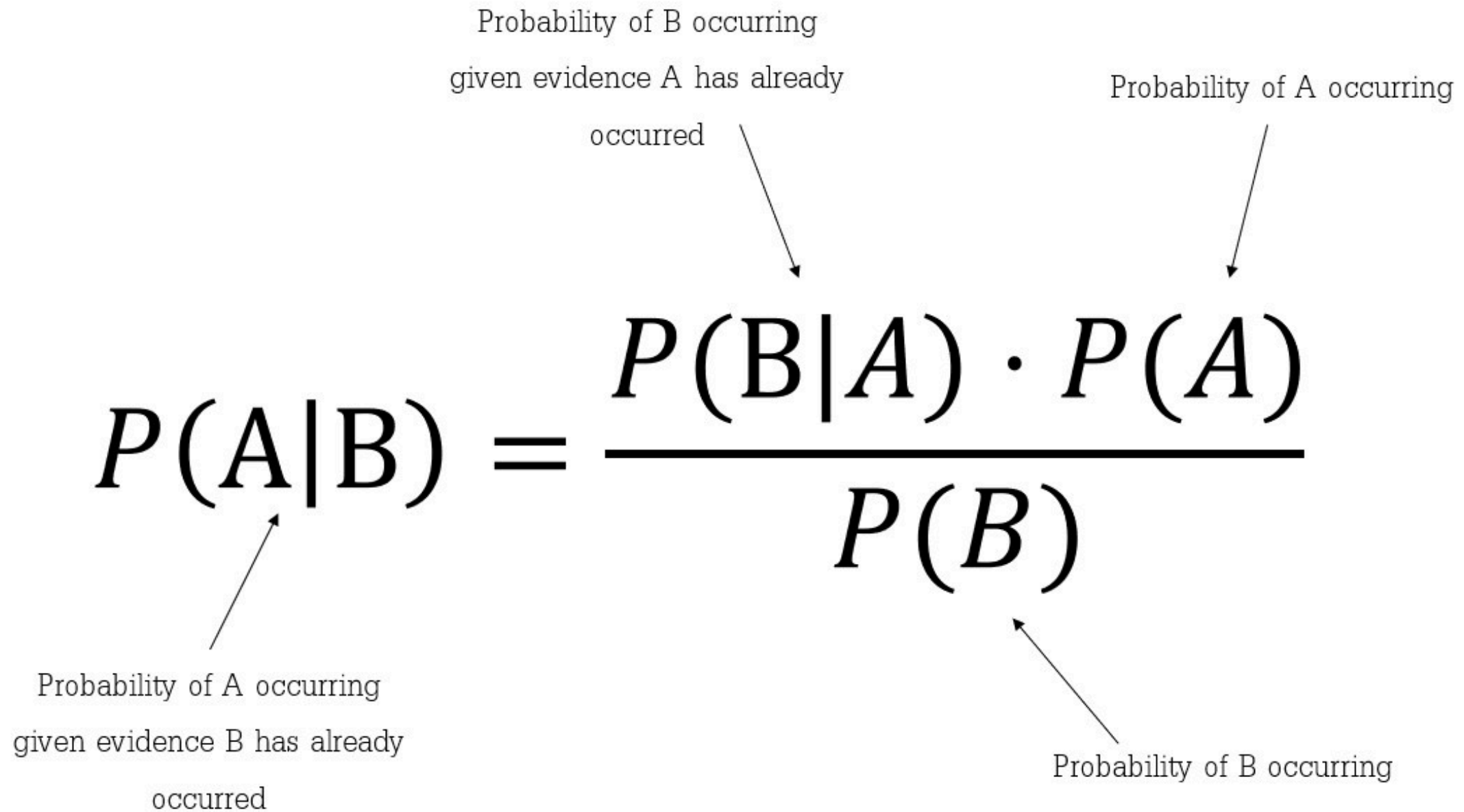
И еще знать это:

$$p(y)$$

Генеративная модель пытается найти $p(y)$, $p(x, y)$ и $p(x|y)$ (второе и третье выражаются друг через друга при условии знания $p(y)$).

$$p(x, y) = p(x | y) \cdot p(y)$$

Наивный байес



Probability of B occurring
given evidence A has already
occurred

Probability of A occurring

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Probability of A occurring
given evidence B has already
occurred

Probability of B occurring

The diagram illustrates the components of Bayes' theorem. The formula $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$ is centered. Four arrows point from descriptive text to parts of the formula: 1. An arrow from the top-left text 'Probability of B occurring given evidence A has already occurred' points to the term $P(B|A)$. 2. An arrow from the top-right text 'Probability of A occurring' points to the term $P(A)$. 3. An arrow from the bottom-left text 'Probability of A occurring given evidence B has already occurred' points to the term $P(A|B)$. 4. An arrow from the bottom-right text 'Probability of B occurring' points to the term $P(B)$ in the denominator.

Наивный Байес

Генеративная модель, нам надо выучить априорные вероятности классов и условные вероятности $p(x|\text{class})$

NAIVE BAYES CLASSIFIER

$$P(\text{class} | \text{data}) = \frac{P(\text{data} | \text{class}) \times p(\text{class})}{P(\text{data})}$$

This is our prior belief

We don't calculate this in naive bayes classifiers

ChrisAlbon

Наивный Байес

**Пусть y объекта есть вектор признаков объекта x .
Тогда получим следующее (из определения условной вероятности)**

$$\begin{aligned} p(x | y = c) &= p(x_1, x_2, \dots, x_n | y = c) = \\ p(x_1 | y = c) &\cdot p(x_2, \dots, x_n | y = c, x_1) = \\ p(x_1 | y = c) &\cdot p(x_2 | y = c, x_1) \cdot p(x_3, \dots, x_n | y = c, x_1, x_2) = \\ p(x_1 | y = c) &\cdot p(x_2 | y = c, x_1) \cdot \dots \cdot p(x_n | y = c, x_1, \dots, x_{n-1}) \end{aligned}$$

**Нужно оценить n условных вероятностей. Это сложно.
Например, для оценки последней вероятности нужно для каждого набора значений $y=c, x_1, \dots$ нужно посчитать, какие при этом наборе встречаются значения x_n**

Наивный Байес

Предположение наивного Байеса - признаки независимы друг от друга. Тогда

$$p(x_1 | x_2) = p(x_1) \quad \text{Из определения условной вероятности}$$

$$\begin{aligned} p(x | y = c) &= p(x_1, x_2, \dots, x_n | y = c) = \\ &= p(x_1 | y = c) \cdot p(x_2, \dots, x_n | y = c, x_1) = \\ &= p(x_1 | y = c) \cdot p(x_2, \dots, x_n | y = c) = \\ &= p(x_1 | y = c) \cdot p(x_2 | y = c) \cdot \dots \cdot p(x_n | y = c) \end{aligned}$$

Нужно оценить n условных вероятностей. Это легче. Просто для каждого значения y оцениваем, как распределены x_i .

Наивный Байес

Хотим предсказывать, текст о спорте или нет?

| Слово | Сколько раз встретило сь слово в тексте о спорте | Сколько раз встретилось слово в тексте не о спорте |
|-------|--|--|
| very | 5 | 0 |
| close | 3 | 10 |
| a | 5 | 5 |
| game | 10 | 2 |
| bed | 0 | 4 |
| Всего | 23 | 21 |



| Слово | Частота | Частота |
|-------|---------|---------|
| very | 5/23 | 0/21 |
| close | 3/23 | 10/21 |
| a | 5/23 | 5/21 |
| game | 10/23 | 2/21 |
| bed | 0/23 | 4/21 |

В чем проблема?

Наивный Байес

Хотим предсказывать, текст о спорте или нет?

| Слово | Сколько раз встретилось слово в тексте о спорте | Сколько раз встретилось слово в тексте не о спорте |
|-------|---|--|
| very | 5 | 0 |
| close | 3 | 10 |
| a | 5 | 5 |
| game | 10 | 2 |
| bed | 0 | 4 |
| Всего | 23 | 21 |



| Слово | Частота | Частота |
|-------|---------|---------|
| very | 5/23 | 0/21 |
| close | 3/23 | 10/21 |
| a | 5/23 | 5/21 |
| game | 10/23 | 2/21 |
| bed | 0/23 | 4/21 |

Вероятность не должна быть равна 0 или 1!

Наивный Байес

Добавляем псевдокаунты. Самое простое - добавить каждого слова по 1 разу

| Слово | Сколько раз встретило сь слово в тексте о спорте | Сколько раз встретилось слово в тексте не о спорте |
|-------|--|--|
| very | 5+1 | 0+1 |
| close | 3+1 | 10+1 |
| a | 5+1 | 5+1 |
| game | 10+1 | 2+1 |
| bed | 0+1 | 4+1 |
| Всего | 23+5 | 21+5 |



| Слово | Вероятность спорт | Вероятность не спорт |
|-------|---------------------|------------------------|
| very | 6/28 | 1/26 |
| close | 4/28 | 11/26 |
| a | 6/28 | 6/26 |
| game | 11/28 | 3/26 |
| bed | 1/28 | 5/26 |

Наивный Байес

Считаем вероятность текста “a very close game” быть про спорт и не про спорт

$$p(sport | text) = \frac{p(text | sport) \cdot p(sport)}{p(text)} =$$

$$\frac{\frac{6}{28} \cdot \frac{6}{28} \cdot \frac{4}{28} \cdot \frac{11}{28} \cdot p(sport)}{p(text)}$$

$$p(\overline{sport} | text) = \frac{p(text | \overline{sport}) \cdot p(\overline{sport})}{p(text)} =$$

$$\frac{\frac{6}{26} \cdot \frac{1}{26} \cdot \frac{11}{26} \cdot \frac{3}{26} \cdot p(\overline{sport})}{p(text)}$$

| Слово | Вероятность спорт | Вероятность не спорт |
|-------|---------------------|------------------------|
| very | 6/28 | 1/26 |
| close | 4/28 | 11/26 |
| a | 6/28 | 6/26 |
| game | 11/28 | 3/26 |
| bed | 1/28 | 5/26 |

Наивный Байес

Считаем вероятность текста “a very close game” быть про спорт и не про спорт

$$p(sport | text) = \frac{p(text | sport) \cdot p(sport)}{p(text)} =$$

$$\frac{\frac{6}{28} \cdot \frac{6}{28} \cdot \frac{4}{28} \cdot \frac{11}{28} \cdot p(sport)}{p(text)}$$

$$p(\overline{sport} | text) = \frac{p(text | \overline{sport}) \cdot p(\overline{sport})}{p(text)} =$$

$$\frac{\frac{6}{26} \cdot \frac{1}{26} \cdot \frac{11}{26} \cdot \frac{3}{26} \cdot p(\overline{sport})}{p(text)}$$

Получается, считать полную вероятность текста, чтобы решить, вероятность какого текста больше - нам не надо.

Но нам нужны априорные вероятности того, что текст про спорт - можем оценить по нашей выборке (просто частота текстов)

Наивный байес

$$\log p(sport | text) = \log \frac{p(text | sport) \cdot p(sport)}{p(text)} = \log p(text | sport) + \log p(sport) - \log(p(text))$$

$$\log p(\overline{sport} | text) = \log \frac{p(text | \overline{sport}) \cdot p(\overline{sport})}{p(text)} = \log p(text | \overline{sport}) + \log p(\overline{sport}) - \log(p(text))$$

Обычно нам интересно отношение вероятностей (odds) или логарифм отношения вероятностей: log-odds

$$\log \frac{p(sport | text)}{p(\overline{sport} | text)} = \log \frac{p(sport)}{p(\overline{sport})} + \log \frac{p(text | sport)}{p(text | \overline{sport})}$$

Апостериорный log-odds

Априорный log-odds

Изменение log-odds
за счет нашего наблюдения

Наивный байес

$$\log p(sport | text) = \log \frac{p(text | sport) \cdot p(sport)}{p(text)} = \log p(text | sport) + \log p(sport) - \log(p(text))$$

$$\log p(\overline{sport} | text) = \log \frac{p(text | \overline{sport}) \cdot p(\overline{sport})}{p(text)} = \log p(text | \overline{sport}) + \log p(\overline{sport}) - \log(p(text))$$

Обычно нам интересно отношение вероятностей (odds) или логарифм отношения вероятностей: log-odds

$$\log \frac{p(sport | text)}{p(\overline{sport} | text)} = \log \frac{p(sport)}{p(\overline{sport})} + \log \frac{p(text | sport)}{p(text | \overline{sport})}$$

Апостериорный log-odds

Априорный log-odds

Изменение log-odds
за счет нашего наблюдения

За счет этой формулы и допущения о том, что у нас все слова появляются независимо, мы можем наблюдать текст по частям.

Наивный Байес

За счет этой формулы и допущения о том, что у нас все слова появляются независимо, мы можем наблюдать текст по частям.

Первая часть текста

$$\log \frac{p(sport)_1}{p(\overline{sport})_1} = \log \frac{p(sport | text_1)}{p(\overline{sport} | text_1)} = \log \frac{p(sport)}{p(\overline{sport})} + \log \frac{p(text_1 | sport)}{p(text_1 | \overline{sport})}$$

Апостериорный log-odds

Априорный log-odds

Изменение log-odds
за счет нашего наблюдения

Вторая часть текста

$$\log \frac{p(sport | text_2)}{p(\overline{sport} | text_2)} = \log \frac{p(sport)_1}{p(\overline{sport})_1} + \log \frac{p(text_2 | sport)}{p(text_2 | \overline{sport})}$$

Апостериорный log-odds

Априорный log-odds,
Просто подставляем
апостериорные
вероятности,
подсчитанные на
основе первой части

Изменение log-odds
за счет нашего наблюдения

Наивный Байес

$$\begin{aligned}\log \frac{p(sport)_1}{p(\overline{sport})_1} &= \log \frac{p(sport | text_1)}{p(\overline{sport} | text_1)} = \log \frac{p(sport)}{p(\overline{sport})} + \log \frac{p(text_1 | sport)}{p(text_1 | \overline{sport})} \\&= \log \frac{p(sport)}{p(\overline{sport})} + \log \prod_{word \in text} \frac{p(word | sport)}{p(word | \overline{sport})} = \\&= \log \frac{p(sport)}{p(\overline{sport})} + \sum_{word \in text} \log \frac{p(word | sport)}{p(word | \overline{sport})}\end{aligned}$$

Априорный log-odds

Каждое слово влияет на наш апостериорный log-odds независимо. Можем вообще процессировать текст по одному слову за раз. Просто добавляя соответствующее отношение логарифмов

Наивный Байес

$$\begin{aligned}\log \frac{p(sport)_1}{p(\overline{sport})_1} &= \log \frac{p(sport | text_1)}{p(\overline{sport} | text_1)} = \log \frac{p(sport)}{p(\overline{sport})} + \log \frac{p(text_1 | sport)}{p(text_1 | \overline{sport})} \\&= \log \frac{p(sport)}{p(\overline{sport})} + \log \prod_{word \in text} \frac{p(word | sport)}{p(word | \overline{sport})} = \\&= \log \frac{p(sport)}{p(\overline{sport})} + \sum_{word \in text} \log \frac{p(word | sport)}{p(word | \overline{sport})}\end{aligned}$$

Априорный log-odds

Каждое слово влияет на наш апостериорный log-odds независимо. Можем вообще процессировать текст по одному слову за раз.

Наивный Байес

$$\log \frac{p(sport)_1}{p(\overline{sport})_1} = \log \frac{p(sport | text_1)}{p(\overline{sport} | text_1)} = \log \frac{p(sport)}{p(\overline{sport})} + \log \frac{p(text_1 | sport)}{p(text_1 | \overline{sport})}$$

$$= \log \frac{p(sport)}{p(\overline{sport})} + \log \prod_{word \in text} \frac{p(word | sport)}{p(word | \overline{sport})} =$$

$$= \log \frac{p(sport)}{p(\overline{sport})} + \sum_{word \in text} \log \frac{p(word | sport)}{p(word | \overline{sport})}$$

Пусть у нас встречается в текстах (в принципе) n различных слов (vocabulary)

$$= \log \frac{p(sport)}{p(\overline{sport})} + \sum_{word \in vocabulary} k_{word} \cdot \log \left(\frac{p(word | sport)}{p(word | \overline{sport})} \right)$$

Априорный log-odds

Каждое слово влияет на наш апостериорный log-odds независимо. Даже если слова одинаковые. Можем просто умножать log-odds каждого слова из словаря на то, сколько раз оно встретилось в тексте.

Наивный Байес

Пусть у нас встречается в текстах (в принципе) n различных слов (vocabulary)

$$= \log \frac{p(sport)}{p(\overline{sport})} + \sum_{word \in vocabulary} k_{word} \cdot \log \left(\frac{p(word | sport)}{p(word | \overline{sport})} \right)$$

Априорный log-odds

Каждое слово влияет на наш апостериорный log-odds независимо. Даже если слова одинаковые. Можем просто умножать log-odds каждого слова из словаря на то, сколько раз оно встретилось в тексте.

$$= a + w^T x$$

a - свободный параметр

x - описание нашего текста в виде (число word_1, число word_2) и т.д

w - вектор log-ods для слов из словаря.

$w^T x$ - скалярное произведение

Наивный Байес

Пусть у нас встречается в текстах (в принципе) n различных слов (vocabulary)

$$= \log \frac{p(sport)}{p(\overline{sport})} + \sum_{word \in vocabulary} k_{word} \cdot \log \left(\frac{p(word | sport)}{p(word | \overline{sport})} \right)$$

Априорный log-odds

Каждое слово влияет на наш апостериорный log-odds независимо. Даже если слова одинаковые. Можем просто умножать log-odds каждого слова из словаря на то, сколько раз оно встретилось в тексте.

$$= a + w^T x$$

Naive Bayes - частый случай generalized additive models
(просто для общего развития)

Наивный Байес

Можно ли использовать численные признаки в этом методе?

Наивный Байес

Можно ли использовать численные признаки в этом методе?

Да, мы можем придумать, как этот признак распределен для каждого класса. Часто предполагают, что $p(\text{численный признак} | y)$ распределен нормально. Тогда достаточно оценить из нашей выборки среднее и стандартное отклонение признака для объектов данного класса