

Введение в анализ данных

Лекция 5. Логистическая регрессия

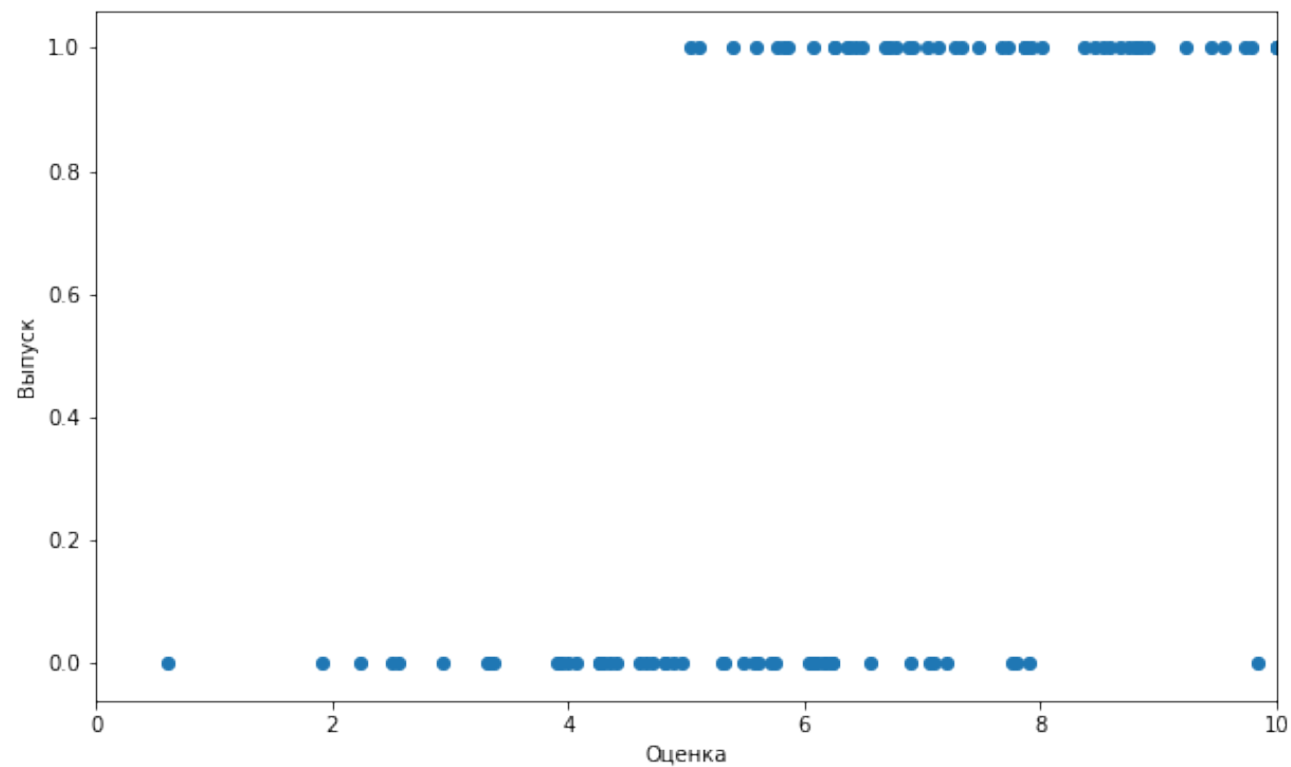
Логистическая регрессия:
простое объяснение

Логистическая регрессия

- Решаем задачу бинарной классификации: $\mathbb{Y} = \{-1, +1\}$
- Минимизация верхней оценки:

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

Предсказание вероятностей



Предсказание вероятностей



Предсказание вероятностей

- Кредитный скоринг
- Стратегия: выдавать кредит только клиентам с $b(x) > 0.9$
- 10% невозвращённых кредитов — нормально

Предсказание вероятностей

- Баннерная реклама
- $b(x)$ — вероятность, что пользователь кликнет по рекламе
- $c(x)$ — прибыль в случае клика
- $c(x)b(x)$ — хотим оптимизировать

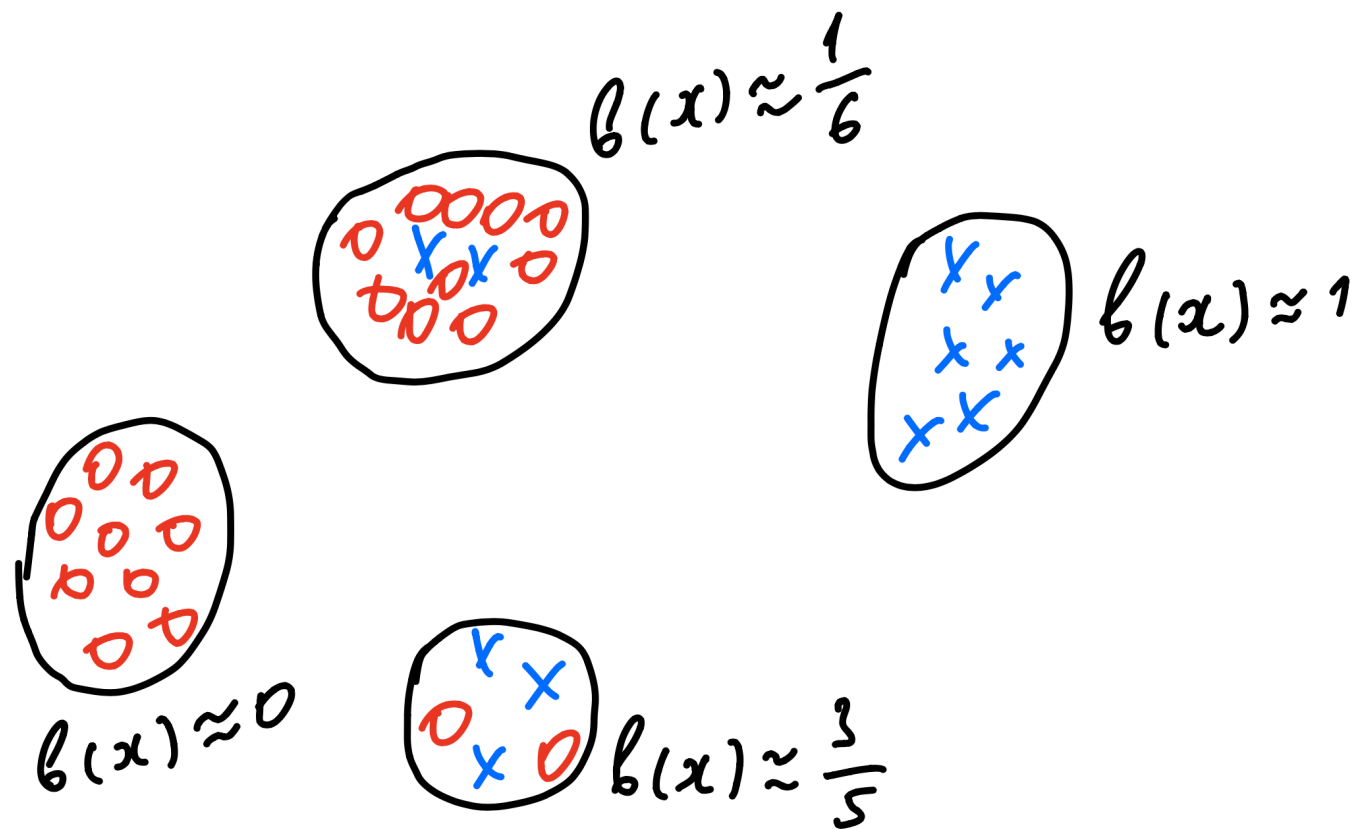
Предсказание вероятностей

- Прогнозирование оттока клиентов
- Медицинская диагностика
- Поисковое ранжирование (насколько веб-страница соответствует запросу?)

Предсказание вероятностей

Будем говорить, что модель $b(x)$ предсказывает вероятности, если среди объектов с $b(x) = p$ доля положительных равна p .

Предсказание вероятностей



Предсказание вероятностей



Линейный классификатор

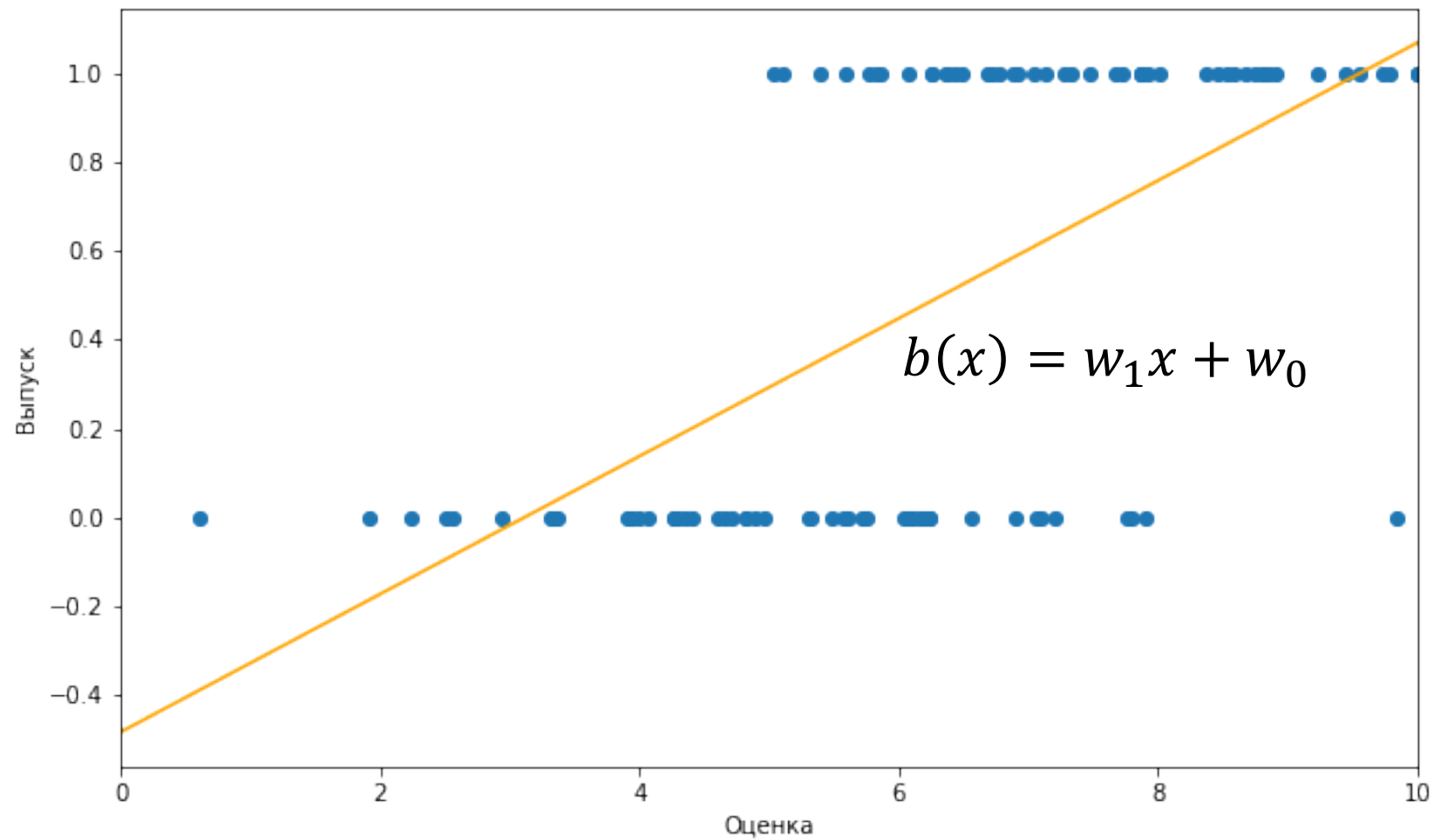
$$a(x) = \text{sign } \langle w, x \rangle$$

- Обучим как-нибудь — например, на логистическую функцию потерь:

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

- Может, $\langle w, x \rangle$ сойдёт за оценку?

Предсказание вероятностей

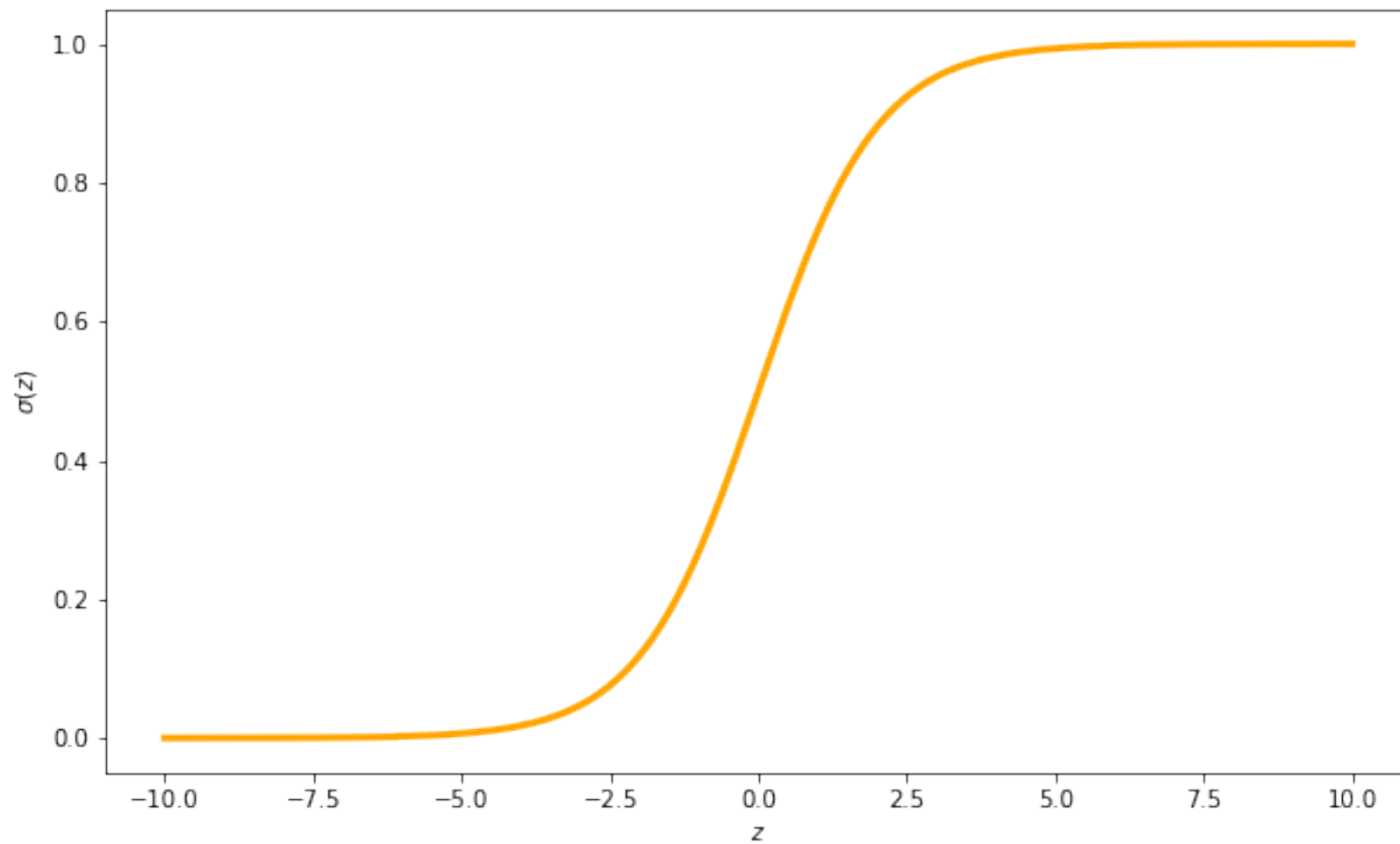


Линейный классификатор

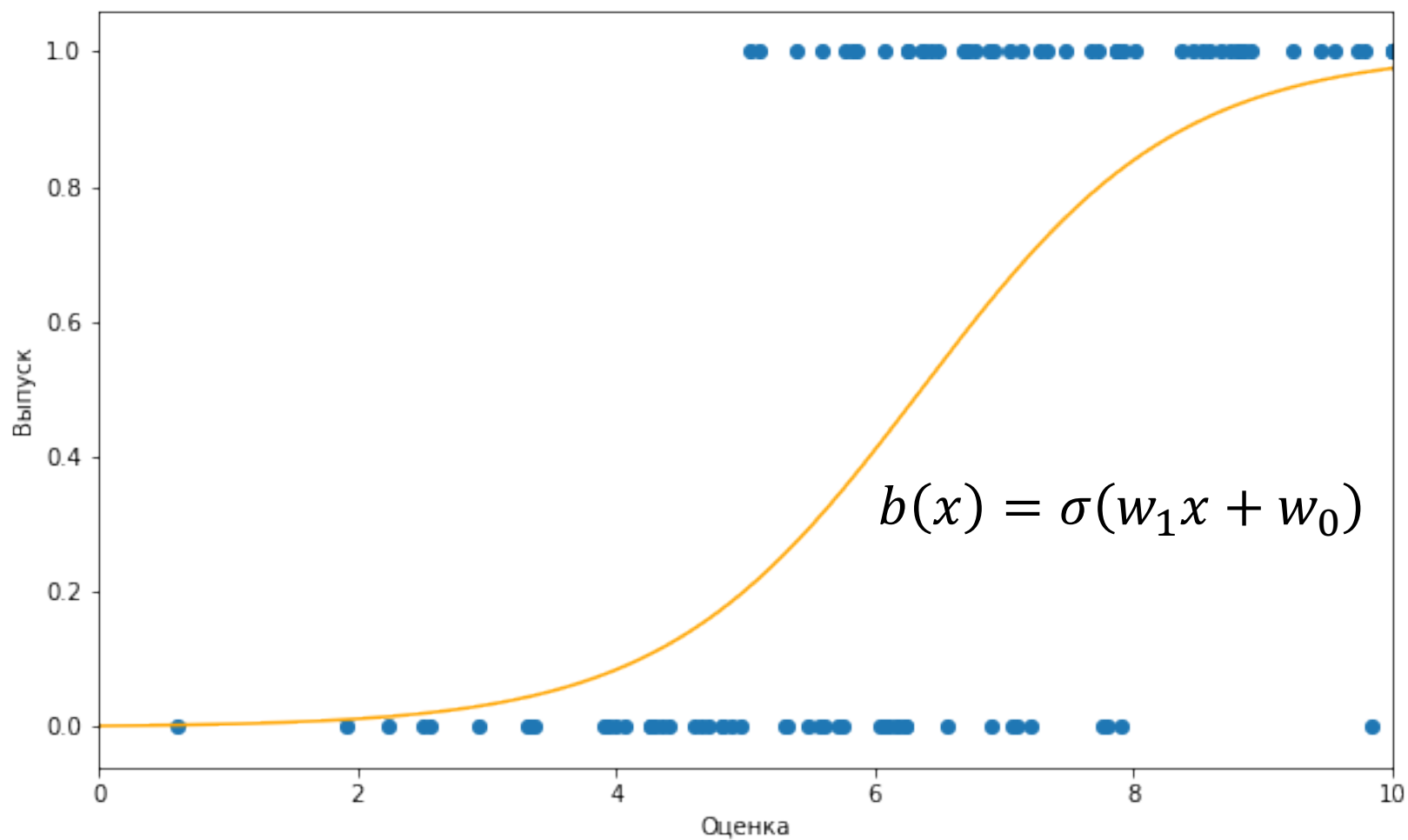
- Переведём выход модели на отрезок $[0, 1]$
- Например, с помощью сигмоиды:

$$\sigma(\langle w, x \rangle) = \frac{1}{1 + \exp(-\langle w, x \rangle)}$$

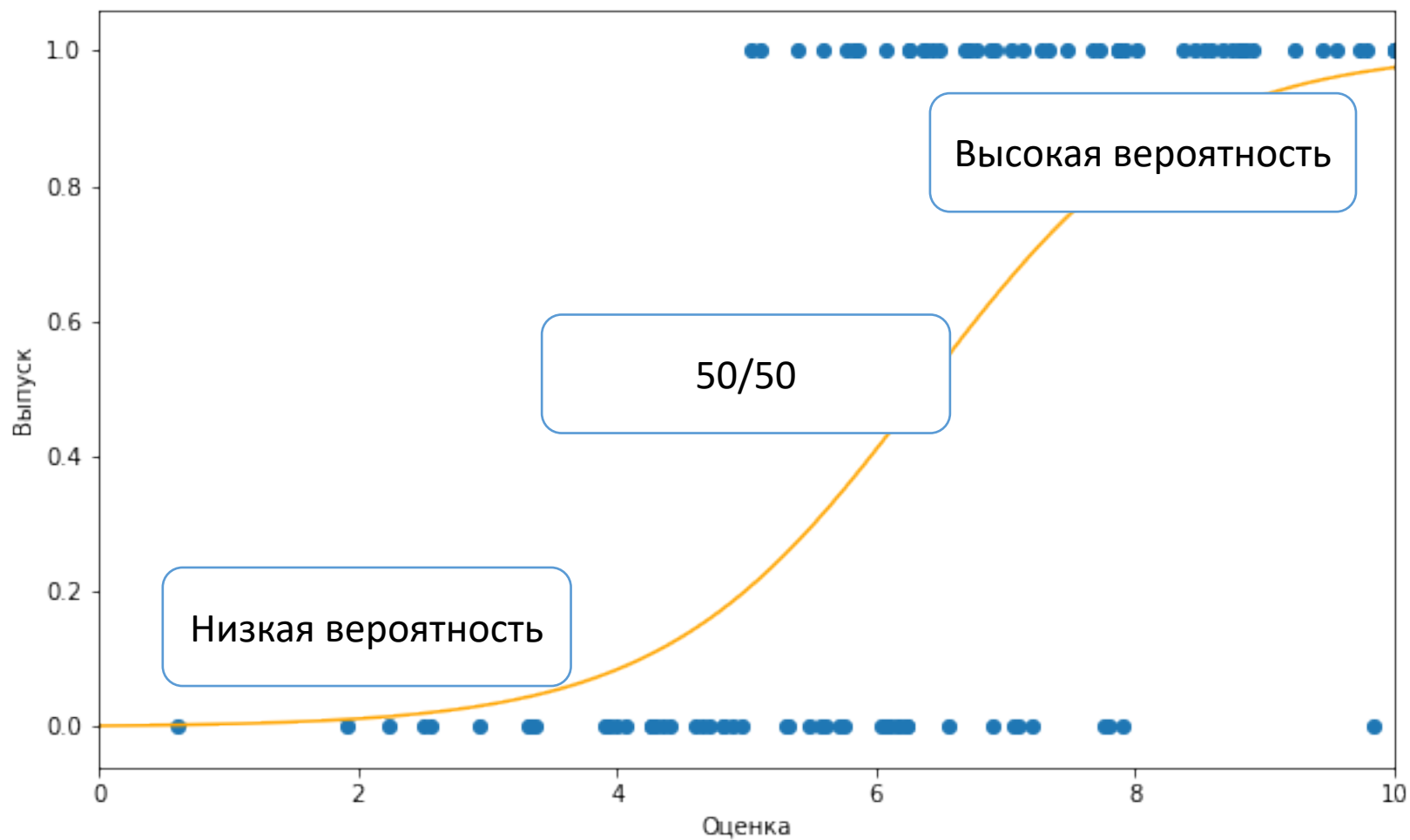
Сигмоида



Предсказание вероятностей



Предсказание вероятностей



Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?

Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 0$

Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 1$ или $\langle w, x_i \rangle \rightarrow +\infty$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 0$ или $\langle w, x_i \rangle \rightarrow -\infty$

Предсказание вероятностей

- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 1$ или $\langle w, x_i \rangle \rightarrow +\infty$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 0$ или $\langle w, x_i \rangle \rightarrow -\infty$
- То есть задача — сделать отступы на всех объектах максимальными

$$y_i \langle w, x_i \rangle \rightarrow \max_w$$

Предсказание вероятностей

- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 0$

$$-\sum_{i=1}^{\ell} \{ [y_i = 1] \sigma(\langle w, x_i \rangle) + [y_i = -1] (1 - \sigma(\langle w, x_i \rangle)) \} \rightarrow \min_w$$

Предсказание вероятностей

$$-\sum_{i=1}^{\ell} \{[y_i = 1]\sigma(\langle w, x_i \rangle) + [y_i = -1](1 - \sigma(\langle w, x_i \rangle))\} \rightarrow \min_w$$

- Если $y_i = +1$ и $\sigma(\langle w, x_i \rangle) = 0$, то штраф равен 1
- Если $y_i = +1$, то заменить $\sigma(\langle w, x_i \rangle) = 1$ на $\sigma(\langle w, x_i \rangle) = 0.5$ так же плохо, как заменить $\sigma(\langle w, x_i \rangle) = 0.5$ на $\sigma(\langle w, x_i \rangle) = 0$
- Надо строже!

Предсказание вероятностей

$$-\sum_{i=1}^{\ell} \{ [y_i = 1] \log \sigma(\langle w, x_i \rangle) + [y_i = -1] \log(1 - \sigma(\langle w, x_i \rangle)) \} \rightarrow \min_w$$

- Если $y_i = +1$ и $\sigma(\langle w, x_i \rangle) = 0$, то штраф равен $-\log 0 = +\infty$
- Достаточно строго
- Функция потерь называется **log-loss**

$$L(y, z) = -[y = 1] \log z - [y = -1] \log(1 - z)$$

Логистическая регрессия

$$\begin{aligned} & - \sum_{i=1}^{\ell} \{ [y_i = 1] \log \sigma(\langle w, x_i \rangle) + [y_i = -1] \log(1 - \sigma(\langle w, x_i \rangle)) \} = \\ & - \sum_{i=1}^{\ell} \left\{ [y_i = 1] \log \frac{1}{1 + \exp(-\langle w, x \rangle)} + [y_i = -1] \log \left(1 - \frac{1}{1 + \exp(-\langle w, x \rangle)} \right) \right\} = \\ & - \sum_{i=1}^{\ell} \left\{ [y_i = 1] \log \frac{1}{1 + \exp(-\langle w, x \rangle)} + [y_i = -1] \log \left(\frac{1}{1 + \exp(\langle w, x \rangle)} \right) \right\} = \\ & \sum_{i=1}^{\ell} \{ [y_i = 1] \log(1 + \exp(-\langle w, x \rangle)) + [y_i = -1] \log(1 + \exp(\langle w, x \rangle)) \} = \\ & \sum_{i=1}^{\ell} \log(1 + \exp(-y_i \langle w, x_i \rangle)) \end{aligned}$$

Логистическая регрессия:
сложное объяснение

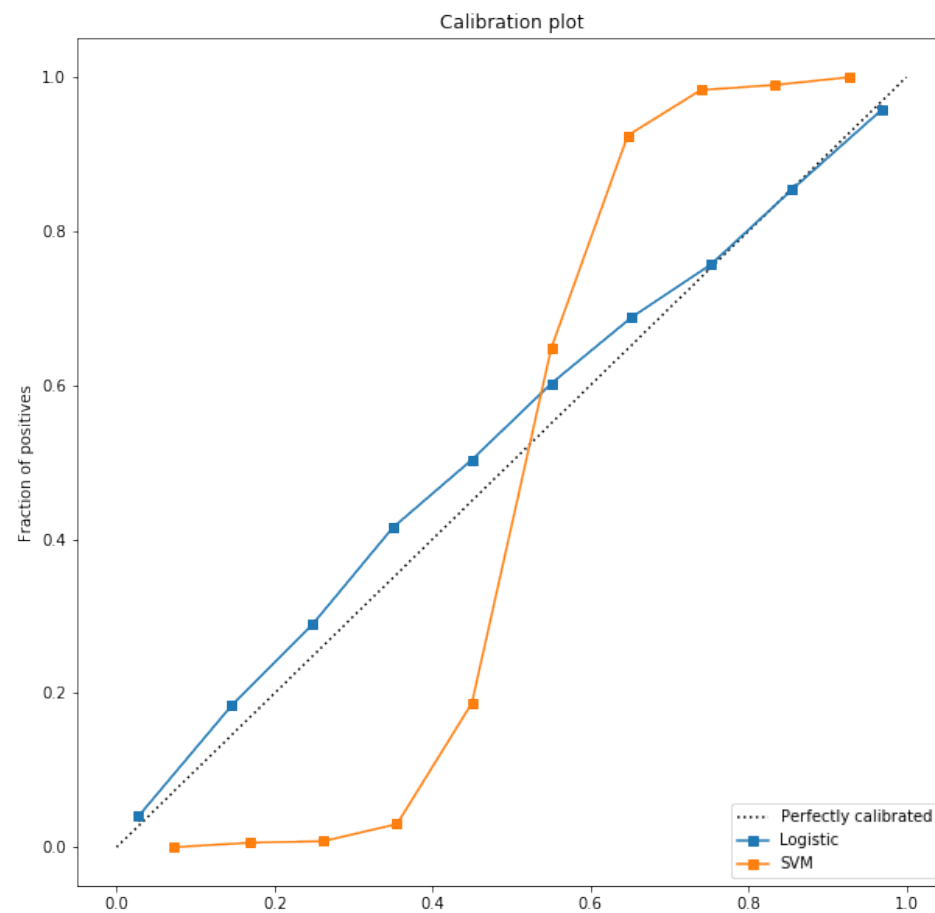
Предсказание вероятностей

Будем говорить, что модель $b(x)$ предсказывает вероятности, если среди объектов с $b(x) = p$ доля положительных равна p .

Калибровочная кривая

- Разобьём отрезок $[0, 1]$ на n корзинок $[0, t_1], [t_1, t_2], \dots, [t_{n-1}, 1]$ — это ось X
- Для каждого отрезка $[t_i, t_{i+1}]$ берём объекты, для которых $b(x) \in [t_i, t_{i+1}]$
- Считаем среди объектов долю положительных, откладываем её на оси Y

Калибровочная кривая



Предсказание вероятностей

- Функционал ошибки:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, b(x_i)) \rightarrow \min_a$$

Предсказание вероятностей

- Функционал ошибки:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, b(x_i)) \rightarrow \min_a$$

- Рассмотрим ошибку только на объектах x_1, \dots, x_n , где модель $b(x)$ выдаёт вероятность около p :

$$\sum_{i=1}^n L(y_i, b(x_i)) = \sum_{i=1}^n L(y_i, p)$$

- А что было бы оптимально выдать на этих объектах?

Предсказание вероятностей

- Рассмотрим ошибку только на объектах x_1, \dots, x_n , где модель $b(x)$ выдаёт вероятность около p :

$$\sum_{i=1}^n L(y_i, b(x_i)) = \sum_{i=1}^n L(y_i, p)$$

- А что было бы оптимально выдать на этих объектах?

$$p_* = \arg \min \sum_{i=1}^n L(y_i, p)$$

- Мы ожидаем, что $p_* = \frac{1}{n} \sum_{i=1}^n [y_i = +1]$

Log-loss

- Рассмотрим ошибку только на объектах, где модель $b(x)$ выдаёт вероятность около p :

$$\sum_{i=1}^n L(y_i, b(x_i)) = \sum_{i=1}^n L(y_i, p)$$

- А что было бы оптимально выдать на этих объектах?

$$p_* = \arg \min \sum_i \{-[y_i = +1] \log p - [y_i = -1] \log(1 - p)\}$$

Log-loss

$$p_* = \arg \min \sum_i \{-[y_i = +1] \log p - [y_i = -1] \log(1 - p)\}$$

- Посчитаем производную по p и приравняем к нулю:

$$\sum_i \left\{ -\frac{[y_i = +1]}{p} + \frac{[y_i = -1]}{1 - p} \right\} = -\frac{n_+}{p} + \frac{n_-}{1 - p} = 0$$

$$p_* = \frac{n_+}{n_+ + n_-} = \frac{1}{n} \sum_{i=1}^n [y_i = +1]$$

Предсказание вероятностей

- Считаем, что модель корректно оценивает вероятности, если для любых $y_1, \dots, y_n \in \mathbb{Y}$

$$\arg \min \sum_{i=1}^n L(y_i, p) = \frac{1}{n} \sum_{i=1}^n [y_i = +1]$$

- Это условие на функцию потерь
- Оно выполнено для log-loss, то есть логистическая регрессия корректно оценивает вероятности
- Значит, для объектов с близкими вероятностями она будет пытаться выдать число, близкое к доле положительных объектов

MSE

$$p_* = \arg \min \sum_{i=1}^n (p - [y_i = +1])^2$$

- Посчитаем производную по p и приравняем к нулю:

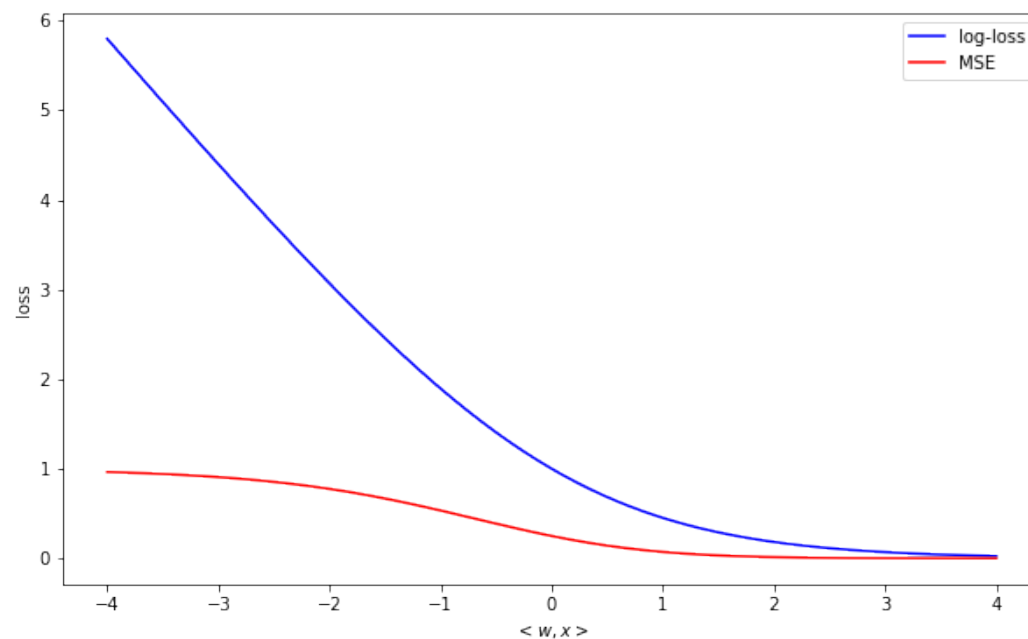
$$2 \sum_{i=1}^n (p - [y_i = +1]) = 0$$

$$p_* = \frac{1}{n} \sum_{i=1}^n [y_i = +1]$$

MSE

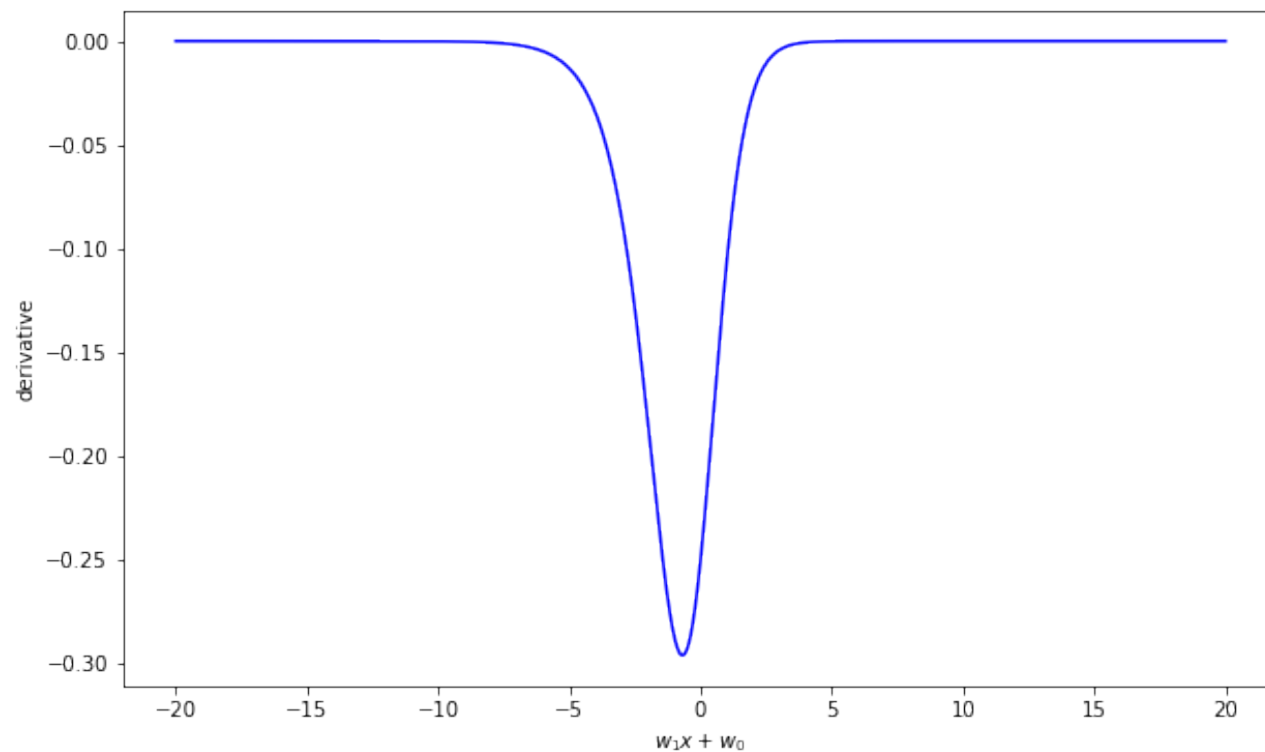
- Почему бы не обучать классификаторы на MSE?

$$\sum_{i=1}^n (\sigma(\langle w, x_i \rangle) - [y_i = +1])^2 \rightarrow \min_w$$



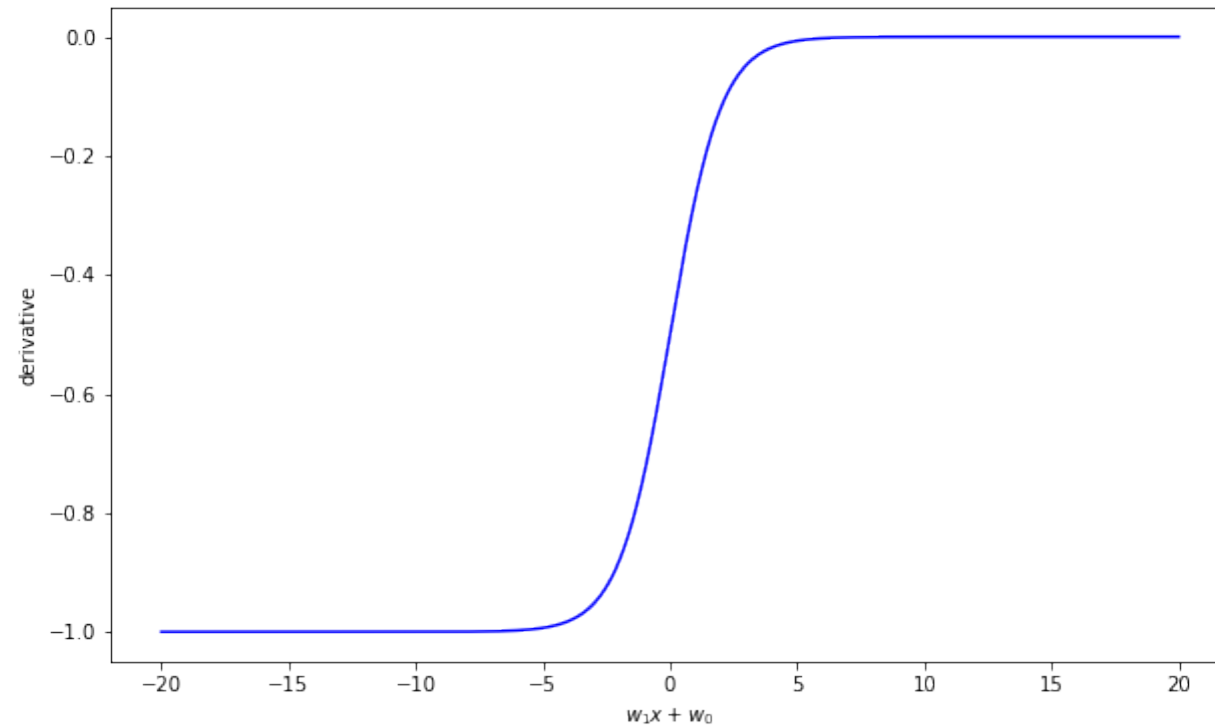
MSE

$$\frac{\partial}{\partial w_1} \left(\frac{1}{1 + e^{-w_1 x - w_0}} - 1 \right)^2 = - \frac{2x e^{w_1 x + w_0}}{(1 + e^{w_1 x + w_0})^3}$$



Log-loss

$$\frac{\partial}{\partial w_1} \left(\log \frac{1}{1 + e^{-w_1 x - w_0}} \right) = \frac{x}{1 + e^{w_1 x + w_0}}$$



MAE

$$p_* = \arg \min \sum_{i=1}^n |p - [y_i = +1]|$$

- Можно показать, что p_* равно либо 0, либо 1