# Машинное обучение

Лекция 3

Линейная классификация

Михаил Гущин

mhushchyn@hse.ru

НИУ ВШЭ, 2023

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

# На прошлой лекции

- Модель линейной регрессии:

$$\hat{y} = Xw$$

- Функция потерь MSE с регуляризацией:

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \alpha R(w)$$

- Мы хотим минимизировать $L$:

$$L \to \min_{w}$$

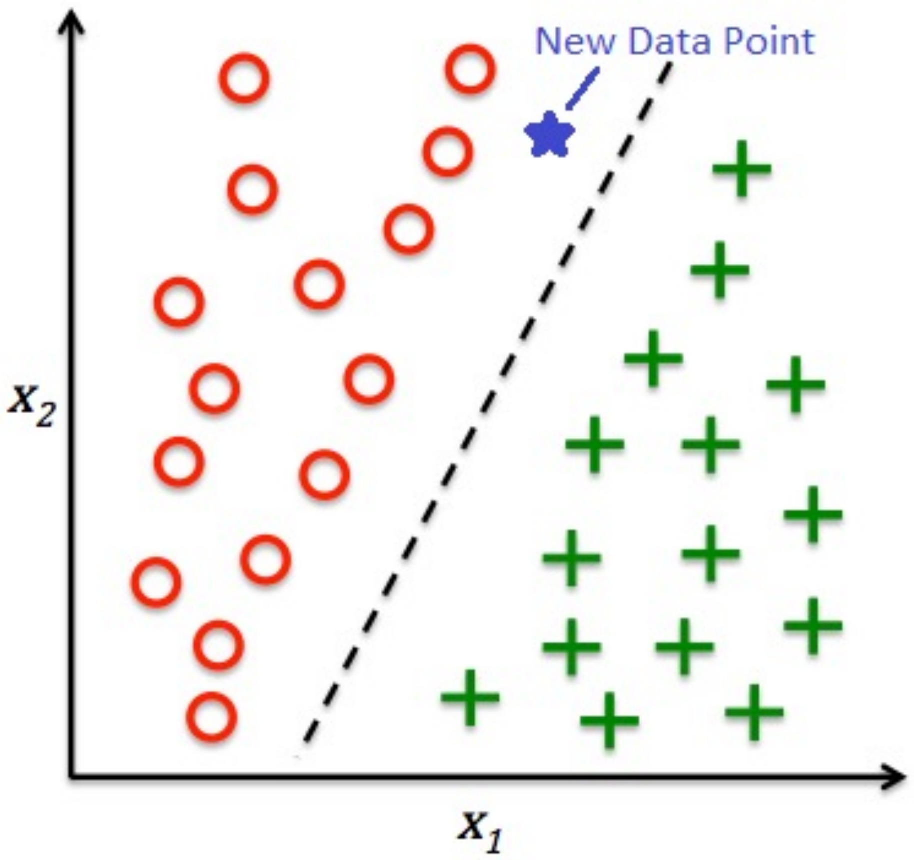
- Градиентный спуск:

$$w^{(k+1)} = w^{(k)} - \eta \nabla L(w^{(k)})$$

# Линейная классификация

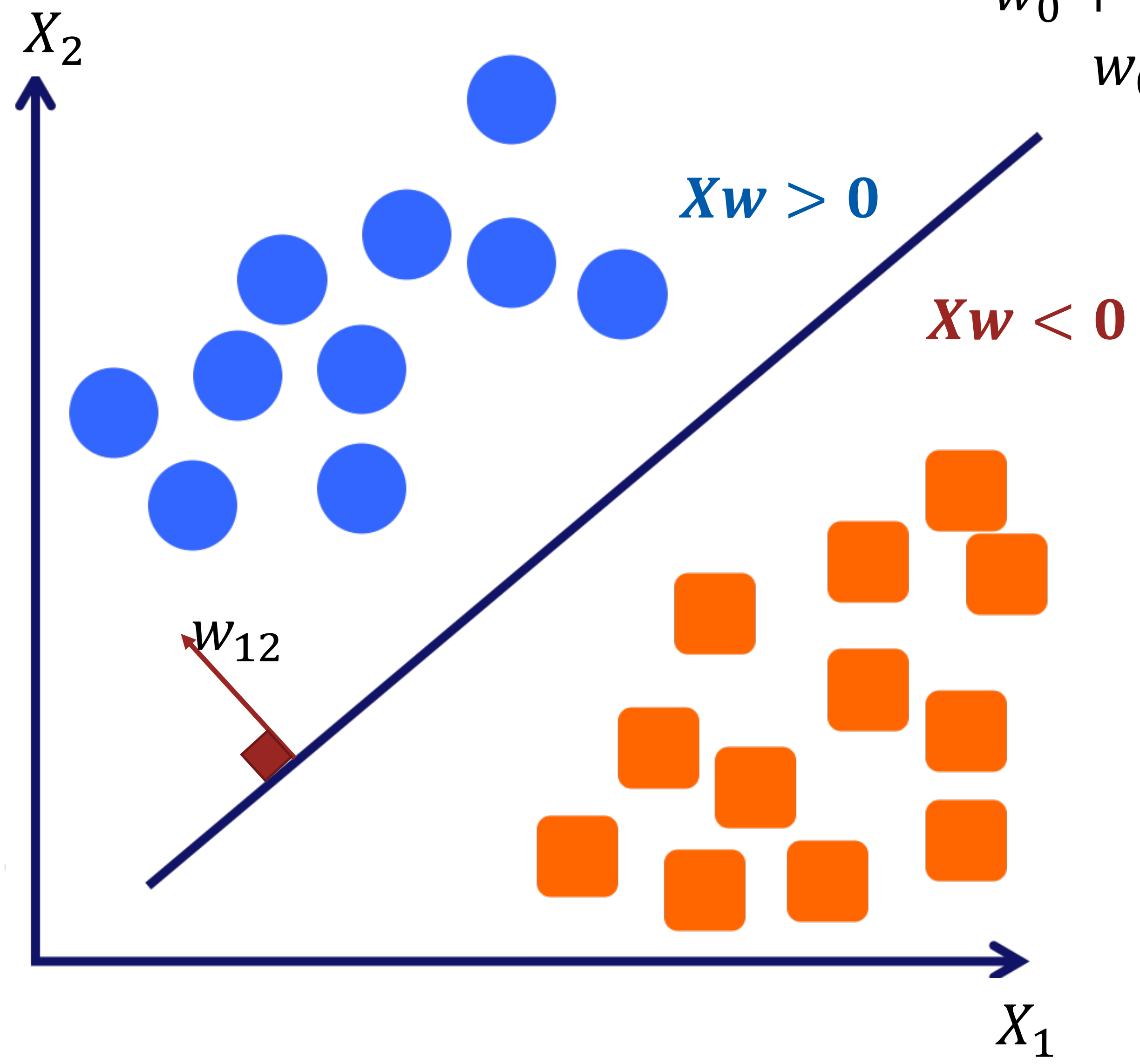
# Задача

- Есть объекты двух классов
- Нужно разделить объекты по классам некоторой **гиперплоскостью**
- Эту гиперплоскость будем называть **линейным классификатором**

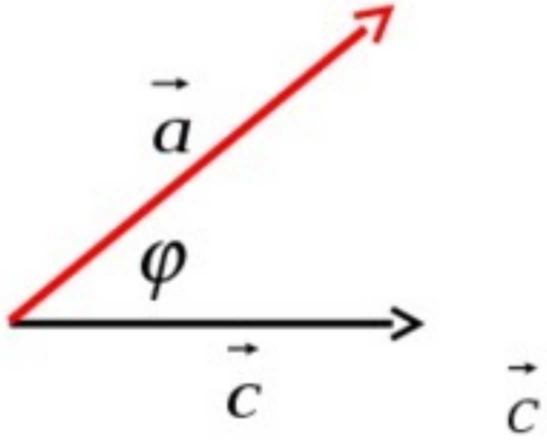


# Гиперплоскость

$$w_0 + w_1 x_1 + w_2 x_2 =$$
$$w_0 + w_{12}^T x =$$
$$Xw = 0$$



# Скалярное произведение

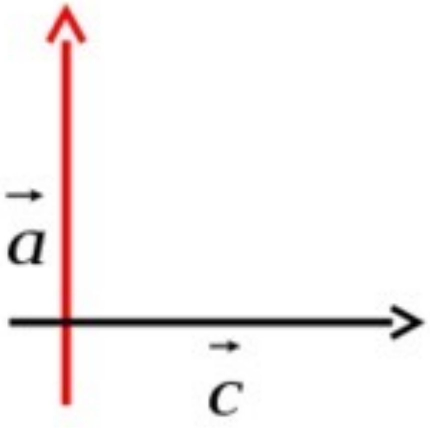



$$\vec{a}\cdot\vec{c}=\left|\vec{a}\right|\cdot\left|\vec{c}\right|\cdot\cos\varphi$$

$$\vec{a}\{x_1;y_1\}$$

$$\vec{c}\{x_2;y_2\}$$

$$\vec{a}\cdot\vec{c}=x_1\cdot x_2+y_1\cdot y_2$$




Если векторы перпендикулярны, то скалярное произведение этих векторов равно 0.

$$\vec{a}\cdot\vec{c}=0$$

# Нормальный вектор к плоскости

- Возьмем такие $x_A, x_B \in \{\, x: w_{12}^T x + w_0 = 0\}$ на гиперплоскости
- Тогда:

$$w_{12}^T x_A + w_0 = 0$$
$$w_{12}^T x_B + w_0 = 0$$

- Найдем разность:

$$w_{12}^T (x_A - x_B) = 0$$

- Поскольку скалярное произведение равно 0, а $(x_A - x_B)$ лежат на гиперплоскости, то вектор $w_{12}$ ортогонален к гиперплоскости

# Расстояние до плоскости

- Расстояние от вектора $x_D$ до плоскости $w_{12}^T x + w_0 = 0$ равно:

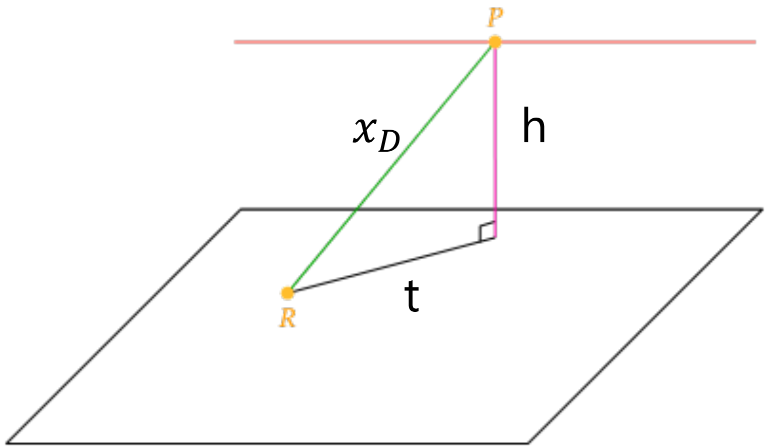
$$\frac{w_{12}^T x_D + w_0}{\|w_{12}\|} \sim Xw$$

- Докажем это. Пусть $x_D = t + h$, где $t$ лежит в плоскости, а $h$ - ортогонален ей. Тогда,

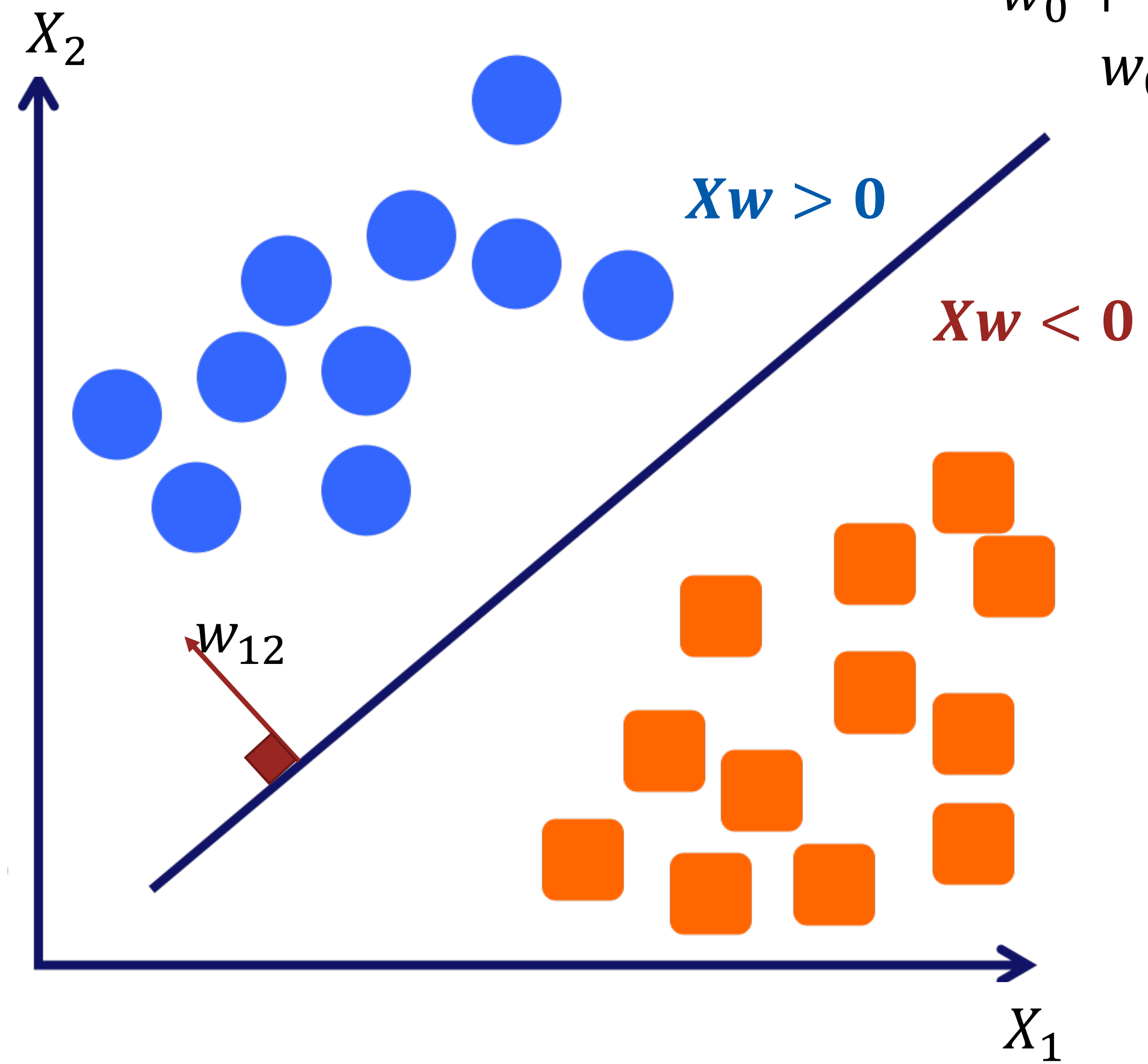
$$w_{12}^T t + w_0 = 0$$
$$w_{12}^T x_D + w_0 = w_{12}^T (t + h) + w_0 = w_{12}^T h$$

- Откуда получаем:

$$\mathrm{h} = \frac{w_{12}^T x_D + w_0}{\|w_{12}\|}$$



# Гиперплоскость




$$w_0 + w_1 x_1 + w_2 x_2 =$$
$$w_0 + w_{12}^T x =$$
$$Xw = 0$$

# Векторная форма

- Пусть дан набор из $n$ точек: $\{x_i, y_i\}_{i=1}^{n}$, где
  - $x_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$ - вектор из $d$ признаков объекта;
  - $y_i = \{-1, +1\}$ – метка класса объекта.
- Модель линейной классификации:

$$\hat{y}_i = sign\left(w_0 + \sum_{j=1}^{d} w_j x_{ij}\right)$$

  - $w_j$ - веса модели;
  - $\hat{y}_i$ - прогноз для объекта;
- Ошибка прогноза модели для объекта: $\hat{y}_i \neq y_i$

# Матричная форма

- Модель линейной классификации:

$$\hat{y} = sign(Xw\ )$$

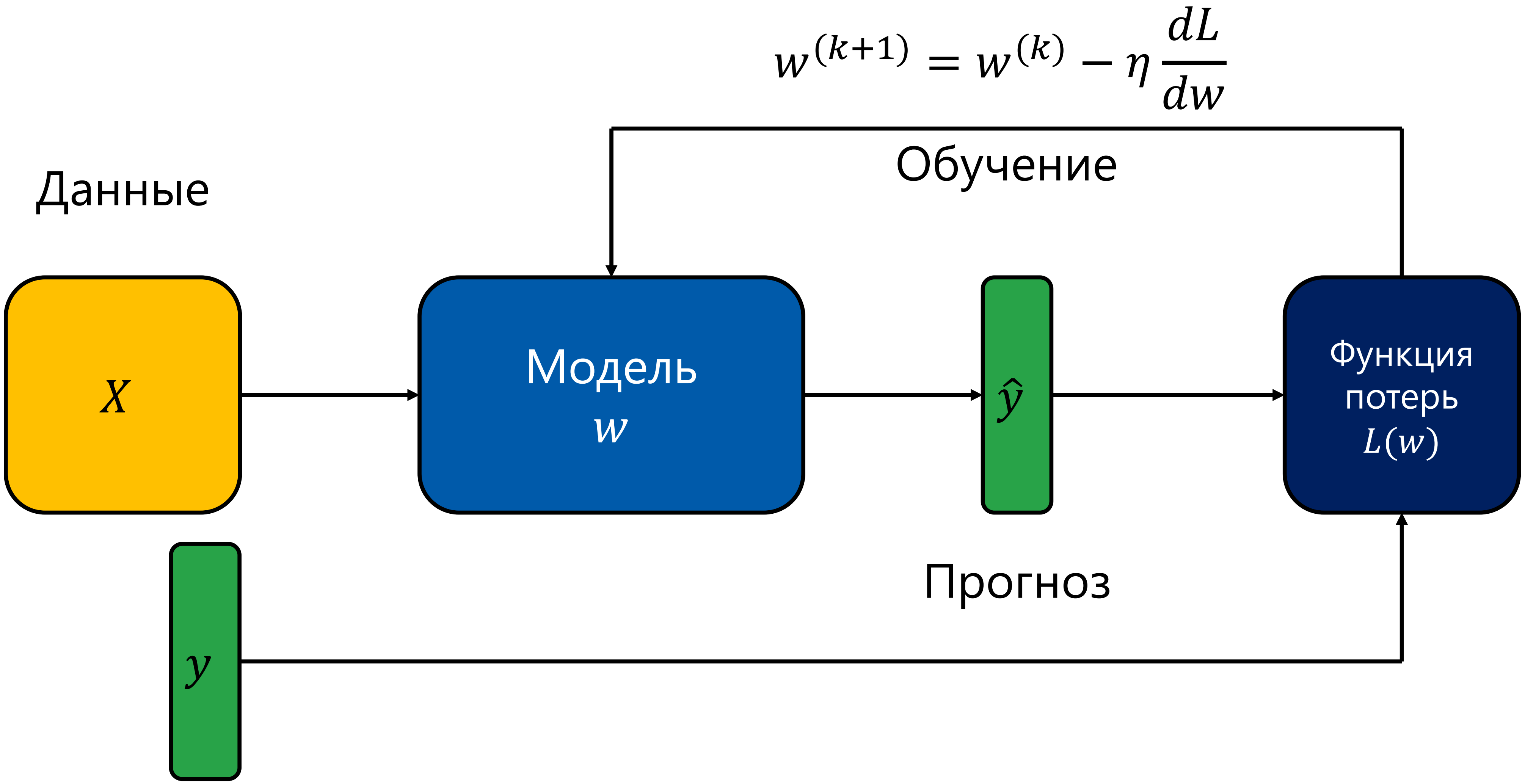
– $X = \begin{pmatrix} \mathbf{1}\ x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ \mathbf{1}\ x_{n1} & \cdots & x_{nd} \end{pmatrix}$ - матрица признаков объектов;

– $w = (w_0, w_1, \dots, w_d)^T$ - вектор $(d + 1)$ весов модели;

– $\hat{y}\ = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)^T$ - вектор прогнозов модели для $(n)$ объектов;

- Вектор ошибок прогнозов модели: $\hat{y}_i \neq y_i$

# Обучение классификатора



# Функция потерь

- Функция потерь (Loss function) для классификации:

$$L = \frac{1}{n}\sum_{i=1}^{n}[\hat{y}_i \neq y_i]$$

- Значение $L$ – доля неправильных ответов
- Мы хотим минимизировать $L$:

$$L \to \min_{w}$$

# Функция потерь

- Дискретная относительно весов модели
- Нет производной (0, либо не определена)
- Не можем использовать градиентный спуск
- Много глобальных минимумов (несколько способов разделить объекты на классы)

# Повтор

- Модель линейной классификации:

$$\hat{y} = sign(Xw\ )$$

- Функция потерь, доля неправильных ответов:

$$L = \frac{1}{n}\sum_{i=1}^{n}[\hat{y}_i \neq y_i]$$

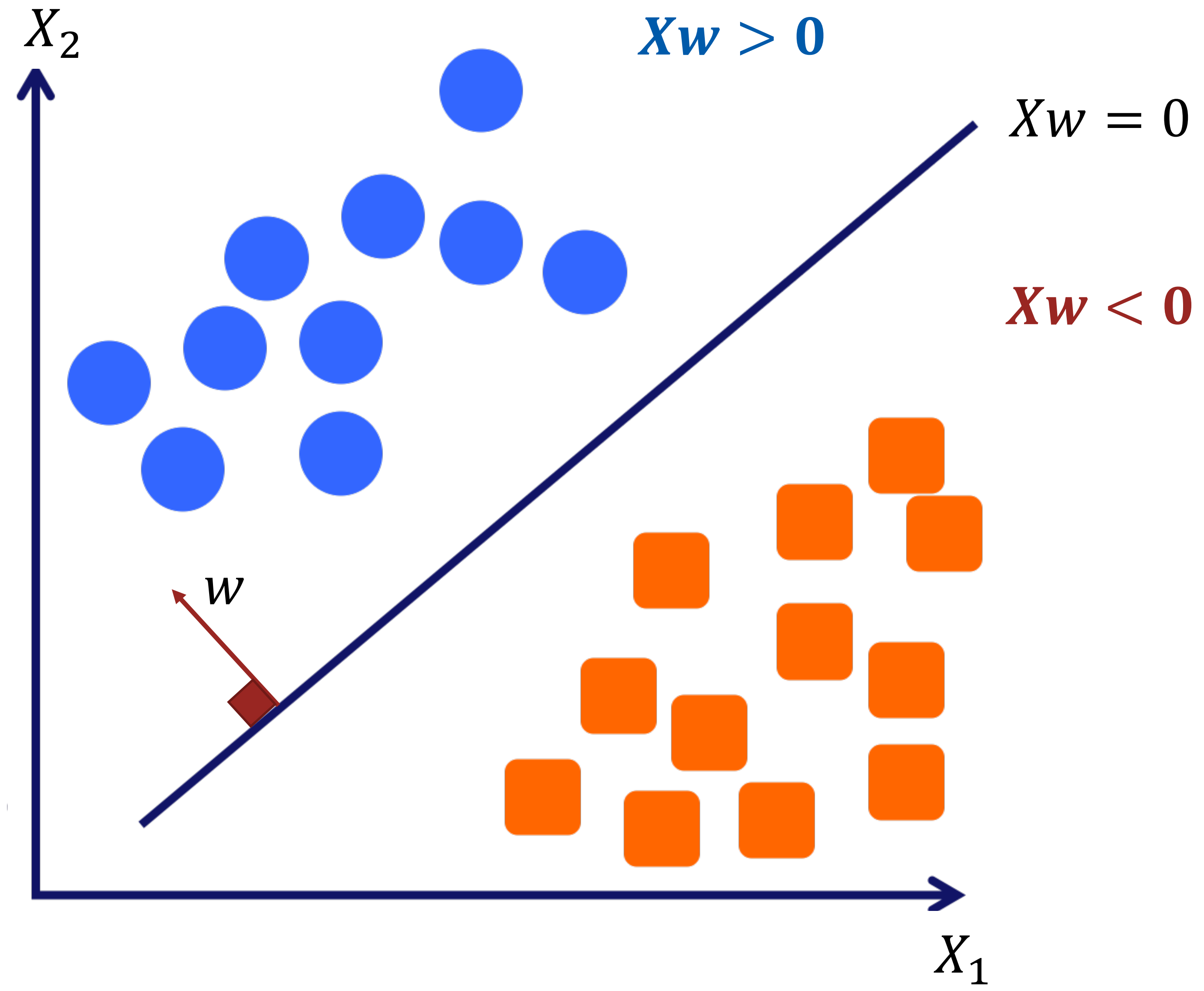
- Мы хотим минимизировать $L$:

$$L \to \min_{w}$$

- Решение: пока не знаем ☺

# Отступы

# Гиперплоскость



# Отступ

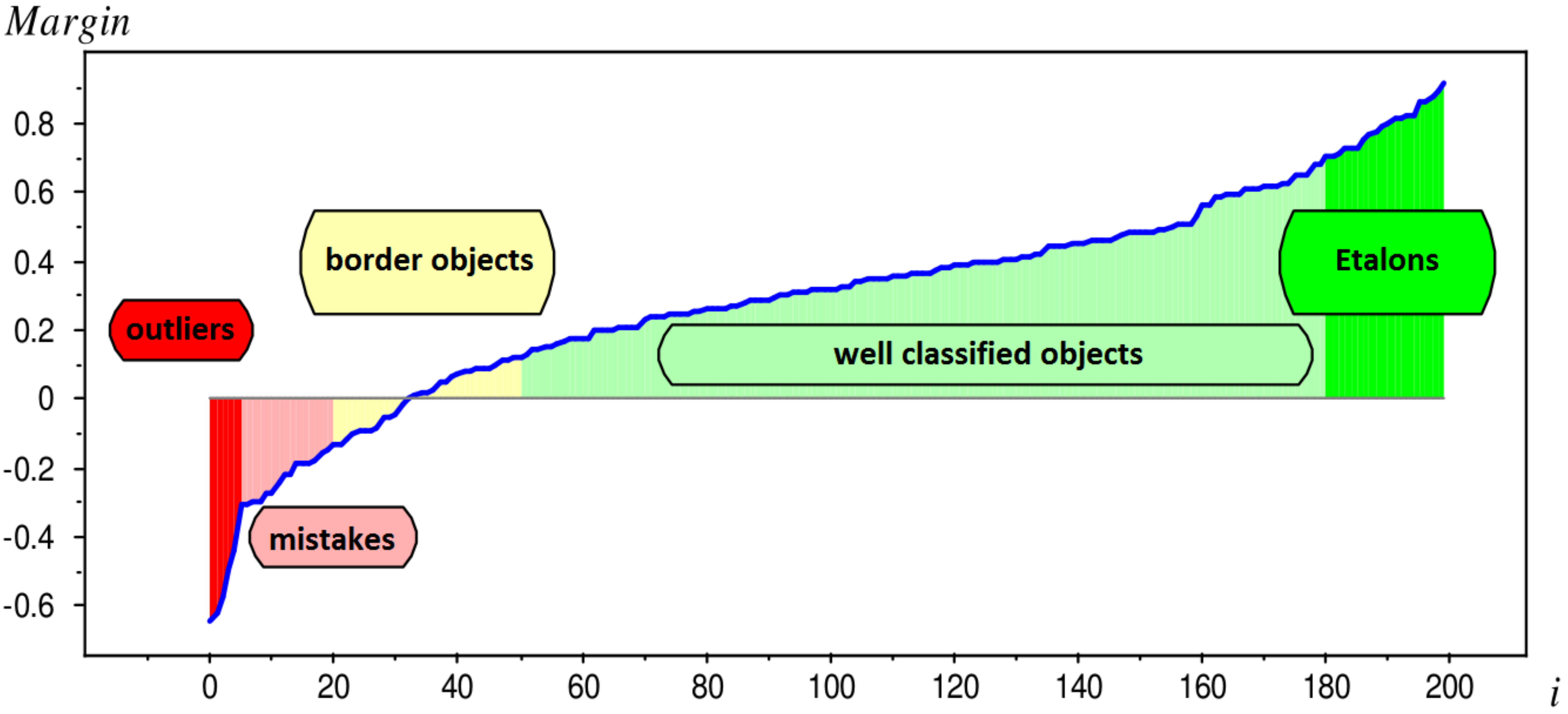
- Отступ (margin) $M$:

$$z = Xw$$

$$\mathrm{M} = yz$$

- Знак отступа говорит о корректности прогноза
  - $M_i > 0$ – верный прогноз
  - $M_i < 0$ – неправильный прогноз
- Абсолютная величина – степень уверенности классификатора
- Чем ближе $M$ к 0, тем ближе объект к границе классов

# Отступ



Margin
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
border objects
outliers
Etalons
well classified objects
mistakes
0
20
40
60
80
100
120
140
160
180
200
i

# Новая функция потерь

- Функция потерь (Loss function) для классификации:

$$L = \frac{1}{n}\sum_{i=1}^{n}[M_i < 0]$$

- Значение $L$ – доля неправильных ответов
- Мы хотим минимизировать $L$:

$$L \to \min_{w}$$

# Верхние оценки

# Задача

- Есть функция потерь для классификации:

$$L = \frac{1}{n}\sum_{i=1}^{n}[M_i < 0]$$

- Не можем использовать для градиентного спуска
- Хотим заменить ее на гладкую функцию

# Верхние оценки

- Есть функция потерь для одного произвольного объекта:

$$L(M_i) = [M_i < 0]$$

- Хотим найти такую $\tilde{L}(M)$, что

$$L(M_i) \leq \tilde{L}(M_i)$$

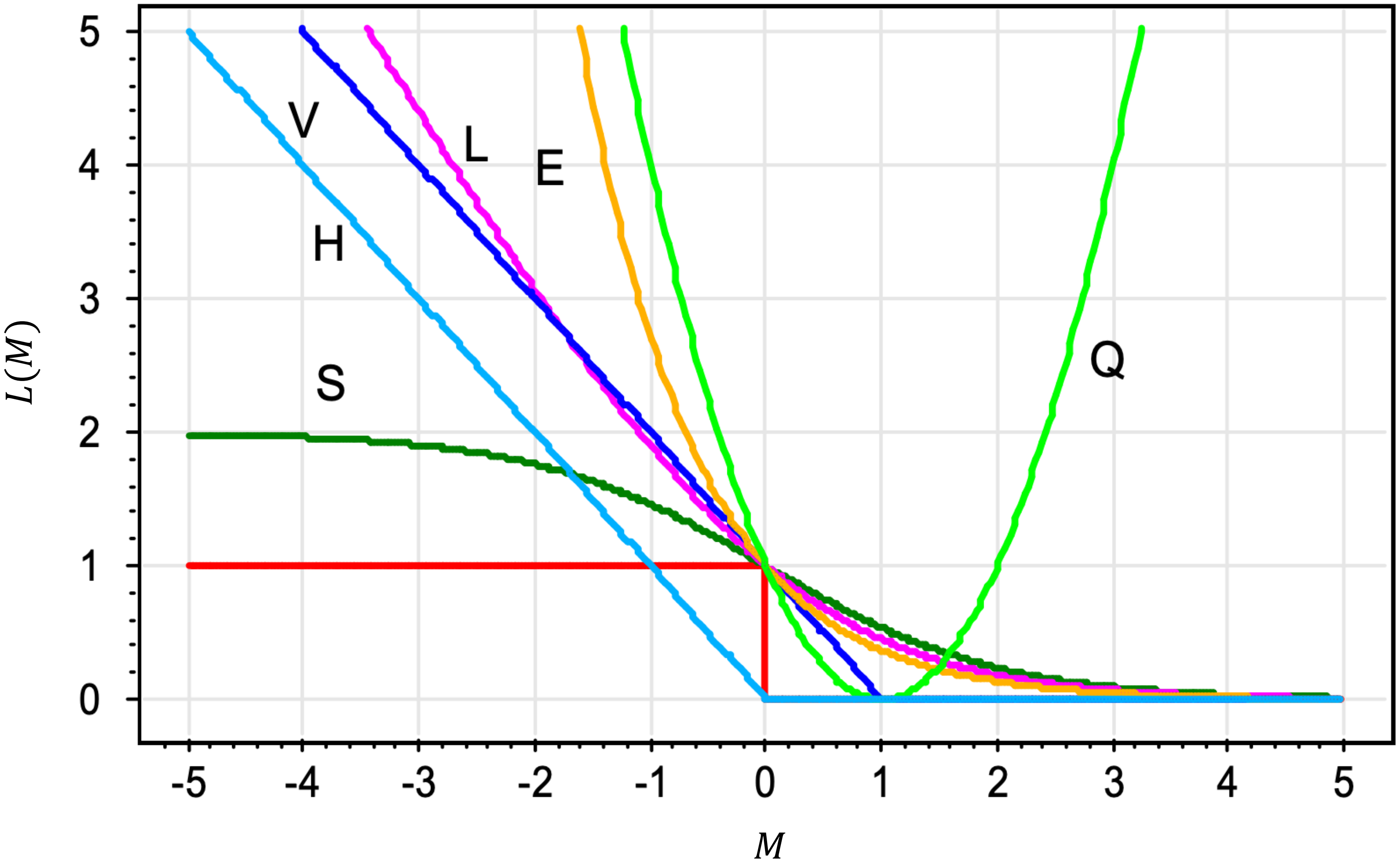
- Тогда верхняя оценка выглядит так:

$$L = \frac{1}{n}\sum_{i=1}^{n}[M_i < 0] \leq \frac{1}{n}\sum_{i=1}^{n}\tilde{L}(M_i) \to \min_w$$

# Примеры

- $\tilde{L}(M) = \log(1 + e^{-M})$ – логистическая функция потерь (рассмотрим подробно далее)
- $\tilde{L}(M) = \max(0, 1 - M)$ – кусочно-линейная функция потерь
- $\tilde{L}(M) = \max(0, -M)$ – кусочно-линейная функция потерь
- $\tilde{L}(M) = e^{-M}$ – экспоненциальная функция потерь
- $\tilde{L}(M) = 2(1 + e^{M})^{-1}$ – сигмоидная функция потерь

# Примеры



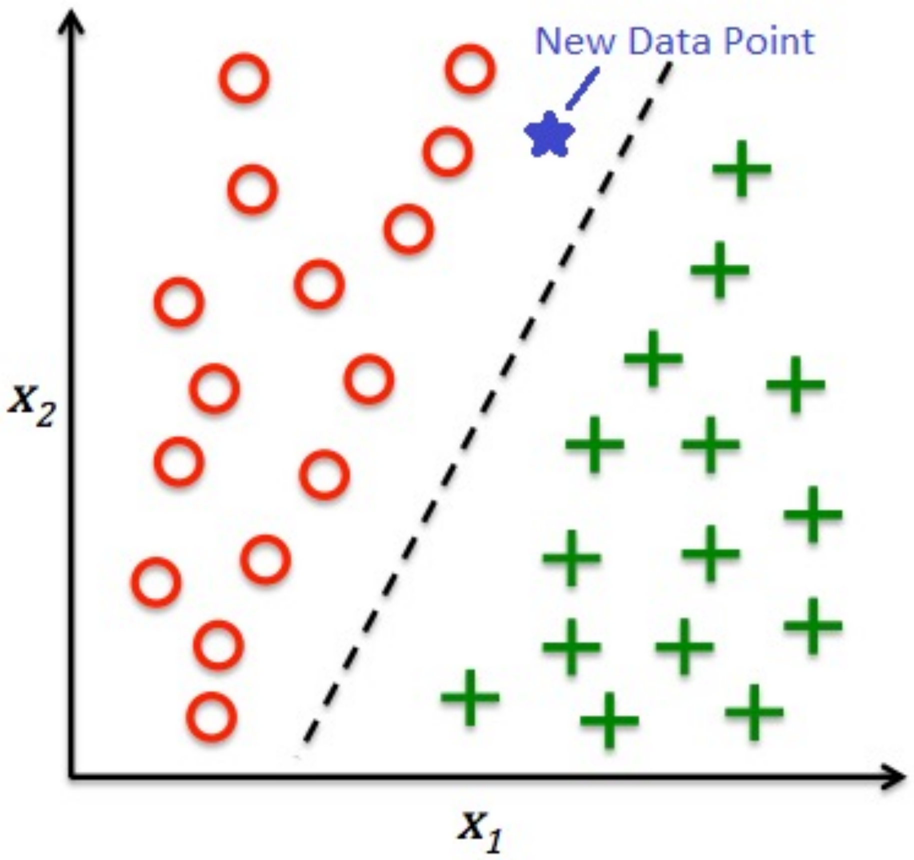
# Логистическая регрессия

# Задача

- Есть объекты двух классов

- Нужно разделить объекты по классам некоторой **гиперплоскостью**

- Эту гиперплоскость будем называть **линейным классификатором**



# Матричная форма

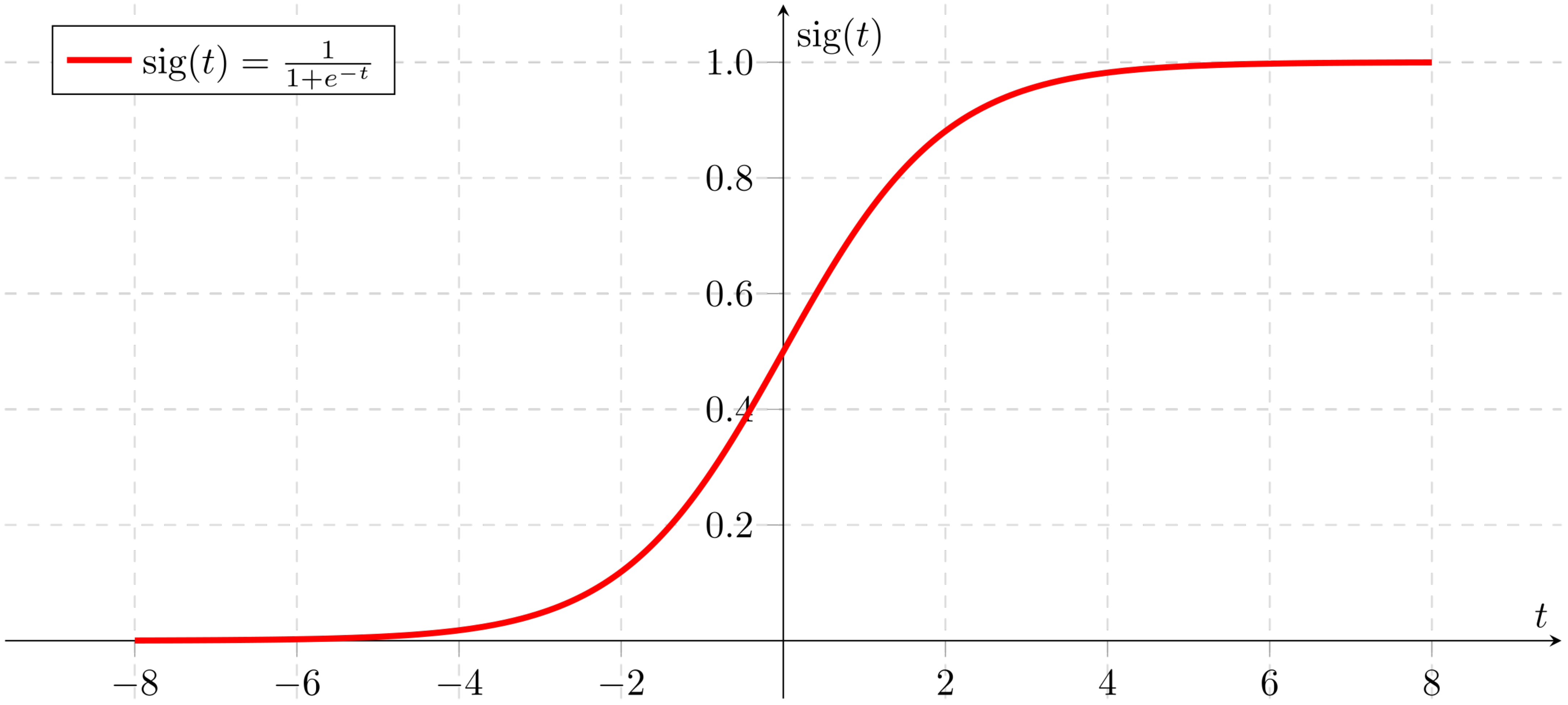
- Пусть дан набор из $n$ точек: $\{x_i, y_i\}_{i=1}^{n}$, где
  - $x_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$ - вектор из $d$ признаков объекта;
  - $y_i = \mathbf{\{0, 1\}}$ – метка класса объекта.
- Модель логистической регрессии:

$$\hat{y}_i = \boldsymbol{\sigma}(Xw)$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

- $\hat{y}_i$ - **вероятность класса 1** для объекта;

# Сигмоида



# Функция потерь

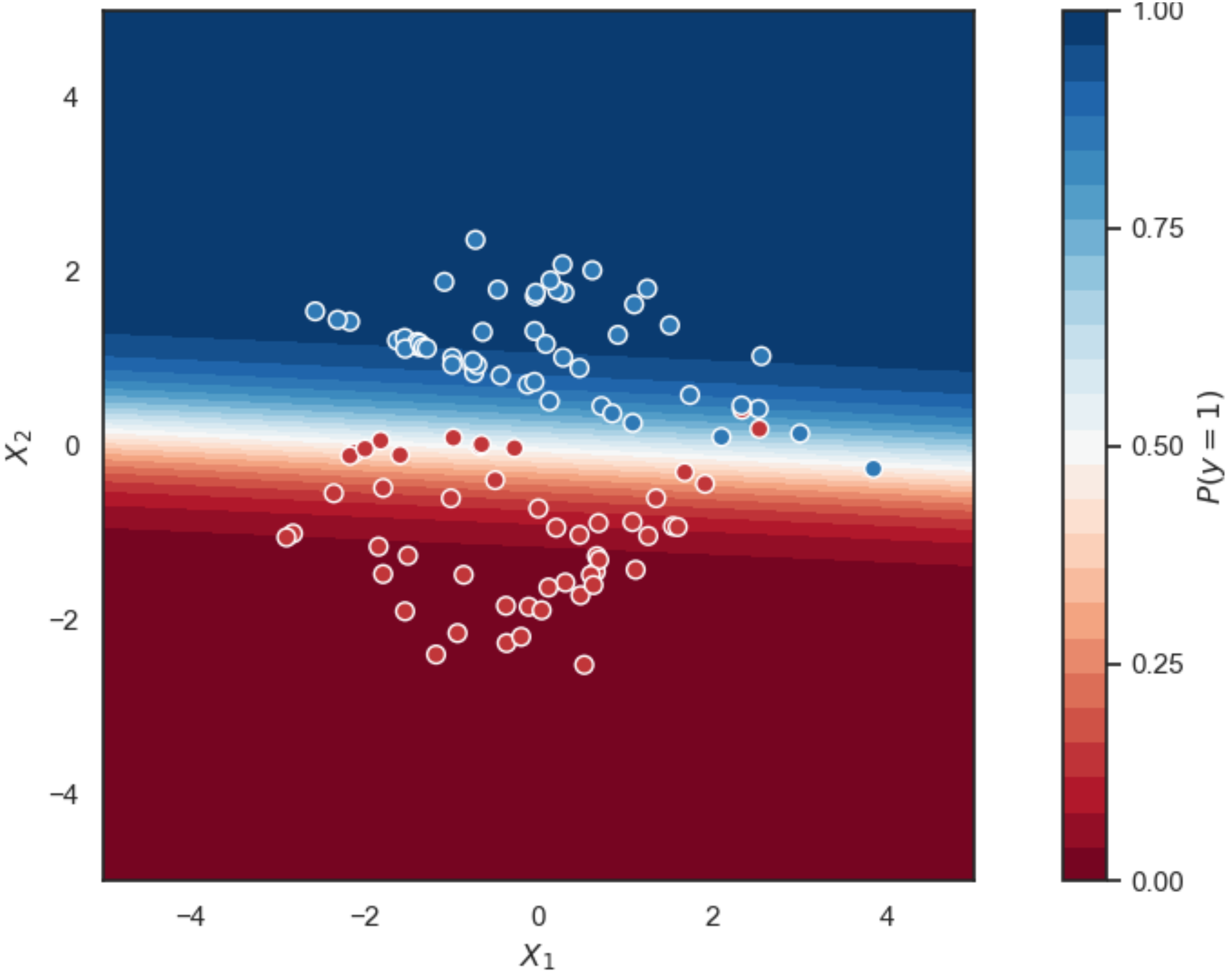
- Функция потерь для логистической регрессии (**log-loss**):

$$L = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right)$$

- Мы хотим минимизировать $L$:

$$L \to \min_{w}$$

# Пример



# Задание

Покажите, что при $y_i = \{\mathbf{0}, \mathbf{1}\}$ функция потерь

$$L = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i))$$

эквивалентна функции потерь

$$L = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-M_i})$$

при $y_i = \{\mathbf{-1}, \mathbf{+1}\}$

# Повтор

- Модель логистической регрессии:

$$\hat{y} = \sigma(Xw)$$

- Функция потерь log-loss:

$$L = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right)$$

- Мы хотим минимизировать $L$:

$$L \to \min_w$$

- Градиентный спуск:

$$w^{(k+1)} = w^{(k)} - \eta \nabla L(w^{(k)})$$

# Вероятностная интерпретация

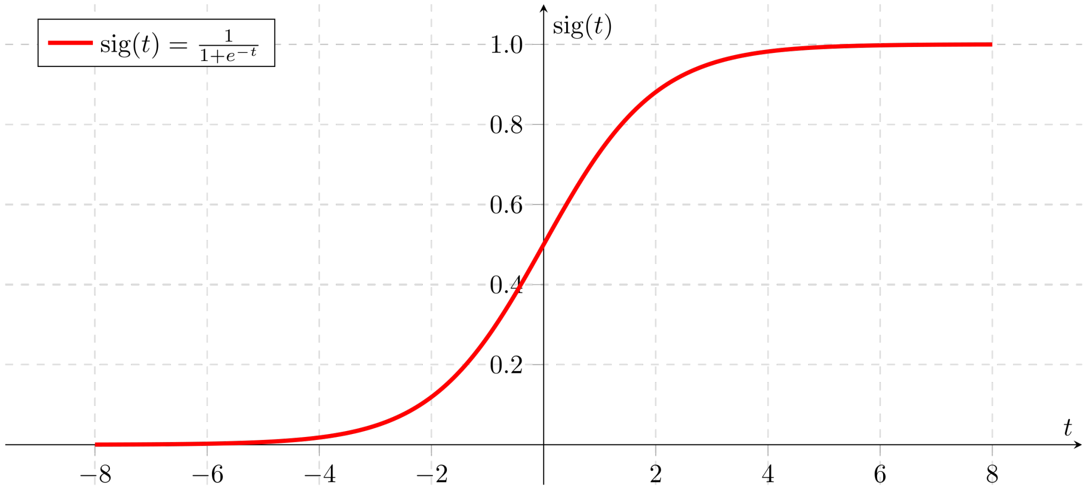
# Логистическая регрессия

- Вероятность класса 1:

$$p(y = \mathbf{1}|x_i) = \sigma\left(x_i^T w\right) = \hat{y}_i$$

- Вероятность класса 0:

$$p(y = \mathbf{0}|x_i) = 1 - \sigma(x_i^T w)$$



# Правдоподобие

- Пусть дан набор из $n$ точек: $\{x_i, y_i\}_{i=1}^{n}$, где
  - $x_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$ - вектор из $d$ признаков объекта;
  - $y_i = \{\mathbf{0}, \mathbf{1}\}$ – метка класса объекта.

- Тогда правдоподобие:

$$\text{Likelihood} = \prod_{i=1}^{n} p(y = 1|x_i)^{[y_i=1]} p(y = 0|x_i)^{[y_i=0]} \to \max_{w}$$

# Логарифм правдоподобия

- Правдоподобие:

$$\text{Likelihood} = \prod_{i=1}^{n} p(y = 1|x_i)^{[y_i=1]} p(y = 0|x_i)^{[y_i=0]}$$

- Логарифм прадоподобия:

$$\text{Log Likelihood} = \sum_{i=1}^{n} y_i \log(p(y = 1|x_i)) + (1 - y_i) \log(p(y = 0|x_i)) =$$

$$= \sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) = -nL$$

# Заключение

# Резюме

- Модель логистической регрессии:

$$\hat{y} = \sigma(Xw)$$

- Функция потерь log-loss:

$$L = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right)$$

- Мы хотим минимизировать $L$:

$$L \to \min_{w}$$

- Градиентный спуск:

$$w^{(k+1)} = w^{(k)} - \eta \nabla L(w^{(k)})$$

# Вопросы

- Запишите формулу для линейной модели классификации. Что такое отступ? Как обучаются линейные классификаторы и для чего нужны верхние оценки пороговой функции потерь?
- Как в логистической регрессии выполняются предсказания для новых объектов? Запишите логистическую функцию потерь. Как она связана с методом максимума правдоподобия?