

# Введение в анализ данных

Лекция 5

Логистическая регрессия и SVM

Ковалев Евгений

[ekovalev@hse.ru](mailto:ekovalev@hse.ru)

НИУ ВШЭ, 2020

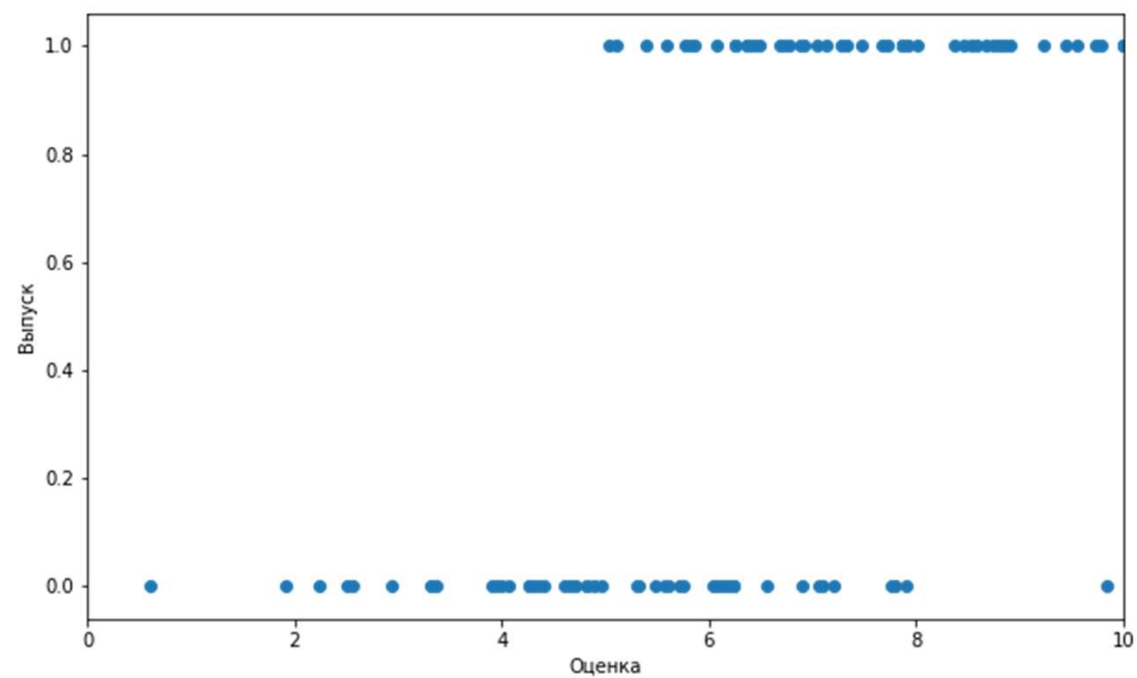
# Логистическая регрессия

# Логистическая регрессия

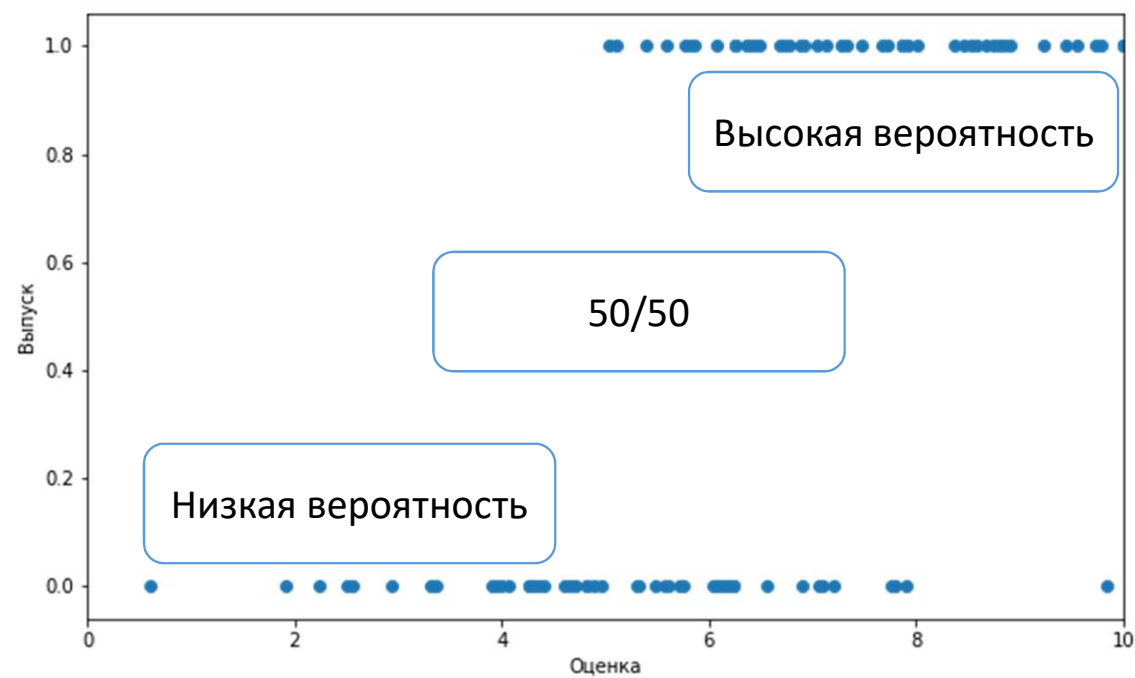
- Решаем задачу бинарной классификации:  $\mathbb{Y} = \{-1, +1\}$
- Минимизация верхней оценки:

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

# 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



# Линейный классификатор

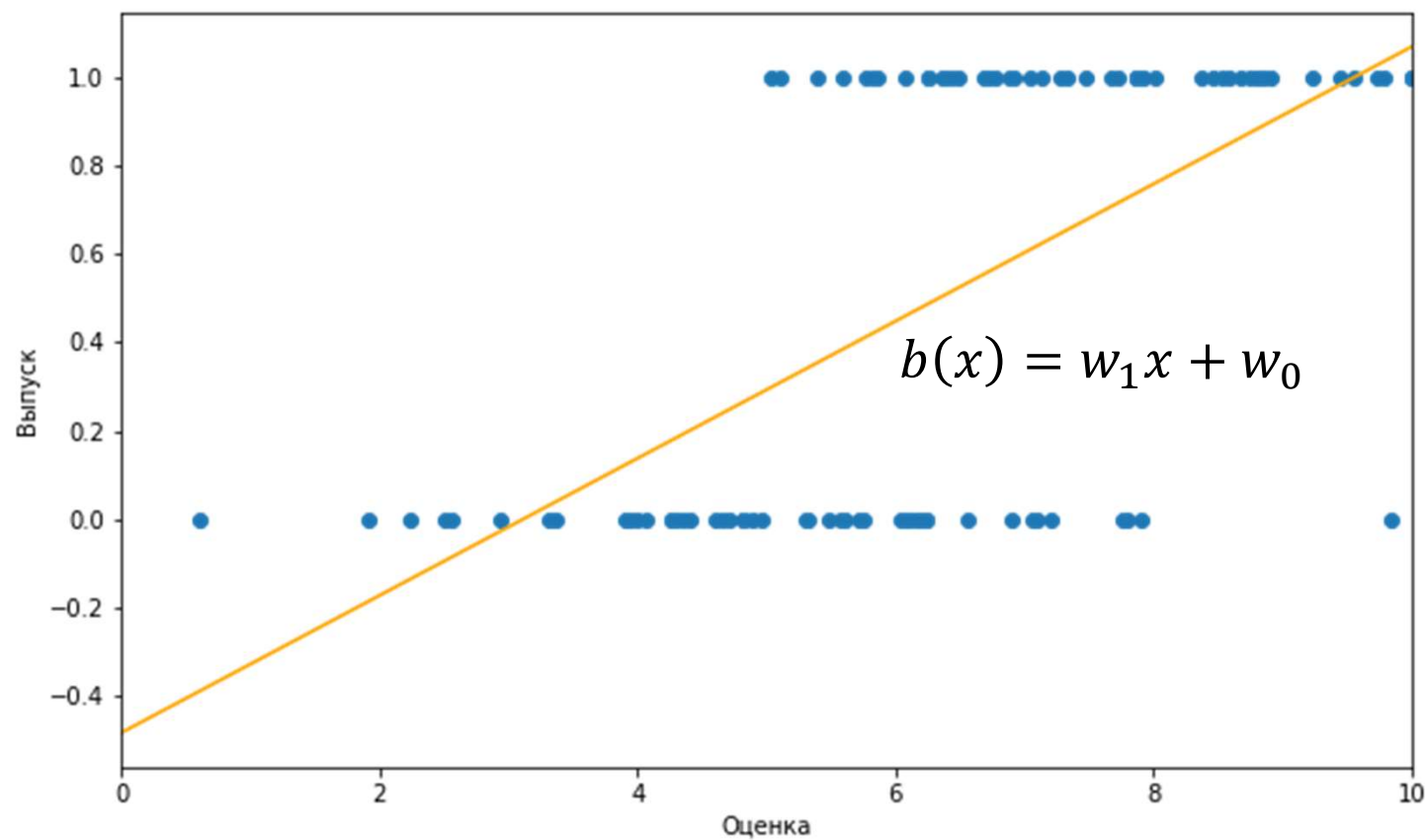
$$a(x) = \text{sign } \langle w, x \rangle$$

- 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

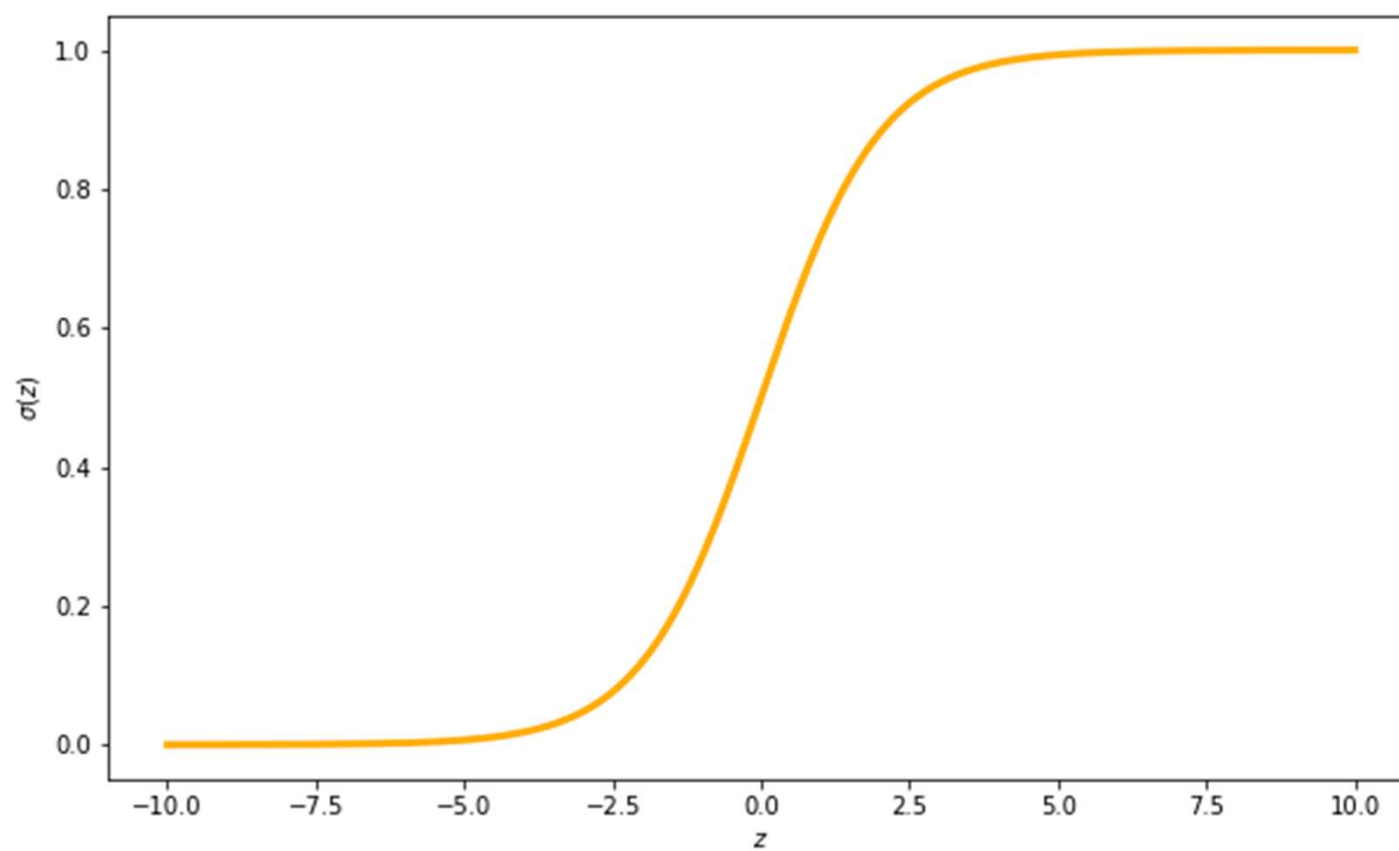


# Линейный классификатор

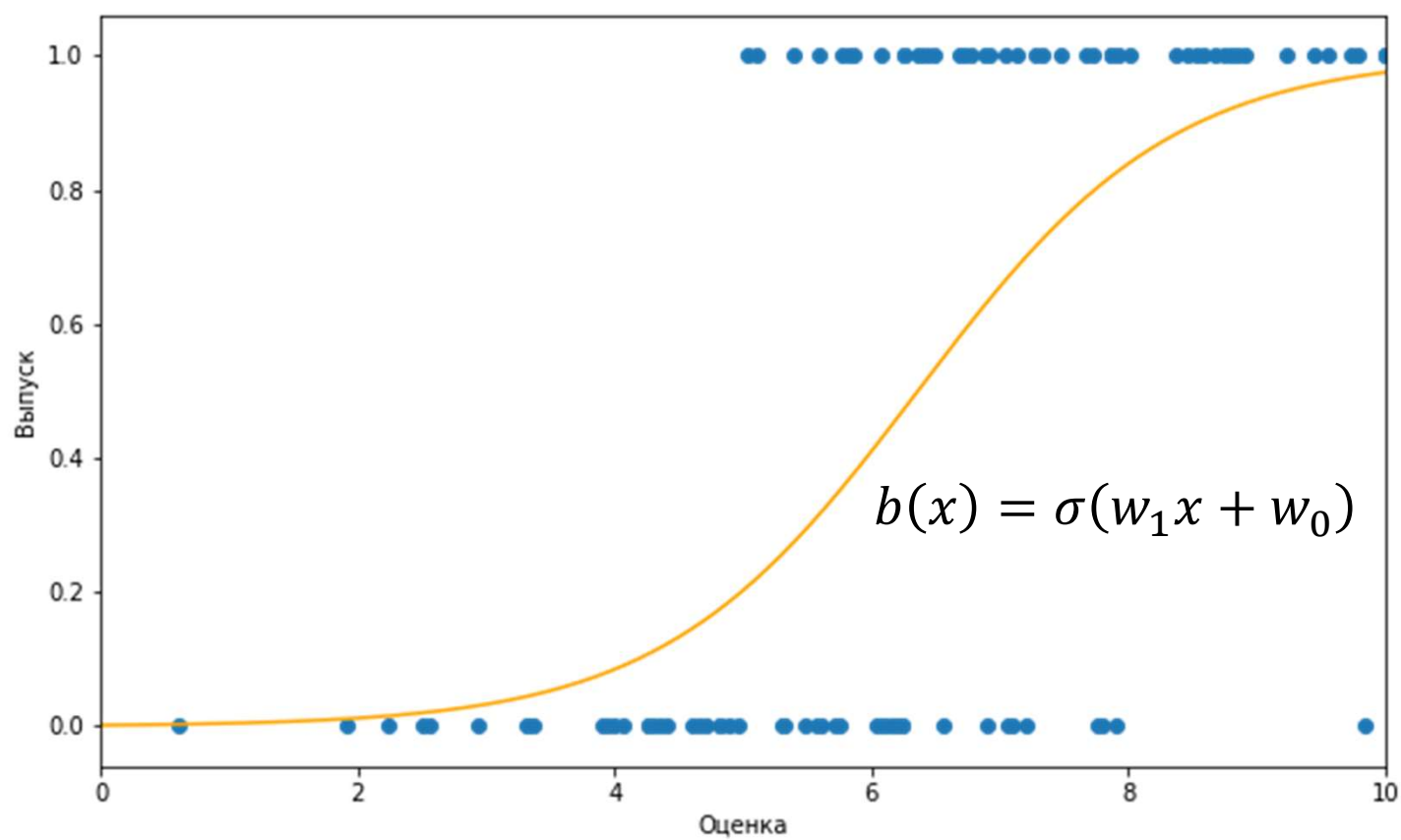
- Переведём выход модели на отрезок  $[0, 1]$
- Например, с помощью сигмоиды:

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

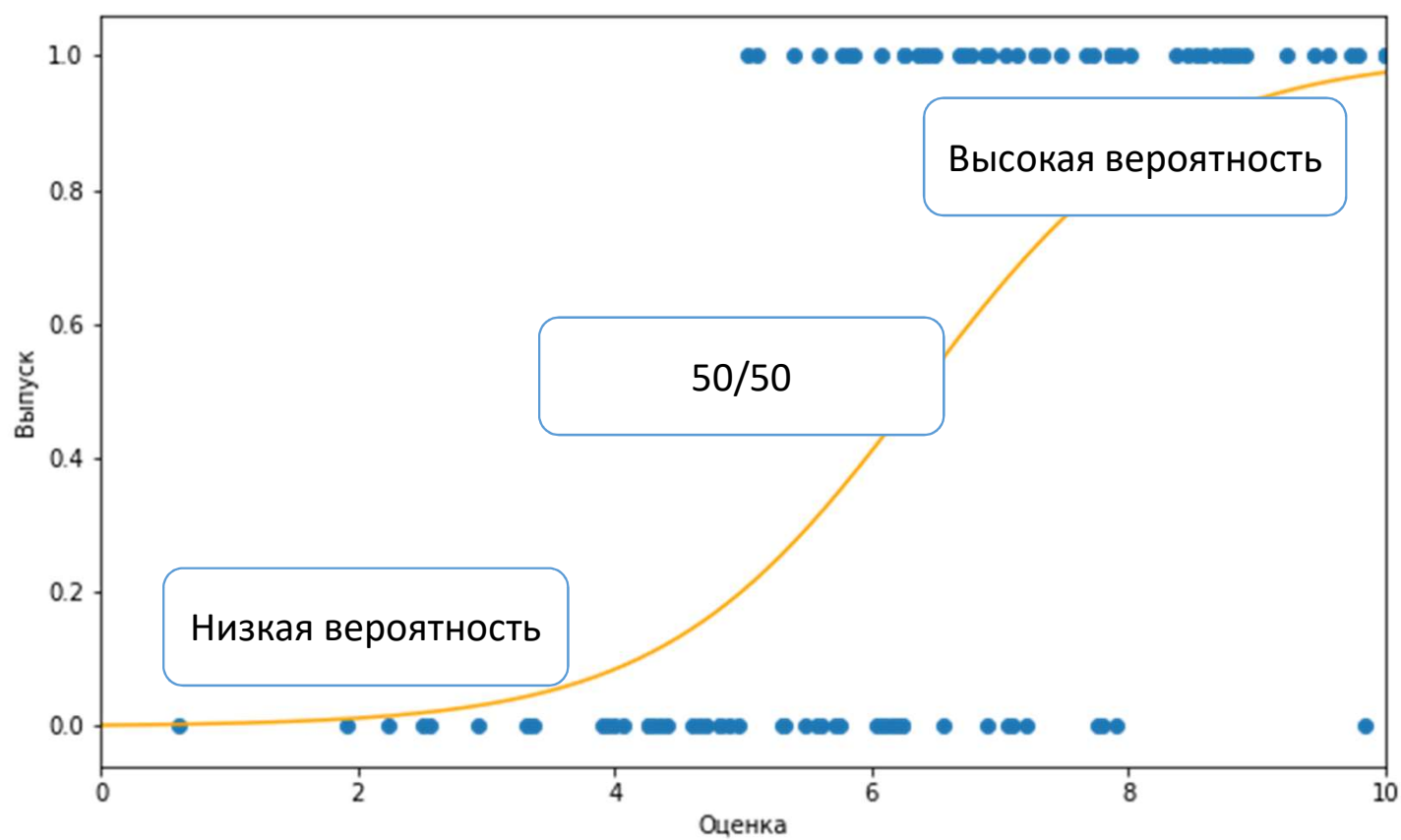
# Сигмоида



# 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
- 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

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

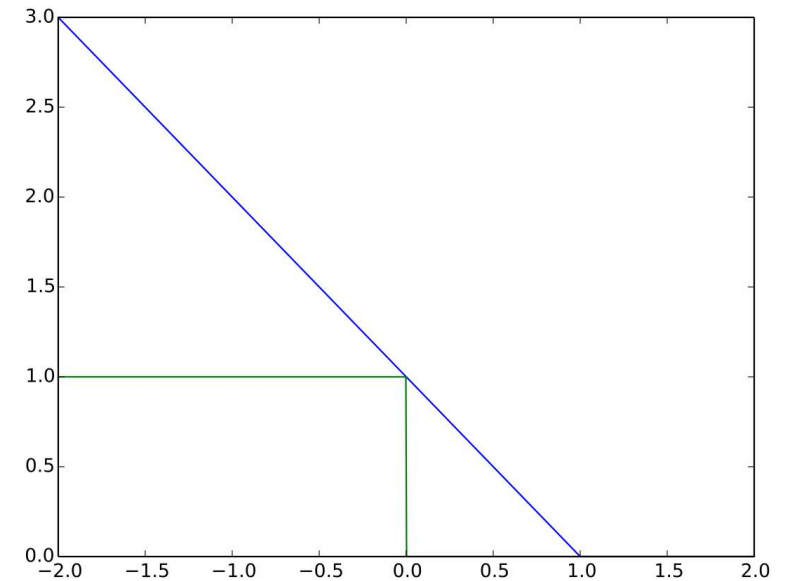


# Метод опорных векторов

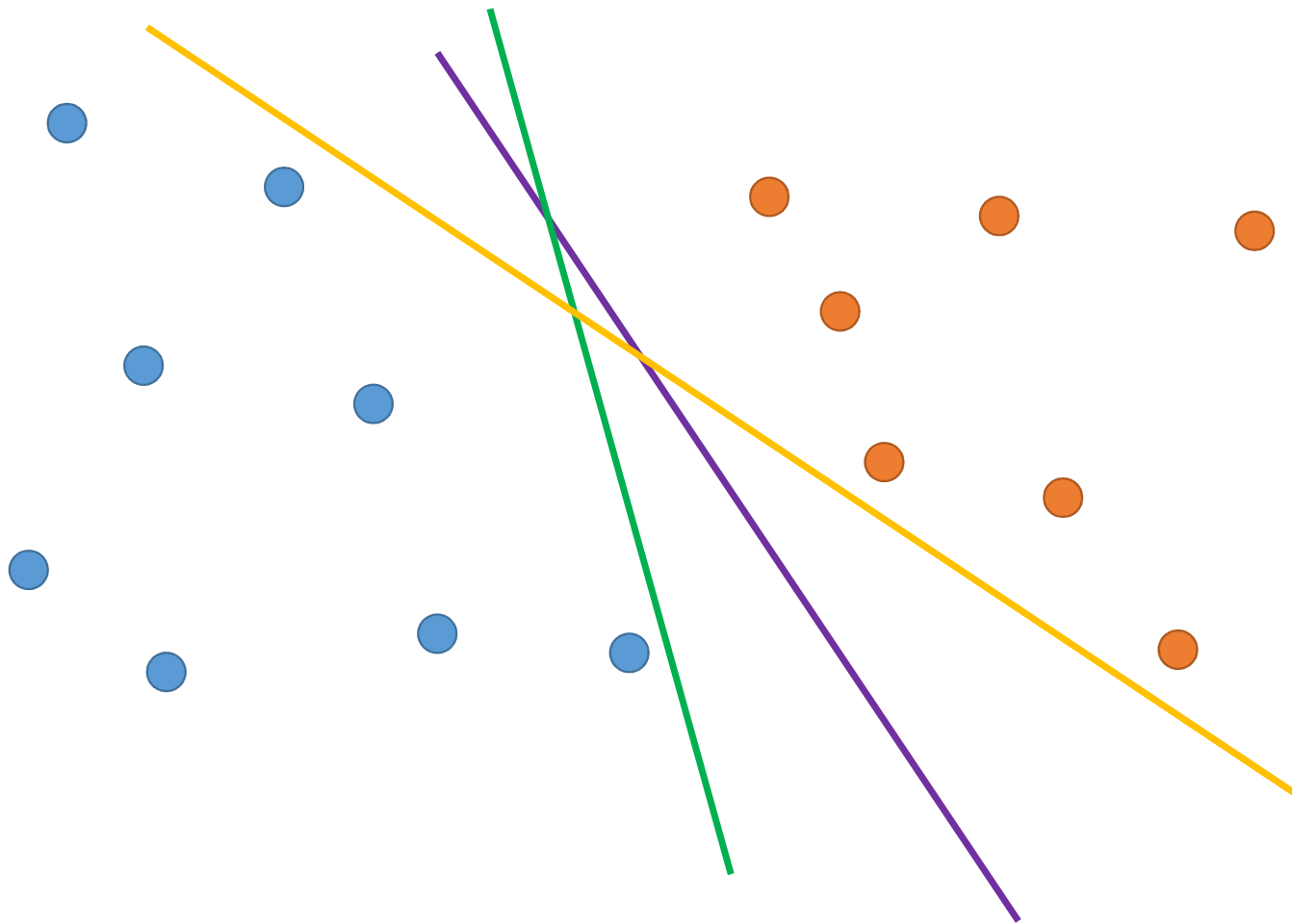
# Hinge loss

- Бинарная классификация:  $\mathbb{Y} = \{-1, +1\}$
- Минимизация верхней оценки:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \max(0, 1 - y_i \langle w, x_i \rangle) \rightarrow \min_w$$



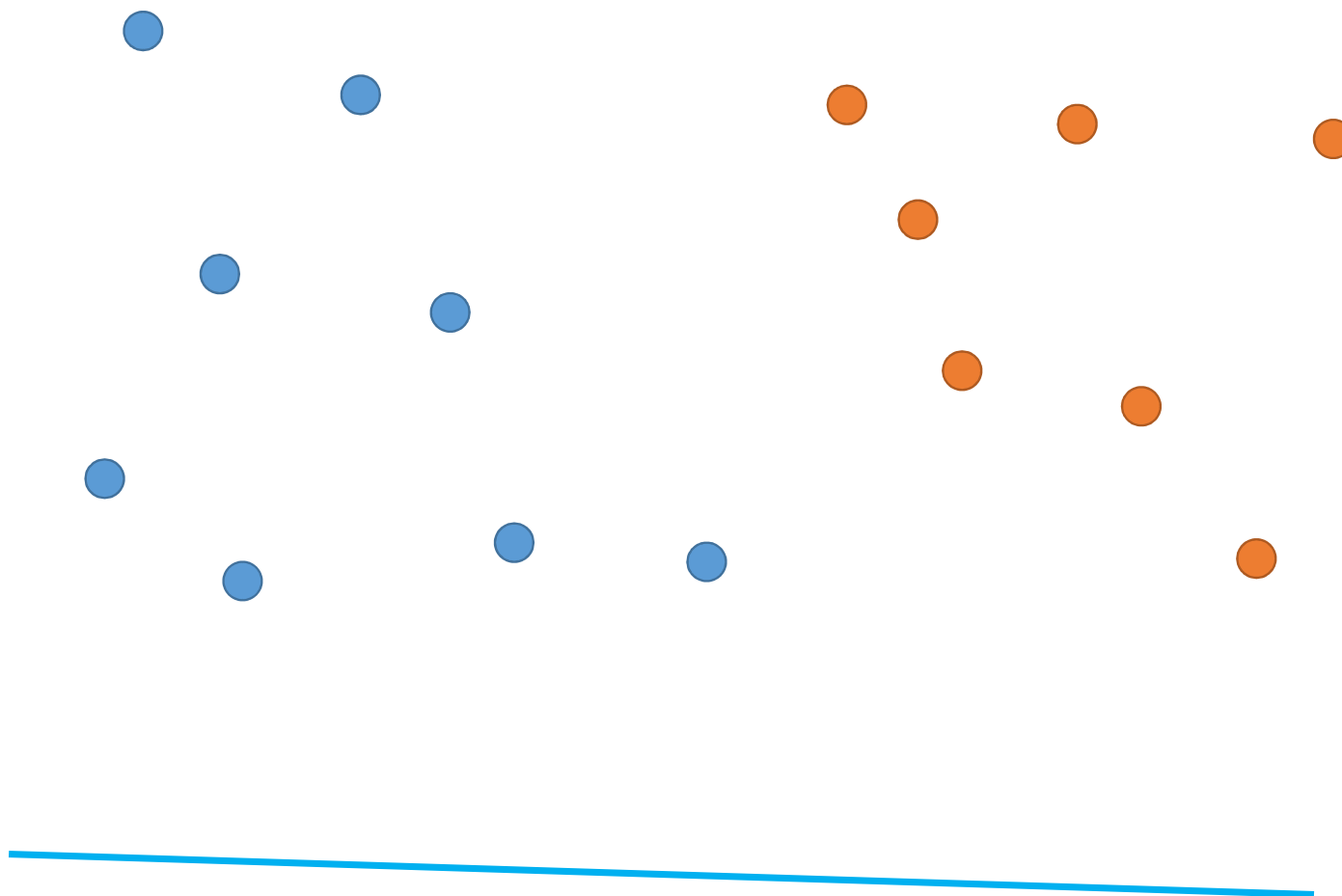
Какой классификатор лучше?



# Отступ классификатора

- Будем максимизировать отступ классификатора — расстояние от гиперплоскости до ближайшего объекта

# Отступ классификатора



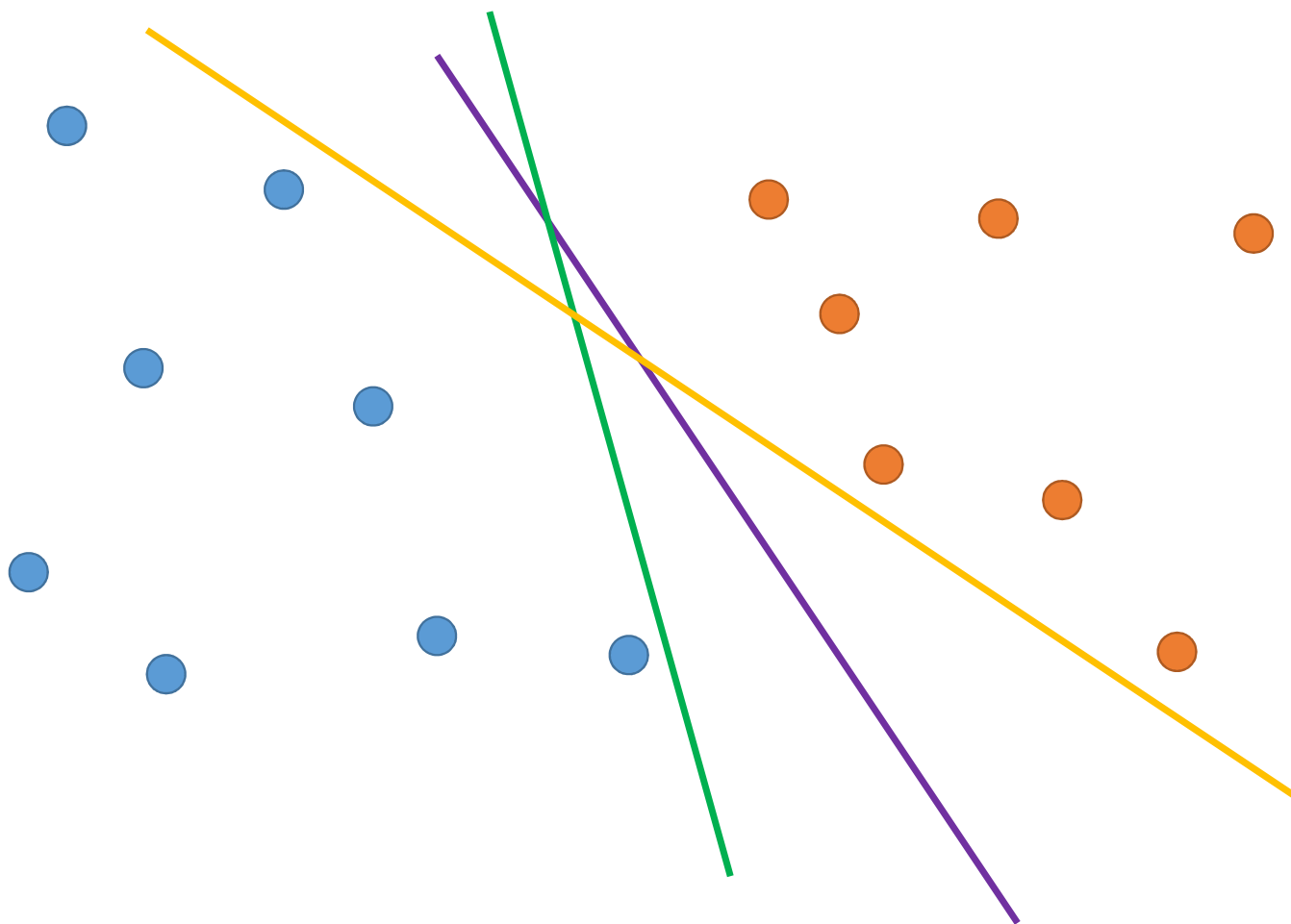
# Отступ классификатора

- Будем максимизировать отступ классификатора — расстояние от гиперплоскости до ближайшего объекта
- При этом будет стараться сделать поменьше ошибок
- По сути, делаем как можно меньше предположений о модели, и верим, что это понизит вероятность переобучения

# Простой случай

- Будем считать, что выборка линейно разделима
- Существует линейный классификатор, не допускающий ни одной ошибки

# Линейно разделимый случай





# Линейно разделимый случай

- **Требование 1:**  $y_i(\langle w, x_i \rangle + w_0) > 0$  для всех  $i = 1, \dots, \ell$
- **Требование 2:** максимальный отступ классификатора

# Отступ классификатора

- Расстояние от точки до гиперплоскости  $\langle w, x \rangle + w_0 = 0$ :

$$\frac{|\langle w, x \rangle + w_0|}{\|w\|}$$

- Отступ классификатора:

$$\min_{i=1, \dots, \ell} \frac{|\langle w, x_i \rangle + w_0|}{\|w\|}$$

# Небольшое предположение

- Линейный классификатор:

$$a(x) = \text{sign} (\langle w, x_i \rangle + w_0)$$

- Если мы поделим  $w$  и  $w_0$  на число  $k > 0$ , то выходы классификатора никак не поменяются:

$$a(x) = \text{sign} \left( \frac{\langle w, x_i \rangle + w_0}{k} \right) = \text{sign} (\langle w, x_i \rangle + w_0)$$

# Небольшое предположение

- Поделим  $w$  и  $w_0$  на  $\min_{i=1,\dots,\ell} |\langle w, x_i \rangle + w_0| > 0$ , после этого будет выполнено

$$\min_{i=1,\dots,\ell} |\langle w, x_i \rangle + w_0| = 1$$

# Отступ классификатора

- Расстояние от точки до гиперплоскости  $\langle w, x \rangle + w_0 = 0$ :

$$\frac{|\langle w, x \rangle + w_0|}{\|w\|}$$

- Отступ классификатора:

$$\min_{i=1, \dots, \ell} \frac{|\langle w, x_i \rangle + w_0|}{\|w\|} = \frac{\min_{i=1, \dots, \ell} |\langle w, x_i \rangle + w_0|}{\|w\|} = \frac{1}{\|w\|}$$

# Линейно разделимый случай

- **Требование 1:**  $y_i(\langle w, x_i \rangle + w_0) > 0$  для всех  $i = 1, \dots, \ell$
- **Требование 2:** максимальный отступ классификатора

$$\frac{1}{\|w\|} \rightarrow \max_w$$

# Линейно разделимый случай

- **Требование 1:**  $y_i(\langle w, x_i \rangle + w_0) > 0$  для всех  $i = 1, \dots, \ell$
- **Требование 2:** максимальный отступ классификатора

$$\frac{1}{\|w\|} \rightarrow \max_w$$

- При условии, что  $\min_{i=1, \dots, \ell} |\langle w, x_i \rangle + w_0| = 1$

# Линейно разделимый случай

- **Требование 1:**  $y_i(\langle w, x_i \rangle + w_0) > 0$  для всех  $i = 1, \dots, \ell$
- **Требование 2:** максимальный отступ классификатора

$$\frac{1}{\|w\|} \rightarrow \max_w$$

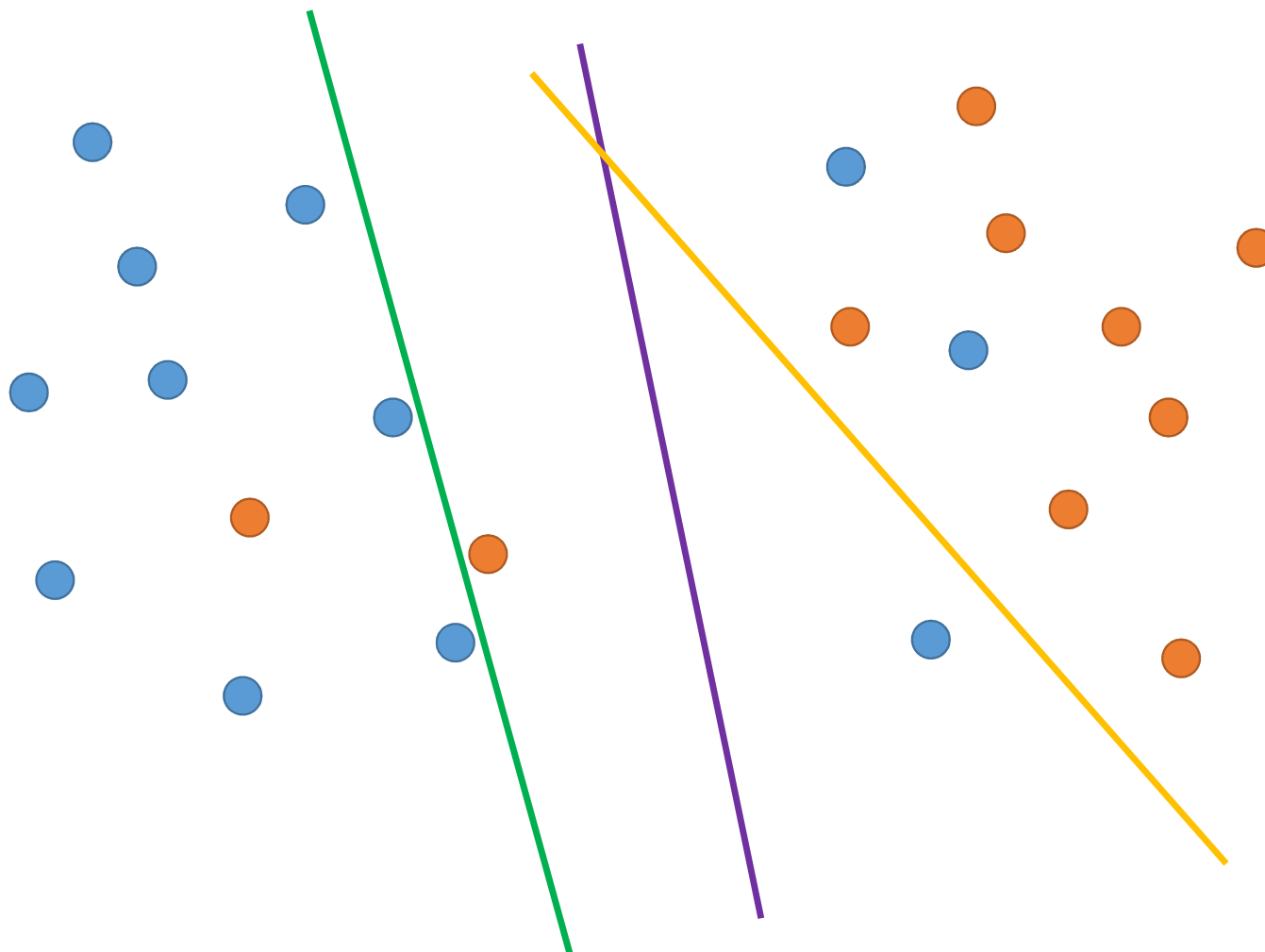
- При условии, что  $|\langle w, x_i \rangle + w_0| \geq 1$
- И мы минимизируем  $\|w\|$  — тогда где-то модуль отступа будет равен 1



# Метод опорных векторов (SVM)

$$\begin{cases} \|w\|^2 \rightarrow \min_{w, w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 \end{cases}$$

# Линейно неразделимый случай



# Линейно неразделимый случай

- Любой линейный классификатор допускает хотя бы одну ошибку

$$\begin{cases} \|w\|^2 \rightarrow \min_{w, w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 \end{cases}$$

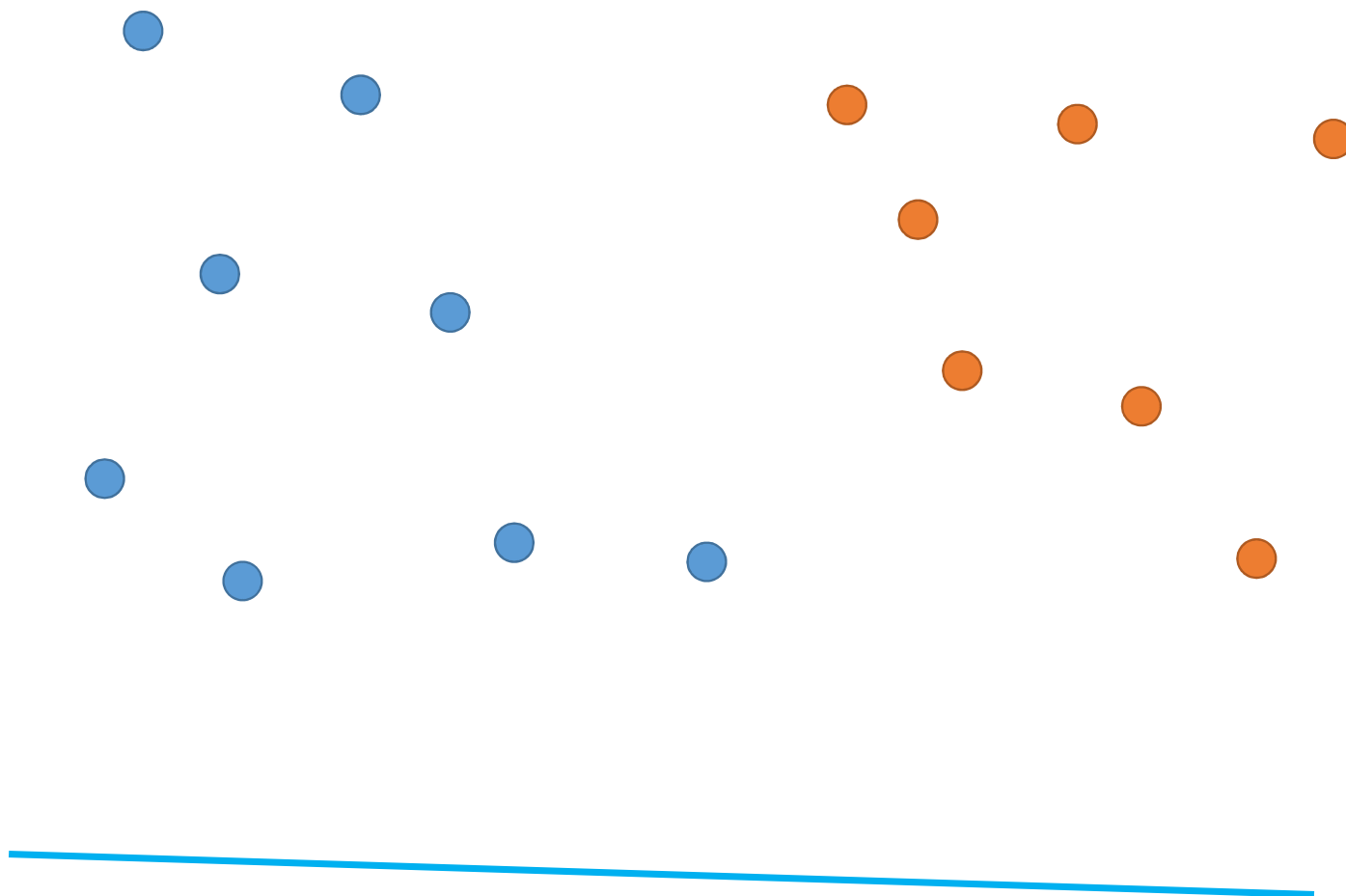
# Линейно неразделимый случай

$$\begin{cases} \|w\|^2 \rightarrow \min_{w, w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

# Линейно неразделимый случай

$$\begin{cases} \|w\|^2 \rightarrow \min_{w, w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - 10^{1000} \end{cases}$$

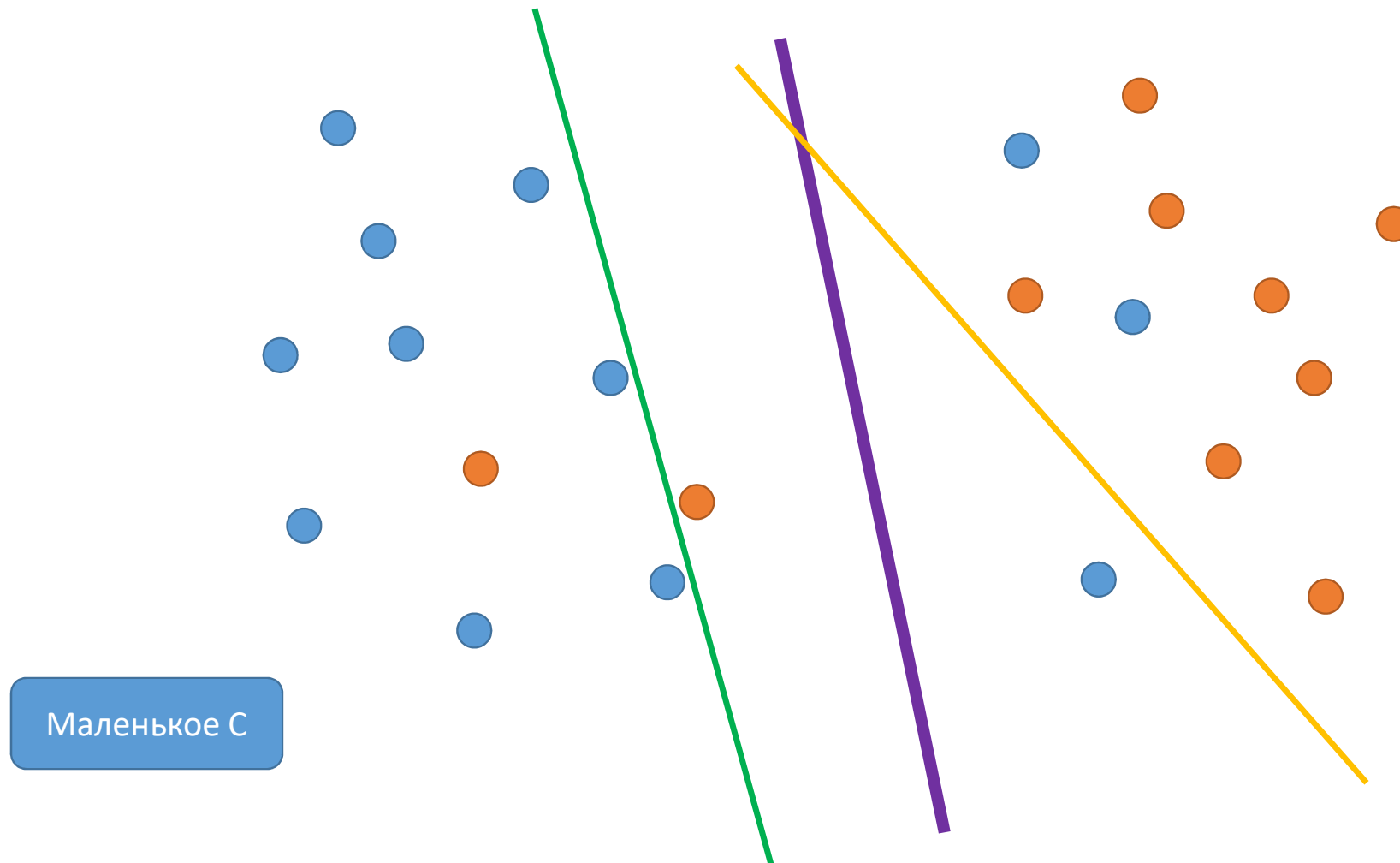
# Отступ классификатора



# Метод опорных векторов

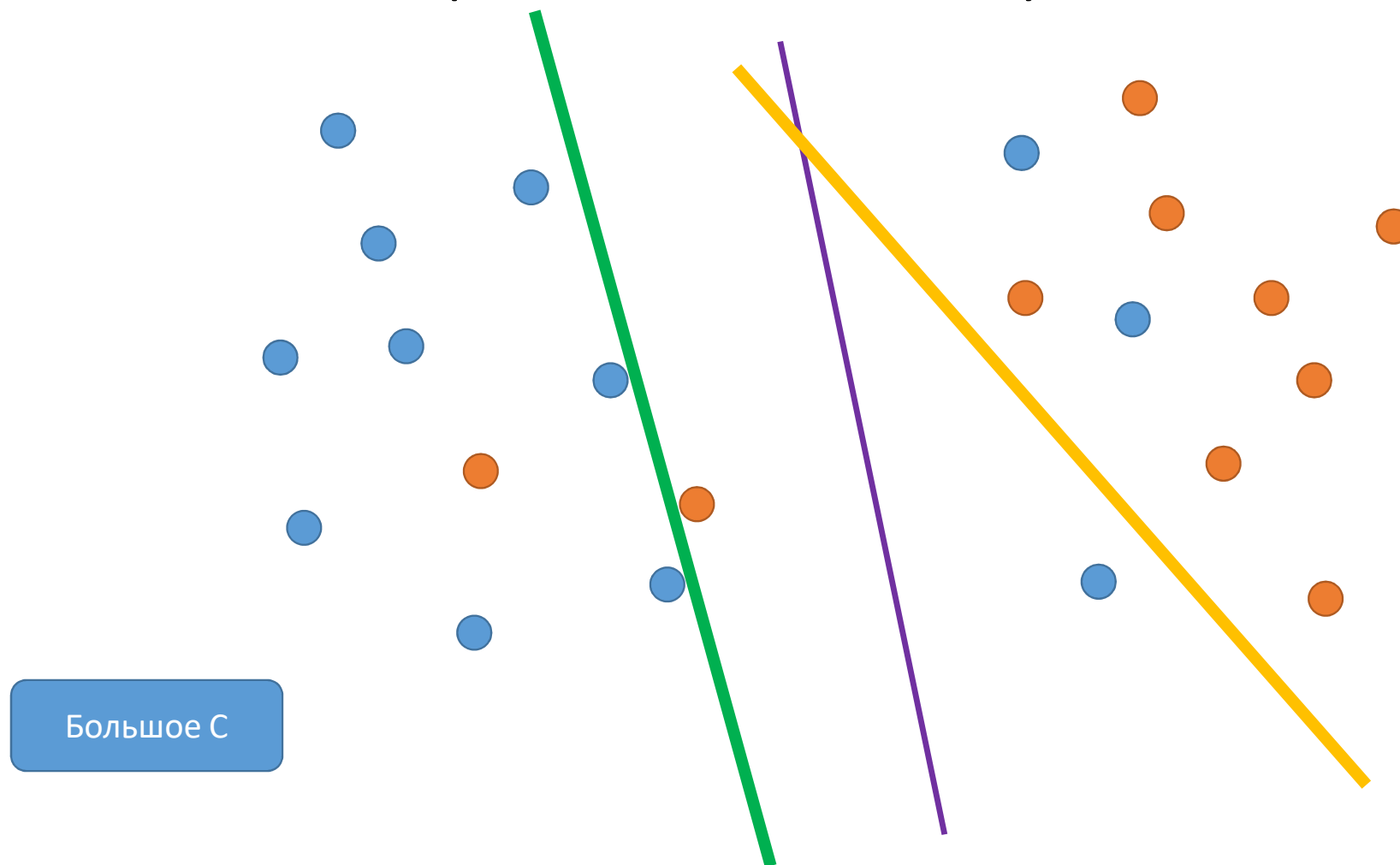
$$\left\{ \begin{array}{l} \|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \rightarrow \min_{w, w_0, \xi_i} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{array} \right.$$

# Линейно неразделимый случай





# Линейно неразделимый случай



# Метод опорных векторов

$$\begin{cases} \|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \rightarrow \min_{w, w_0, \xi_i} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

- Объединим ограничения:

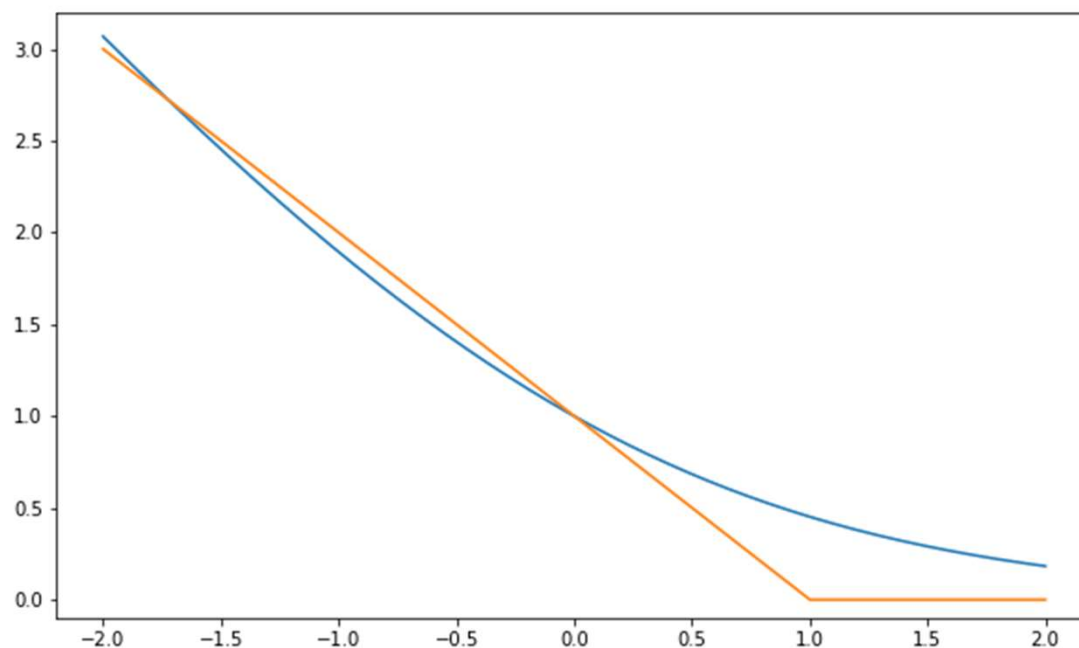
$$\xi_i \geq \max(0, 1 - y_i(\langle w, x_i \rangle + w_0))$$

# Метод опорных векторов

$$C \sum_{i=1}^{\ell} \max(0, 1 - y_i(\langle w, x_i \rangle + w_0)) + \|w\|^2 \rightarrow \min_{w, w_0}$$

- Функция потерь (hinge loss) + регуляризация

# Сравнение логистической регрессии и SVM



# Резюме

- Логистическая регрессия минимизирует логистические потери
- Метод опорных векторов основан на идее максимизации отступа классификатора