Машинное обучение

Лекция 5

Логистическая регрессия и SVM

План лекции

- Бинарная классификация
- Предсказание вероятностей
- Логистическая регрессия
- Метод опорных векторов

Линейная классификация

- Решаем задачу бинарной классификации:

$$\mathbb{Y} = \{-1, +1\}$$

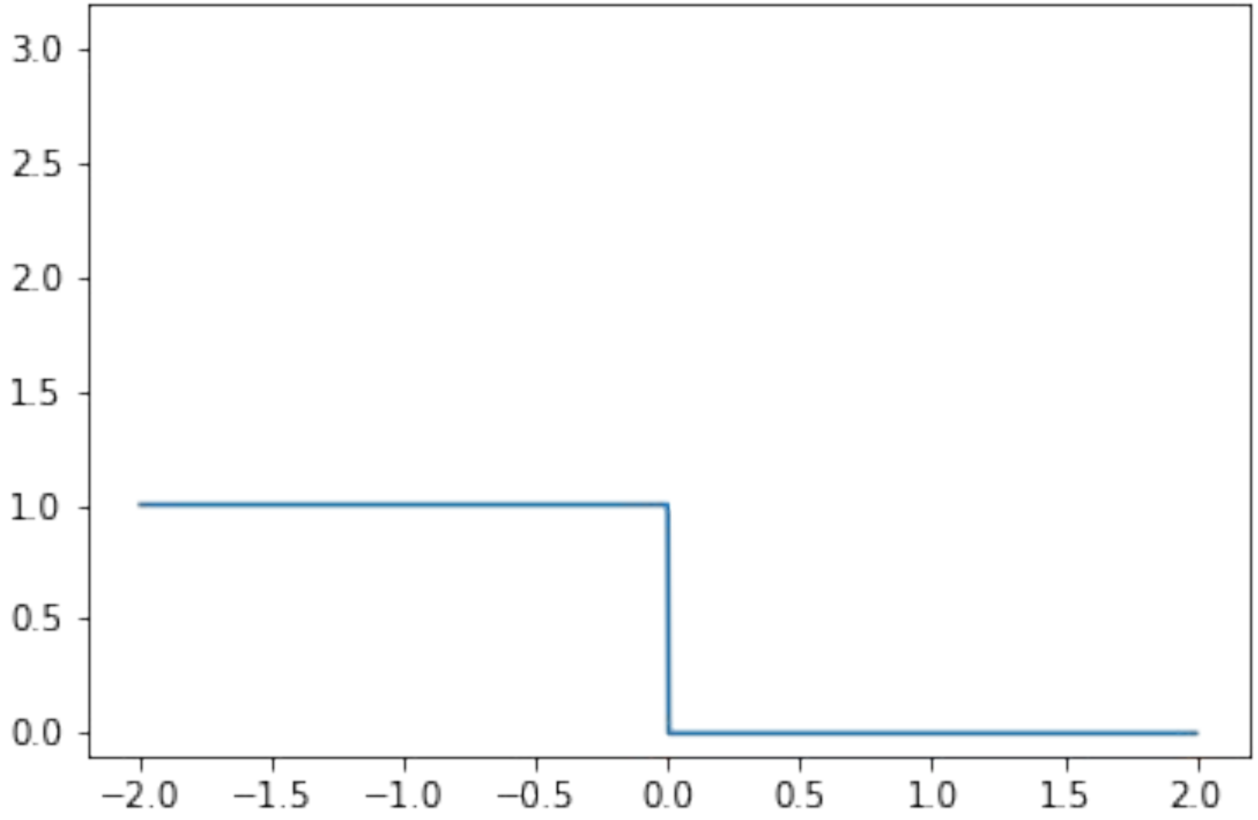
- Линейная модель:

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

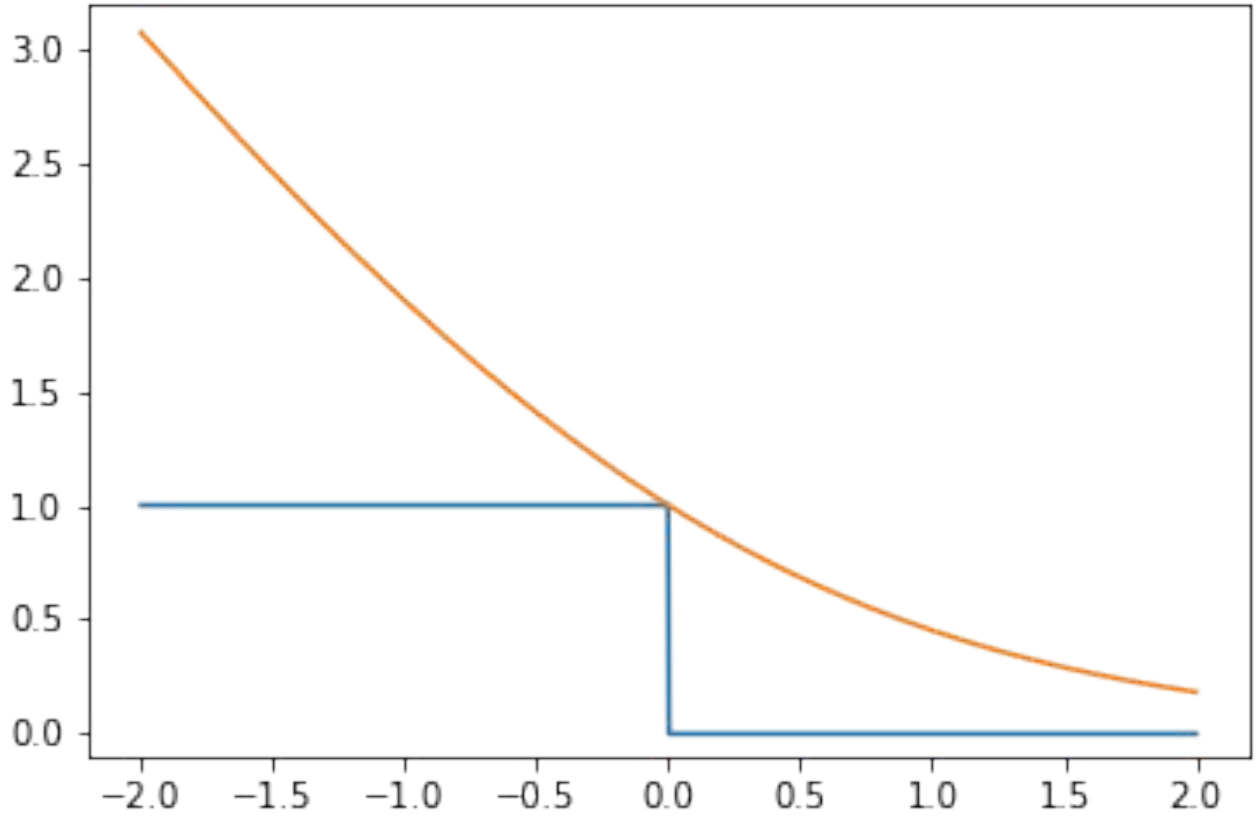
- Функция потерь:

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

Линейная классификация

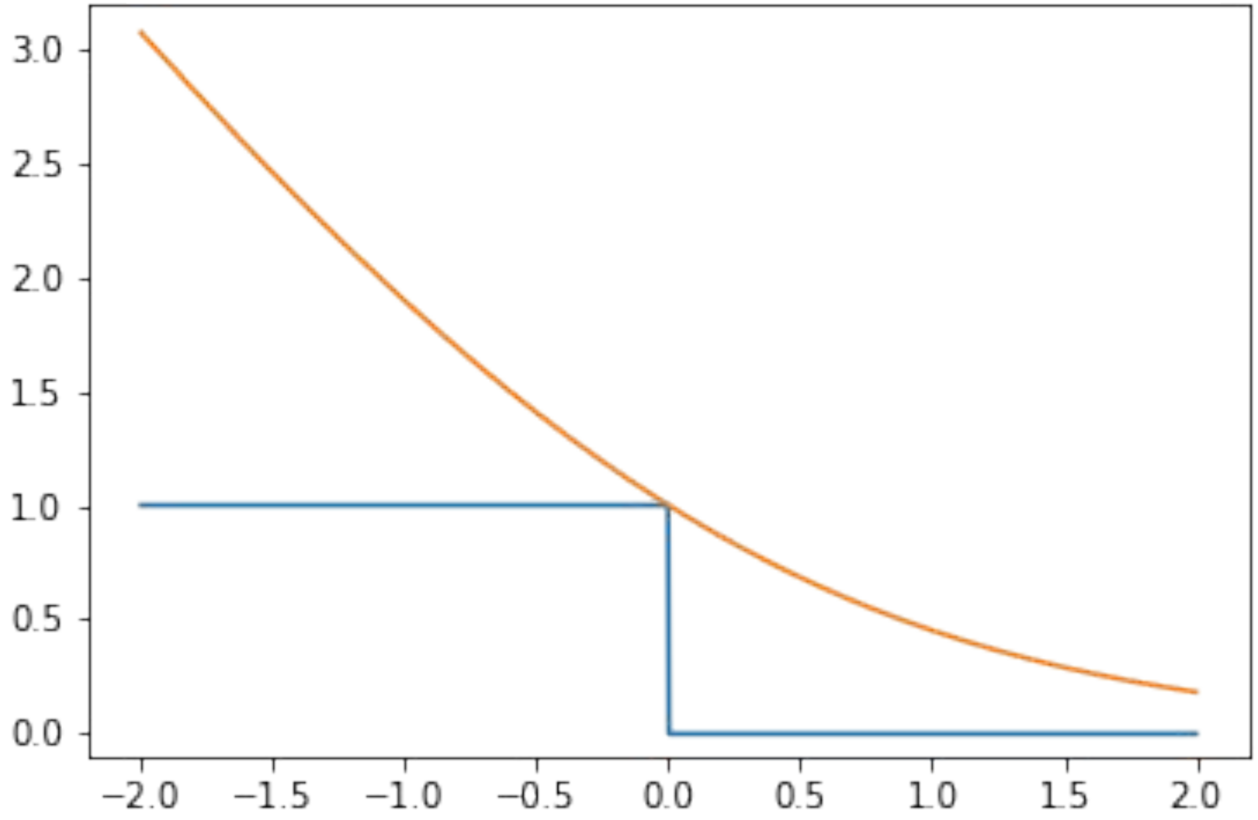


Линейная классификация



Линейная классификация

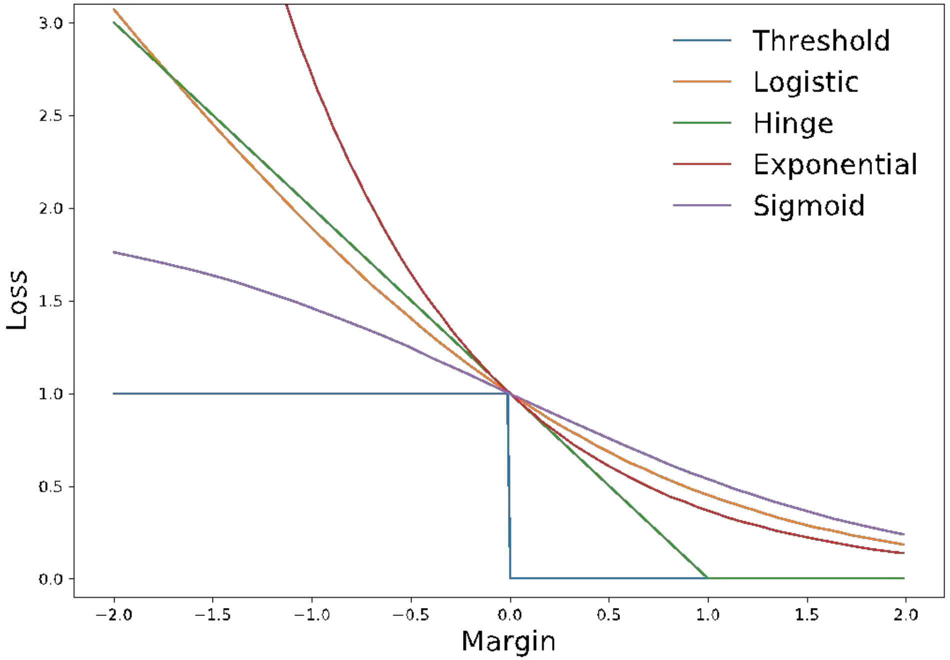
$$0 \le \frac{1}{\ell}\sum_{i=1}^{\ell}[y_i\langle w, x_i\rangle < 0] \le \frac{1}{\ell}\sum_{i=1}^{\ell}\tilde{L}(y_i\langle w, x_i\rangle) \to \min_w$$



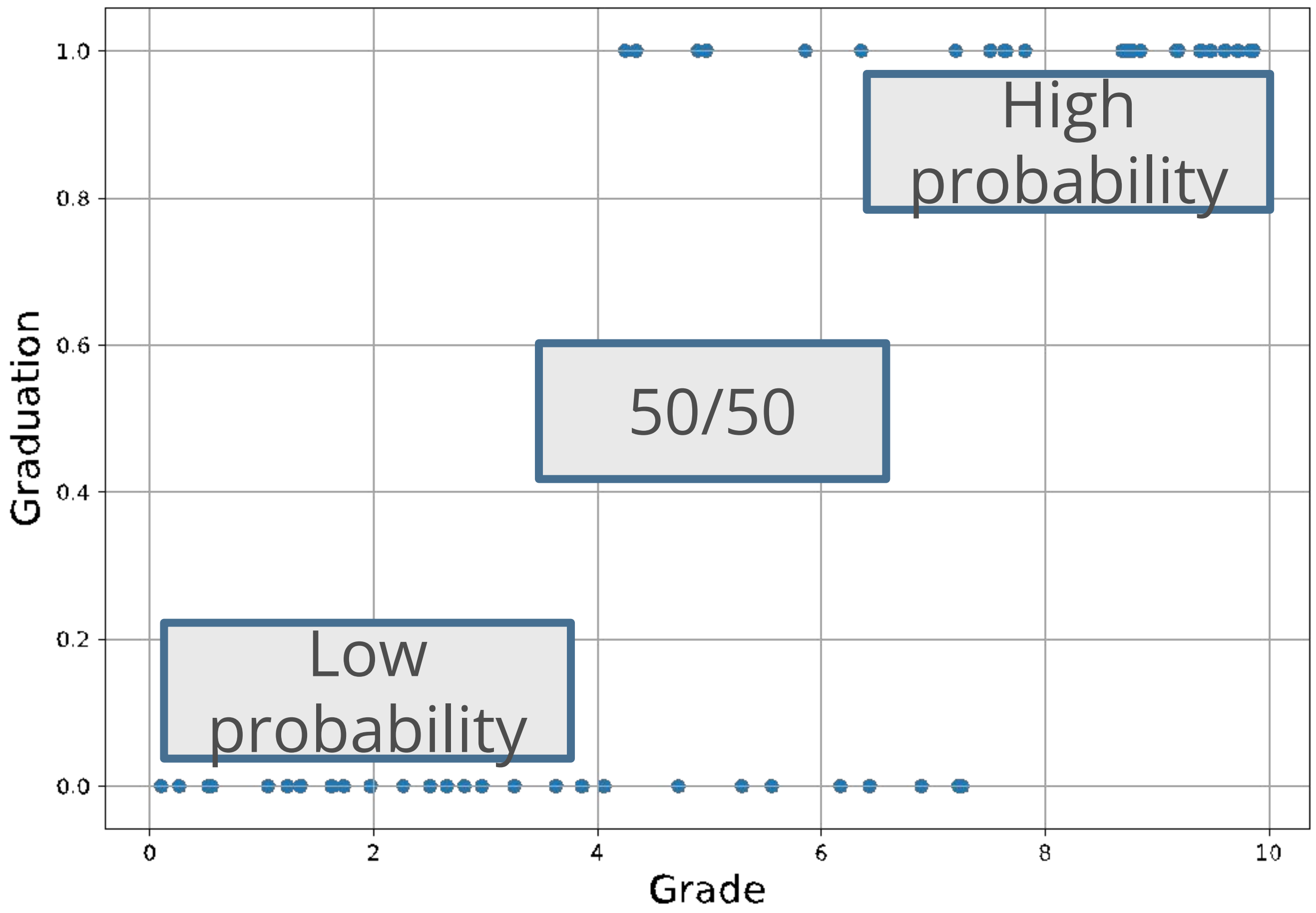
Линейная классификация

- Разные верхние оценки => разные модели

- Свойства модели => верхняя оценка



Предсказание вероятностей



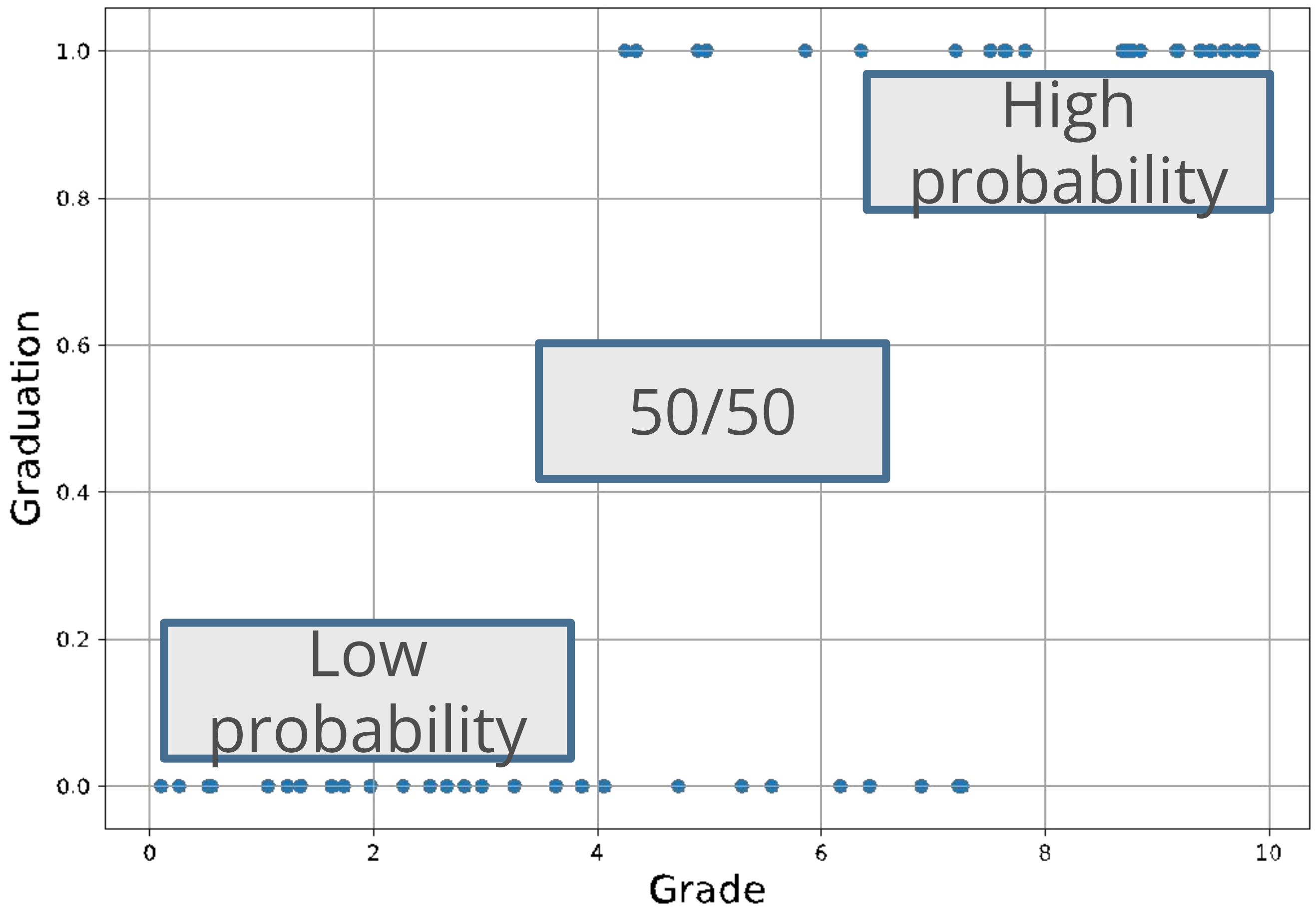
Предсказание вероятностей

- Кредитный скоринг
- Стратегия: выдавать кредит только клиентам с вероятностью возврата > 0.9
- 10% невозвращенных кредитов - нормально

Предсказание вероятностей

- Реклама в интернете:
  - $b(x)$ — вероятность клика
  - $c(x)$ — прибыль в случае клика
  - $c(x)b(x)$— хотим оптимизировать

Предсказание вероятностей



Логистическая регрессия

Логистическая регрессия

- Решаем задачу бинарной классификации:

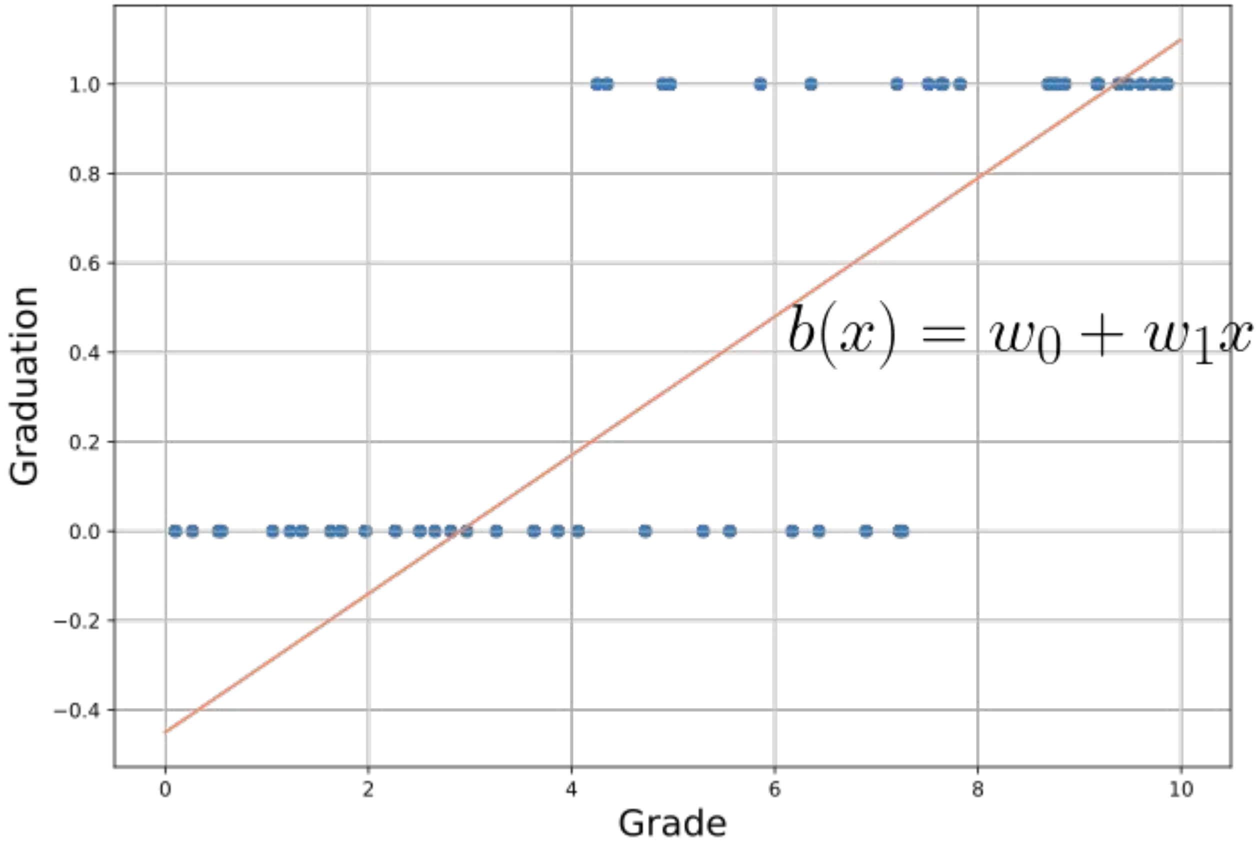
$$\mathbb{Y} = \{-1, +1\}$$

- Линейная модель:

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

- Может использовать b(x)=⟨w, x⟩ как оценку вероятности?

Логистическая регрессия



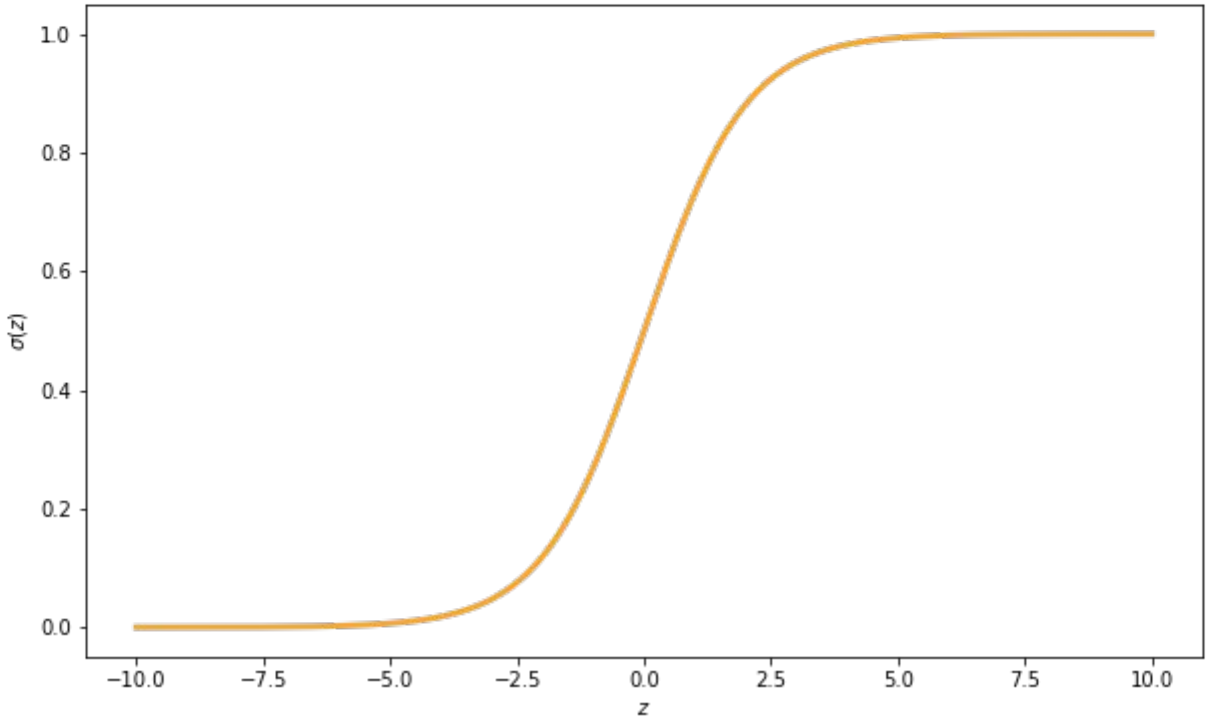
Логистическая регрессия

- Давайте переведем выход модели на отрезок [0, 1]

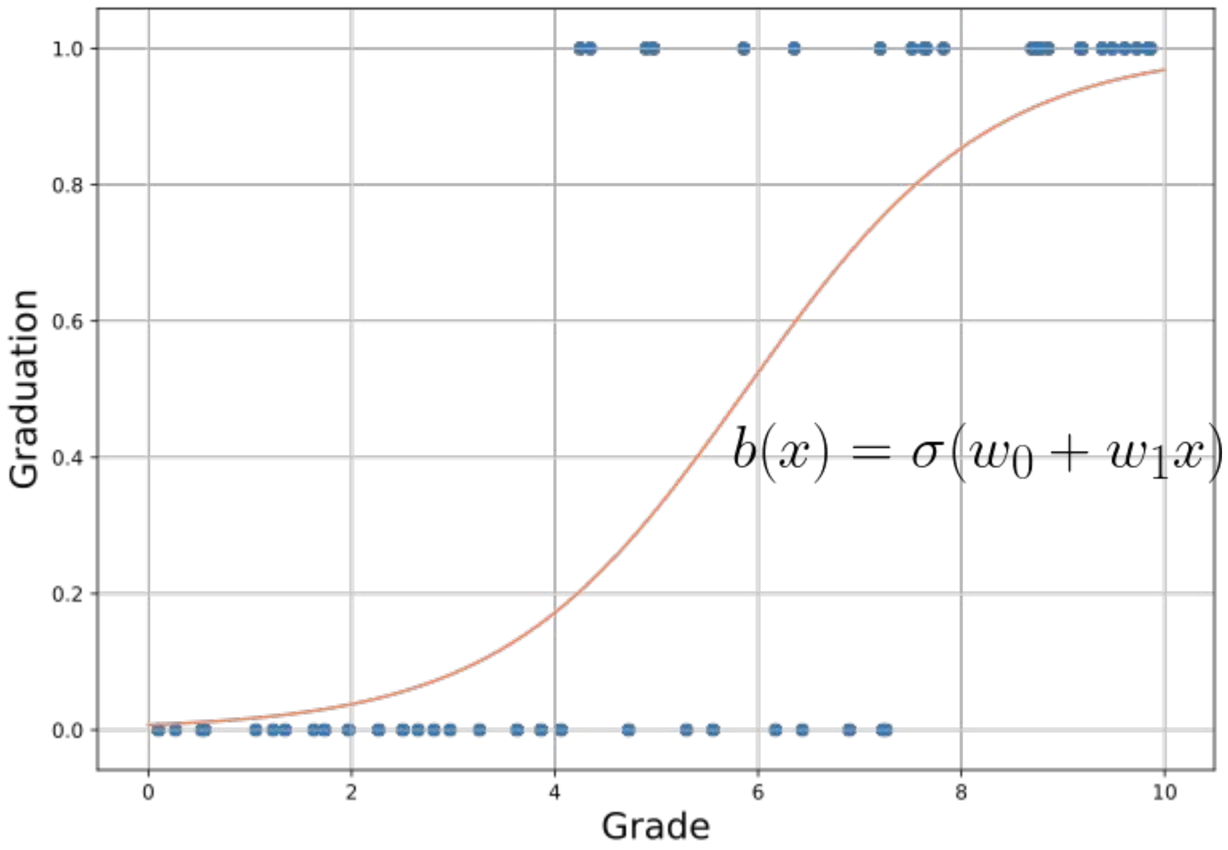
- Например, с помощью сигмоиды

$$\sigma(\langle w, x\rangle) = 1/(1+\exp(-\langle w, x\rangle))$$

Сигмоида



Предсказание вероятностей



Логистическая регрессия

- Задача бинарной классификации: $\mathbb{Y}=\{-1, +1\}$

- Предсказание вероятностей

$$P(y\_i=1)=\sigma(\langle w, x\_\mathrm{i}\rangle)$$

Теперь мы можем использовать метод максимального правдоподобия

Логистическая регрессия

Логистическая регрессия

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

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

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

$$\sum_{i=1}^{\ell}\{[y_i = 1]\log(1 + \exp(-\langle w, x\rangle)) + [y_i = -1]\log(1 + \exp(\langle w, x\rangle))\} =$$

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

Логистическая регрессия

- Немало задач, где нужно предсказывать вероятности

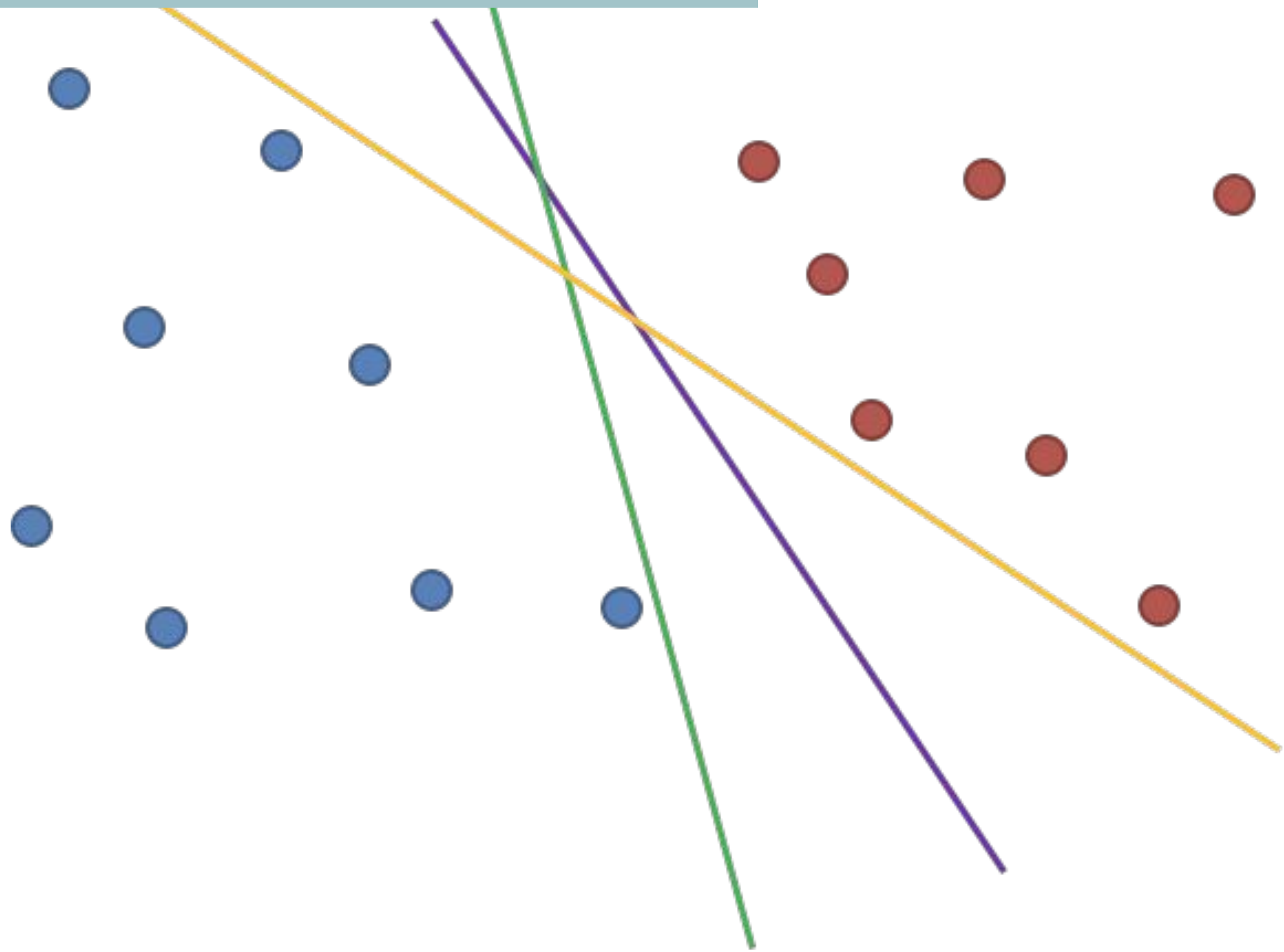
- Можно применить сигмоиду к линейной модели, чтобы получать числа от 0 до 1

- Логистическая функция максимизирует отступы — то есть приближает вероятности к 1 или 0 в зависимости от класса

Логистическая регрессия и друзья

Метод опорных векторов

Какой классификатор лучше?



Отступ классификатора

Вспомним, что линейный классификатор задает гиперплоскость

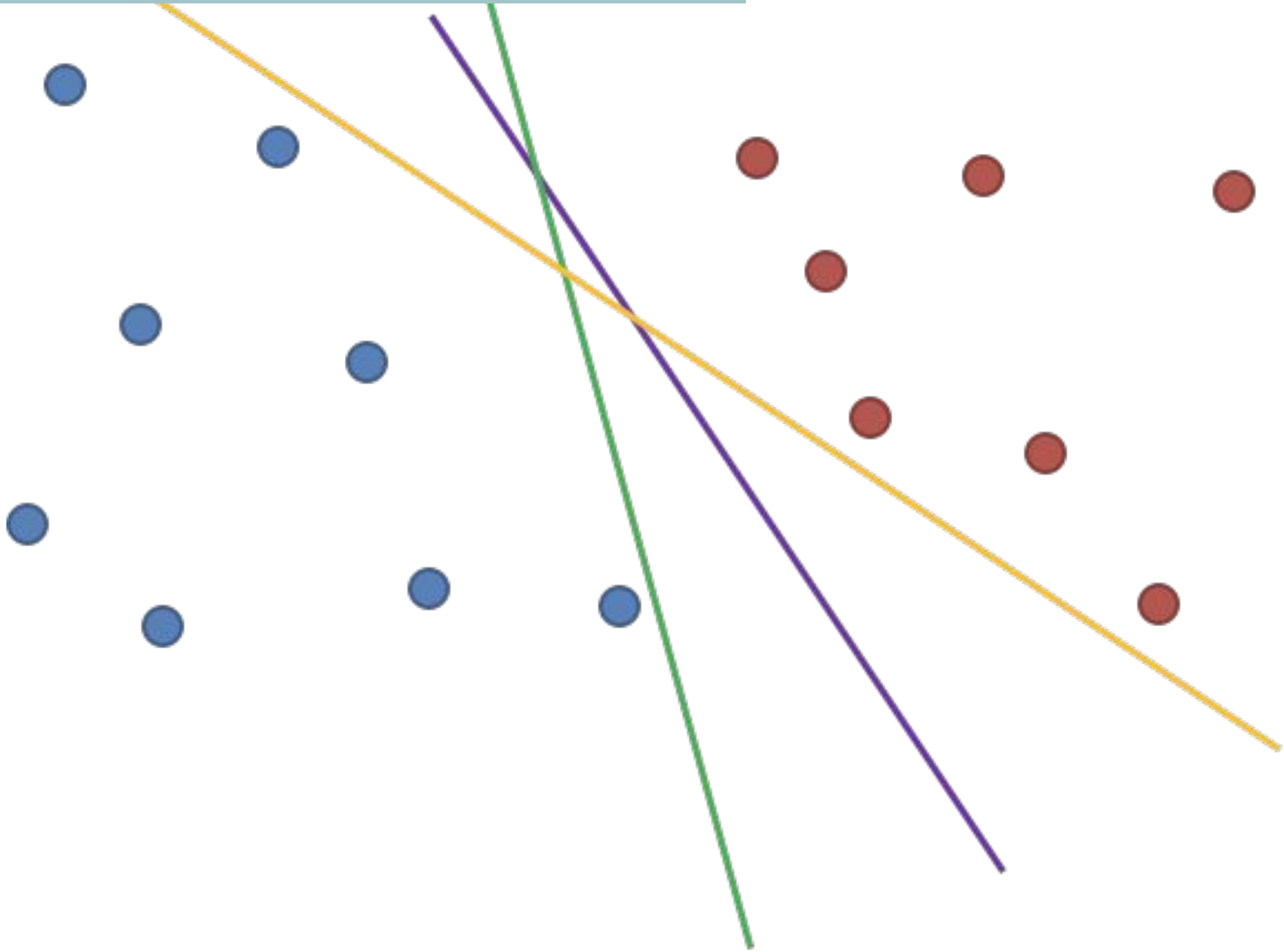
$$\langle w, x\rangle = 0$$

Расстояние от любой точки x до этой гиперплоскости

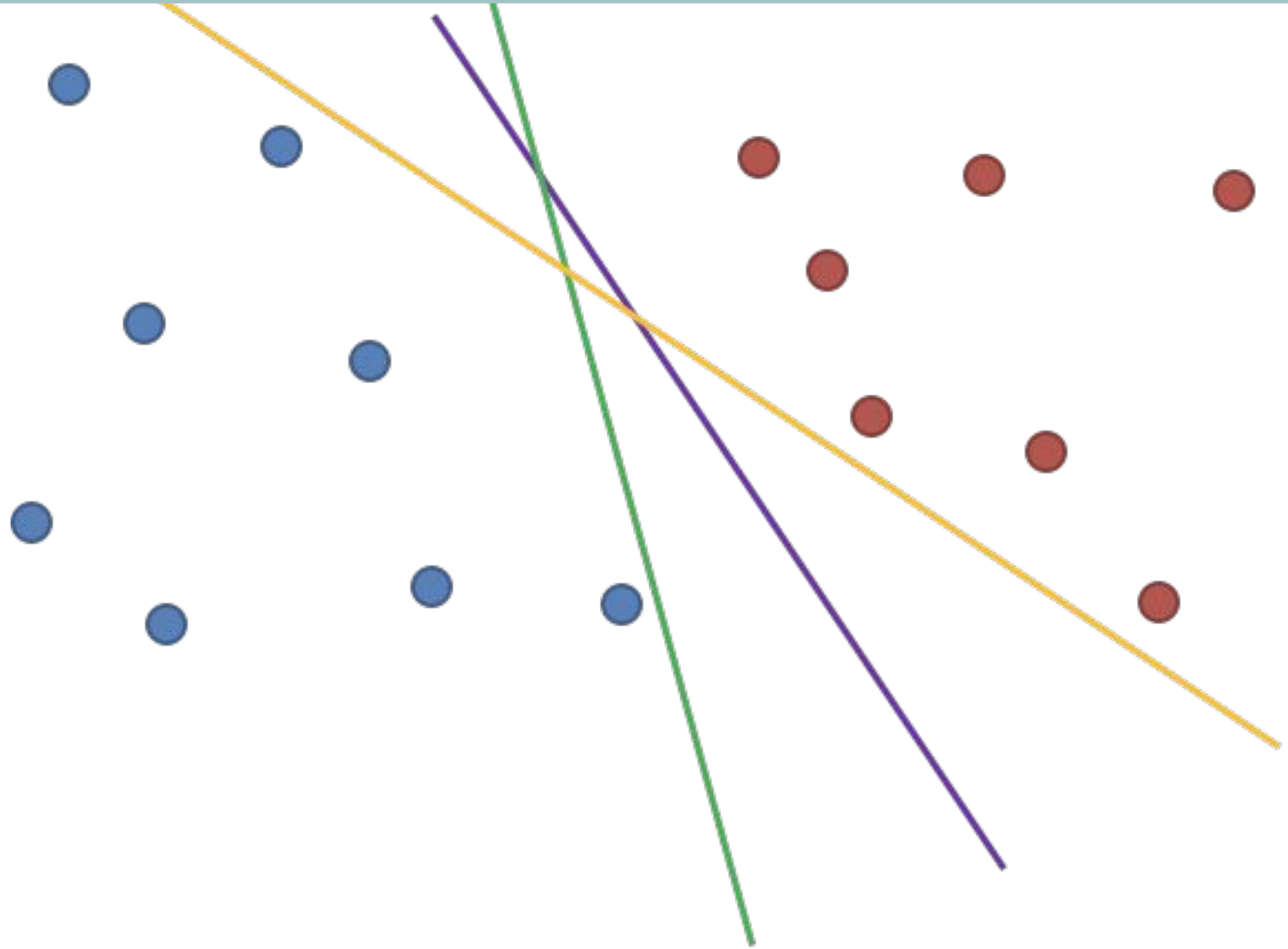
$$\frac{|\langle w, x\rangle|}{\|w\|}$$

Отступ классификатора — расстояние от гиперплоскости до ближайшего объекта

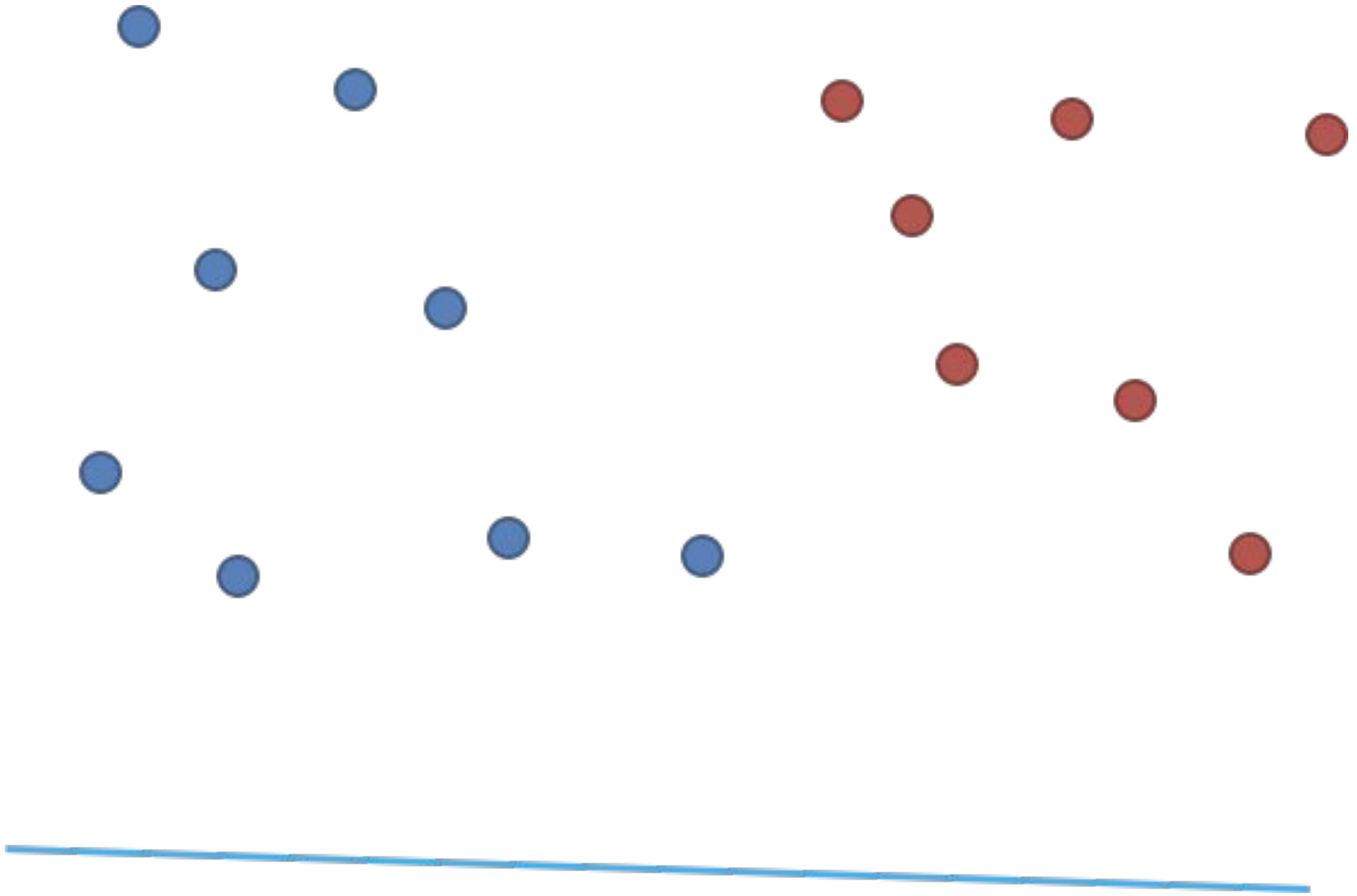
Какой классификатор лучше?



Какой классификатор ~~лучше?~~ имеет больший отступ?



Какой классификатор ~~лучше?~~ имеет больший отступ?



Отступ классификатора

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

- При этом будет стараться сделать поменьше ошибок

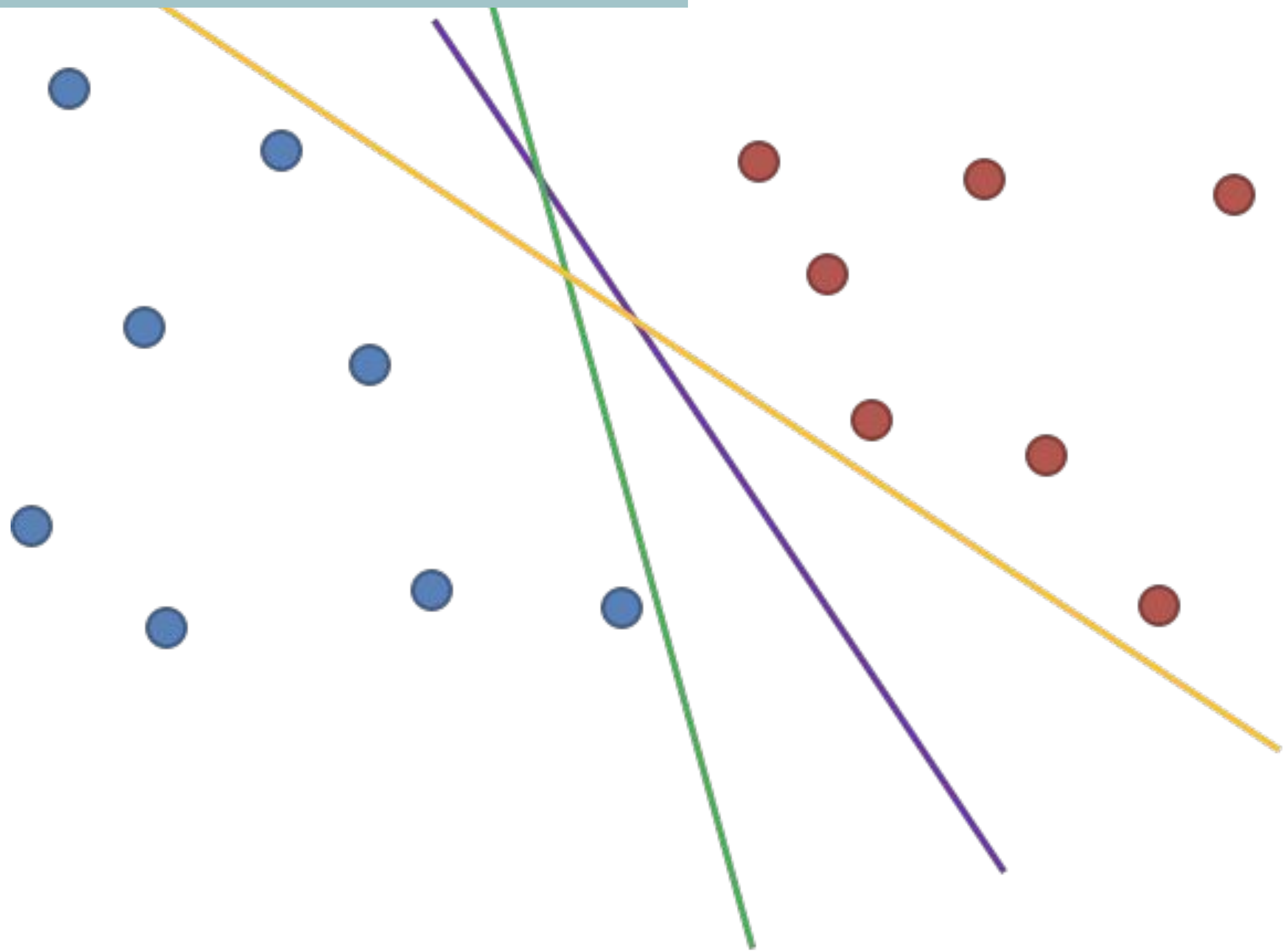
- По сути, делаем как можно меньше предположений о модели, и верим, что это снизит вероятность переобучения

Простой случай

- Будем считать, что выборка линейно разделима

- Существует линейный классификатор, не допускающий ни одной ошибки

Линейно разделимый случай



Линейно разделимый случай

Требование 1: максимальный отступ

$$\min_i \frac{|\langle w, x_i \rangle|}{\|w\|}$$

Требование 2: Нет ошибок

$$y_i \langle w, x_i \rangle > 0, i = 1 \dots l$$

Небольшое предположение

Линейный классификатор:

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

Если мы поделим w на число a > 0, то выходы классификатора никак не поменяются:

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

Небольшое предположение

Если мы поделим w на число a > 0, то выходы классификатора никак не поменяются:

$$a(x_i) = \operatorname{sign}\left(\frac{\langle w, x_i \rangle}{a}\right) = \operatorname{sign}\left(\langle w, x_i \rangle\right)$$

Давайте поделим на $\min\limits_i |\langle w, x_i \rangle|$

$$\tilde{w} = \frac{w}{\min_i |\langle w, x_i \rangle|}$$

$$\min_i |\langle w, x_i \rangle| = 1$$

Небольшое предположение

Давайте поделим на $\min_i |\langle w, x_i \rangle|$

$$\min_i |\langle \tilde{w}, x_i \rangle| = 1$$

Отступ классификатора:

$$\min_i \frac{|\langle \tilde{w}, x_i \rangle|}{\|\tilde{w}\|} = \frac{\min_i |\langle \tilde{w}, x_i \rangle|}{\|\tilde{w}\|} = \frac{1}{\|\tilde{w}\|}$$

Линейно разделимый случай

Требование 1: максимальный отступ

$$\frac{1}{\|w\|} \to \max_{w}$$

Требование 2: Нет ошибок

$$y_i \langle w, x_i \rangle > 0, i = 1 \dots l$$

Требование 3:

Линейно разделимый случай

Требование 1: максимальный отступ

$$\frac{1}{\|w\|} \to \max_{w}$$

Требование 2: Нет ошибок

$$y_i\langle w, x_i\rangle > 0, i = 1 \dots l$$

Требование 3:

$$\min_{i} |\langle w, x_i\rangle| = 1$$

Линейно разделимый случай

Требование 1: максимальный отступ

$$\frac{1}{\|w\|} \to \max_{w}$$

Требование 2: Нет ошибок

$$y_i \langle w, x_i \rangle > 0, i = 1 \ldots l$$

Требование 3:

$$|\langle w, x_i \rangle| \geq 1, \quad i = 1 \ldots l$$

Линейно разделимый случай

Требование 1: максимальный отступ

$$\frac{1}{\|w\|} \to \max_{w}$$

Требование 2: $y_i\langle w, x_i\rangle > 0, i = 1 \dots l$

Требование 3: $|\langle w, x_i\rangle| \geq 1, \quad i = 1 \dots l$

Требование 2 + 3:

Линейно разделимый случай

Требование 1: максимальный отступ

$$\frac{1}{\|w\|} \to \max_{w}$$

Требование 2: $y_i\langle w, x_i\rangle > 0, i = 1 \dots l$

Требование 3: $|\langle w, x_i\rangle| \geq 1, \quad i = 1 \dots l$

Требование 2 + 3:

$$y_i\langle w, x_i\rangle \geq 1, \quad i = 1 \dots l$$

Линейно разделимый случай

Требование 1: максимальный отступ

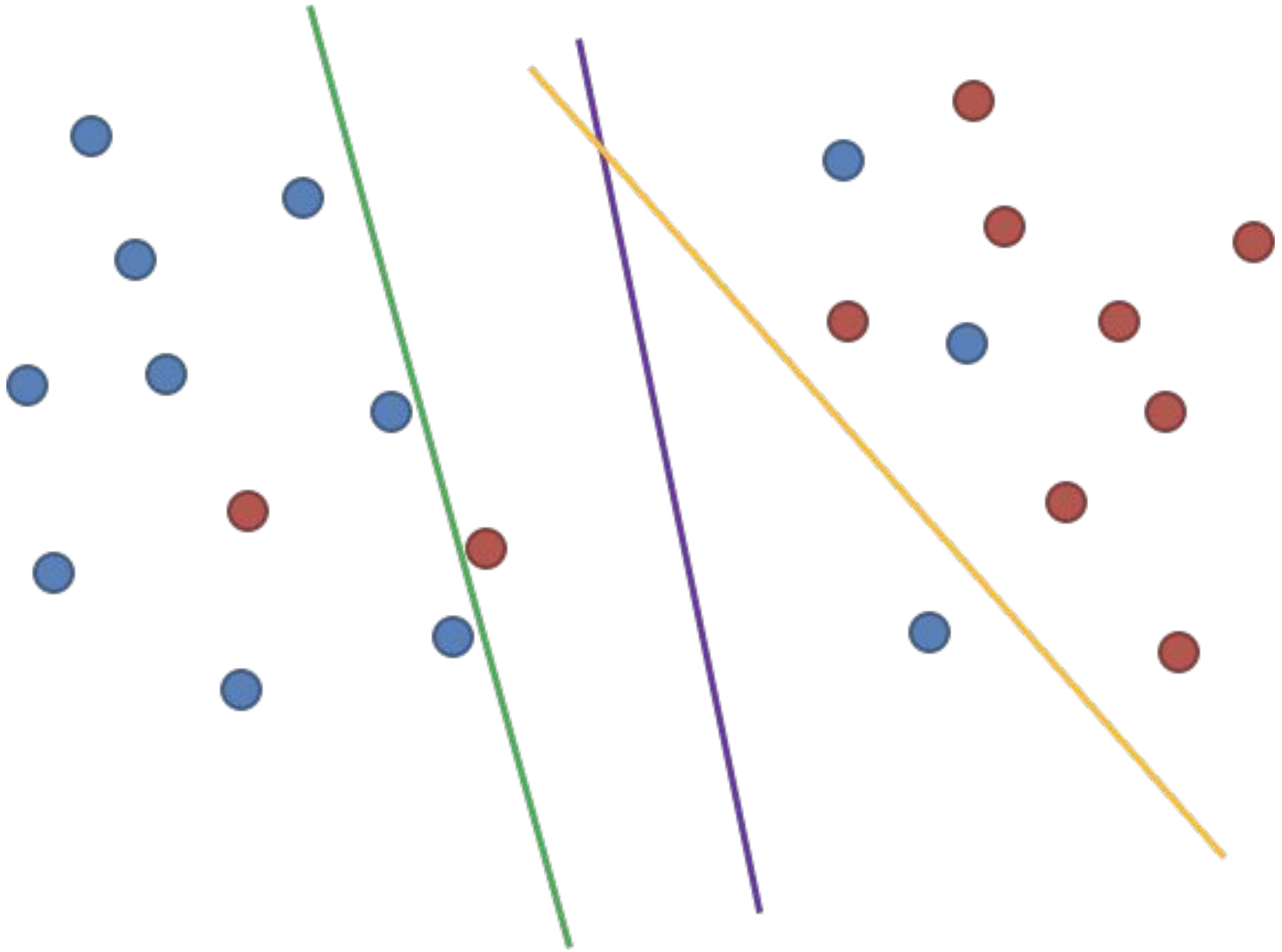
$$\frac{1}{\|w\|} \to \max_w$$

Требование 2 + 3: $y_i \langle w, x_i \rangle \geq 1, \quad i = 1 \dots l$

Итого:

$$\begin{cases} \min_w \|w\| \\ y_i \langle w, x_i \rangle \geq 1 \end{cases}$$

Линейно неразделимый случай



Линейно неразделимый случай

- Любой линейный классификатор допускает хотя бы одну ошибку

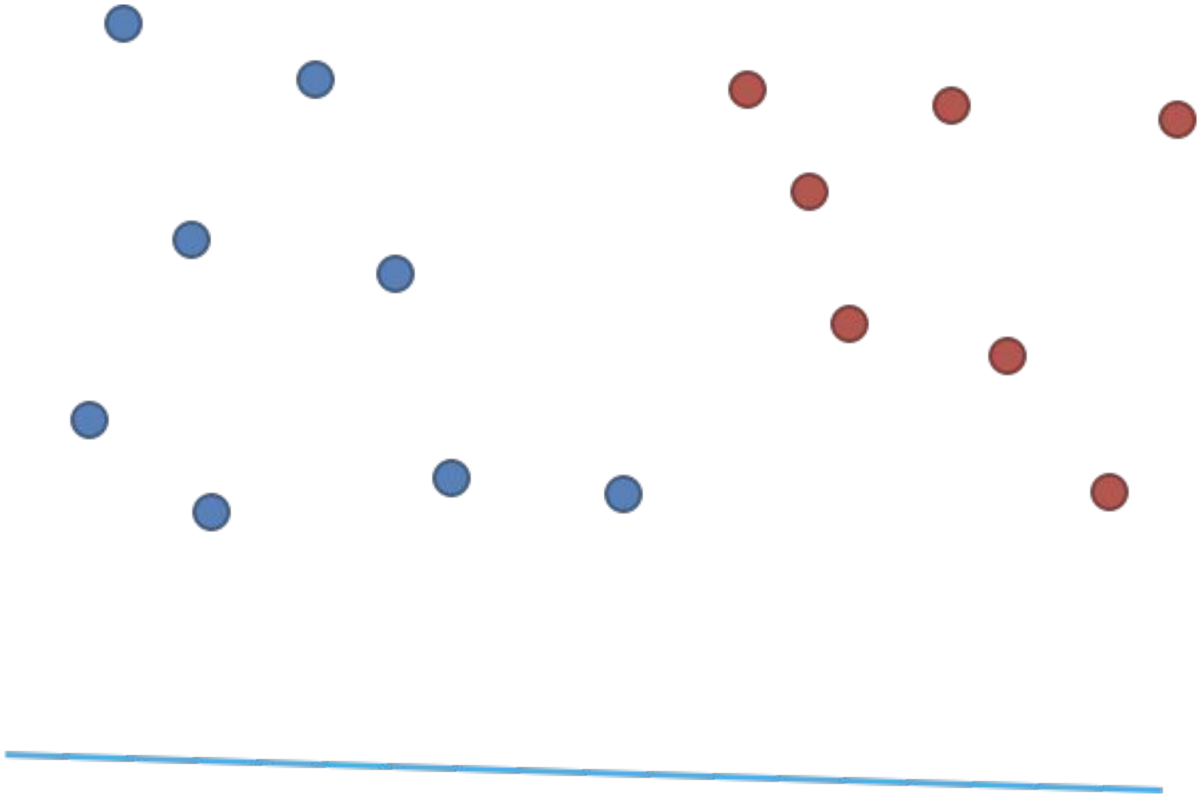
$$\begin{cases} \min_w \|w\| \\ \boxed{y_i \langle w, x_i \rangle \geq 1} \end{cases}$$

- Давайте смягчим условия

$$\begin{cases} \min_w \|w\| \\ y_i \langle w, x_i \rangle \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

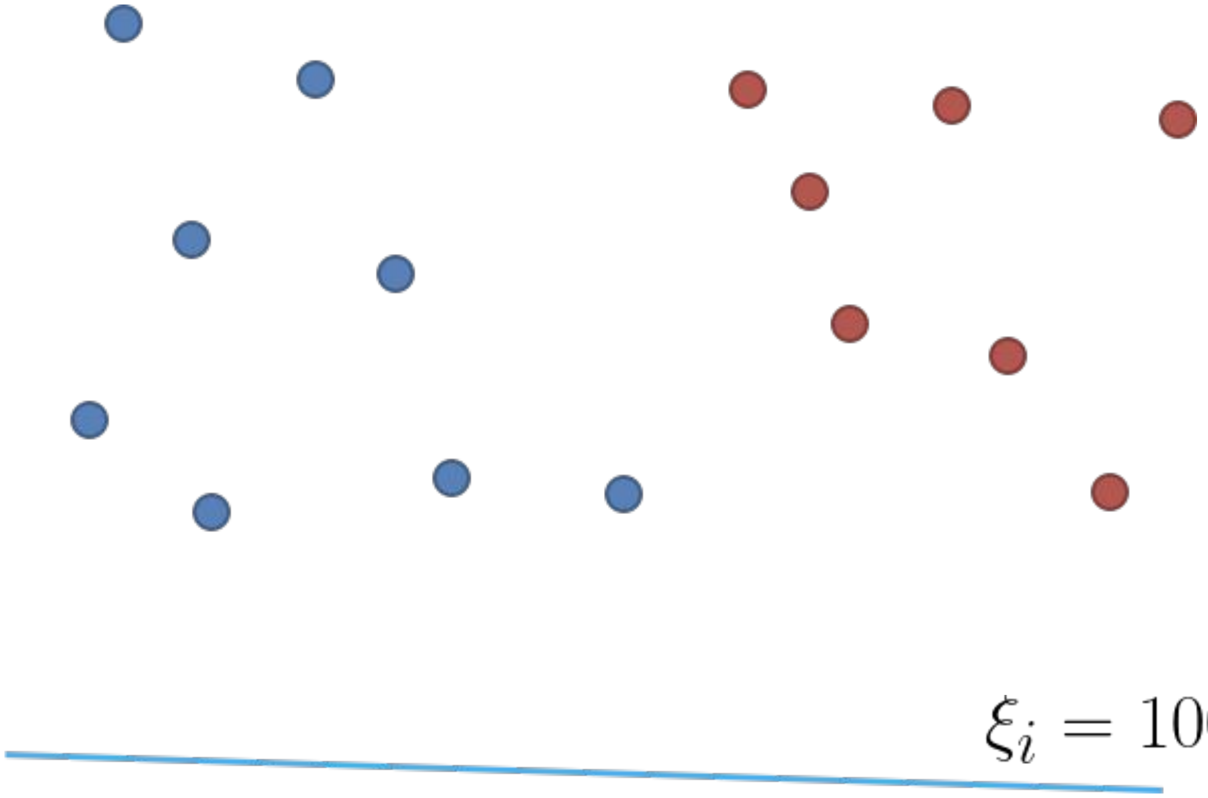
Линейно неразделимый случай

$$\begin{cases} \min_w \|w\| \\ y_i \langle w, x_i \rangle \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$



Линейно неразделимый случай

$$
\begin{cases}
\min_w \|w\| \\
y_i \langle w, x_i \rangle \geq 1 - \xi_i \\
\xi_i \geq 0
\end{cases}
$$



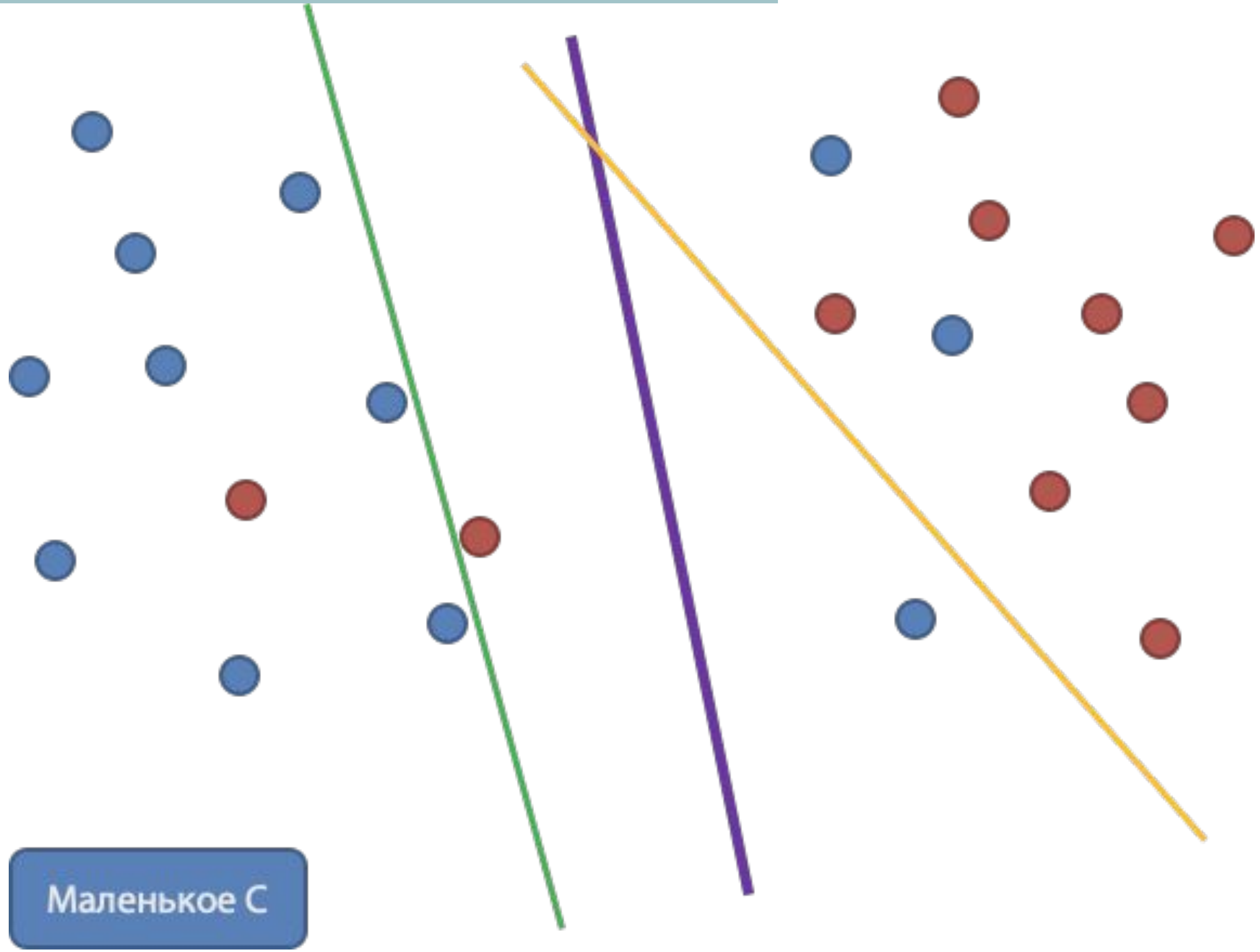
$$\xi_i = 100000000$$

Линейно неразделимый случай

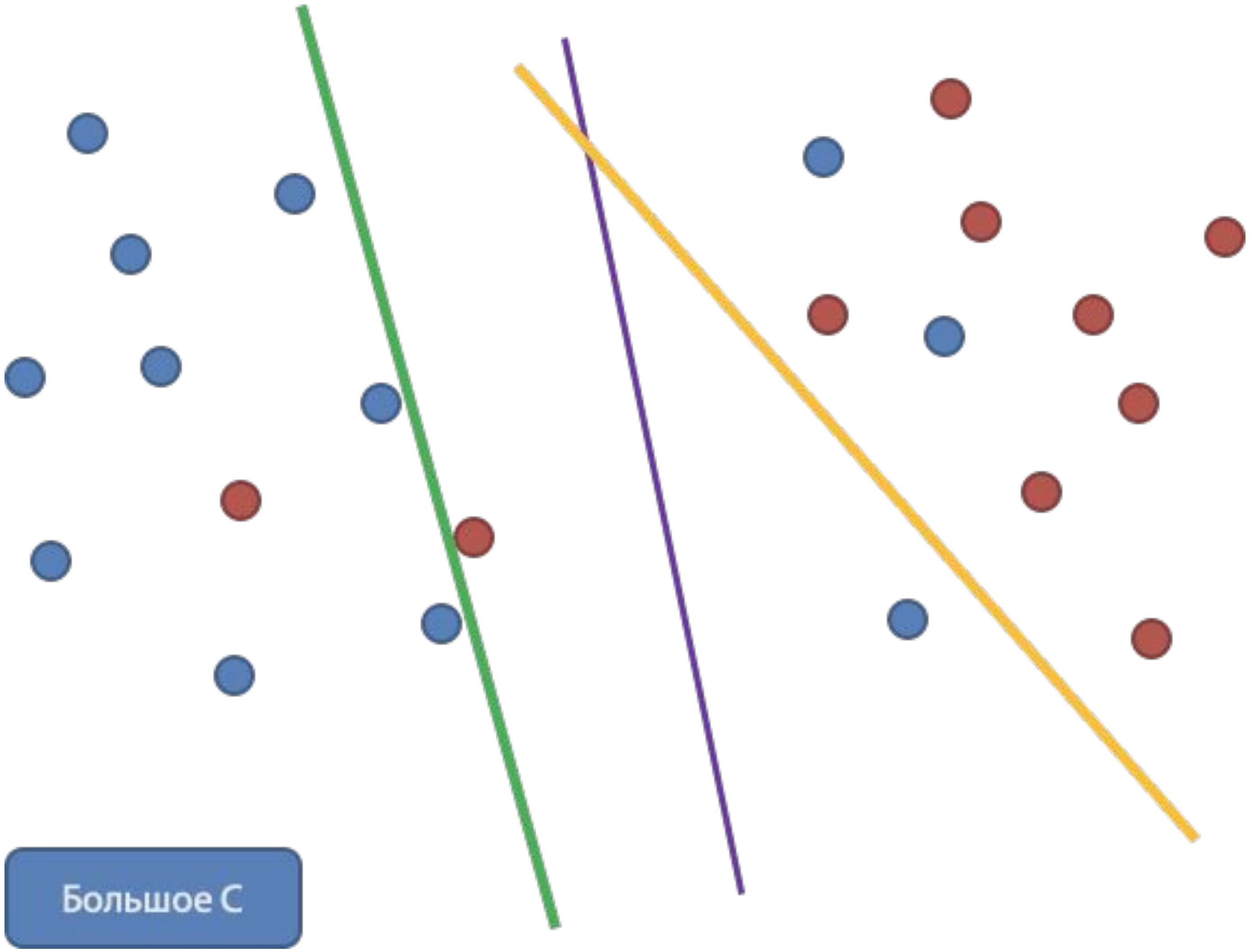
$$\begin{cases} \min_w \|w\| \\ y_i\langle w, x_i\rangle \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

$$\begin{cases} \min_{w,\xi} \|w\| + \boxed{C \sum_i \xi_i} \\ y_i\langle w, x_i\rangle \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Линейно неразделимый случай



Линейно неразделимый случай



Метод опорных векторов

$$\begin{cases} \min_{w,\xi} \|w\| + C\sum_i \xi_i \\ y_i\langle w, x_i\rangle \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Объединим ограничения

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

Итоговая задача оптимизации

$$\min_w \left(\|w\| + C\sum_i \max(0, 1 - y_i\langle w, x_i\rangle)) \right)$$

Метод опорных векторов

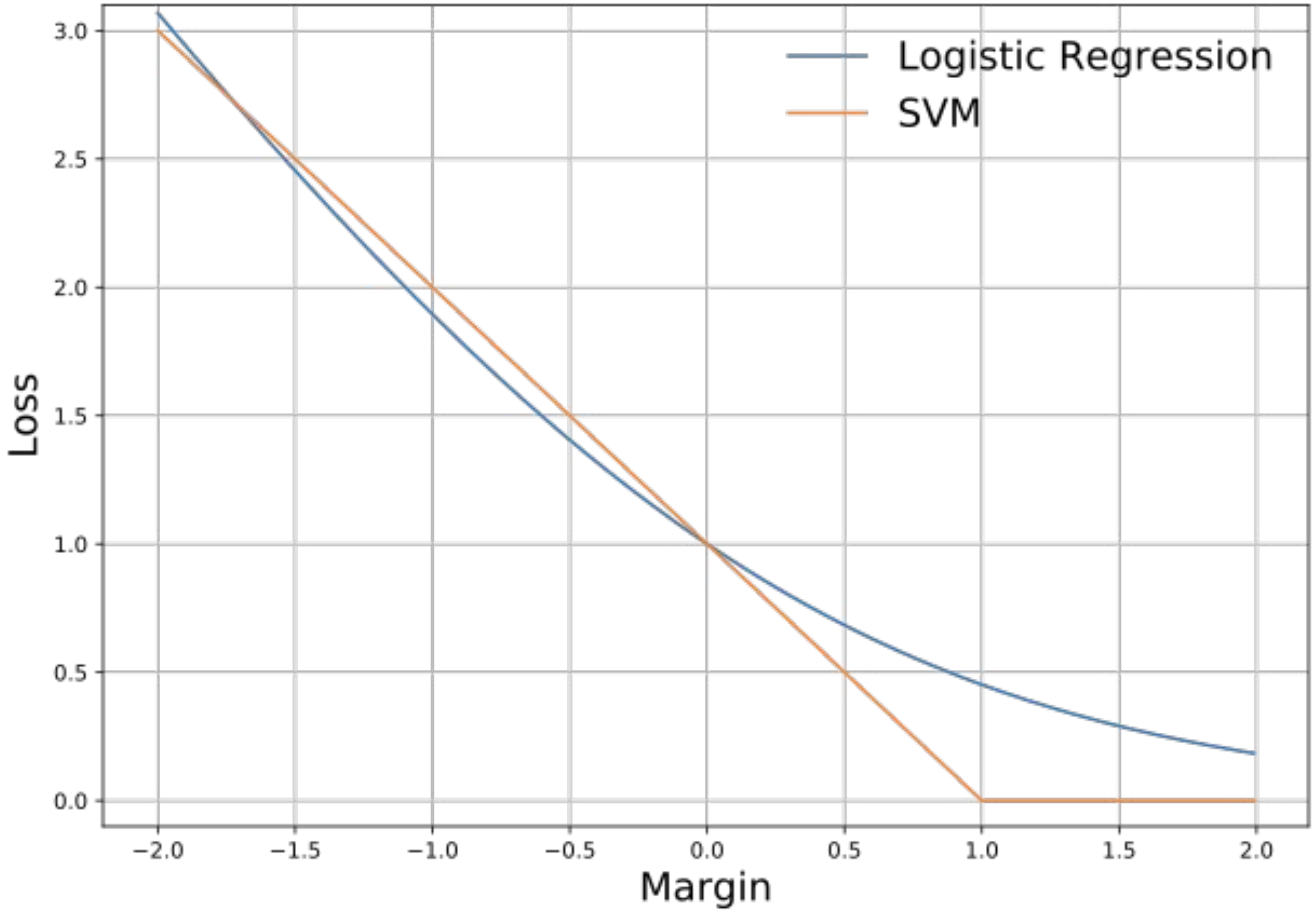
Итоговая задача оптимизации

$$\min_w \left(\|w\| + C \sum_i \max(0, 1 - y_i \langle w, x_i \rangle)) \right)$$

Hingle Loss (верхняя оценка) + L2 регуляризация

$$\min_w \left(\sum_i \max(0, 1 - y_i \langle w, x_i \rangle) + \lambda \|w\| \right)$$

Сравнение логистической регрессии и SVM



Итого

- Метод опорных векторов основан на идее максимизации отступа классификатора
- В линейно неразделимом случае мы должны выбрать, что важнее, — ширина зазора или число ошибок
- По сути, SVM — это определённая функция потерь и регуляризация