Основы машинного обучения

Лекция 10.5

Логистическая регрессия. Метод опорных векторов.

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2022

Логистическая регрессия: сложное объяснение

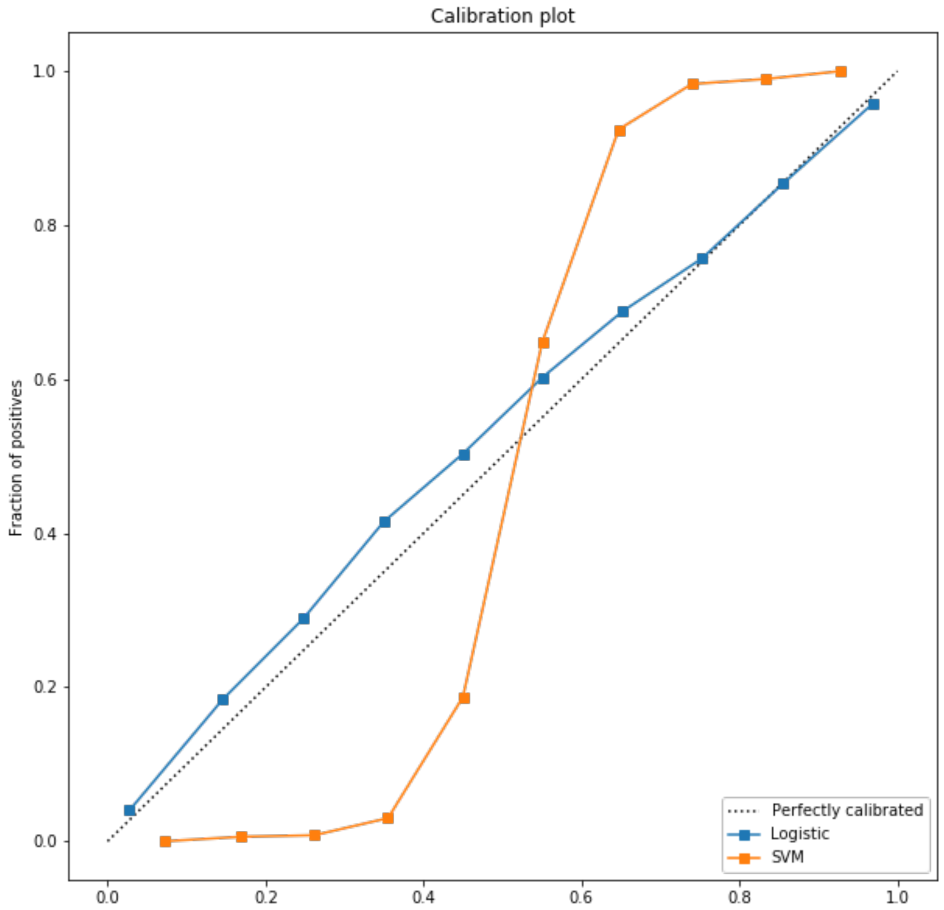
Предсказание вероятностей

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:

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

Предсказание вероятностей

- Функционал ошибки:

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

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

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

- 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:

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

- 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:

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

- 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:

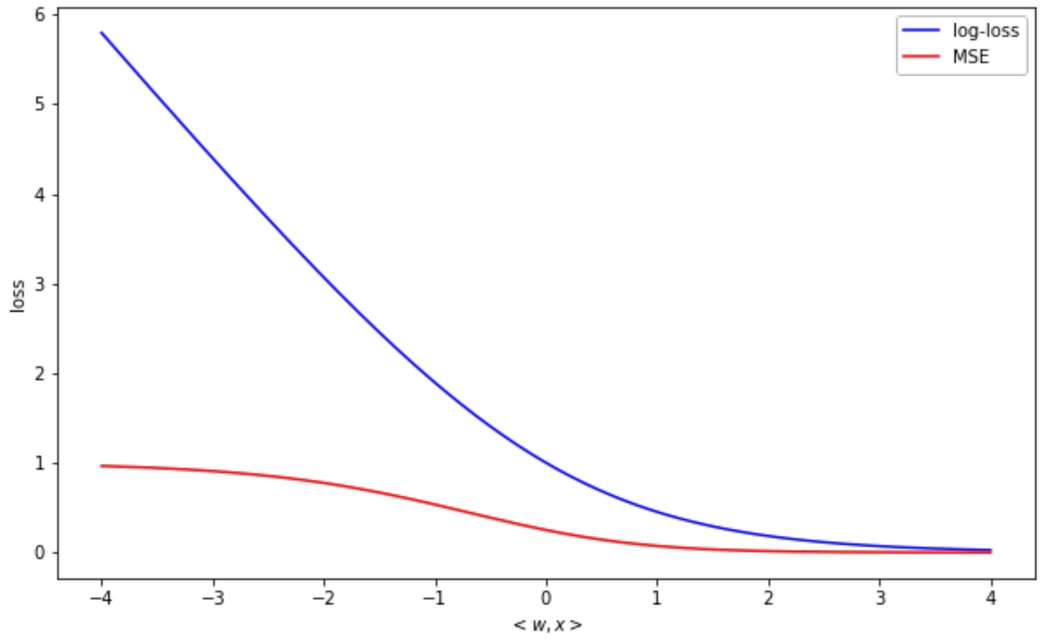
$$2 \sum_{i=1}^{n} (p - [y_i = +1]) \quad = 0$$

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

MSE

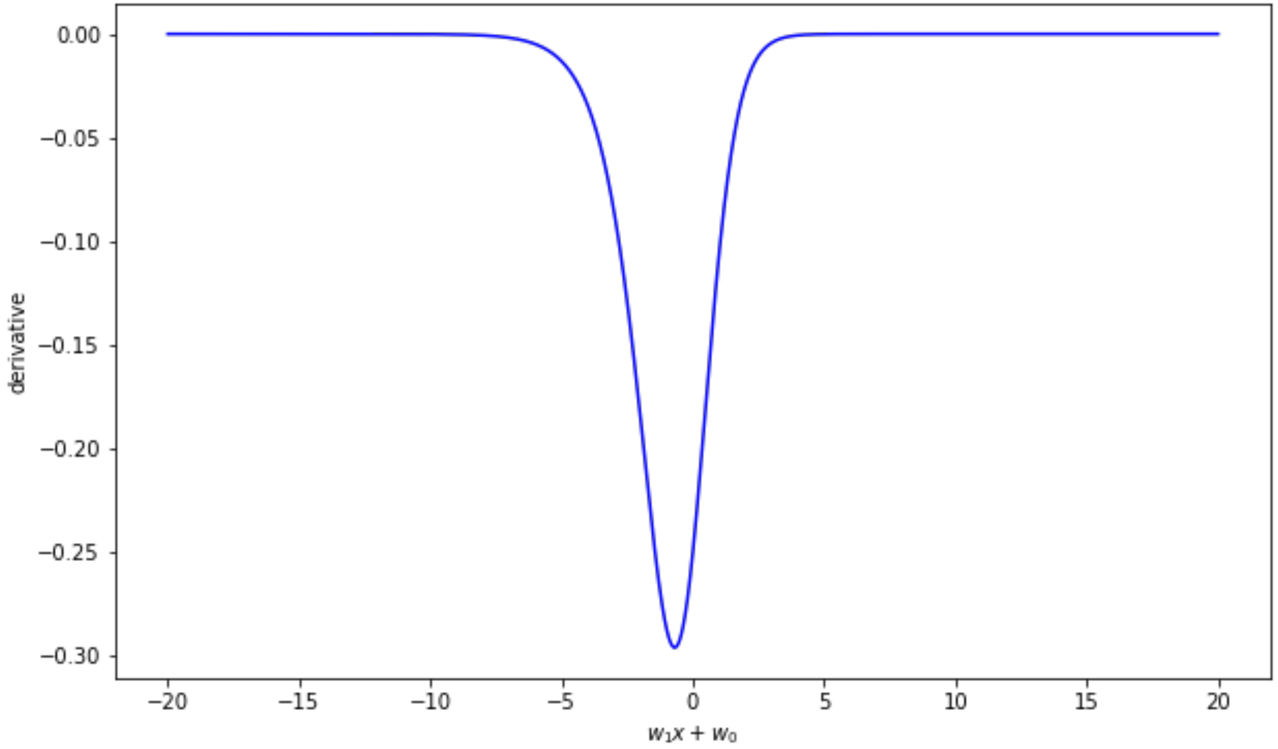
- Почему бы не обучать классификаторы на MSE?

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



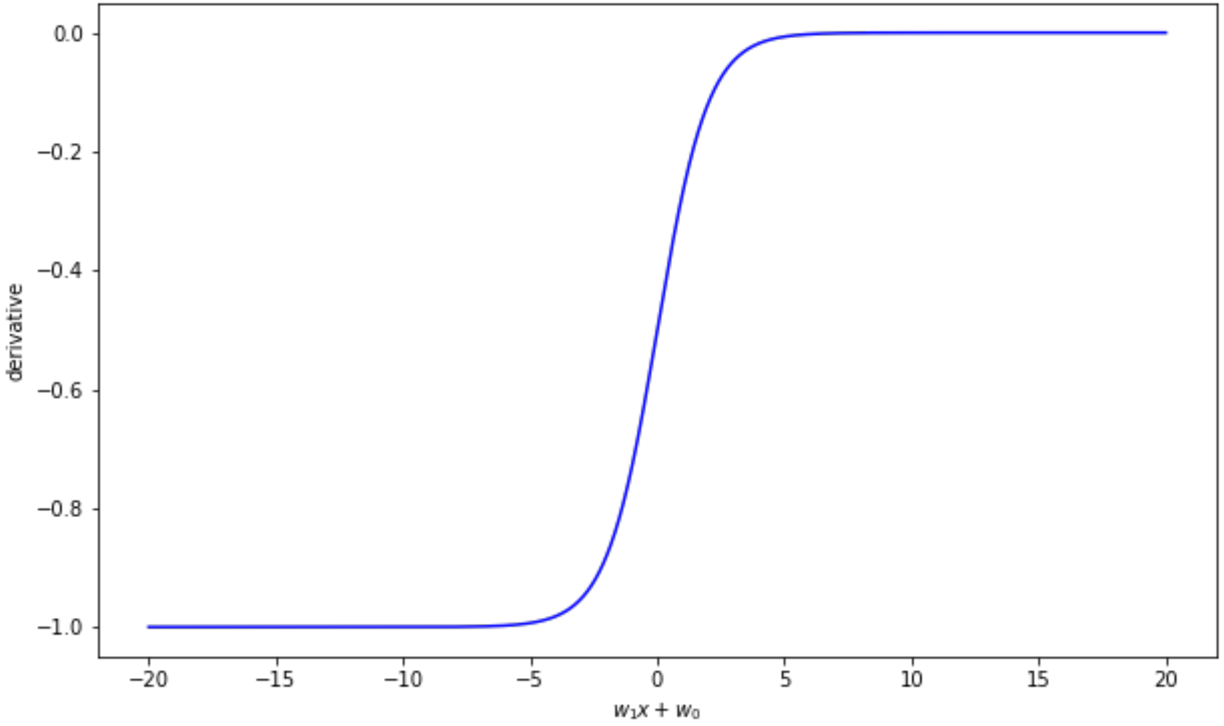
MSE

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



Log-loss

$$\frac{\partial}{\partial w_1}\left(\log\frac{1}{1+e^{-w_1x-w_0}}\right) = \frac{x}{1+e^{w_1x+w_0}}$$



MAE

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

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

Метод опорных векторов

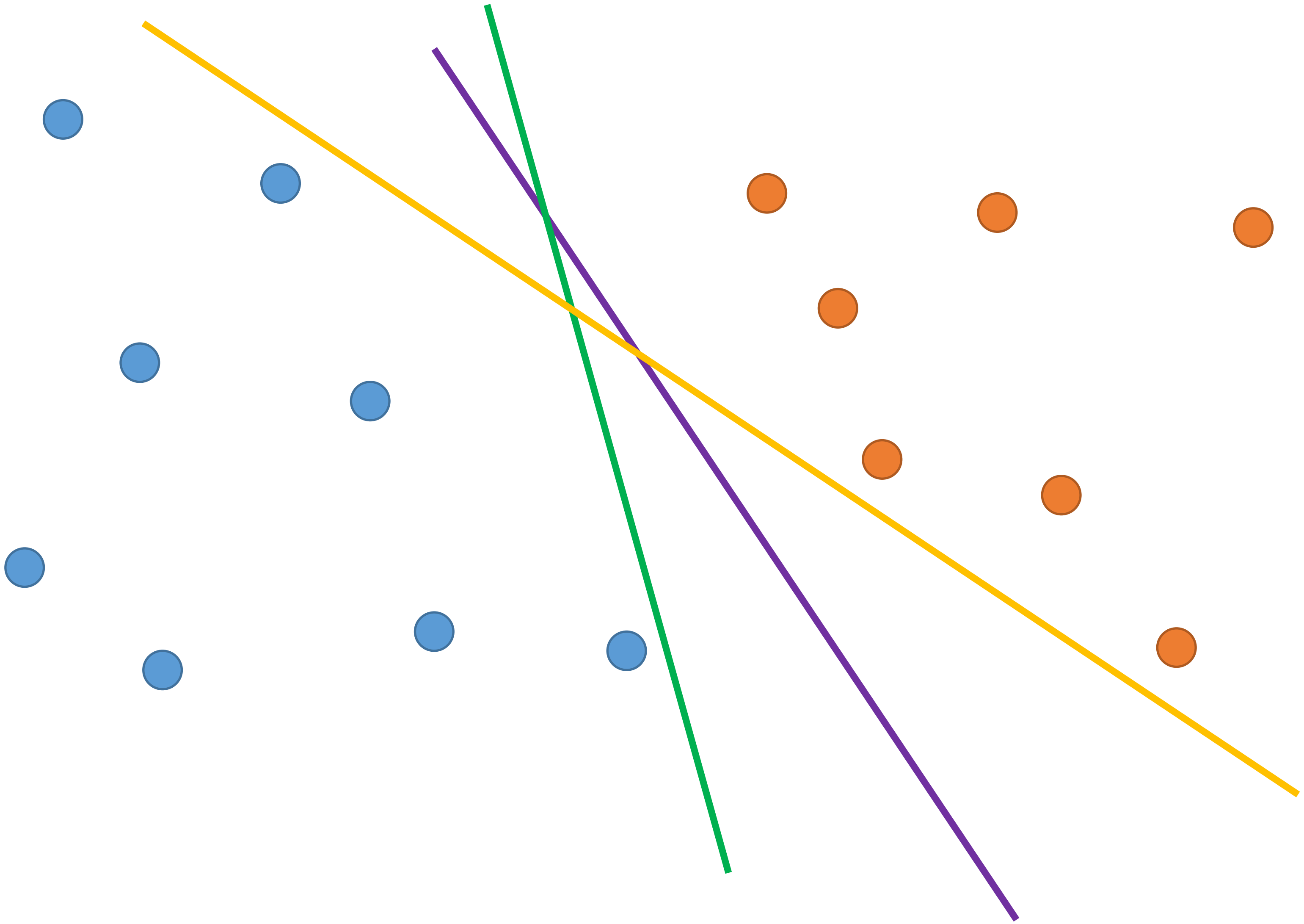
Hinge loss

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

- Минимизация верхней оценки:

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

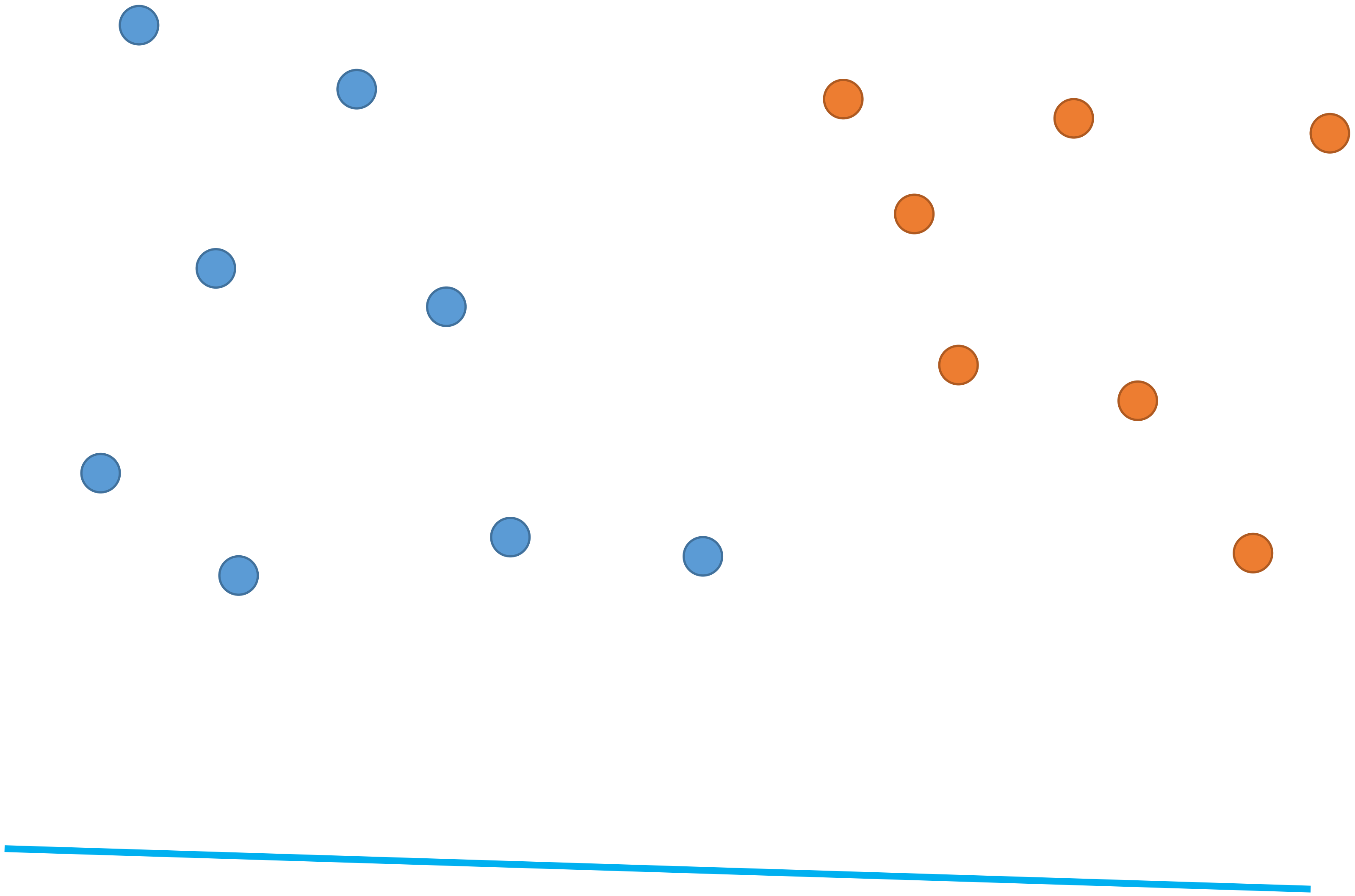
Какой классификатор лучше?



Отступ классификатора

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

Отступ классификатора



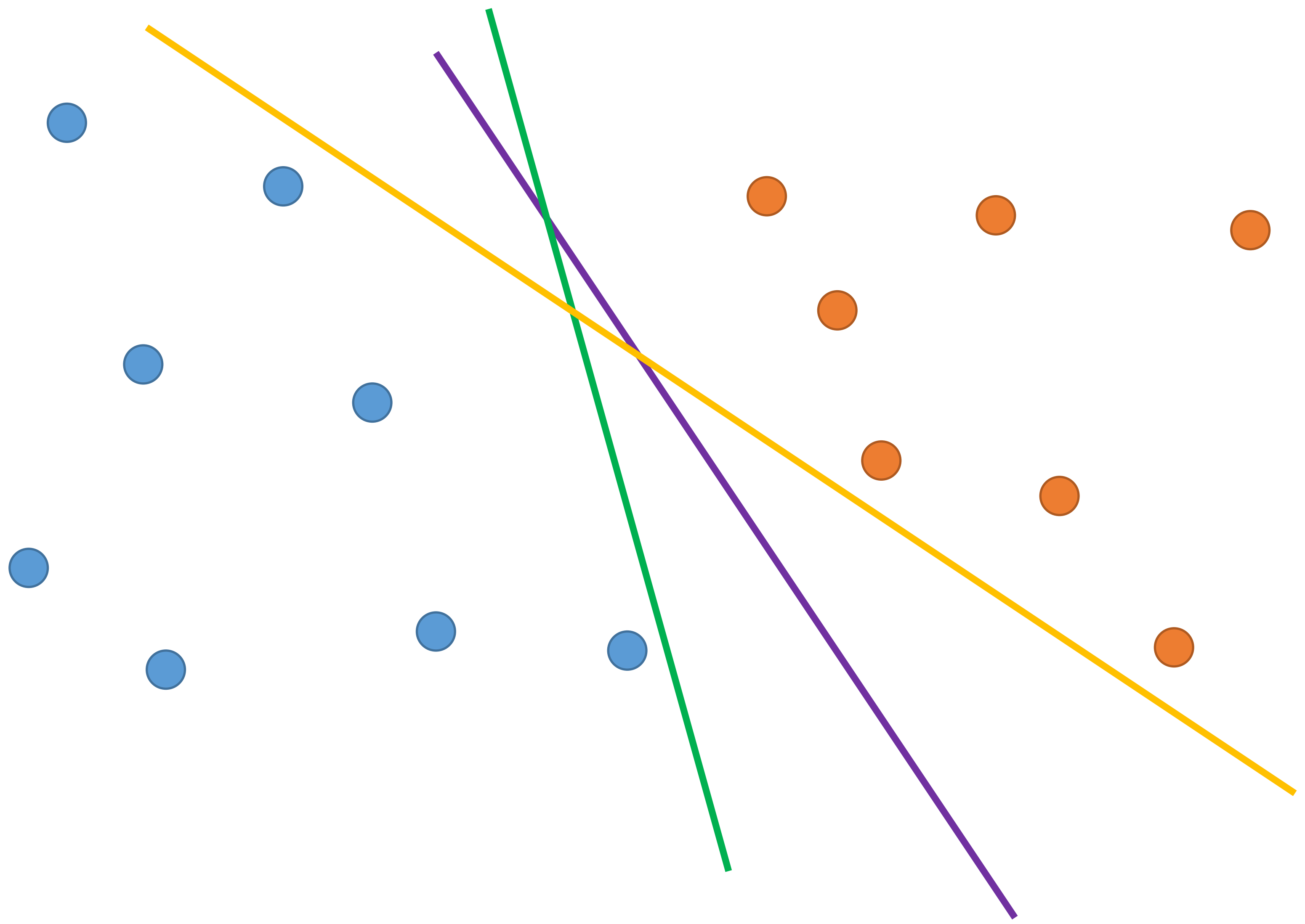
Отступ классификатора

- 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:

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

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

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

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:

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

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

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

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

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

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

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

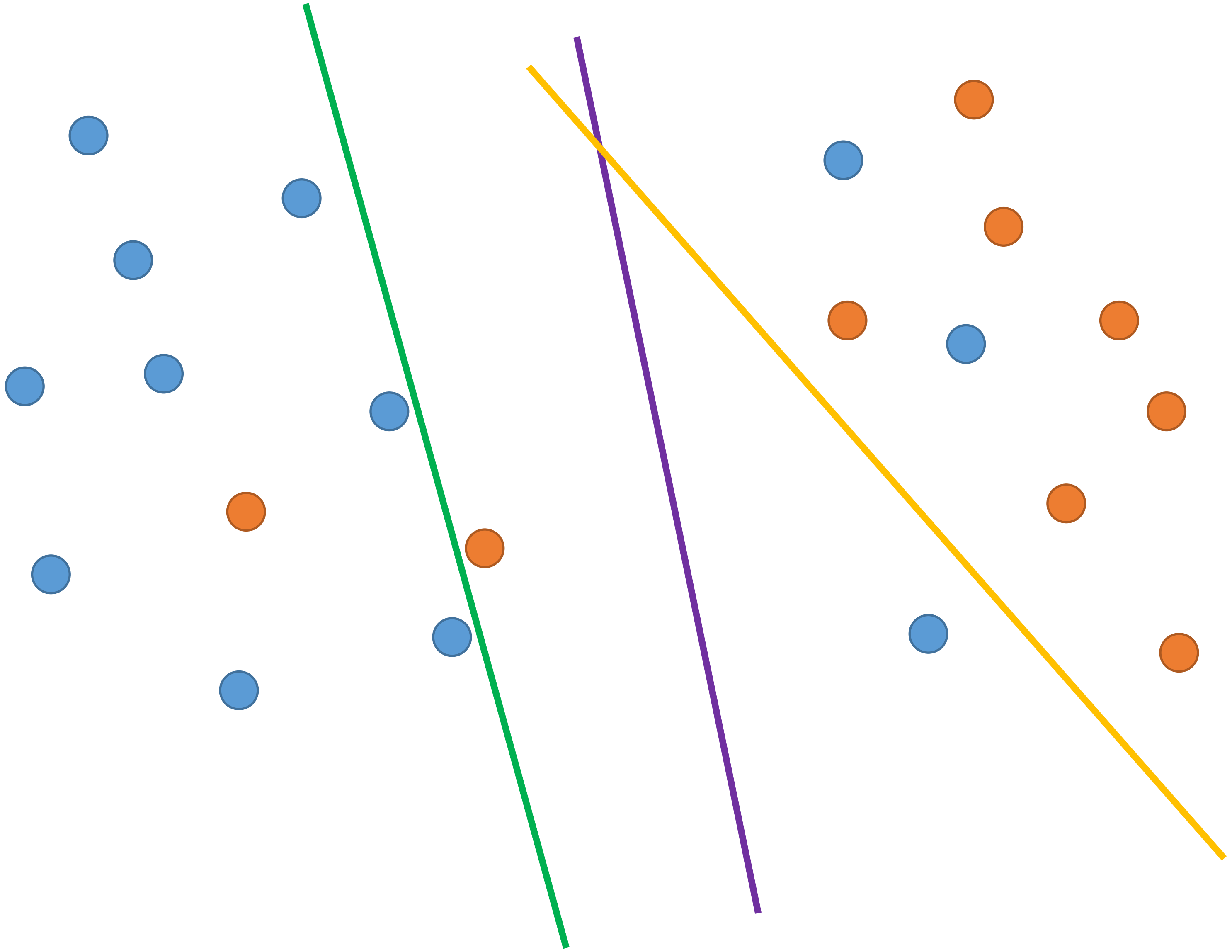
$$\frac{1}{\|w\|} \to \max_w$$

- 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)

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

Линейно неразделимый случай



Линейно неразделимый случай

- Любой линейный классификатор допускает хотя бы одну ошибку

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

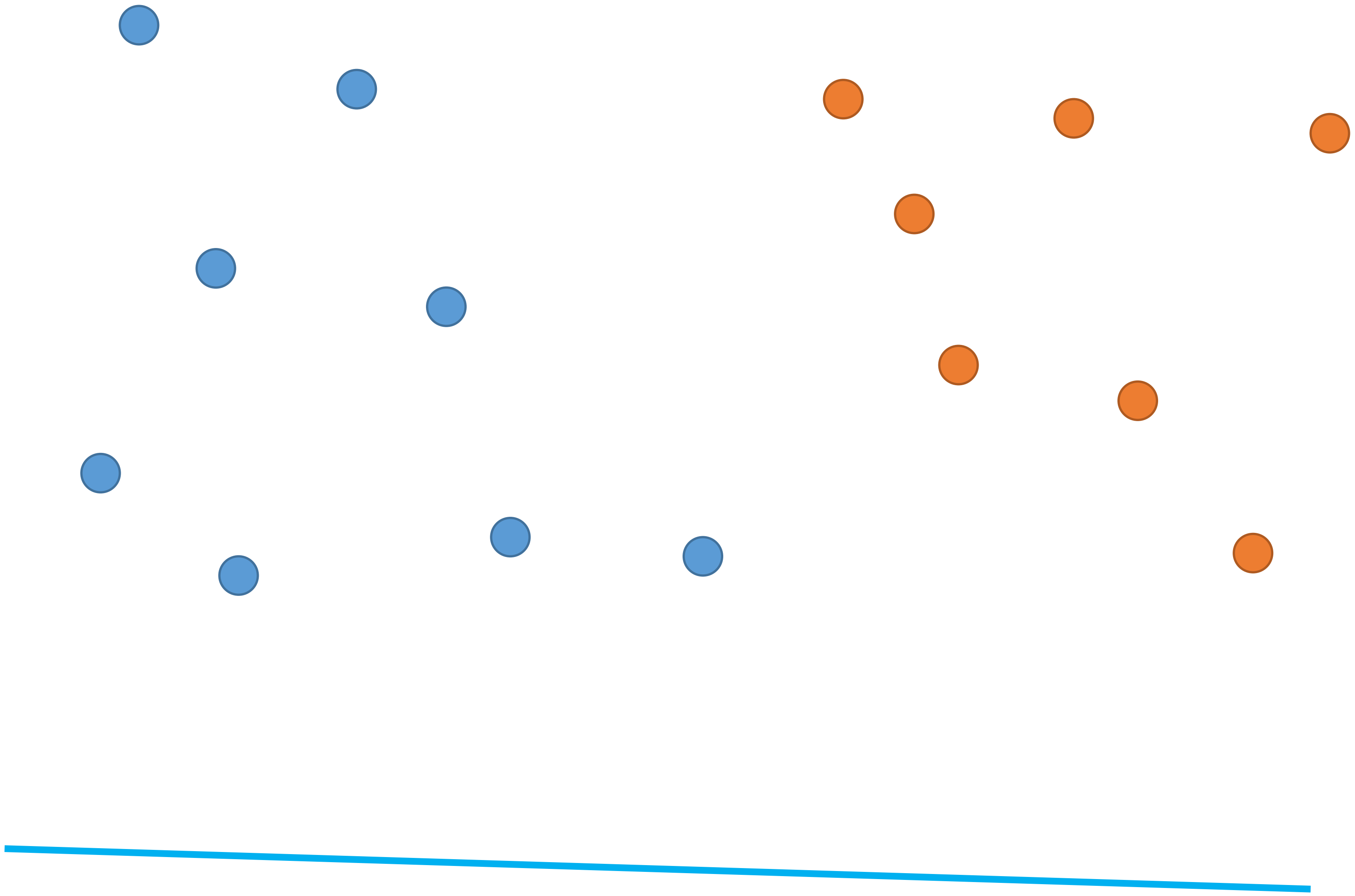
Линейно неразделимый случай

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

Линейно неразделимый случай

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

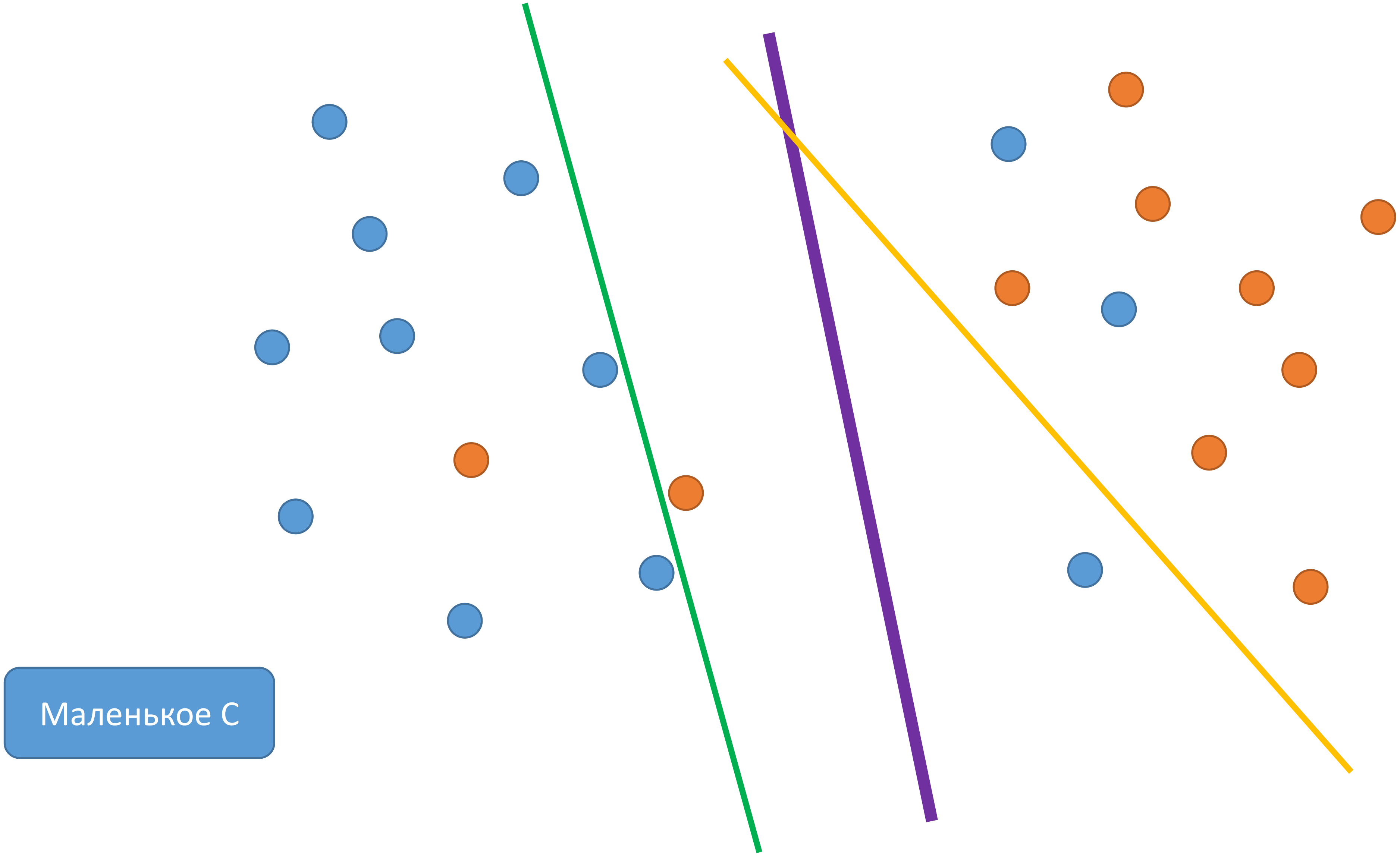
Отступ классификатора



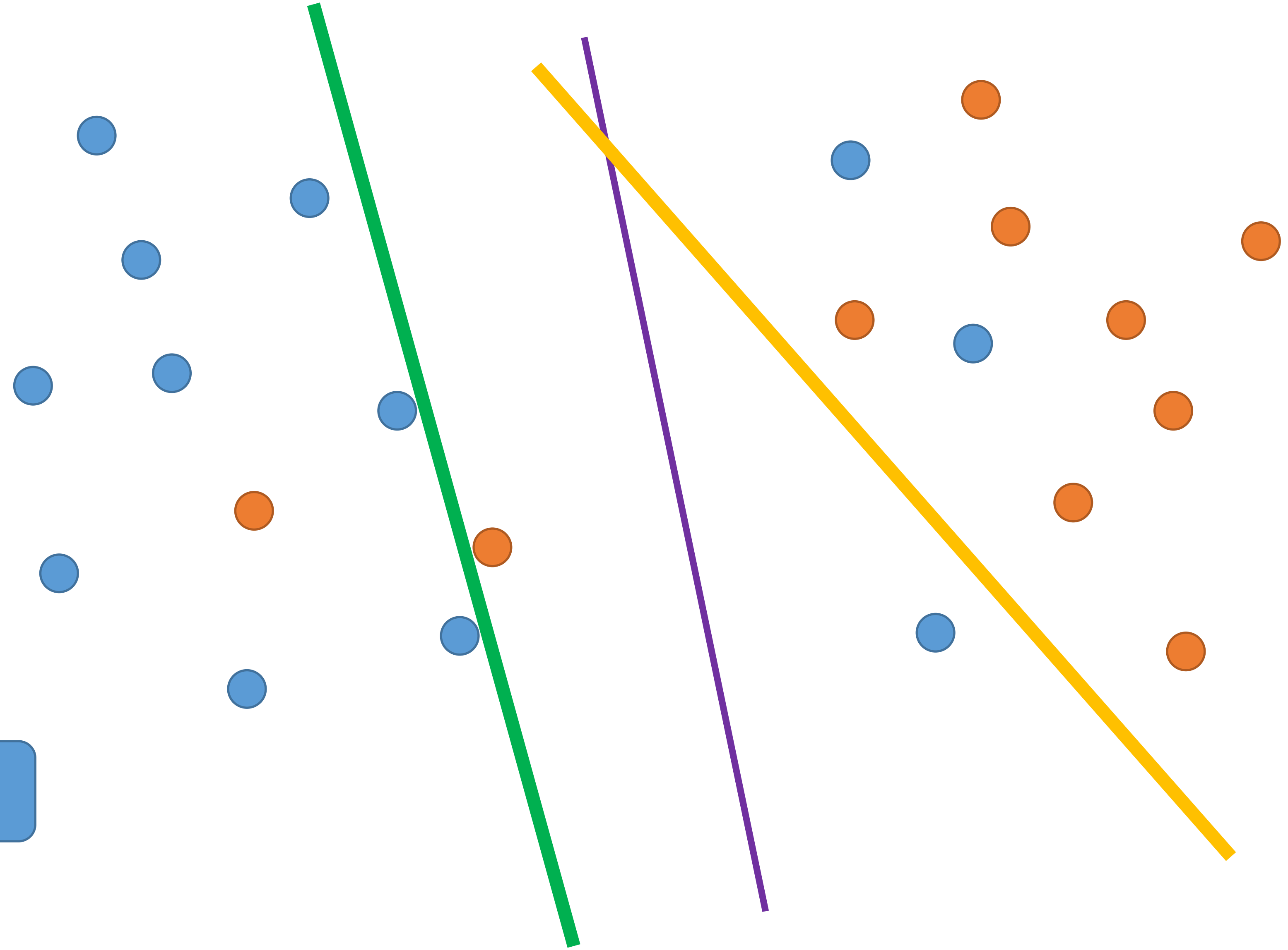
Метод опорных векторов

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

Линейно неразделимый случай



Линейно неразделимый случай



Большое C

Метод опорных векторов

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

- Объединим ограничения:

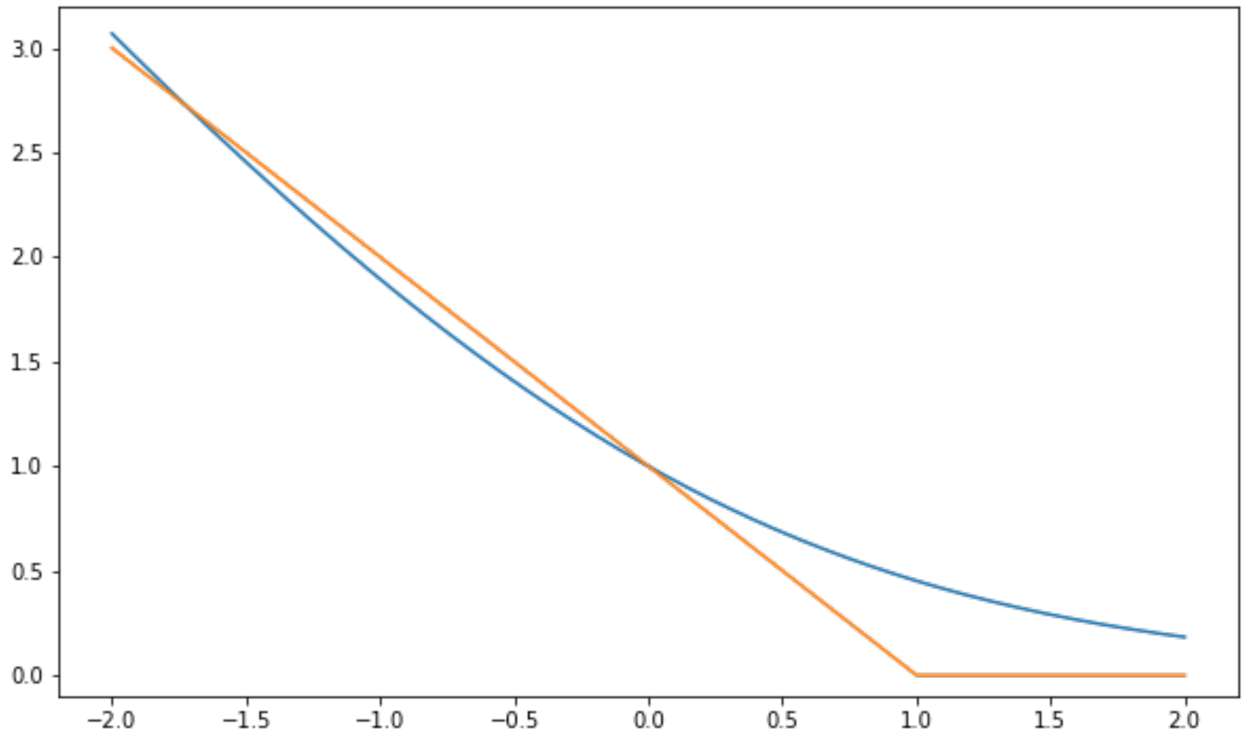
$$\xi_i \geq \max\big(0, 1 - y_i(\langle w, x_i \rangle + w_0)\big)$$

Метод опорных векторов

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

- Функция потерь (hinge loss) + регуляризация

Сравнение логистической регрессии и SVM



Резюме

- Логистическая регрессия — обучение модели так, что на объектах с близкими прогнозами эти прогнозы стремятся к доле положительных объектов

- Метод опорных векторов основан на идее максимизации отступа классификатора