Введение в анализ данных

Лекция 8

Линейная классификация

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2020

Метод опорных векторов

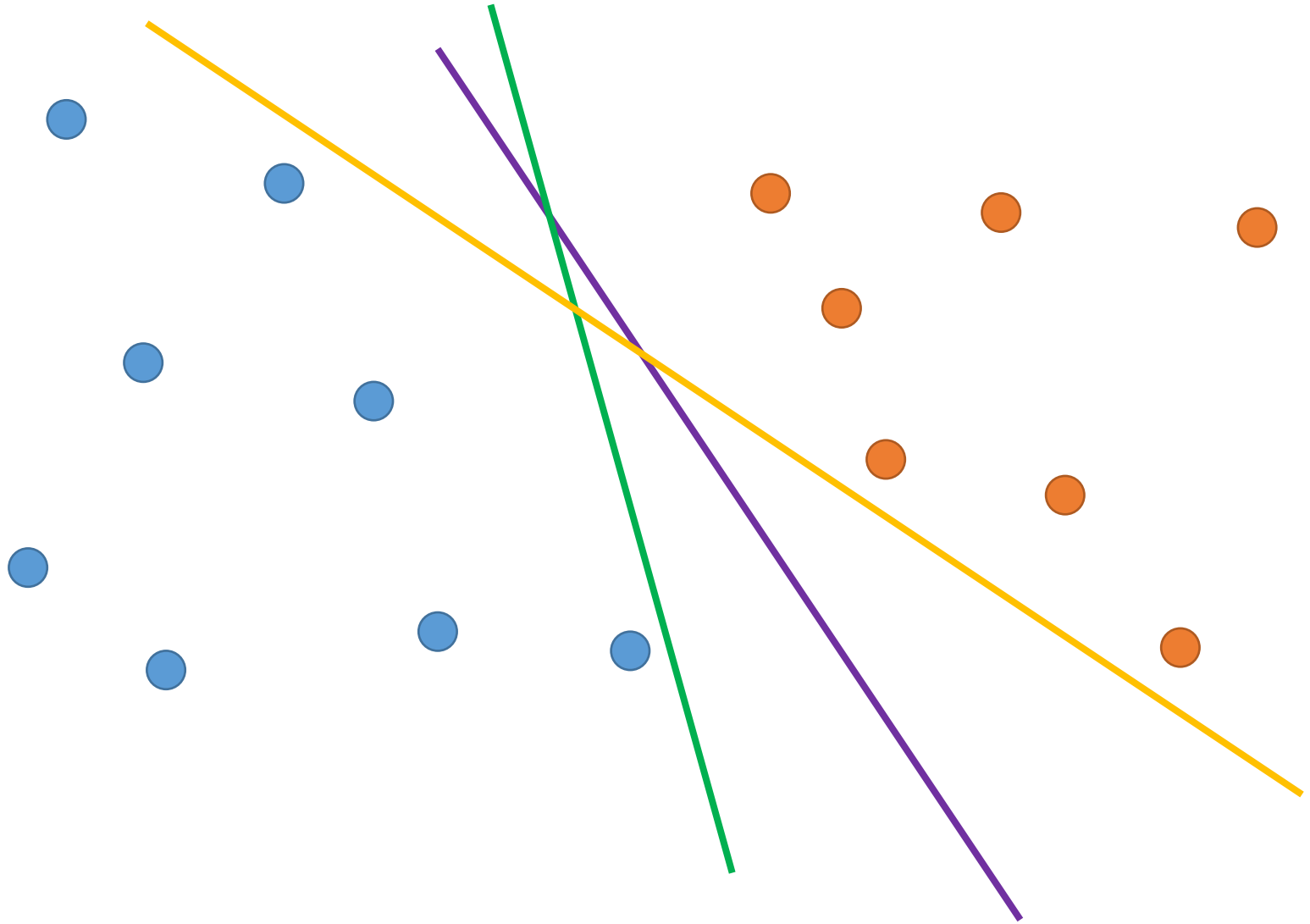
Hinge loss

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

- Минимизация верхней оценки:

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

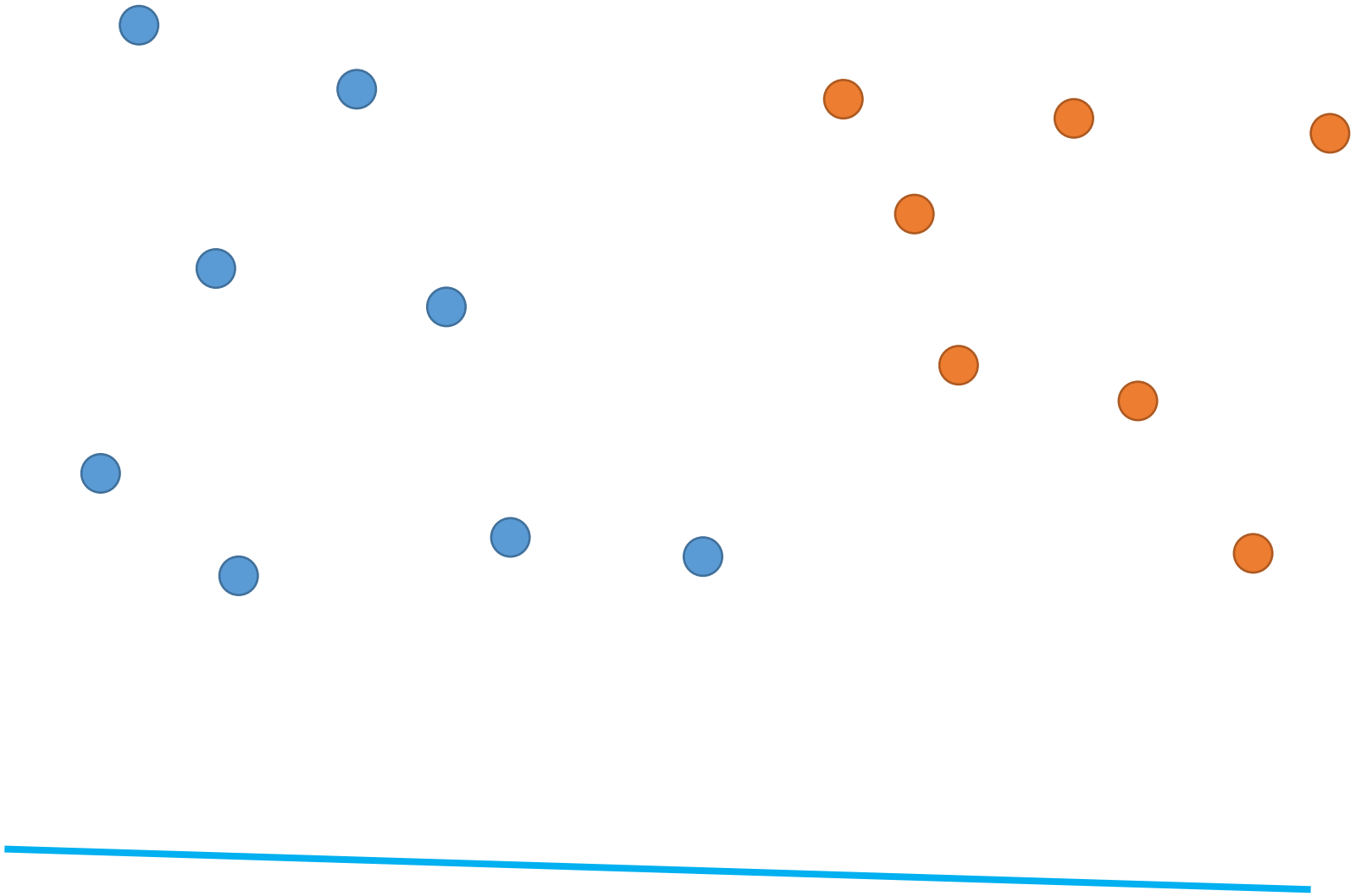
Какой классификатор лучше?



Отступ классификатора

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

Отступ классификатора



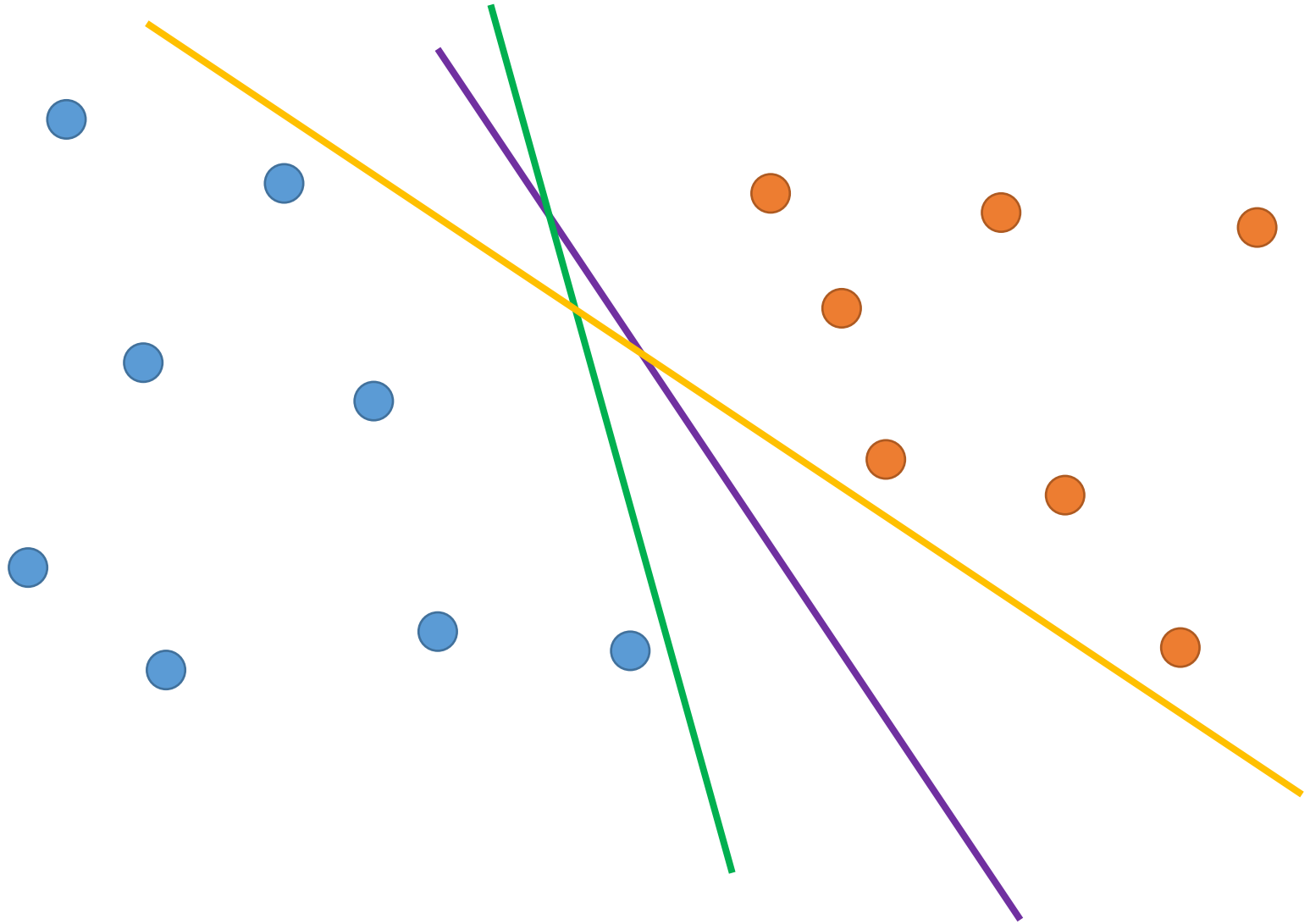
Отступ классификатора

- 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:

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

Небольшое предположение

- Линейный классификатор:

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

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

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

Небольшое предположение

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

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

Отступ классификатора

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

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

- Отступ классификатора:

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

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

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

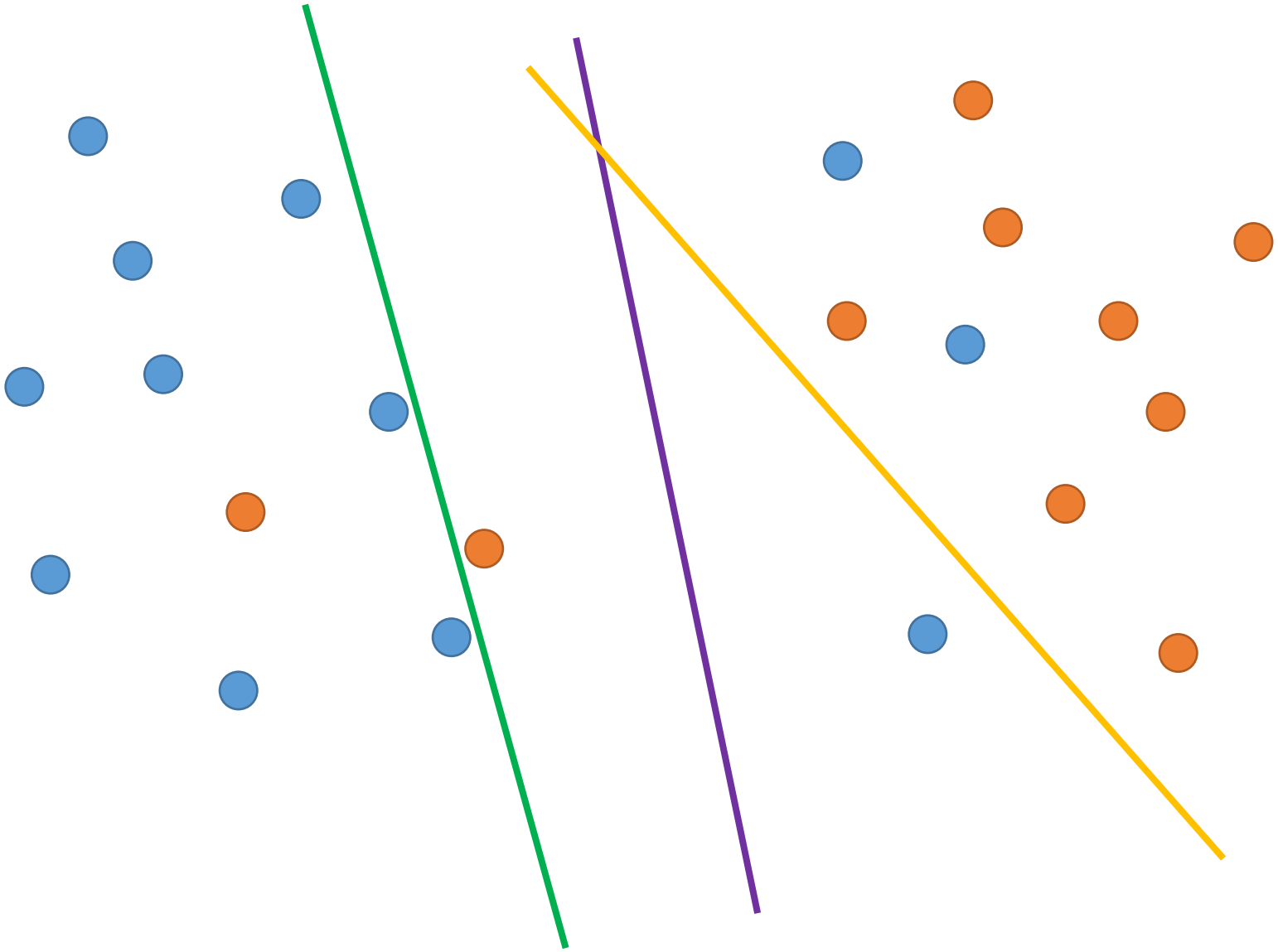
$$\frac{1}{\|w\|} \to \max_w$$

- 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

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

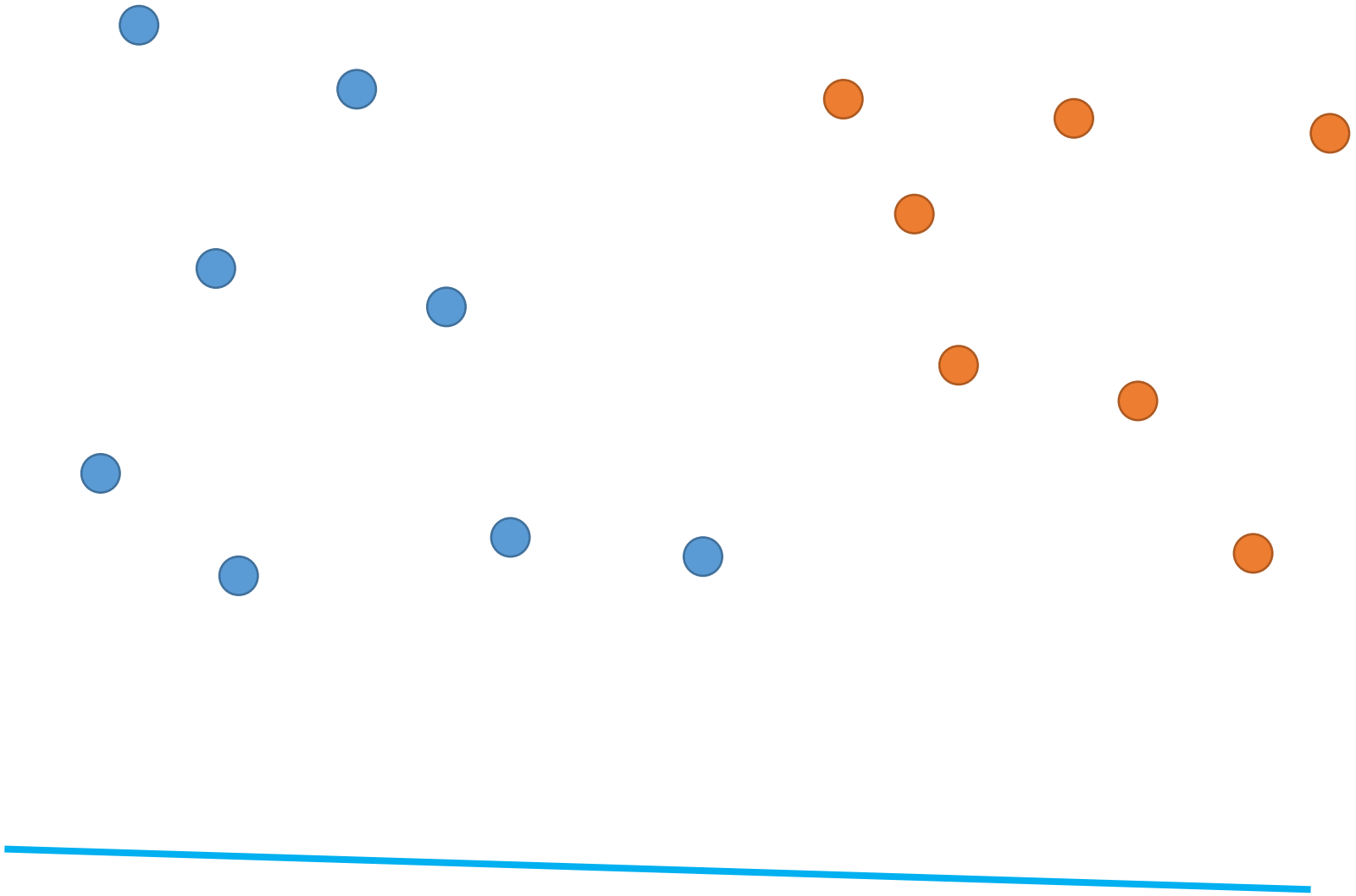
Линейно неразделимый случай

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

Линейно неразделимый случай

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

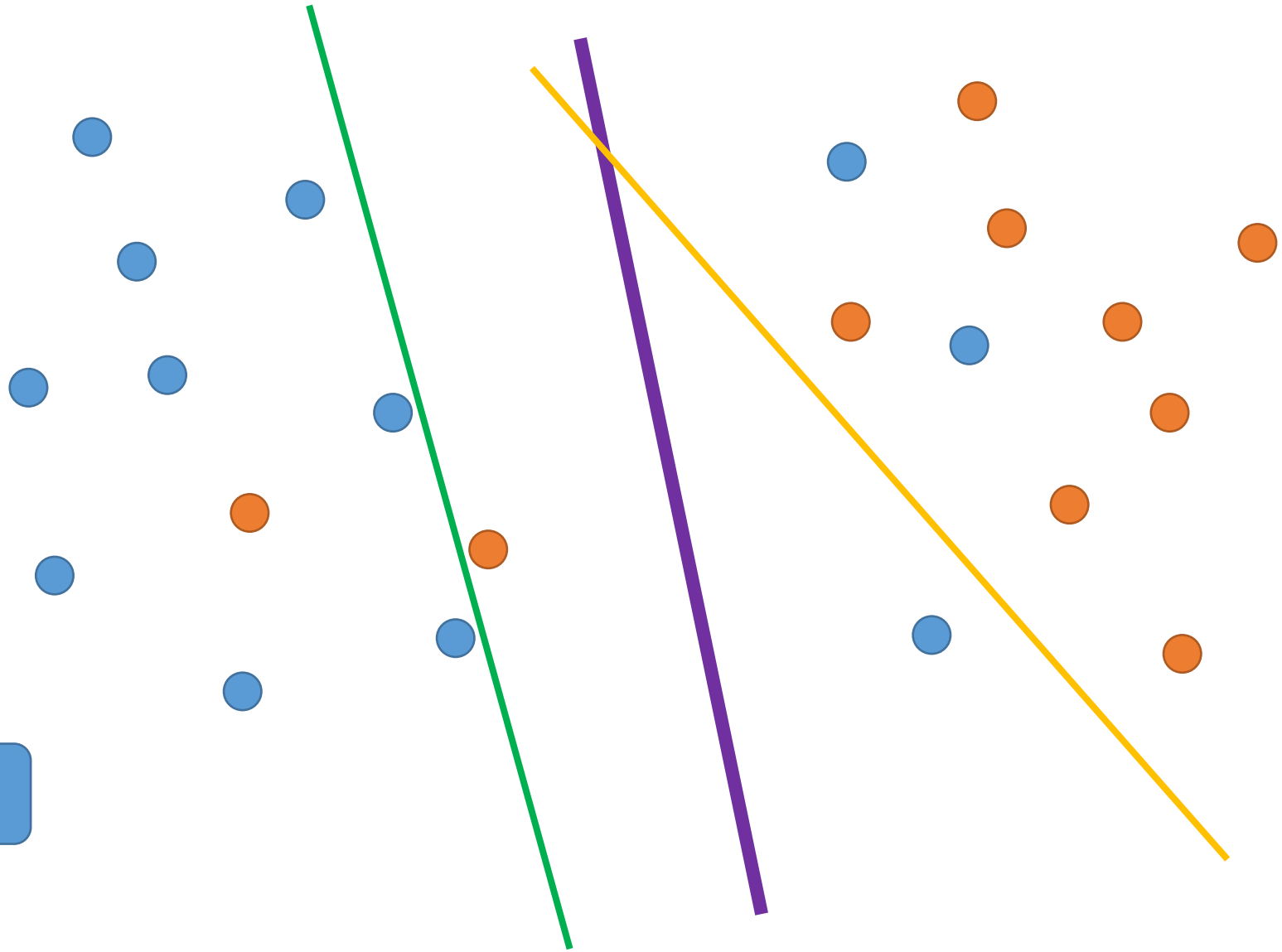
Отступ классификатора



Метод опорных векторов

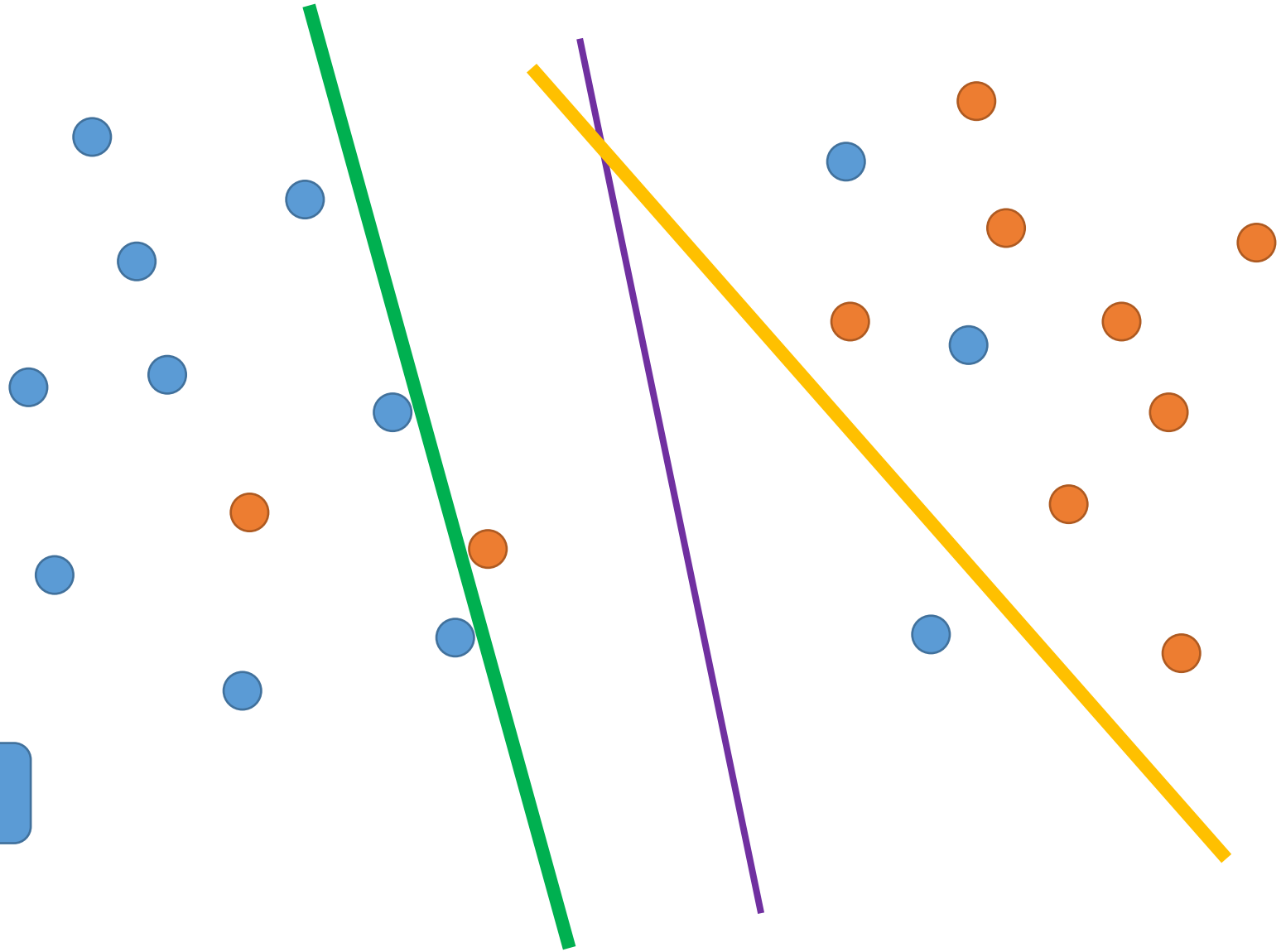
$$\begin{cases} \|w\|^2 + C\sum_{i=1}^{\ell} \xi_i \to \min\limits_{w, w_0, \xi_i} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Линейно неразделимый случай



Маленькое C

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:

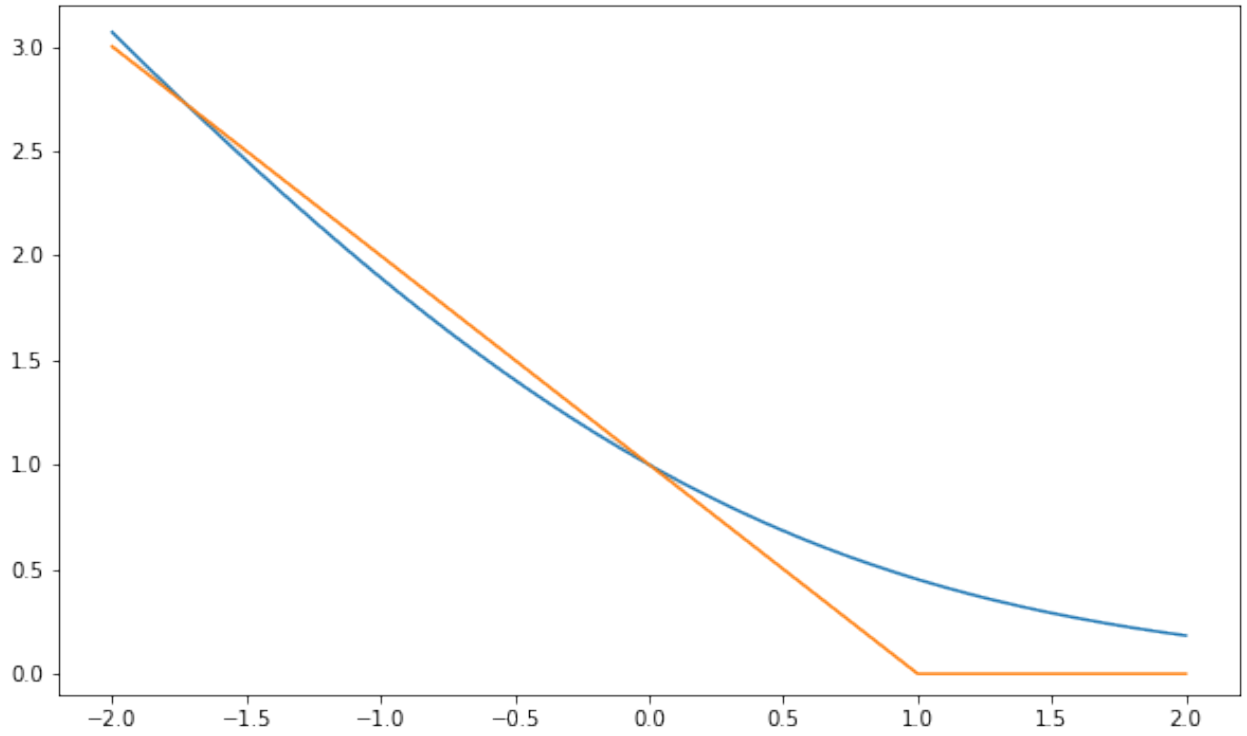
$$\xi_i \geq \max\big(0, 1 - y_i(\langle w, x_i \rangle + w_0)\big)$$

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

Калибровка вероятностей

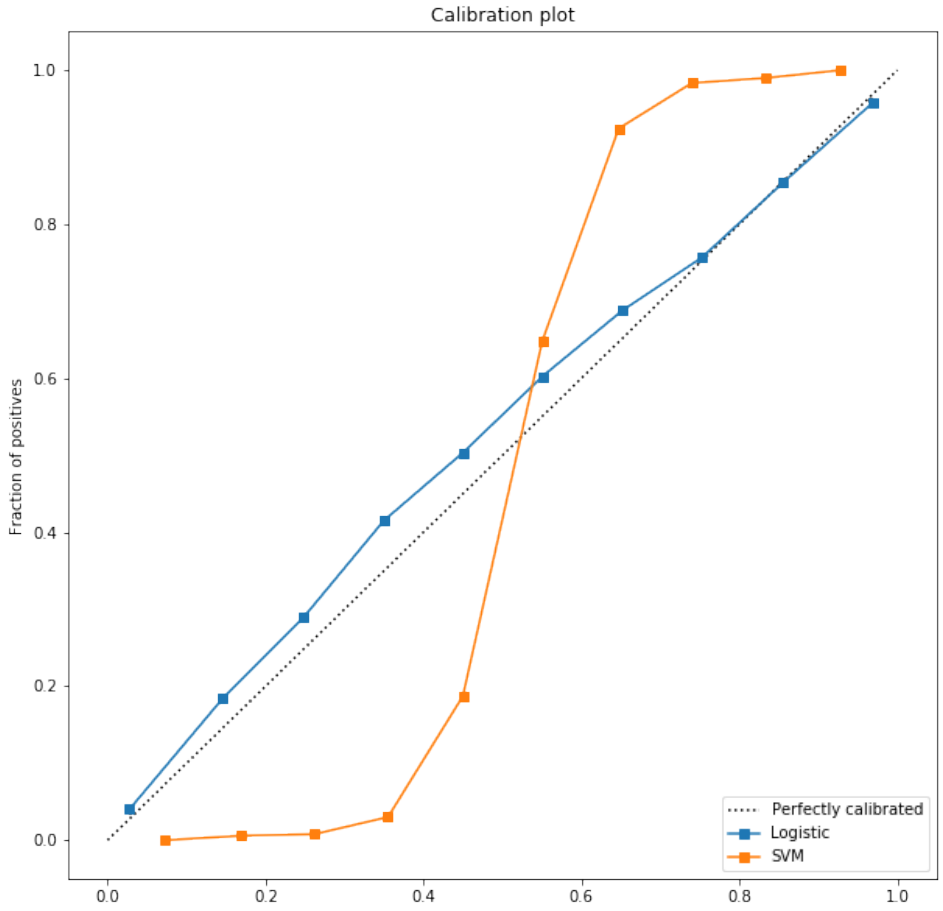
Предсказание вероятностей

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 $[0, t_1], [t_1, t_2], \ldots, [t_{n-1}, 1]$ — 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



Калибровка модели

- Задача: найти преобразование $c\big(b(x)\big)$, которое «выпрямляет» калибровочную кривую
- Два подхода: изотонная регрессия и калибровка Платта

Калибровка Платта

- Обучающая выборка: $(b(x_i), y_i)_{i=1}^{\ell}$
- Один признак
- Будем использовать log-loss — мы знаем, что он позволяет корректно оценивать вероятности

Калибровка Платта

- Обучающая выборка: $(b(x_i), y_i)_{i=1}^{\ell}$
- Один признак
- Будем использовать log-loss — мы знаем, что он позволяет корректно оценивать вероятности

$$c\big(b(x)\big) = \frac{1}{1 + \exp(p * b(x) + q)}$$

$$-\sum_{i=1}^{\ell} \Big([y_i = +1] \log c\big(b(x_i)\big) + [y_i = -1] \log \Big(1 - c\big(b(x_i)\big)\Big)\Big) \to \min_{p,q}$$

Калибровка Платта

- Обучающая выборка: $(b(x_i), y_i)_{i=1}^{\ell}$
- Строить калибровку на тех же данных, на которых обучалась модель $b(x)$ — плохая идея
- На обучающей выборке $b(x)$ неплохо приближает y
- На новых данных у $b(x)$ другое распределение

Калибровка Платта

- Нужно использовать кросс-валидацию
- Строим $b(x)$ на обучающем множестве, подбираем параметры $c(x)$ на тестовом множестве
- Получаем столько моделей, сколько блоков в кросс-валидации — можно их усреднить

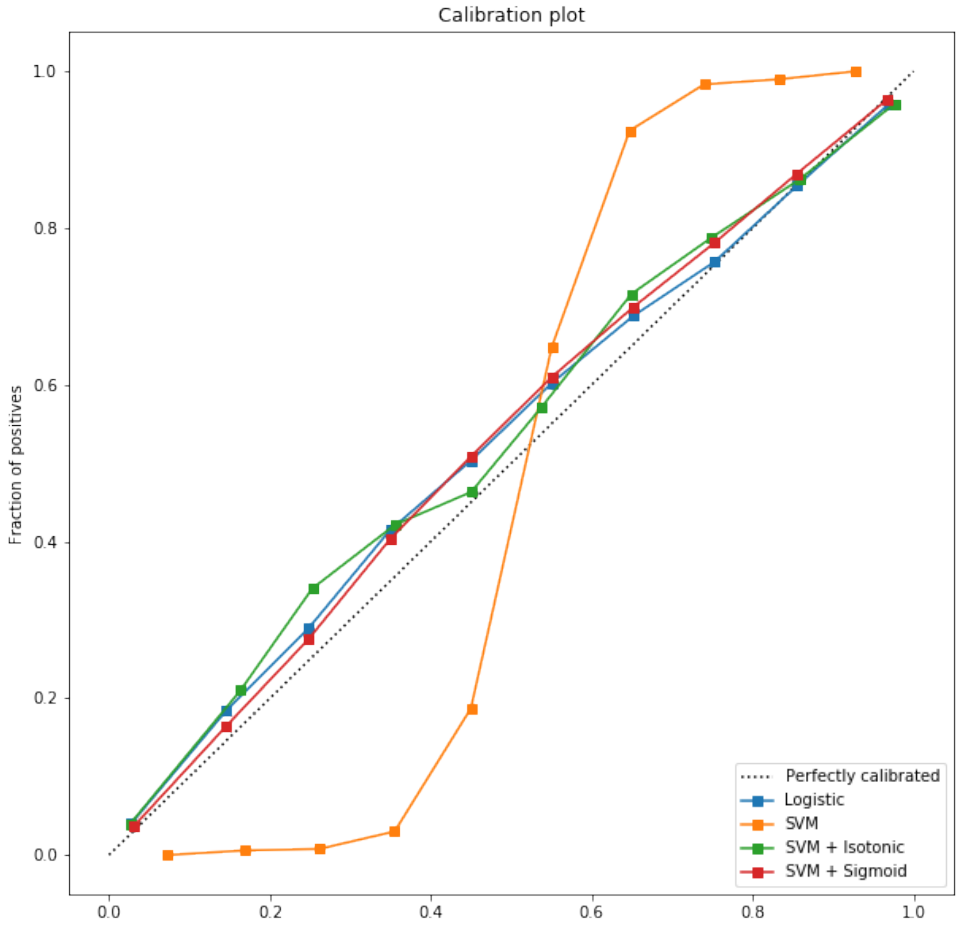
Изотонная регрессия

- Обучающая выборка: $(b(x_i), y_i)_{i=1}^{\ell}$
- Один признак
- Подбираем такую функцию $c\big(b(x)\big)$, что для $b(x_1) < b(x_2)$ выполнено $c\big(b(x_1)\big) \leq c\big(b(x_2)\big)$
- Тоже надо подбирать на отложенной выборке или кросс-валидации

Изотонная регрессия

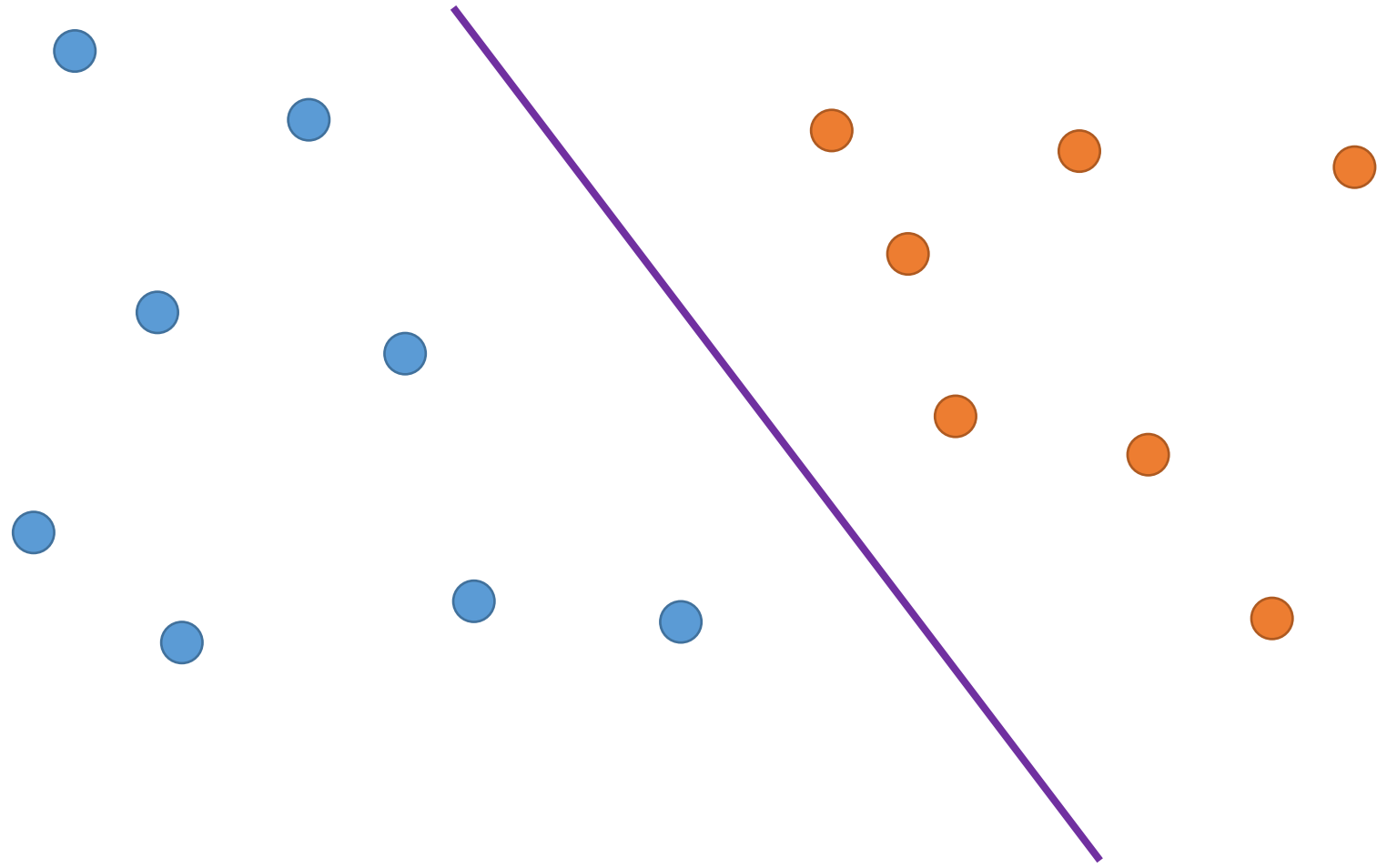
- Обучающая выборка: $(b(x_i), y_i)_{i=1}^{\ell}$
- Один признак
- Подбираем такую функцию $c\big(b(x)\big)$, что для $b(x_1) < b(x_2)$ выполнено $c\big(b(x_1)\big) \leq c\big(b(x_2)\big)$
- Тоже надо подбирать на отложенной выборке или кросс-валидации

Калибровка вероятностей

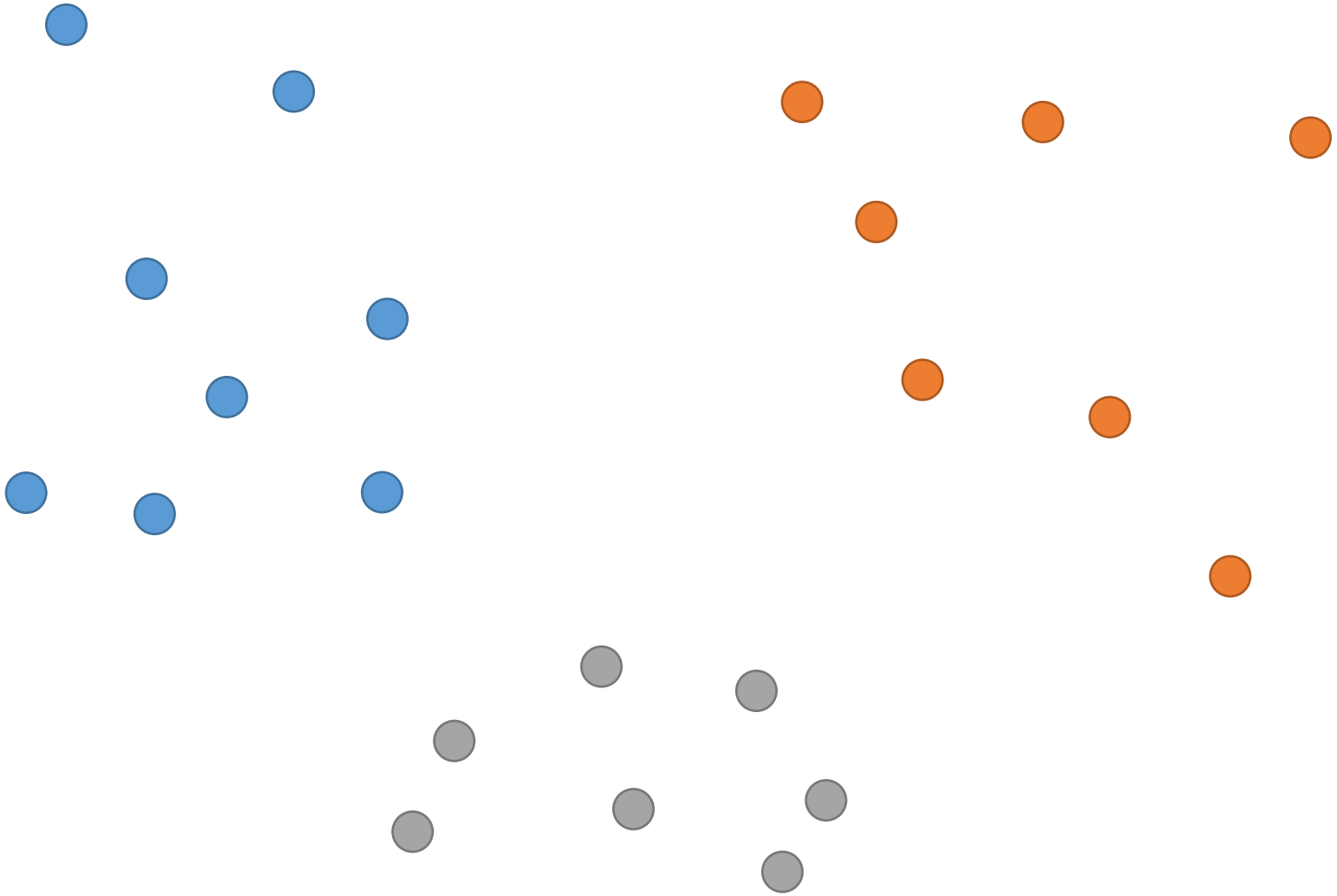


Многоклассовая классификация

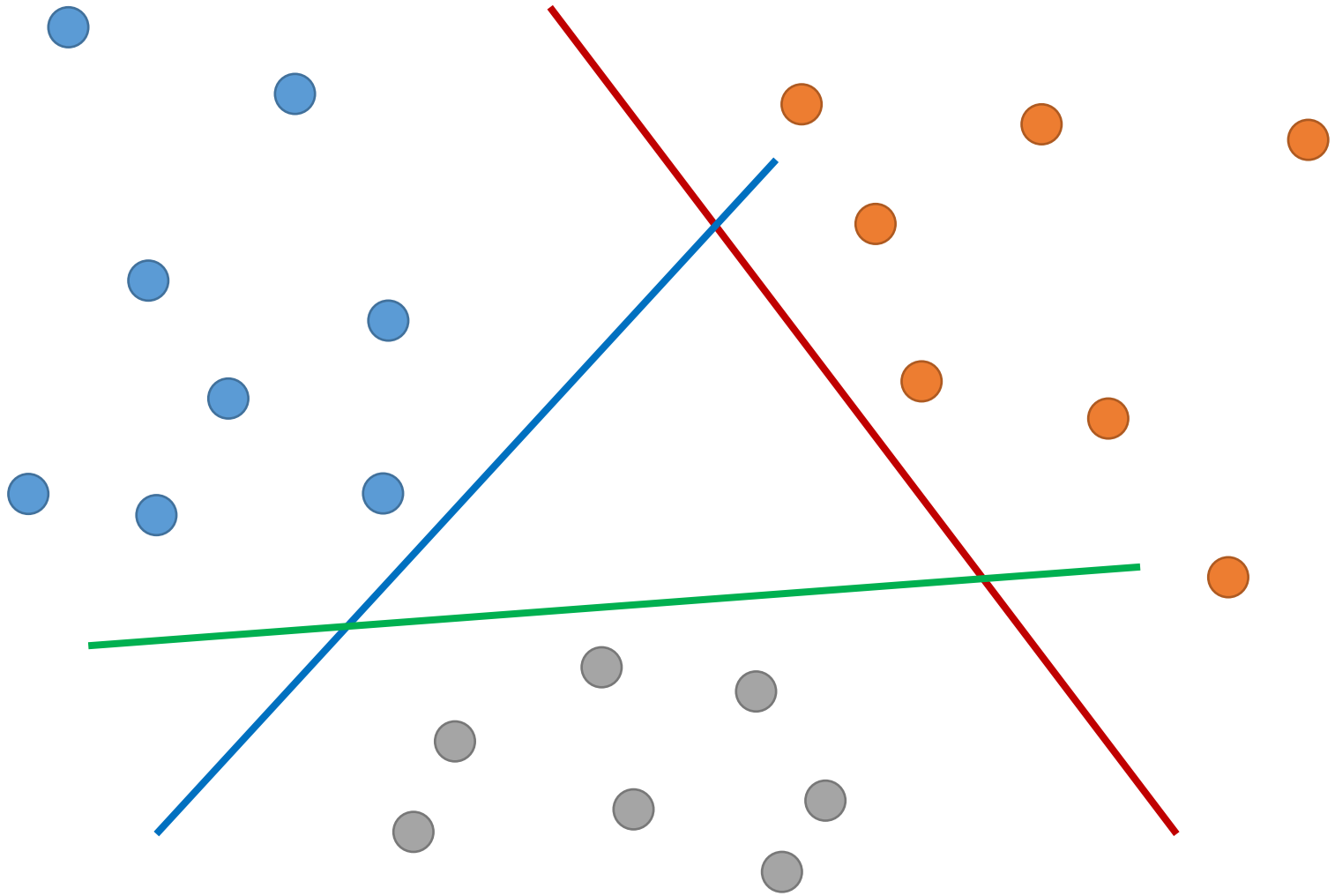
Бинарная классификация



Многоклассовая классификация



Многоклассовая классификация



One-vs-all

- K классов: $\mathbb{Y} = \{1, \dots, K\}$
- $X_k = (x_i, [y_i = k])_{i=1}^{\ell}$
- Обучаем $a_k(x)$ на X_k, $k = 1, \dots, K$
- $a_k(x)$ должен выдавать оценки принадлежности классу (например, $\langle w, x \rangle$ или $\sigma(\langle w, x \rangle)$)
- Итоговая модель:

$$a(x) = \arg \max_{k=1,\dots K} a_k(x)$$

One-vs-all

- Модель $a_k(x)$ при обучении не знает, что её выходы будут сравнивать с выходами других моделей
- Нужно обучать K моделей

All-vs-all

- $X_{km} = \{(x_i, y_i) \in X \mid y_i = k \text{ или } y_i = m\}$
- Обучаем $a_{km}(x)$ на X_{km}
- Итоговая модель:

$$a(x) = \arg \max_{k \in \{1,\ldots,K\}} \sum_{m=1}^{K} [a_{km}(x) = k]$$

All-vs-all

- Нужно обучать порядка K^2 моделей
- Зато каждую обучаем на небольшой выборке

Доля ошибок

- Функционал ошибки — доля ошибок (error rate)

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i]$$

- Нередко измеряют долю верных ответов (accuracy):

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i]$$

- Подходит для многоклассового случая!

Общие подходы

Микро-усреднение	Макро-усреднение
Вычисляем $\mathrm{TP}_k, \mathrm{FP}_k, \mathrm{FN}_k, \mathrm{TN}_k$ для каждого класса	Вычисляем нужную метрику для каждого класса (например, $\mathrm{precision}_1, \ldots, \mathrm{precision}_K$)
Суммируем по всем классам, получаем $\mathrm{TP}, \mathrm{FP}, \mathrm{FN}, \mathrm{TN}$	Усредняем по всем классам
Подставляем их в формулу для precision/recall/…	
Крупные классы вносят больший вклад	Игнорирует размеры классов

Работа с категориальными признаками

Кодирование категориальных признаков

- Значения признака «район»: $U = \{u_1, \dots, u_m\}$
- Новые признаки вместо x_j: $[x_j = u_1], \dots, [x_j = u_m]$
- One-hot кодирование

Кодирование категориальных признаков

Район
ЦАО
ЮАО
ЦАО
САО
ЮАО

ЦАО	ЮАО	САО
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0

Кодирование категориальных признаков

Район	Цена
ЦАО	10.000.000
ЮАО	4.000.000
ЦАО	9.000.000
САО	7.000.000
ЮАО	5.000.000

Счётчики

- Значения признака x_j: $U_j = \{u_1, \ldots, u_m\}$
- Посчитаем все категории в обучающей выборке:

$$\text{count}(j, u_p) = \sum_{i=1}^{\ell} [x_{ij} = u_p]$$

Счётчики

- Значения признака x_j: $U_j = \{u_1, \ldots, u_m\}$
- Для регрессии посчитаем суммарный ответ в категории:

$$\text{target}(j, u_p) = \sum_{i=1}^{\ell} [x_{ij} = u_p] \, y_i$$

Счётчики

- Значения признака x_j: $U_j = \{u_1, \ldots, u_m\}$
- Для классификации посчитаем классы в категории:

$$\text{target}_k(j, u_p) = \sum_{i=1}^{\ell} [x_{ij} = u_p] [y_i = k]$$

Счётчики

- Mean-target encoding
- Задача регрессии
- Заменим категориальный признак на числовой:

$$\widetilde{x_{ij}} = \frac{\mathrm{target}\left(j, x_{ij}\right)}{\mathrm{count}\left(j, x_{ij}\right)}$$

Счётчики

- Mean-target encoding
- Задача классификации
- Заменим категориальный признак на K числовых:

$$\widetilde{x_{ij}} = \left(\frac{\text{target}_1\left(j, x_{ij}\right)}{\text{count}\left(j, x_{ij}\right)}, \dots, \frac{\text{target}_K\left(j, x_{ij}\right)}{\text{count}\left(j, x_{ij}\right)} \right)$$

Кодирование категориальных признаков

Район	Счётчик	Цена
ЦАО	9.500.000	10.000.000
ЮАО	4.500.000	4.000.000
ЦАО	9.500.000	9.000.000
САО	7.000.000	7.000.000
ЮАО	4.500.000	5.000.000

Борьба с переобучением в счётчиках

- Проблема в основном с редкими категориями
- Решение 1: добавление шума

Район	Счётчик	Цена
ЦАО	9.130.000	10.000.000
ЮАО	4.023.000	4.000.000
ЦАО	10.124.000	9.000.000
САО	7.942.000	7.000.000
ЮАО	4.728.000	5.000.000

Борьба с переобучением в счётчиках

- Проблема в основном с редкими категориями
- Решение 2: добавление априорных величин в счётчики

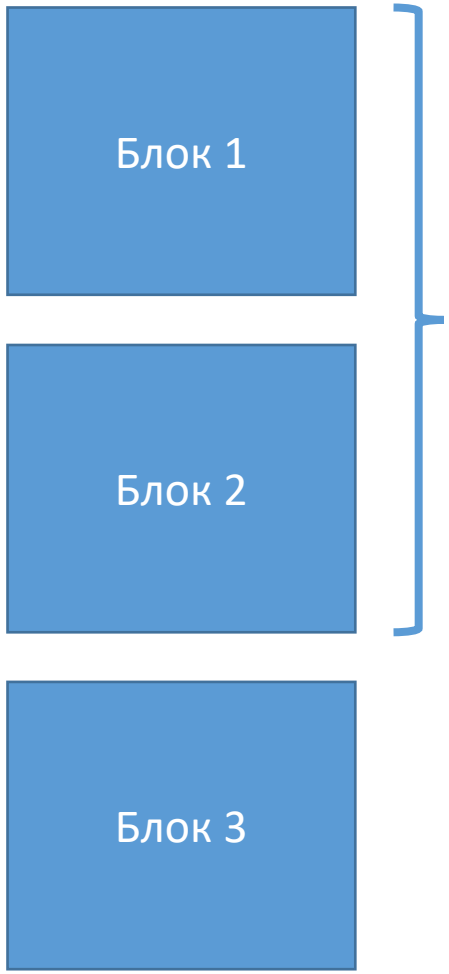
$$\widetilde{x_{ij}} = \frac{\text{target}\left(j, x_{ij}\right) + a}{\text{count}\left(j, x_{ij}\right) + b}$$

Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков

Борьба с переобучением в счётчиках

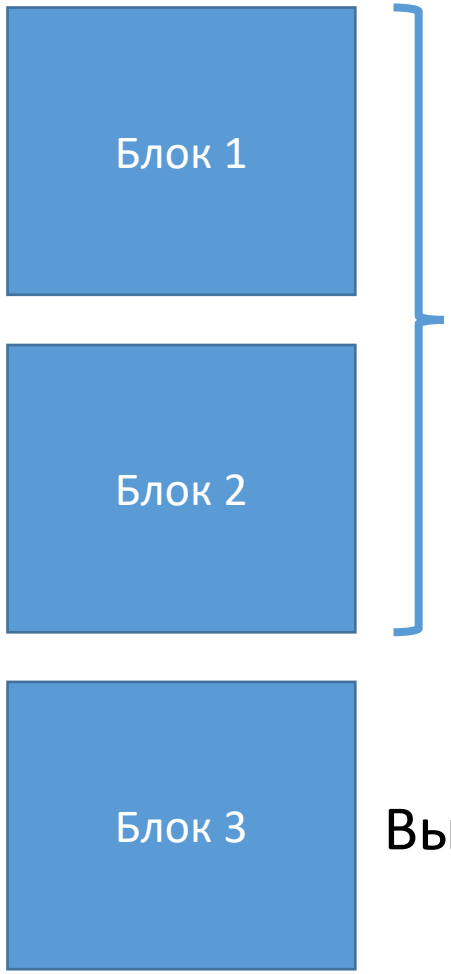
- Решение 3: кросс-валидация счётчиков



Считаем $\mathrm{count}(j, u_p)$ и $\mathrm{target}(j, u_p)$

Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков

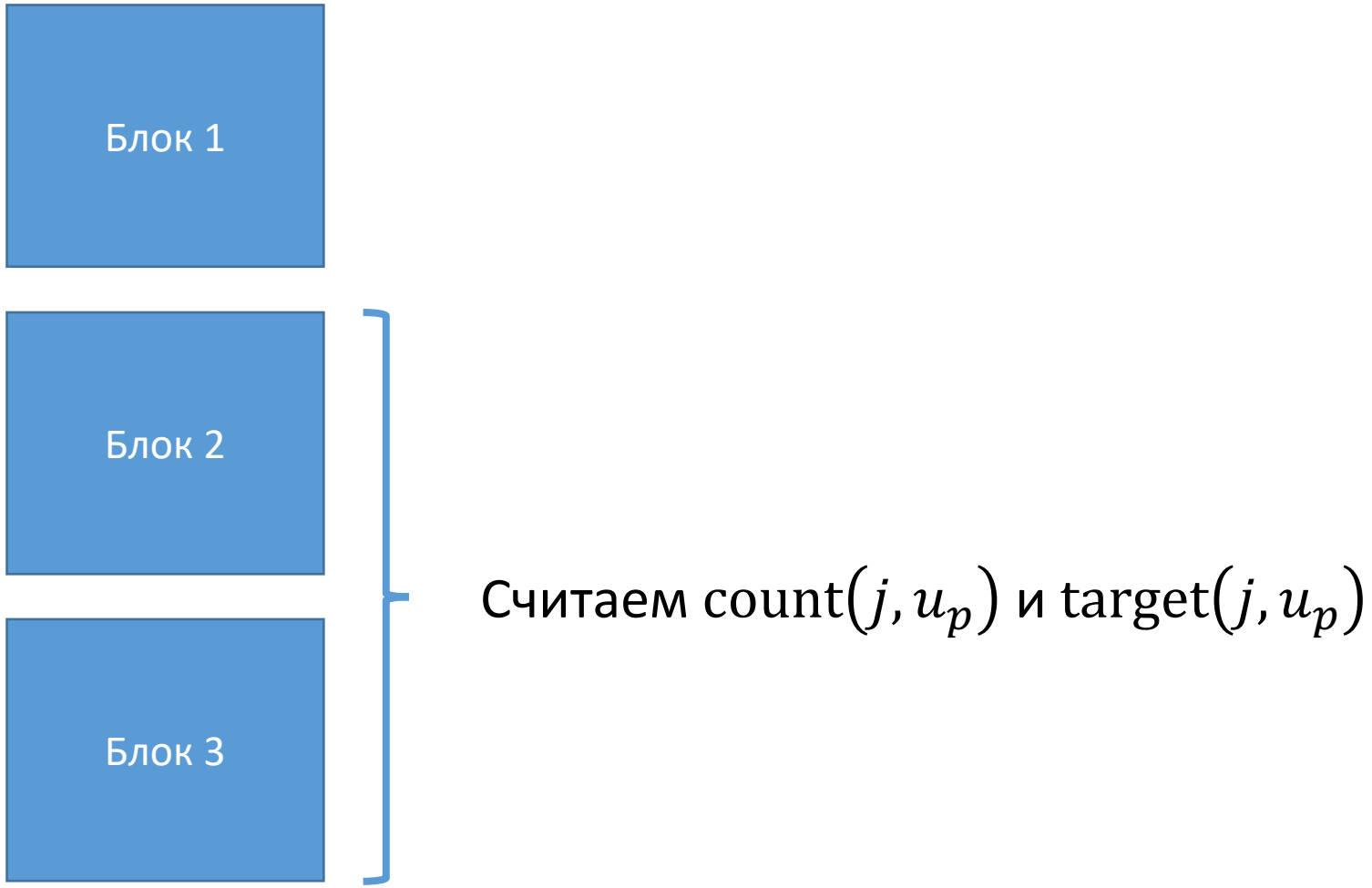


Считаем $\mathrm{count}(j, u_p)$ и $\mathrm{target}(j, u_p)$

Вычисляем признаки: $\widetilde{x_{ij}} = \frac{\mathrm{target}(j, x_{ij})}{\mathrm{count}(j, x_{ij})}$

Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков



Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков

Вычисляем признаки: $\widetilde{x_{ij}} = \frac{\text{target}(j, x_{ij})}{\text{count}(j, x_{ij})}$

Считаем $\text{count}(j, u_p)$ и $\text{target}(j, u_p)$

Резюме

- Счётчики позволяют заменить категориальный признак на один числовой
- Могут привести к переобучению
- Можно бороться с ним через добавление шума, априорных значений или кросс-валидацию