Введение в Компьютерное Зрение
Лекция №7, осень 2020

Введение в линейные модели

ЛИТПМ

МФТИ

Кафедра технологий проектирования сложных технических систем

Линейные модели

<u>Линейная модель</u> - взвешенная сумма признаков и член смещения (*bias term*), который также называют свободным членом (intercept term)

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

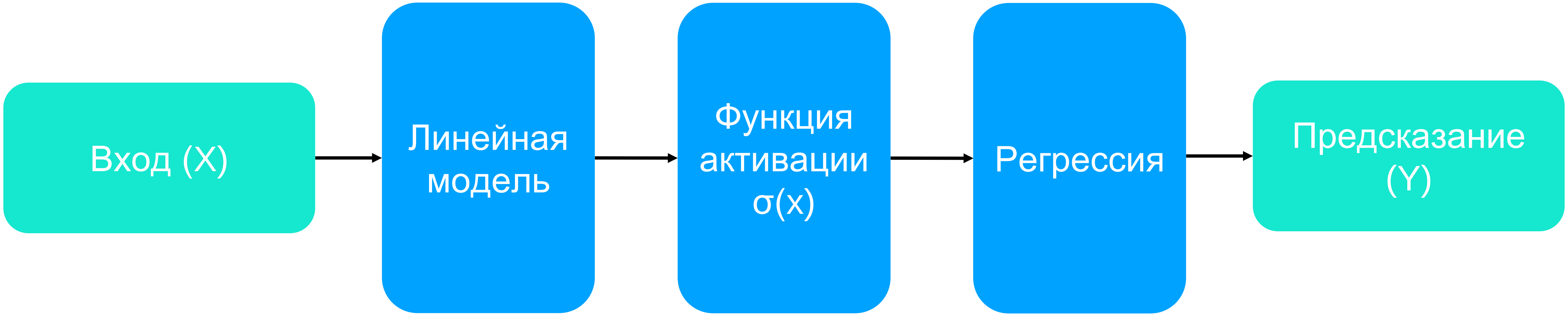
$\hat{y}$	Предсказываемое значение
n	Количество признаков
$x_1, x_2, \ldots, x_n$	Значения признаков
$\theta_0, \theta_1, \theta_2, \ldots, \theta_n$	Веса признаков и свободный член

Векторная форма:

$$\hat{y} = h_\theta(x) = \theta^T \cdot x$$

θ	Вектор весов и свободный член
x	Вектор значений признаков примера, где $x_0 = 0$

Линейные модели



Вообще-то, это уже нейронная сеть. Вы еще встретитесь с этим.

Метод наименьших квадратов

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

Средняя квадратичная ошибка (Mean Squared Error, MSE):

$$MSE = \frac{1}{l}\sum_{i=1}^{l}(f(x_i) - y_i)^2$$

$$MSE = \frac{1}{l}\sum_{i=1}^{l}(\hat{y}^{(i)} - y_i)^2 \rightarrow \frac{1}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y_i)^2 \rightarrow min$$

$$\frac{1}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y_i)^2 = \frac{1}{l}\|(Y - X\theta)\| = \frac{1}{l}(Y - X\theta)^T * (Y - X\theta)$$

Метод наименьших квадратов

$$\frac{1}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y_i)^2 = \frac{1}{l}\|(Y - X\theta)\| = \frac{1}{l}(Y - X\theta)^T * (Y - X\theta)$$

$$\frac{d}{d\theta}\left[\frac{1}{l}(Y - X\theta)^T * (Y - X\theta)\right] = \frac{1}{l}\frac{d}{d\theta}[Y^T Y - 2Y^T X\theta + \theta^T X^T X\theta] = -2X^T Y + 2X^T X\theta = 0$$

$$-2X^T Y + 2X^T X\theta = 0$$

$$2X^T Y = 2X^T X\theta$$

$$(X^T X)^{-1} X^T Y = \theta$$

Аналитический способ поиска оптимальных весов (нормальное уравнение):

$$\hat{\theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot y$$

BLUE (Best Linear Unbiased Estimator)

X Матрица объектов признаков

y Вектор целевой переменной

$\hat{\theta}$ Оптимальный вектор весов, который сводит к минимуму MSE

Метод наименьших квадратов

Теорема Гаусса — Маркова

оценки метода наименьших квадратов оптимальны в классе линейных несмещённых оценок

Условия для парной регрессии:

1. *модель данных правильно специфицирована. Нет лишних переменных, и учтены все важные* $Y = \theta_0 + \theta \cdot X + \epsilon$
2. *все X* детерминированы и не все равны между собой. Иными словами, переменные не должны быть постоянными.
3. Ошибки не носят систематического характера, то есть $E(\epsilon_i) = 0 \forall i$
4. Дисперсия ошибок одинакова (гомоскедастичность) и равна некоторой $\sigma^2 = const$
5. Ошибки некоррелированы, то есть $cov(\epsilon_i, \epsilon_j) = 0 \forall i, j$

Условия для Множественной регрессии:

1. *модель данных правильно специфицирована. Нет лишних переменных, и учтены все важные*
2. $rang(X) = \mathrm{n}$
3. $E(\epsilon_i) = 0 \forall i$
4. $cov(\epsilon_i, \epsilon_j) = 0 \forall i, j$

Градиентный спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}(\hat{y}^{(i)} - y_i)^2 \rightarrow \frac{1}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y_i)^2 \rightarrow min$$

Градиентый спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}(\hat{y}^{(i)} - y_i)^2 \rightarrow \frac{1}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y_i)^2 \rightarrow min$$

Градиент - Вектор указывающий направление наибольшего возрастания функции, компоненты которого равны частным производным по всем её аргументам.

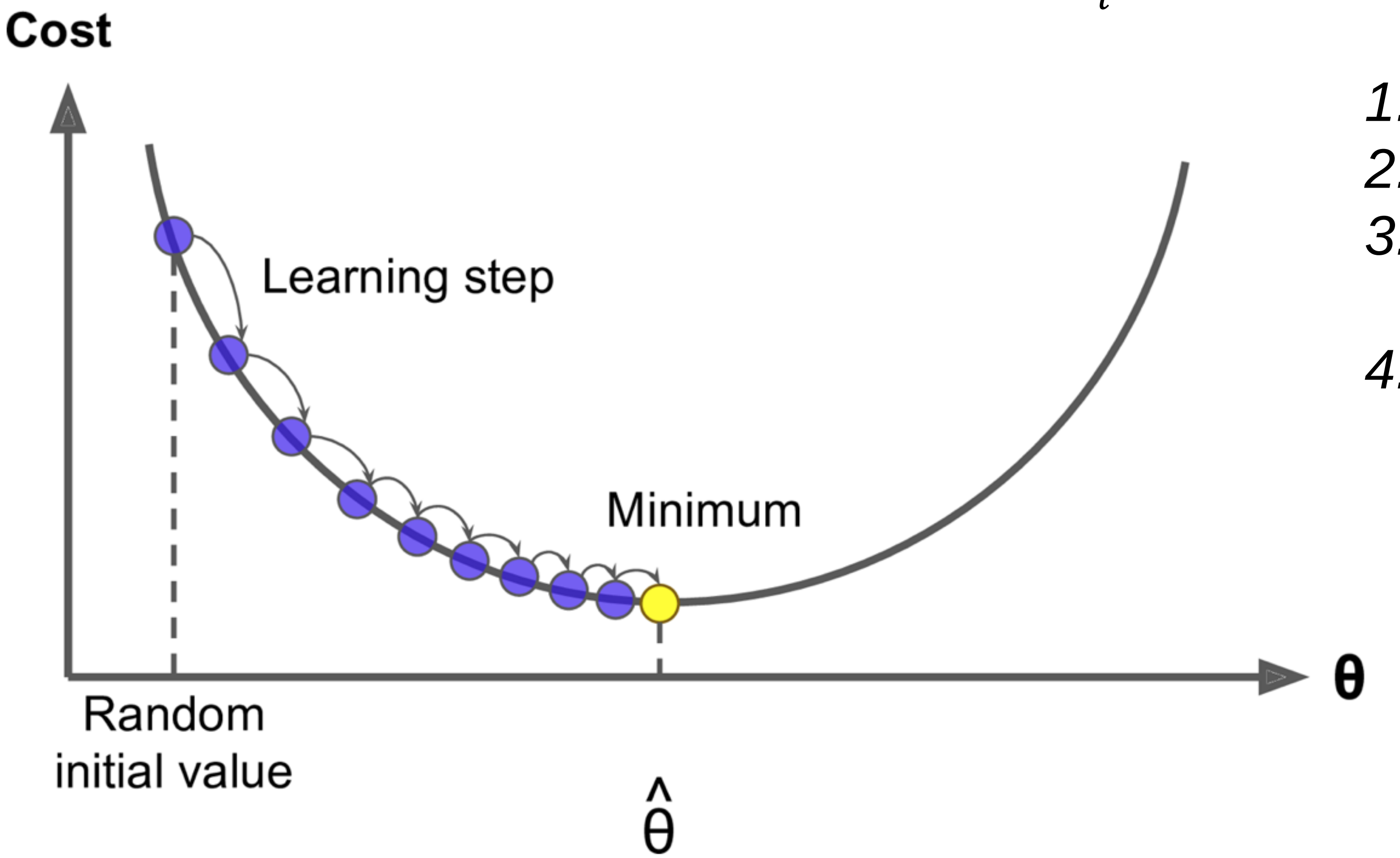
$$\nabla\varphi = (\frac{\partial\varphi}{\partial x_1}, \frac{\partial\varphi}{\partial x_2}, \dots, \frac{\partial\varphi}{\partial x_n},)$$

Градиентый спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}(\hat{y}^{(i)} - y_i)^2 \to \frac{1}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y_i)^2 \to min$$

Градиент - Вектор указывающий направление наибольшего возрастания функции, компоненты которого равны частным производным по всем её аргументам.

$$\nabla MSE(\theta) = \frac{2}{l}X^T \cdot (X \cdot \theta - y)$$

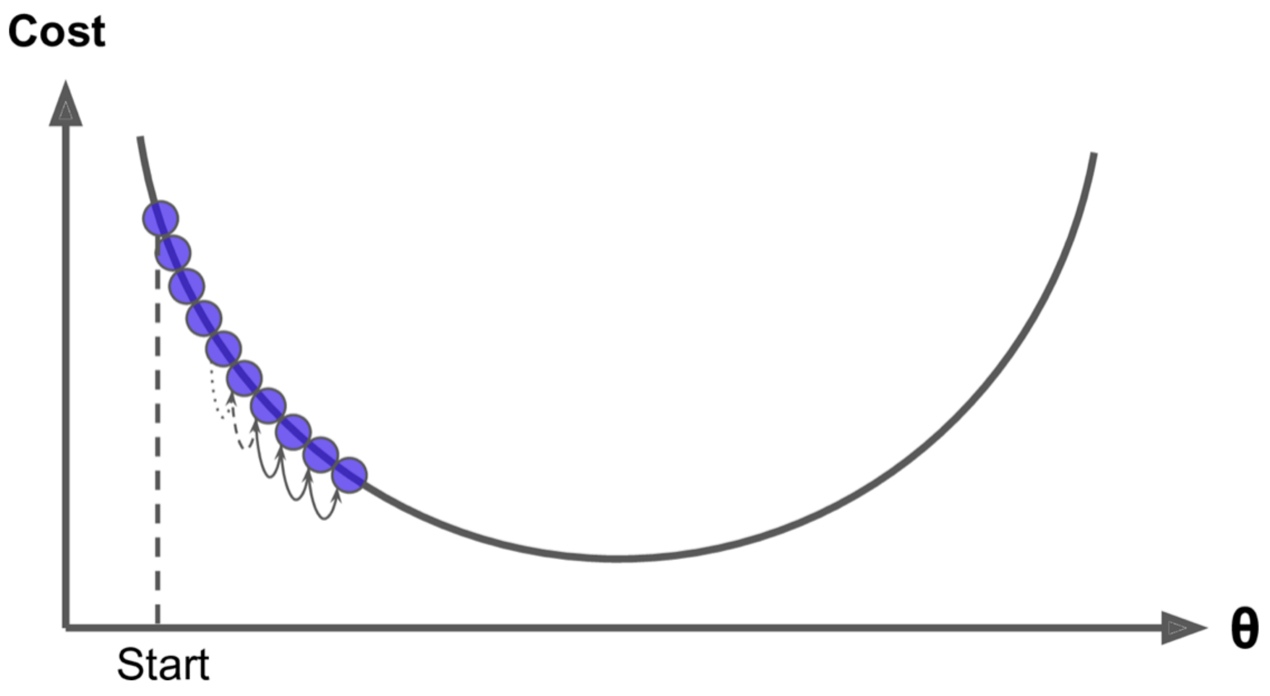


1. *Случайно задаем веса*
2. *Считаем градиент в точке*
3. *Изменяем веса путем вычитания градиента*
4. *Повторяем п.2*

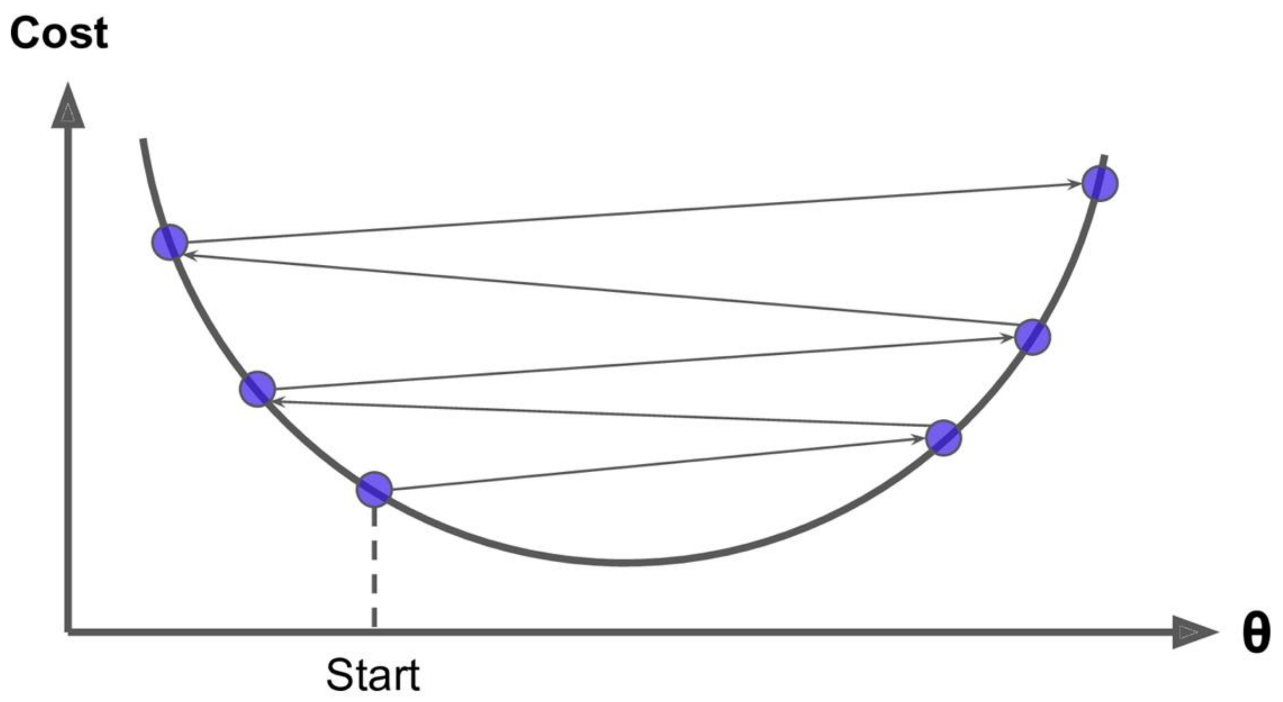
** Мы можем контролировать скорость обучения (learning rate) умножая градиент на шаг обучения*

Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

Выбор шага обучения градиентного спуска



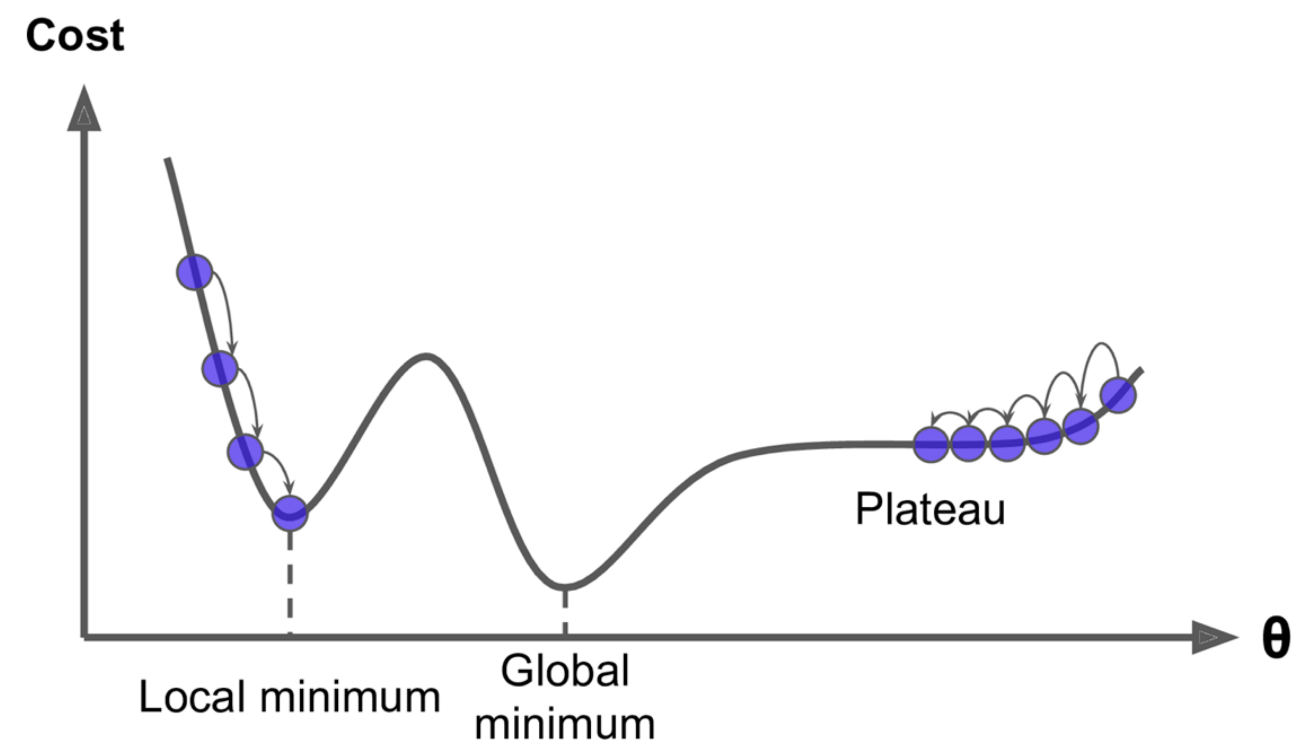
Слишком маленький шаг обучения
рискуем не дойти до минимума



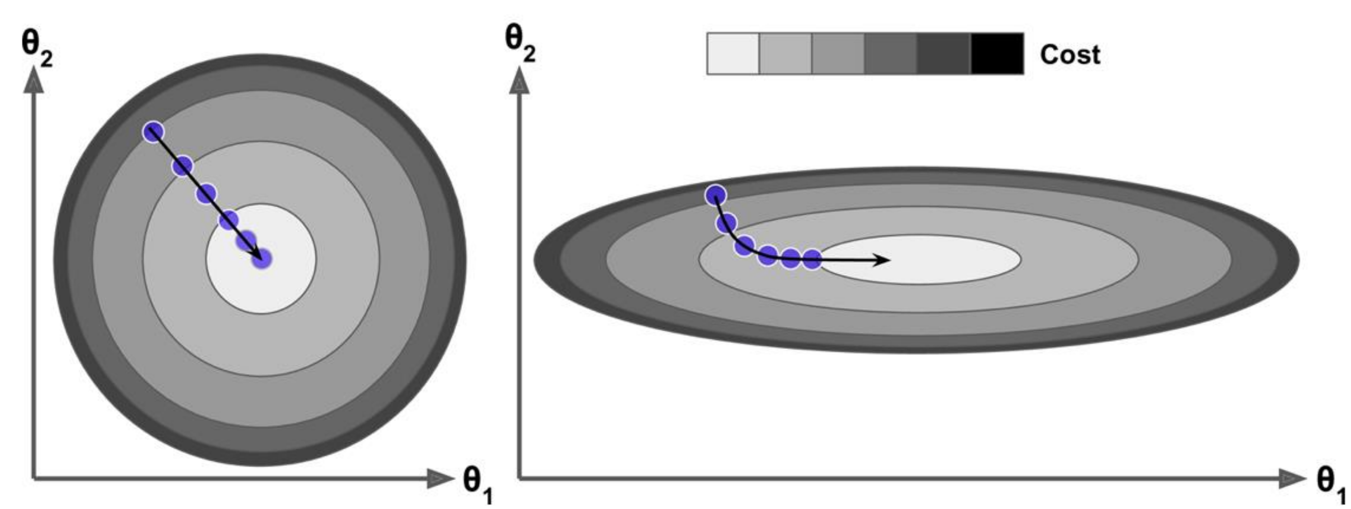
Слишком большой шаг обучения
рискуем проскочить минимум

Другие проблемы градиентного спуска

Не все функции потерь имеют форму чаши (параболоид)



Важно масштабировать данные



Стохастический градиентный спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}(\hat{y}^{(i)} - y_i)^2 \rightarrow \frac{1}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y_i)^2 \rightarrow min$$

$$\frac{\partial MSE(\theta)}{\partial \theta_j} = \frac{2}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

Стохастический градиентный спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}(\hat{y}^{(i)} - y_i)^2 \to \frac{1}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y_i)^2 \to min$$

Пакетный градиентный спуск

$$\frac{\partial MSE(\theta)}{\partial \theta_j} = \frac{2}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

$$\frac{\partial MSE(\theta)}{\partial \theta_j} = (\theta^T \cdot x^{(i)} - y^{(i)}) \cdot x^{(i)}$$

$$\nabla MSE(\theta) = x_i^T \cdot (x_i \cdot \theta - y)$$

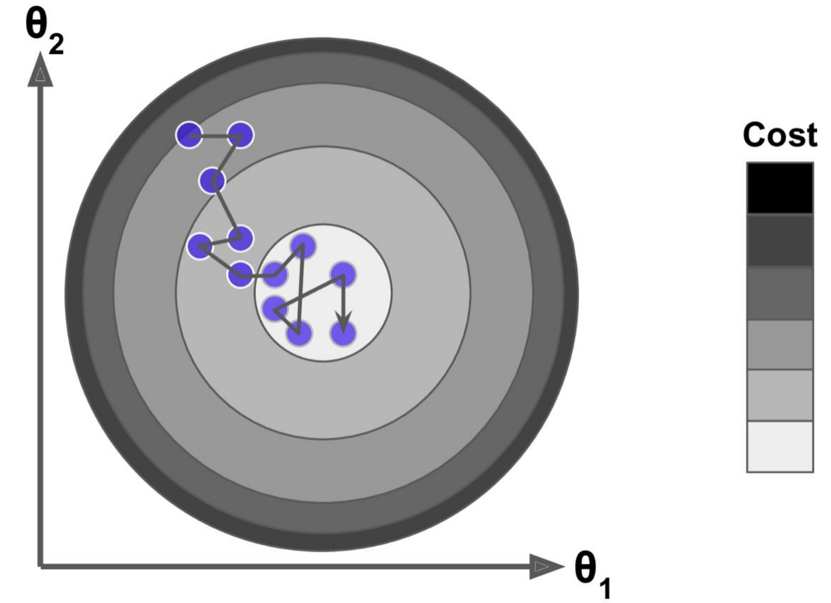
Стохастический градиентный спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}(\hat{y}^{(i)} - y_i)^2 \to \frac{1}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y_i)^2 \to min$$

$$\frac{\partial MSE(\theta)}{\partial \theta_j} = \frac{2}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

$$\frac{\partial MSE(\theta)}{\partial \theta_j} = (\theta^T \cdot x^{(i)} - y^{(i)}) \cdot x^{(i)}$$

$$\nabla MSE(\theta) = x_i^T \cdot (x_i \cdot \theta - y)$$

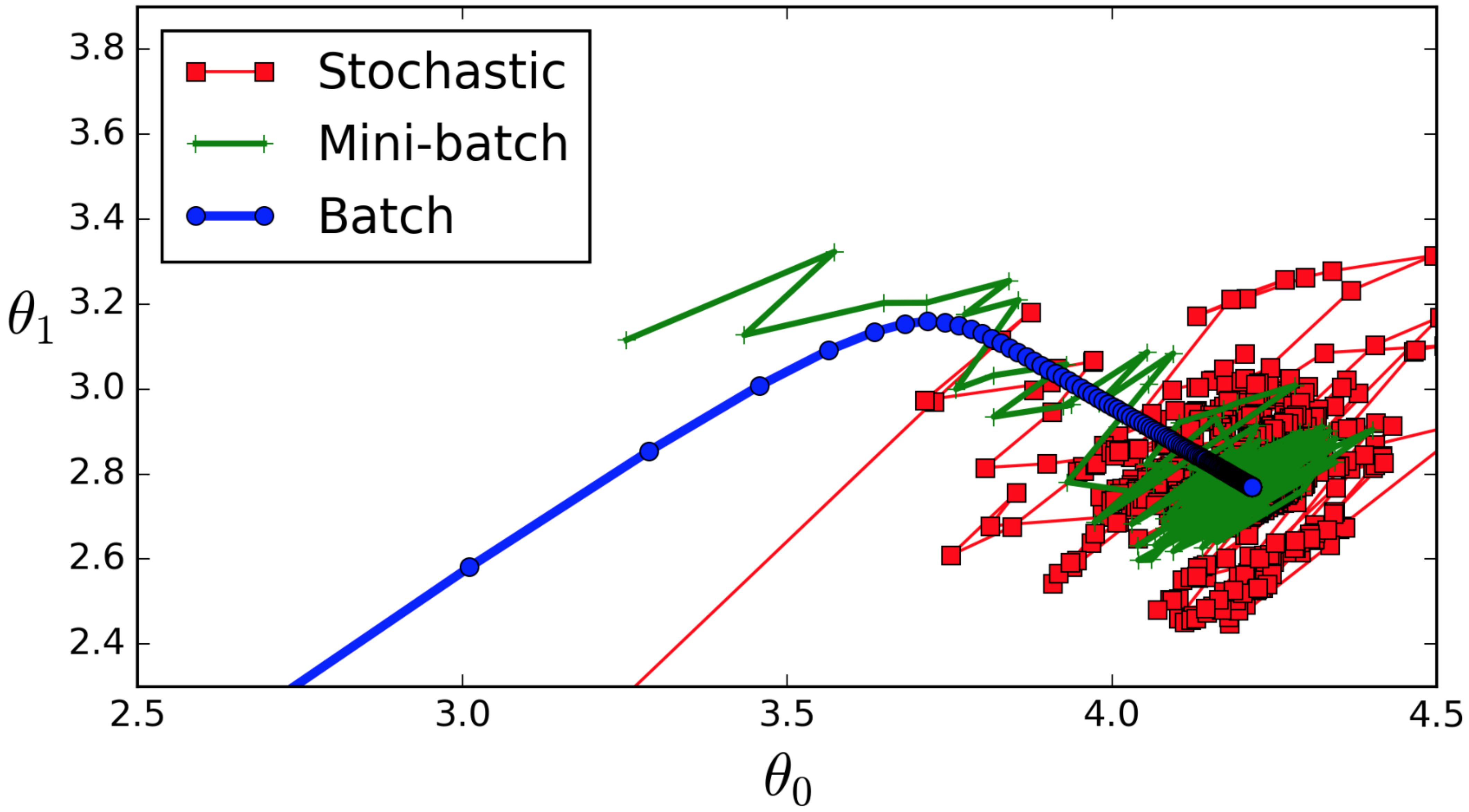


Случайно выбираем объект, и двигаемся к минимуму.
Каждый объект выборки может прогоняться несколько раз или быть не выбран вообще

Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

Выбор метода градиентного спуска

Можно скрестить два похода: Стохастический и пакетный и выбирать случайные подборки например из 100 объектов, тогда получится: метод называемый Mini-batch



Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

Проблемы нормального уравнения

$$\hat{\theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot y$$

$(X^T \cdot X)$ *– матрица размером n x n, где n – количество признаков*

Вычислительная сложность для обратной матрицы: O(n^3)

Регуляризация линейных моделей

Одно из условий Гауса-Маркова: $rang(X) = \text{n}$

Если оно не выполняется, то решение МНК $\hat{\theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot y$ не существует

так как Матрица $X^T \cdot X$ сингулярна (вырождена)

Регуляризация линейных моделей

Одно из условий Гауса-Маркова: $rang(X) = \text{n}$

Если оно не выполняется, то решение МНК $\hat{\theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot \text{Y}$ не существует

так как Матрица $X^T \cdot X$ сингулярна (вырождена)

Чтобы сделать матрицу невырожденной, мы можем добавить диагональные элементы:

$$\hat{\theta} = (X^T \cdot X + \lambda I)^{-1} \cdot X^T \cdot Y$$

Где $I = diag(1 \dots 1)$

В общем случае - это решение следующей задачи минимизации:

$$\text{Q} = \|(Y - X\theta)\|^{2} + \lambda^2\|\theta\|^{2}$$

Регуляризация линейных моделей

L2 - регуляризация, Ridge

$$MSE(\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2 \rightarrow min$$

L1 - регуляризация, Lasso (Least absolute shrinkage and selection operator)

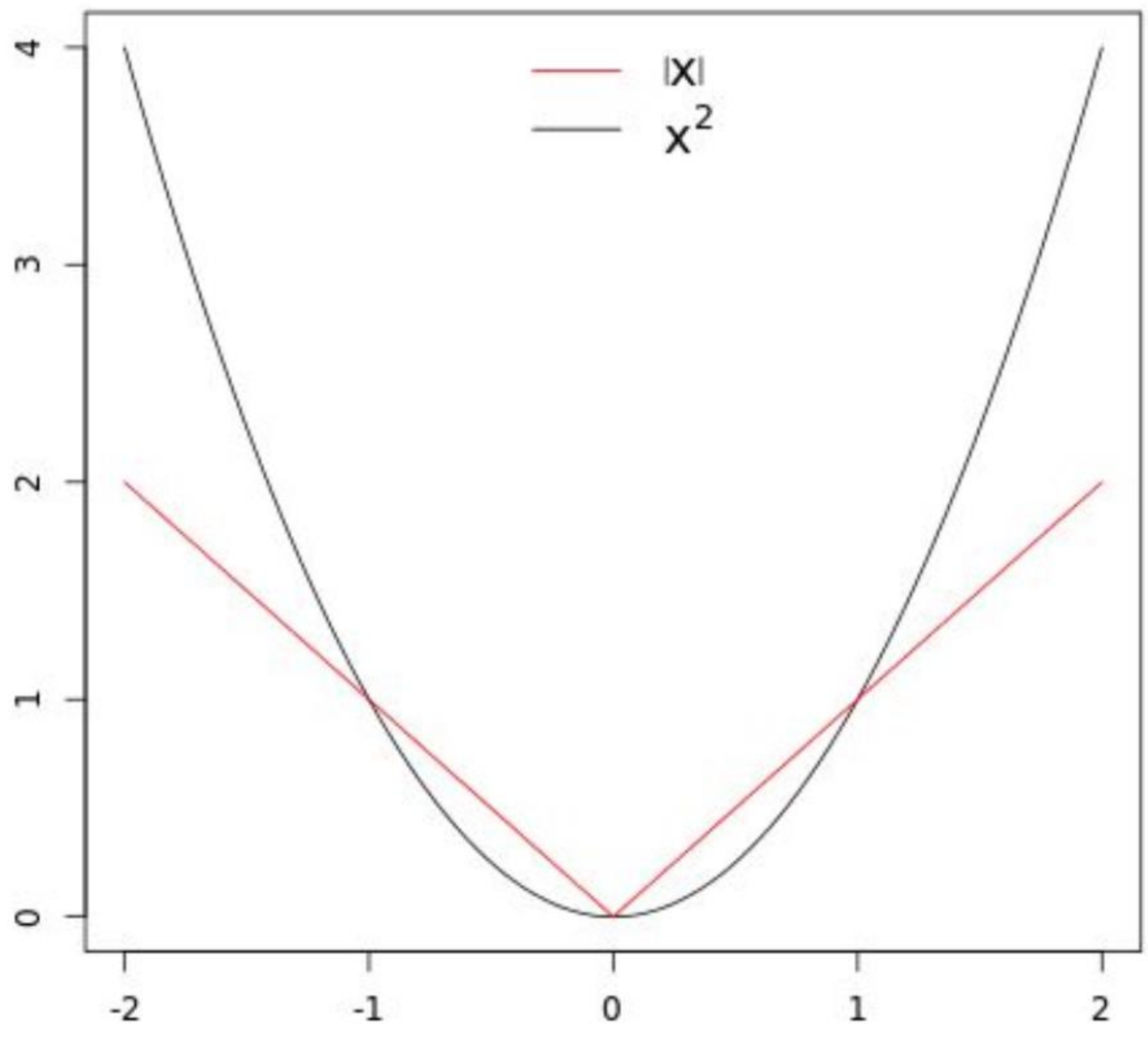
$$MSE(\theta) + \alpha \sum_{i=1}^{n} |\theta_i| \rightarrow min$$

введение ограничений на норму вектора коэффициентов модели приводит к обращению в 0 некоторых коэффициентов модели.

ElasticNet

$$MSE(\theta) + \alpha r \sum_{i=1}^{n} |\theta_i| + \alpha \frac{1-r}{2} \sum_{i=1}^{n} \theta_i^2 \rightarrow min$$

Регуляризация линейных моделей



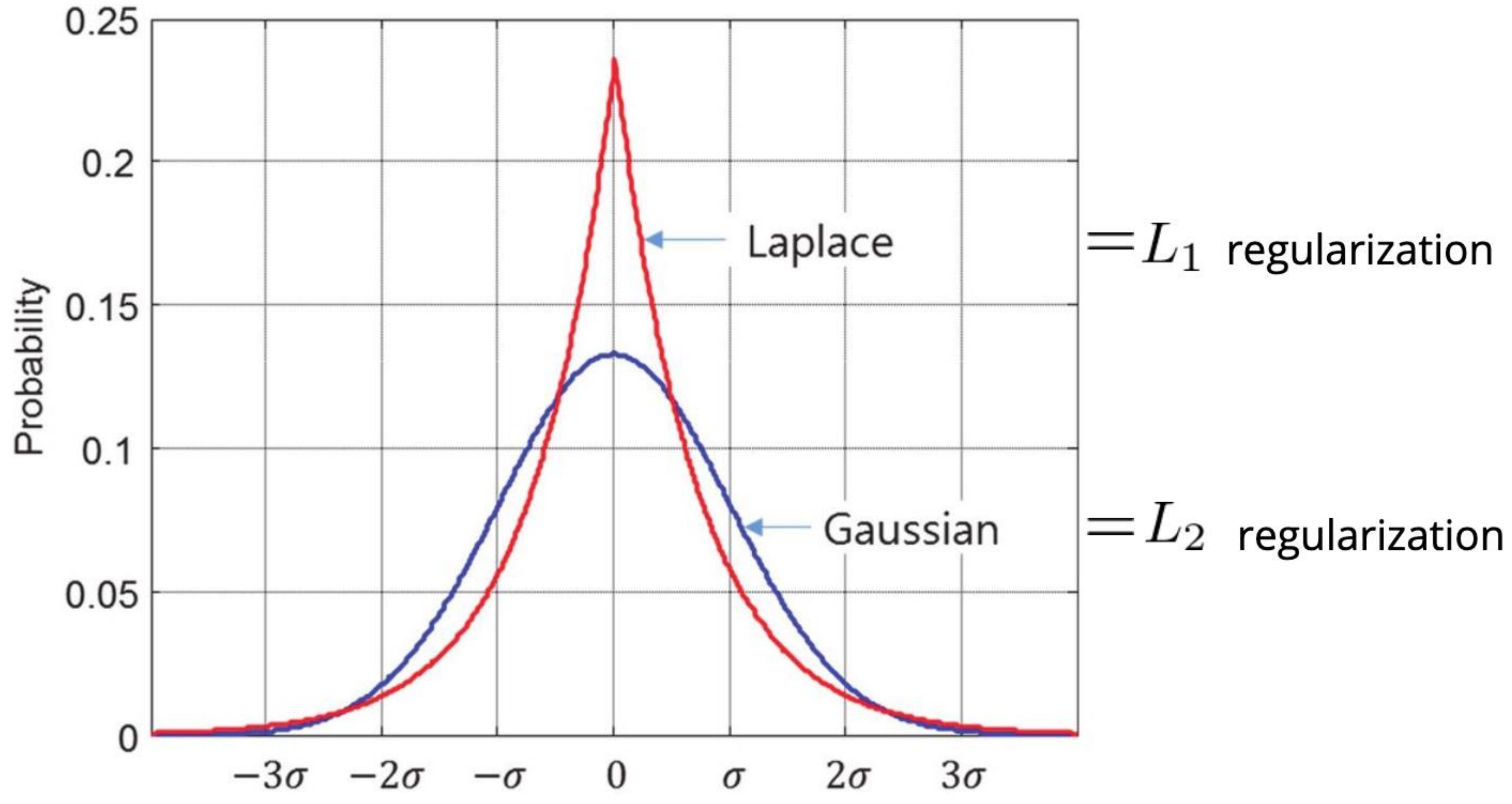
MSE
- дифференцируемая
- чувствительна к шуму
- BLUE

MAE
- отбираем фичи
- не дифференцируемая (вообще-то дифференцируема)
- устойчивее к шуму

L2
- сильнее ограничивает веса = большая обобщающая способность = более стабильна
- Дифференцируемая

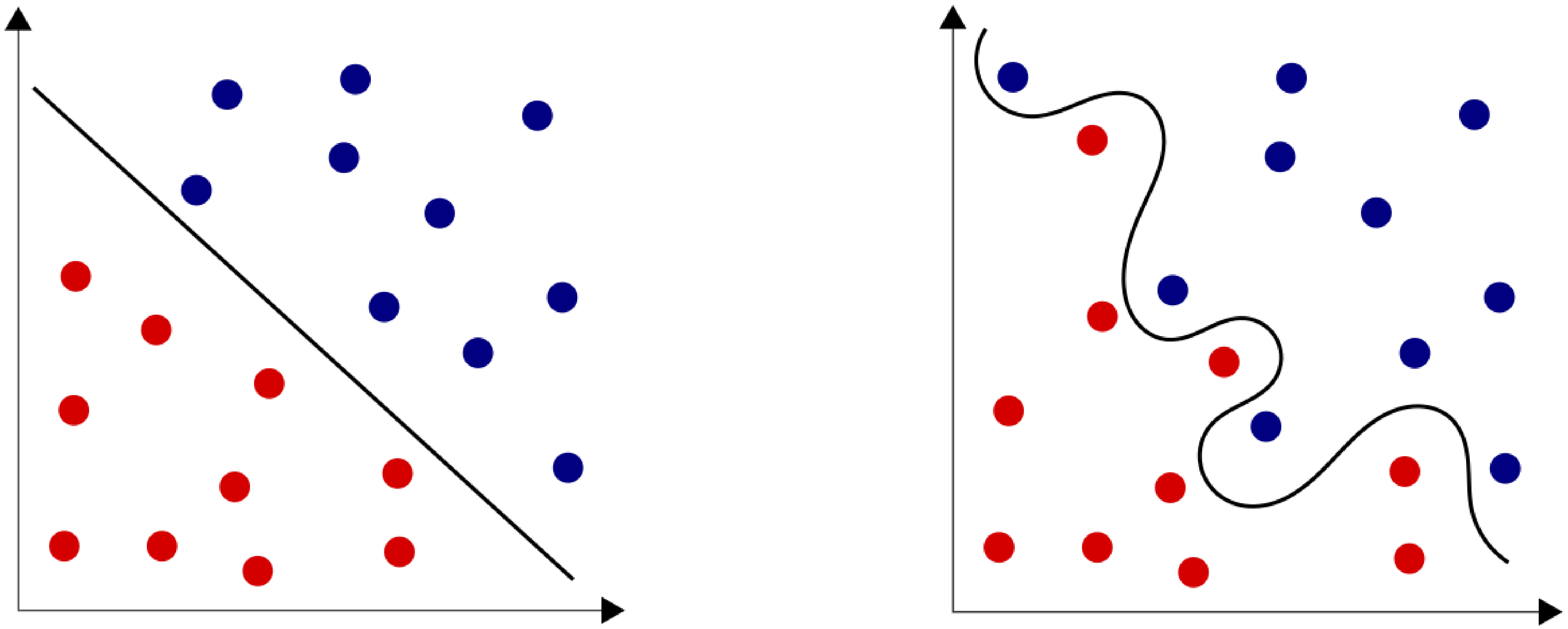
L1
- отбираем фичи
- не дифференцируемая (вообще-то дифференцируема)

Регуляризация линейных моделей



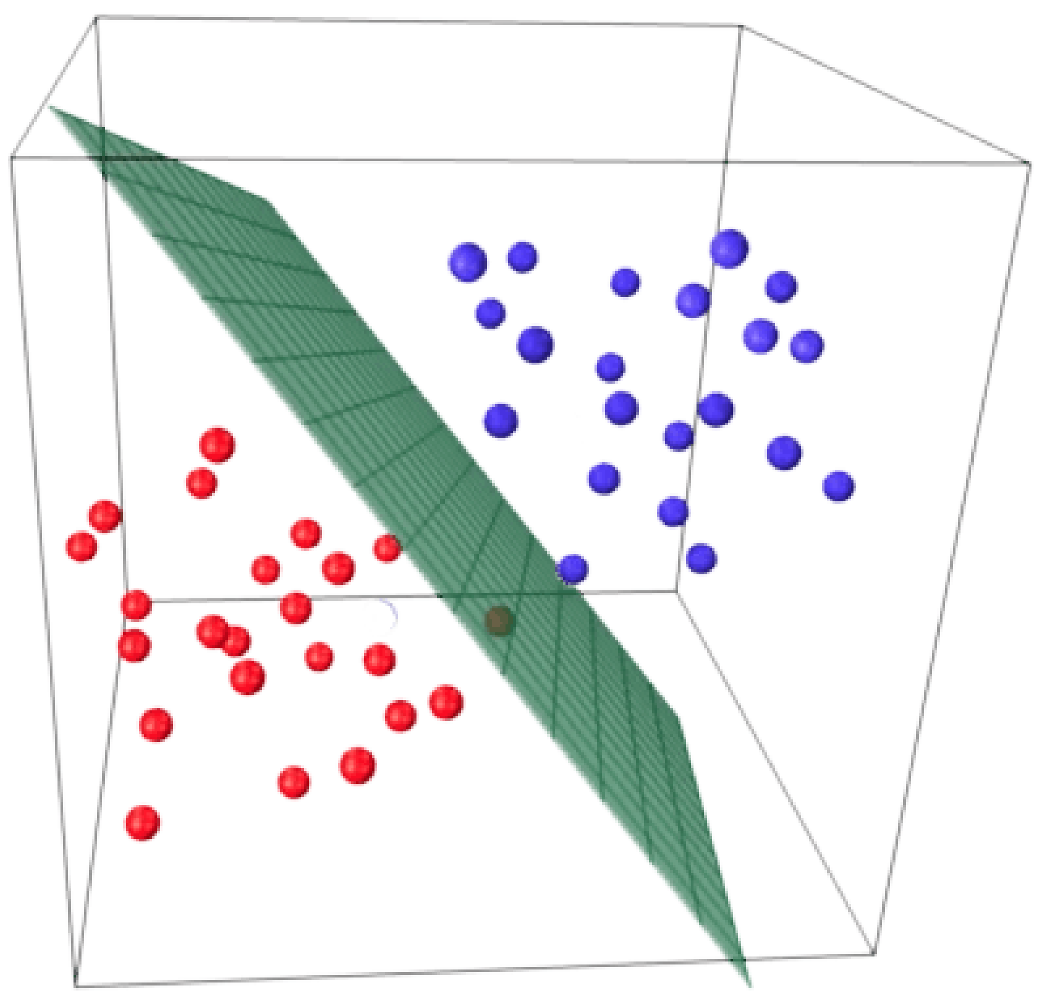
Линейные модели в задаче классификации

Основная идея:
Предполагаем, что существует такая гиперплоскость, которая делит пространство на два полупространства в каждом из которых одно из двух значений целевого класса.



Если существует гиперплоскость которой можно разделить пространство на два класса без ошибок, то обучающая выборка называется *линейно разделимой*

Линейные модели в задаче классификации



Дана обучающая выборка:

$$X_l = \{ (x_1 , y_1), \dots , (x_l , y_l) \}$$

Для задачи классификации - Целевая переменная задана конечным числом меток

$$(x_1 , y_1) \in \mathbb{R}^m \times \mathbb{Y},\ \mathbb{Y} = \{-1; 1\}$$

Простейший классификатор:

$$a(x) = sign(\langle w, x\rangle + x_0) = sign(\vec{w}^T \cdot x)$$

$\vec{w}$ – нормаль гиперплоскости

$\vec{w}^T \cdot x_i$ – расстояние от гиперплоскости до x_i, знак показывает отношение к классу

Линейные модели в задаче классификации

$$a(x) = sign(\langle w, x \rangle + x_0) = sign(\vec{w}^T \cdot x)$$

доля правильных ответов (accuracy):

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} [a(x_i) = y_i] \rightarrow max$$

Линейные модели в задаче классификации

$$a(x) = sign(\langle w, x\rangle + x_0) = sign(\vec{w}^T \cdot x)$$

доля правильных ответов (accuracy):

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) = y_i] \to max$$

доля неправильных ответов:

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) \neq y_i] = \frac{1}{l}\sum_{i=1}^{l}[sign(\langle w, x_i\rangle) \neq y_i] \to min$$

<u>Проблемы:</u>

1. *Функционал дискретный относительно весов ⇒ мы не сможем искать минимум с помощью градиентных методов.*
2. *Функционал может иметь несколько глобальных минимумов ⇒ может быть много способов добиться оптимального количества ошибок.*

Линейные модели в задаче классификации

$$a(x) = sign(\langle w, x\rangle + x_0) = sign(\vec{w}^T \cdot x)$$

доля правильных ответов (accuracy):

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) = y_i] \to max$$

доля неправильных ответов:

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) \neq y_i] = \frac{1}{l}\sum_{i=1}^{l}[sign(\langle w, x_i\rangle) \neq y_i] \to min$$

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}[\underbrace{y_i\langle w, x_i\rangle}_{M_i} < 0] \to min$$

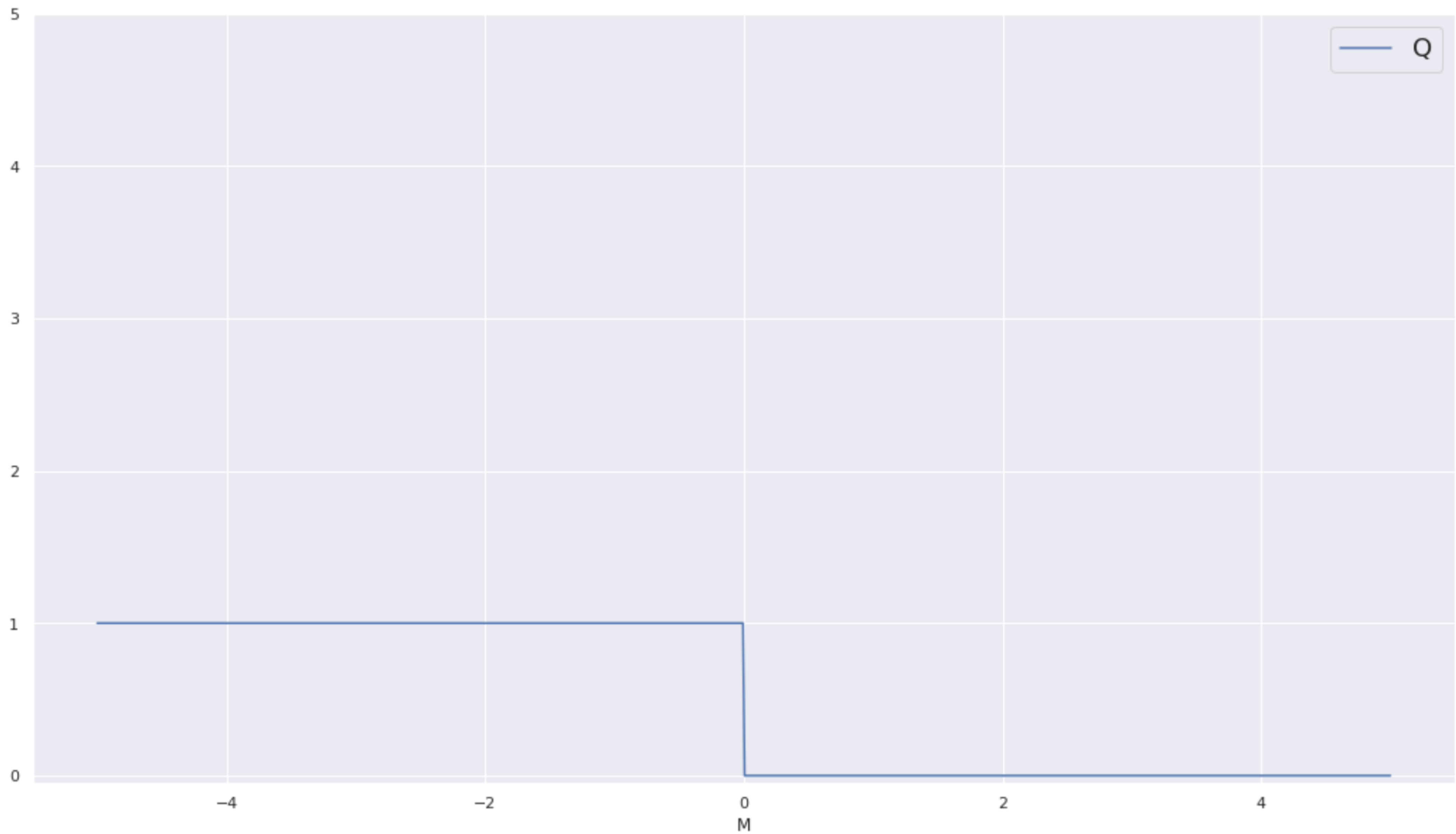
$M_i = y_i\langle w, x_i\rangle$ – отступ (margin)

Знак отступа говорит о корректности ответа классификатора (положительный отступ соответствует правильному ответу, отрицательный неправильному)
абсолютная величина M – характеризует степень уверенности классификатора в своём ответе.

Линейные модели в задаче классификации

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}[\underbrace{y_i\langle w, x_i\rangle}_{M_i} < 0] \rightarrow min$$

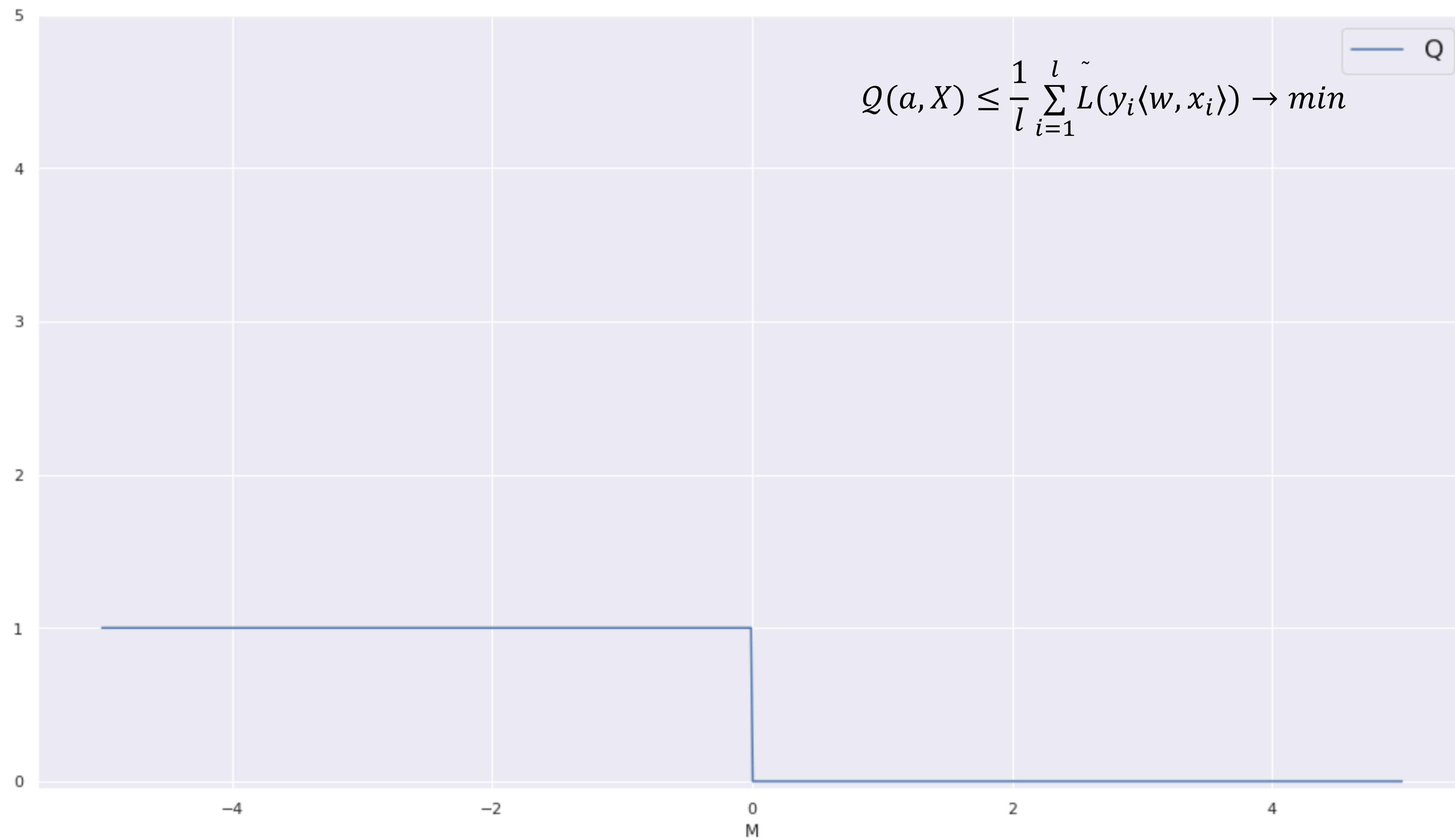
$L(M) = [M < 0]$ *– пороговая функции потерь*



Линейные модели в задаче классификации

$$Q(a,X) = \frac{1}{l}\sum_{i=1}^{l}[\underbrace{y_i\langle w,x_i\rangle}_{M_i} < 0] = \frac{1}{l}\sum_{i=1}^{l}[M_i(w) < 0] \le \breve{Q}(w) = \frac{1}{l}\sum_{i=1}^{l} L(M_i(w)) \to min$$

Пороговая функции потерь *Эмпирический риск* *Функция потерь (верхняя оценка)*



$$Q(a,X) \le \frac{1}{l}\sum_{i=1}^{l}\tilde{L}(y_i\langle w,x_i\rangle) \to min$$

Линейные модели в задаче классификации

$\tilde{L}(M) = (1 - M)^2$

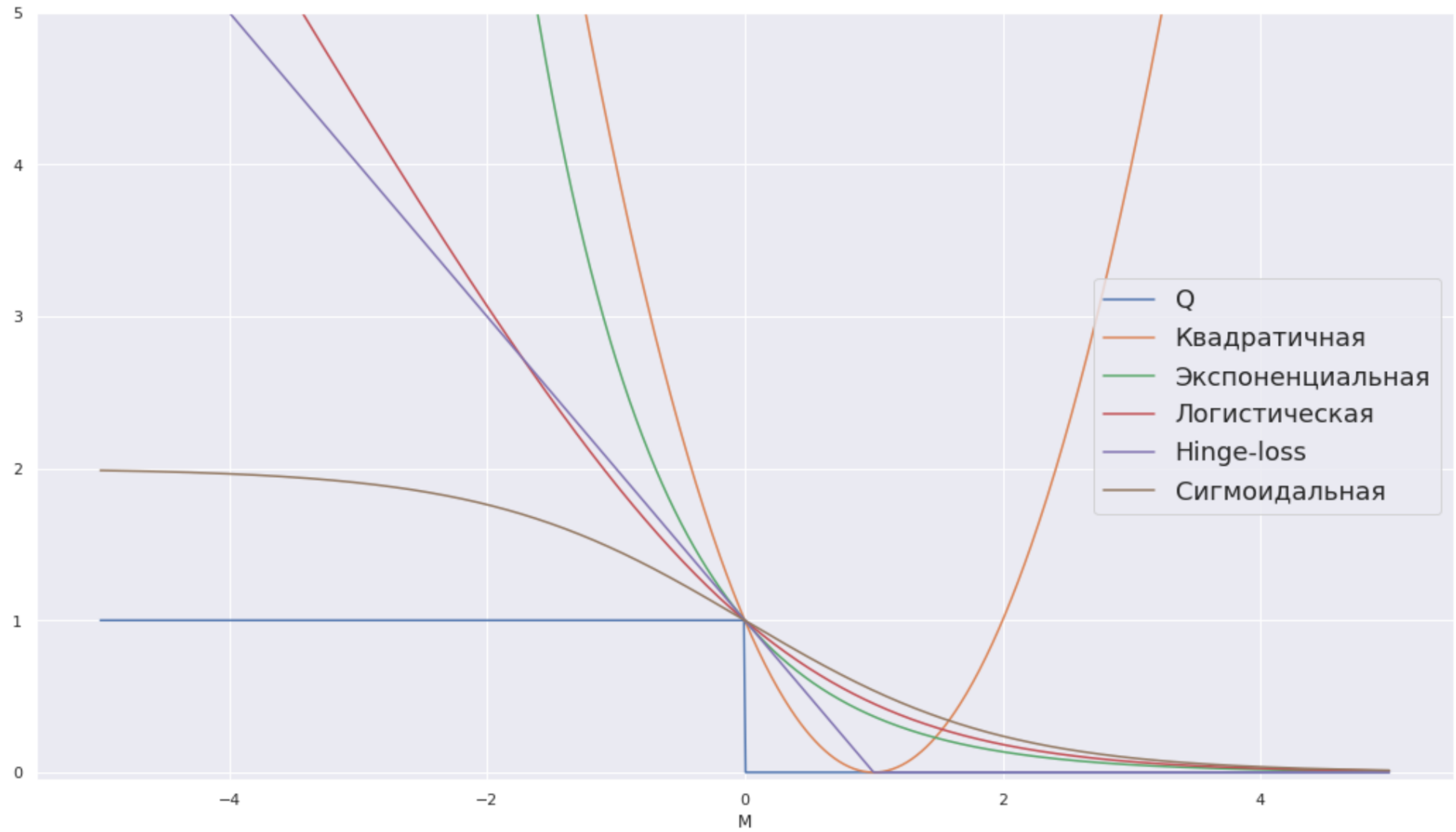
$\tilde{L}(M) = e^{-M}$

$\tilde{L}(M) = \log(1 + e^{-M})$

$\tilde{L}(M) = (1 - M)_+ = max(0{,}1 - M)$

$\tilde{L}(M) = \dfrac{2}{1 + e^{-M}}$

$L(M) = [M < 0]$ *– пороговая функции потерь*



Линейные модели в задаче классификации

$\tilde{L}(M) = (1 - M)^2$

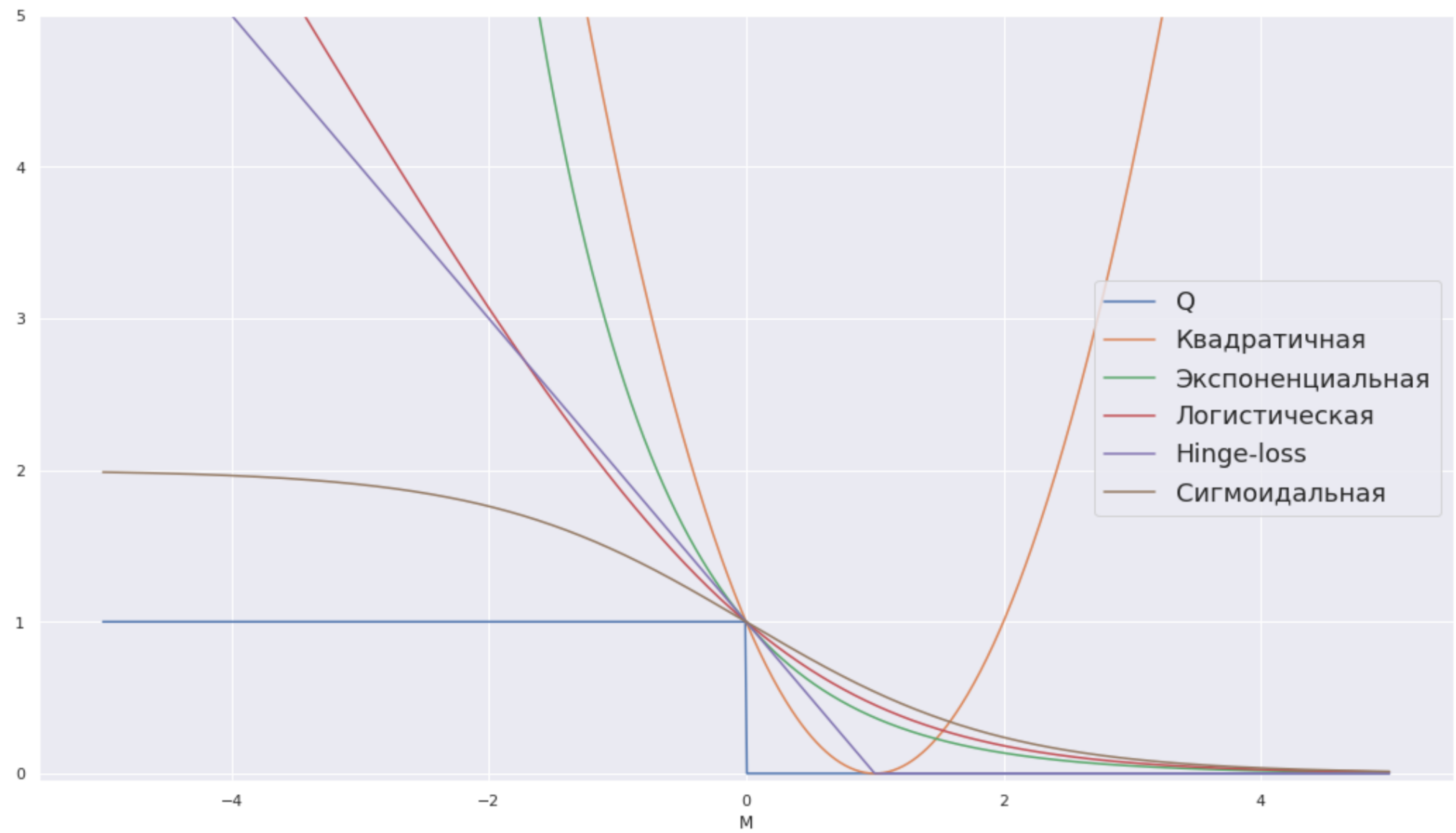
$\tilde{L}(M) = e^{-M}$

$\tilde{L}(M) = \log(1 + e^{-M})$

$\tilde{L}(M) = (1 - M)_+ = max(0,1 - M)$

$\tilde{L}(M) = \dfrac{2}{1 + e^{-M}}$

$L(M) = [M < 0]$ *– пороговая функции потерь*



Линейные модели в задаче классификации

$P(y_i = 1|x_i, w) = \sigma(w^T x)$

$P(y_i = -1|x_i, w) = \sigma(-w^T x)$

$P(y = y_i|x_i, w) = \sigma(y_i w^T x)$

Функция правдоподобия (вероятность наблюдать вектор *y* при заданных значениях *X* и *w*)

<u>Делаем предположение: объекты приходят независимо, из одного распределения</u>

$$P(\vec{y}|X, w) = \prod_{i=1}^{l} P(y = y_i|x_i, w) \to max$$

Линейные модели в задаче классификации

$P(y_i = 1|x_i, w) = \sigma(w^T x)$

$P(y_i = -1|x_i, w) = \sigma(-w^T x)$

$$P(y = y_i|x_i, w) = \sigma(y_i w^T x)$$

Правдоподобие (вероятность наблюдать вектор *y* при заданных значениях *X* и *w*)

Делаем предположение: объекты приходят независимо, из одного распределения

$$P(\vec{y}|X, w) = \prod_{i=1}^{l} P(y = y_i|x_i, w) \to max$$

Так как логарифм монотонно возрастающая функция, то оценка *w* максимизирующая логарифм, будет максимизировать и само правдоподобие

$$\log P(\vec{y}|X, w) = \sum_{i=1}^{l} \log\sigma(y_i w^T x) = \sum_{i=1}^{l} \log \frac{1}{1 + e^{-y_i\langle w, x_i\rangle}} = -\sum_{i=1}^{l} \log(1 + e^{-y_i\langle w, x_i\rangle})$$

$$-\log P(\vec{y}|X, w) = Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} \log(1 + e^{-y_i\langle w, x_i\rangle}) \to min$$

Максимизация правдоподобия = Минимизация эмпирического риска

Линейные модели в задаче классификации

$$P(\vec{y}|X,w) = \prod_{i=1}^{l} P(y = y_i|x_i, w)$$

$$P(\vec{y}|X,w) = P_{\text{ист}} * P_{pred} = \prod_{i=1}^{l} P_i * P_{pred\_i} = \prod_{i=1}^{l} P(y = y_i|x_i, w) = \prod_{i=1}^{l} p_i^{y_i}(1-p_i)^{(1-y_i)}$$

Формула Бернулли

$$P(\vec{y}|X,w) = \sum_{i=1}^{l} \log p_i^{y_i}(1-p_i)^{(1-y_i)} = \sum_{i=1}^{l} y_i \log p_i + (1-y_i)\log(1-p_i) \to max$$

$$Q(p,X) = \frac{1}{l}\sum_{i=1}^{l} - y_i \log p_i - (1-y_i)\log(1-p_i) \to min$$

Истинные значения   предсказанные

http://cs229.stanford.edu/notes/cs229-notes1.pdf

Линейные модели в задаче классификации

Шансы

$$odds = \frac{p}{1-p} \in [0; \infty]$$

$$\ln(odds) \in R$$

$$\vec{w}^T \cdot x \in R$$

где p вероятность, что событие состоится (в нашем случае, что класс примет значение 1):

Линейные модели в задаче классификации

Шансы

$$odds = \frac{p}{1-p} \in [0;\infty]$$

$$\ln(odds) \in R$$

$$\vec{w}^T \cdot x \in R$$

где p вероятность, что событие состоится (в нашем случае, что класс примет значение 1):

$$\ln(odds_+) = ln(\frac{p}{1-p}) = ln(p) - ln(1-p)$$

$$\ln(odds_-) = ln(\frac{1-p}{p}) = ln(1-p) - ln(p)$$

$$\ln(odds_+) = -\ln(odds_-) = w^T \cdot x$$

Линейные модели в задаче классификации

Шансы

$$odds = \frac{p}{1-p} \in [0; \infty] \qquad \ln(odds) \in R$$

$$\vec{w}^T \cdot x \in R$$

где p вероятность, что событие состоится (в нашем случае, что класс примет значение 1):

$$\ln(odds_+) = ln(\frac{p}{1-p}) = ln(p) - ln(1-p)$$

$$\ln(odds_-) = ln(\frac{1-p}{p}) = ln(1-p) - ln(p)$$

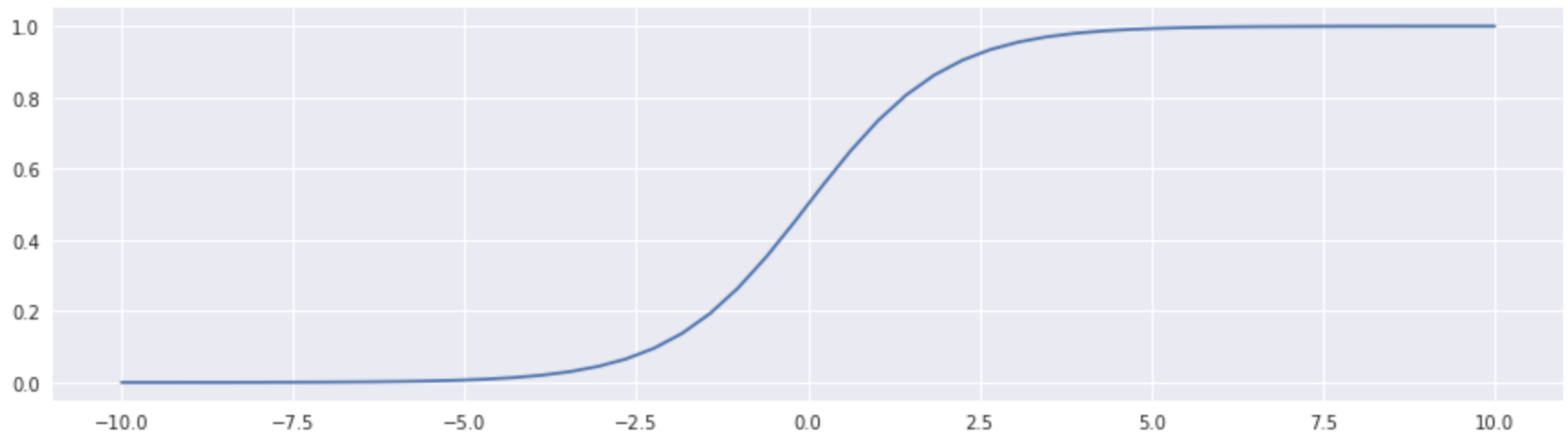
$$\ln(odds_+) = -\ln(odds_-) = w^T \cdot x$$

$$odds = e^{w^T x} \quad \Rightarrow \quad p = \frac{e^{w^T x}}{1 + e^{w^T x}}$$

Линейные модели в задаче классификации

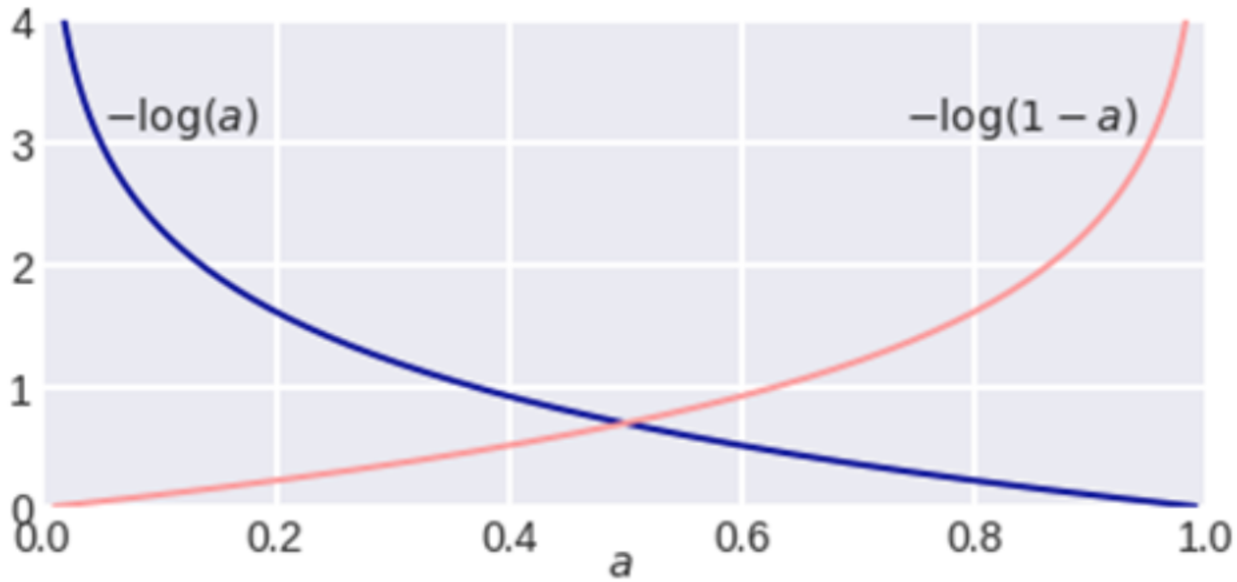
$$p = \frac{e^{w^T x}}{1 + e^{w^T x}} = \frac{1}{1 + e^{-w^T x}}$$

$$f(x_i, w) = \sigma(z) = \frac{1}{1 + e^{-z}} \in [0; 1]$$



Линейные модели в задаче классификации

$$-\begin{cases} \log a_i, & y_i = 1, \\ \log(1 - a_i), & y_i = 0. \end{cases}$$



Если для объекта 1го класса мы предсказываем нулевую вероятность принадлежности к этому классу или, наоборот, для объекта 0го – единичную вероятность принадлежности к классу 1, то ошибка равна бесконечности! Таким образом, грубая ошибка на одном объекте сразу делает алгоритм бесполезным.

$$Q(p,X) = \frac{1}{l}\sum_{i=1}^{l} -y_i \log p_i - (1-y_i)\log(1-p_i) = \frac{1}{l}\sum_{i=1}^{l} -y_i \log\left(\frac{1}{1+e^{-w^T x}}\right) - (1-y_i)\log\left(\frac{1}{1+e^{w^T x}}\right) =$$

$$p = \frac{1}{1+e^{-w^T x}}$$

$$\begin{cases} \log\left(1+e^{-w^T x}\right), y_i = 1 \\ \log\left(1+e^{w^T x}\right), y_i = 0 \end{cases} \Rightarrow Q(p,X) = \frac{1}{l}\sum_{i=1}^{l} \log\left(1+e^{-y_i \langle w, x_i \rangle}\right) \to min$$

https://dyakonov.org/2018/03/12/логистическая-функция-ошибки/

Обобщение для многомерного случая

Sigmoid

$$\sigma(z) = \frac{1}{1 + e^{-z}} \in [0; 1]$$

$$\mathcal{L} = \frac{1}{l}\sum_{i=1}^{l} - y_i \log a_i - (1 - y_i)\log(1 - a_i)$$

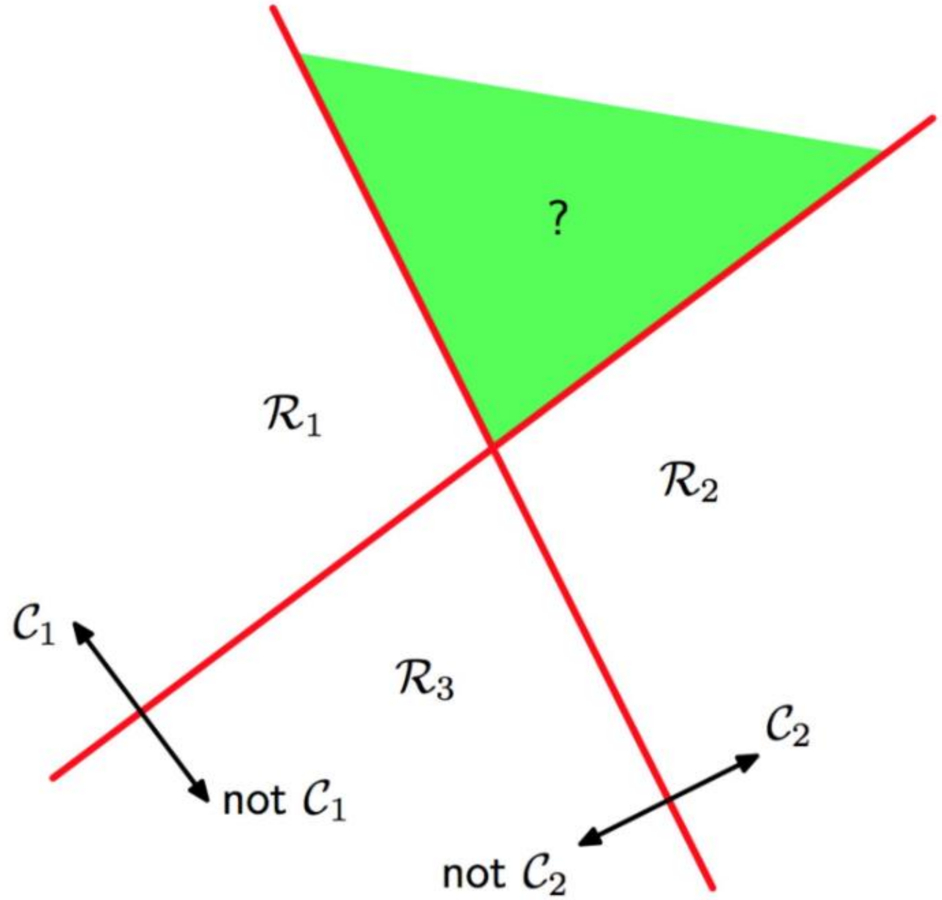
Softmax

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \in [0; 1]$$

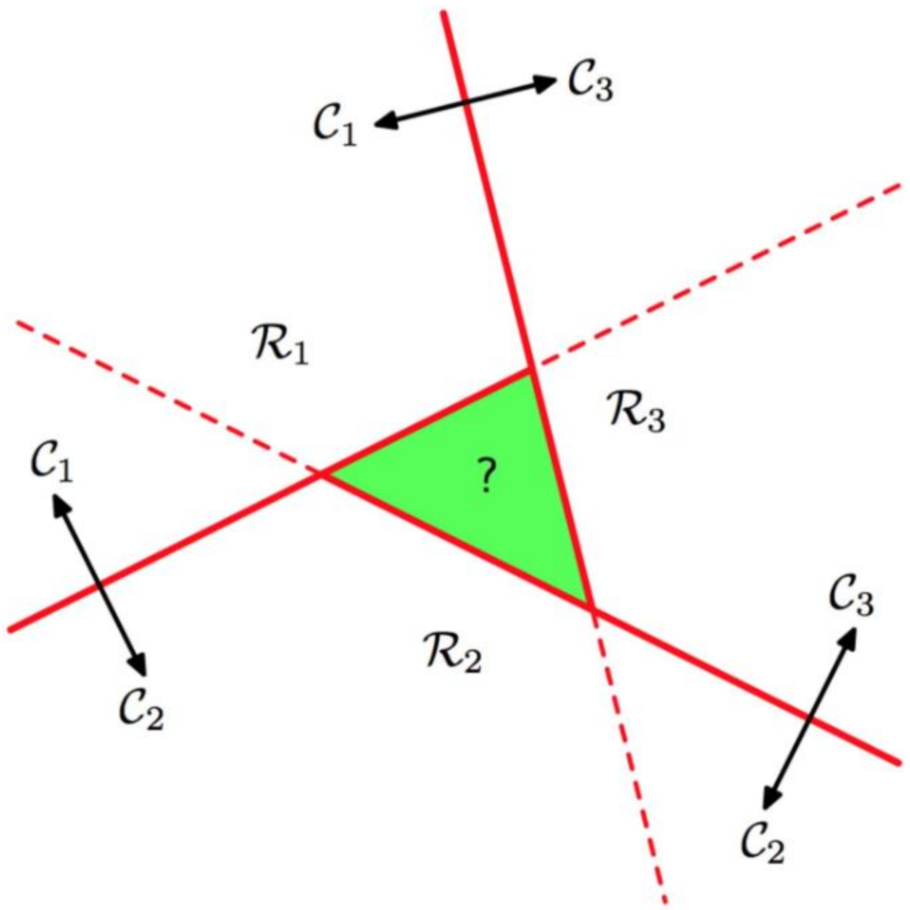
$$\sum_{k=1}^{K} \sigma(z)_k = 1$$

$$\mathcal{L} = \frac{1}{l}\sum_{i=1}^{l}\sum_{k=1}^{K} y_{ik} \log a_{ik}$$

Обобщение для случая множества классов



One vs Rest



One vs One