Введение в анализ данных

Лекция 10

Линейная классификация

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2021

Метод опорных векторов

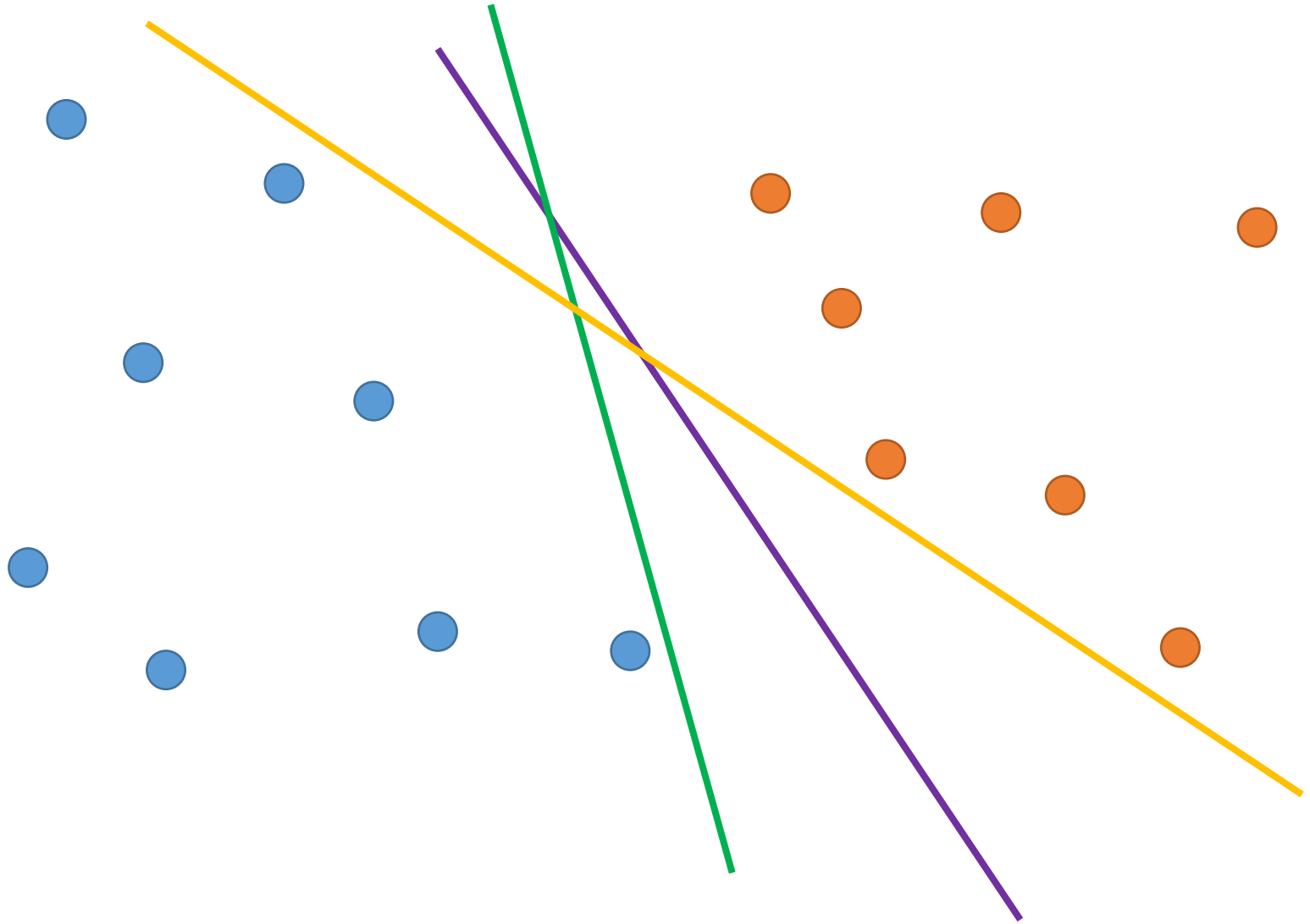
Hinge loss

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

- Минимизация верхней оценки:

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

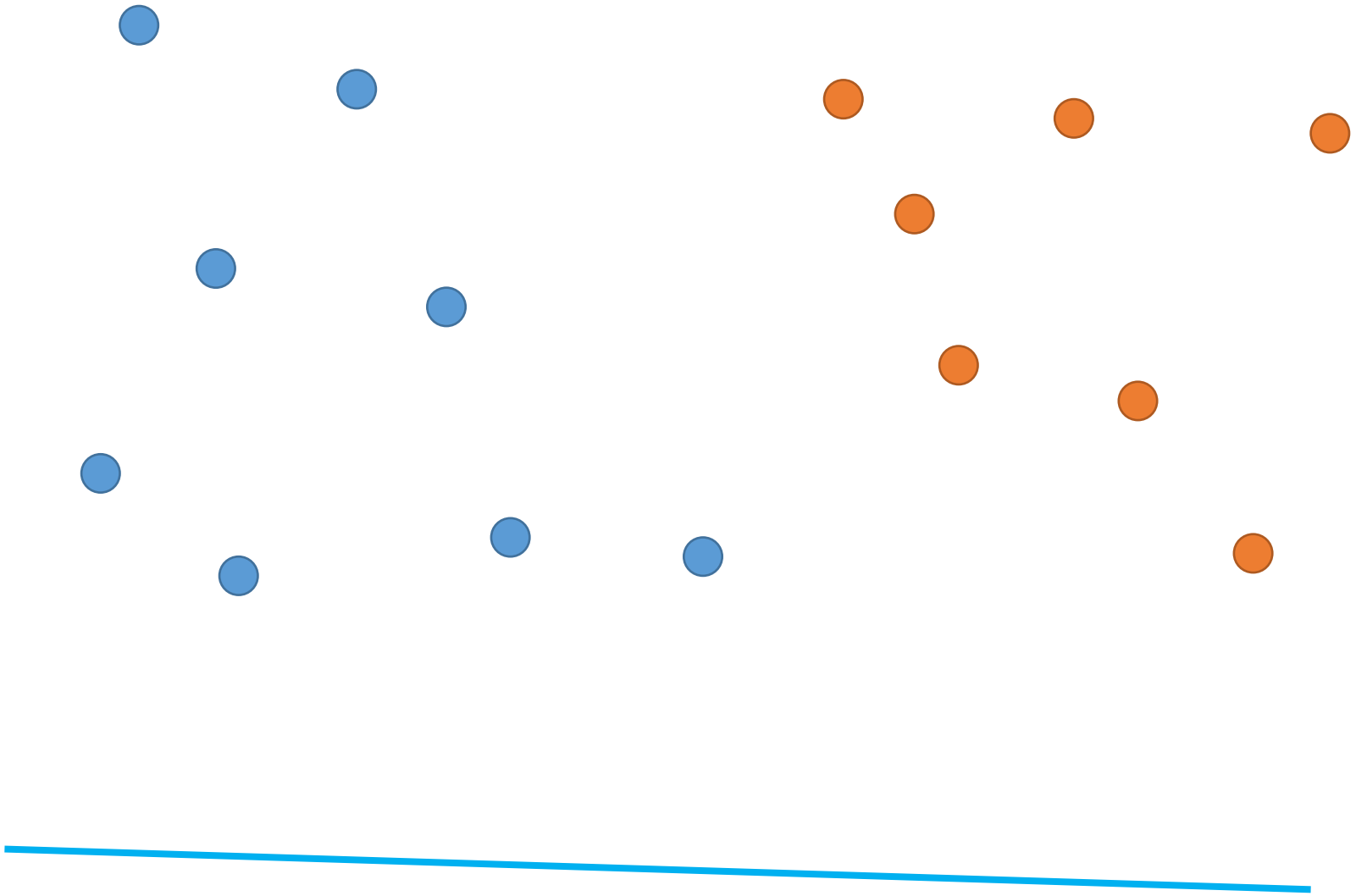
Какой классификатор лучше?



Отступ классификатора

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

Отступ классификатора



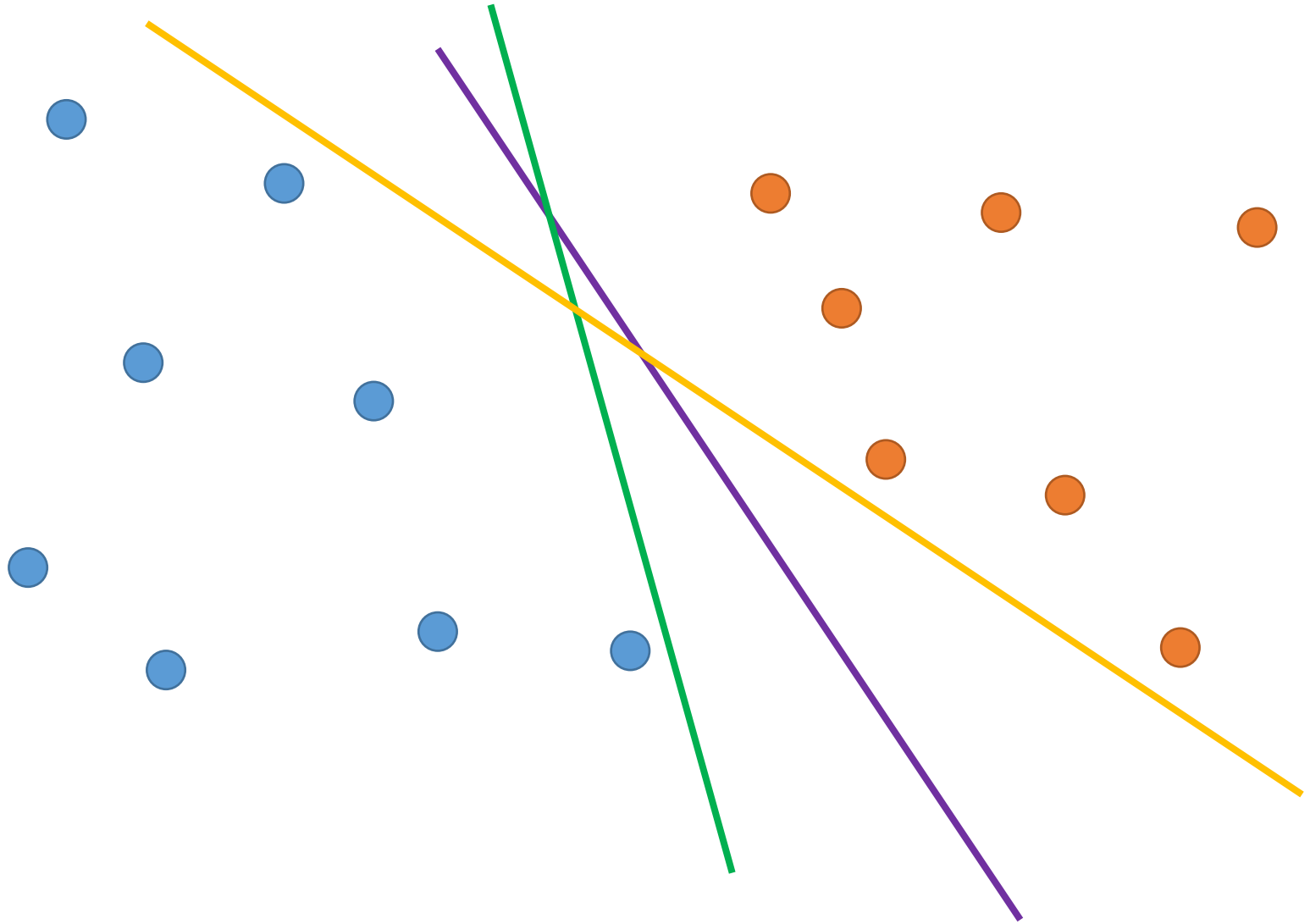
Отступ классификатора

- 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

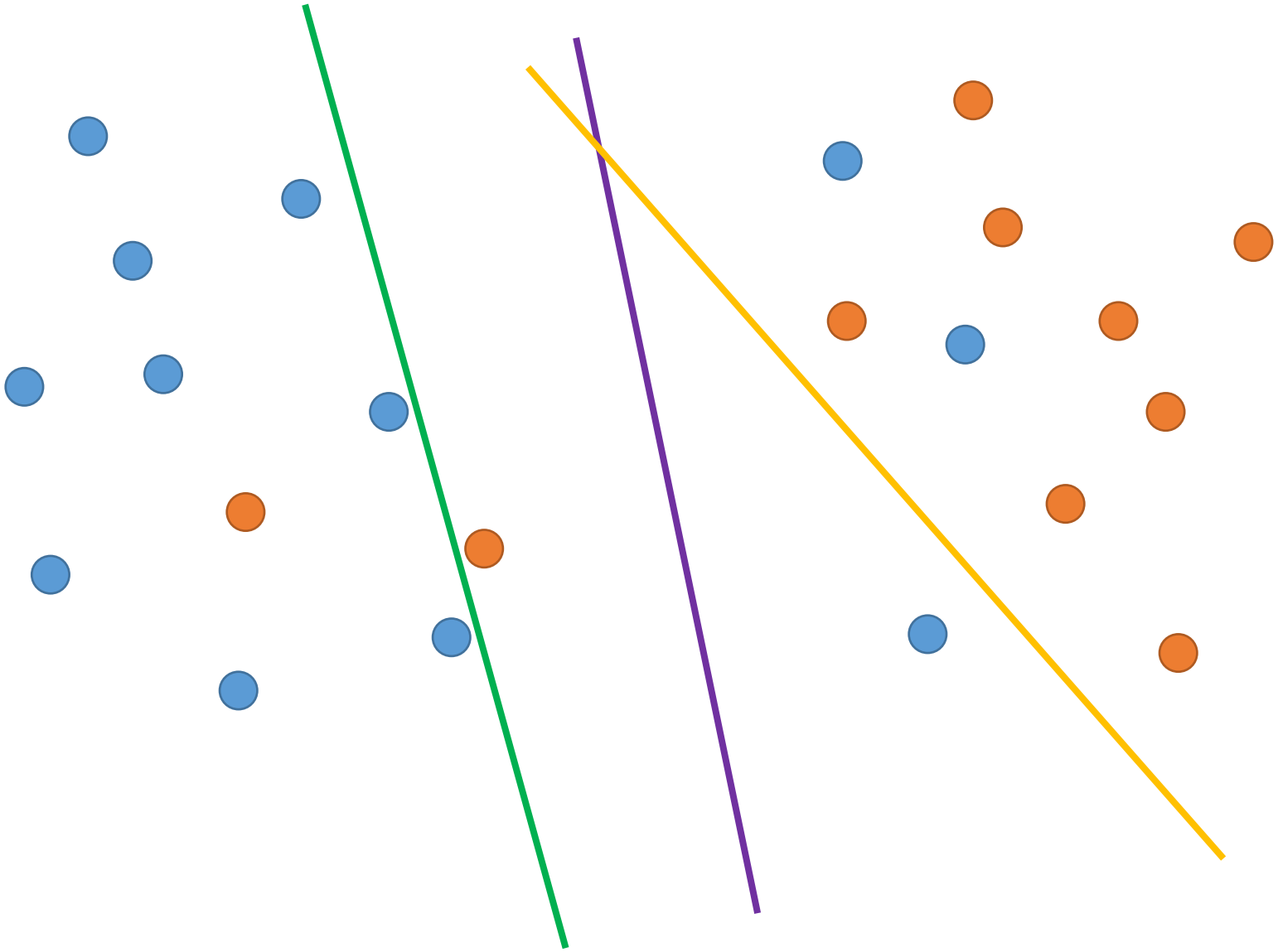
$$\frac{1}{\|w\|} \to \max_w$$

- 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

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

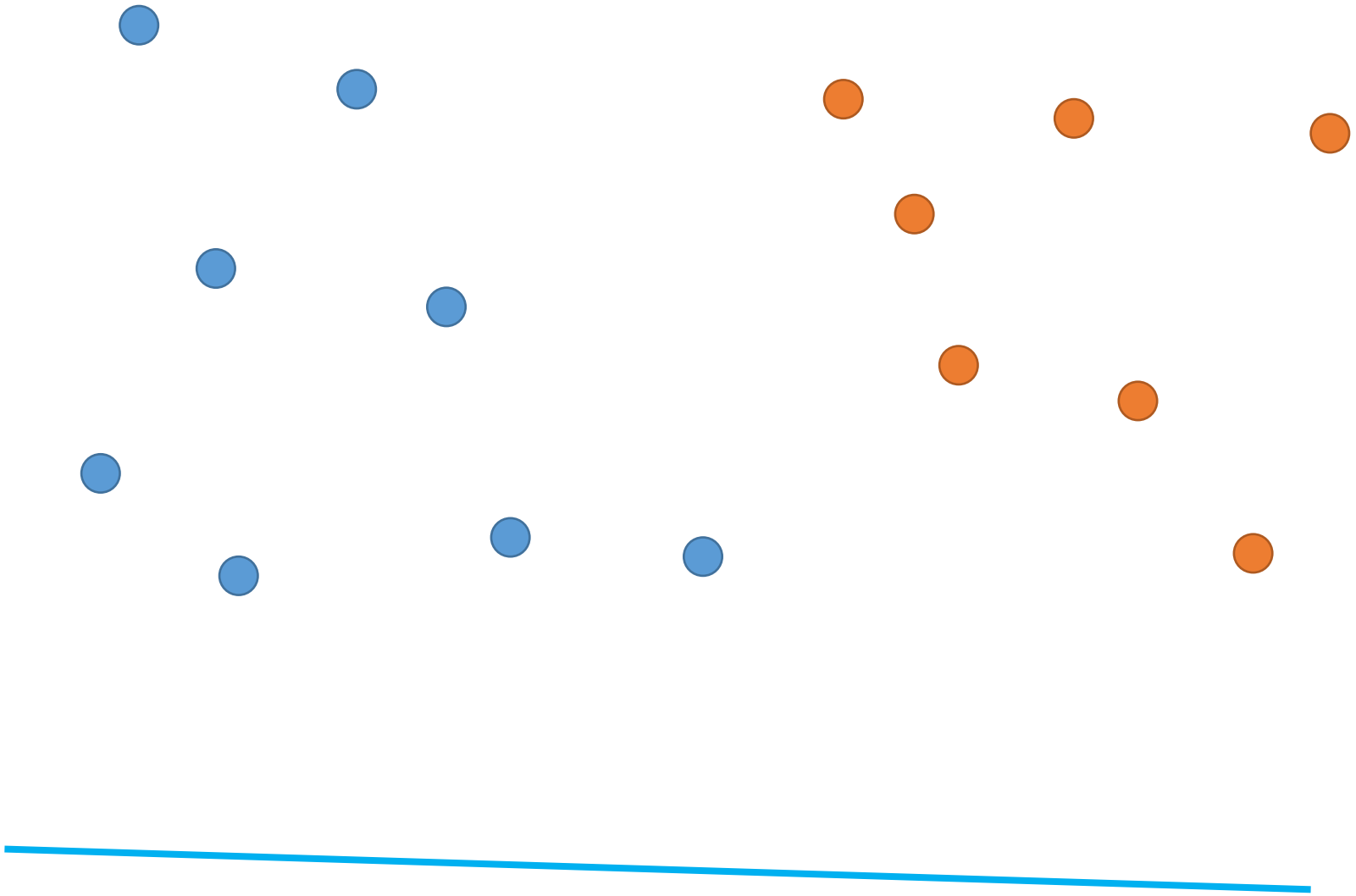
Линейно неразделимый случай

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

Линейно неразделимый случай

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

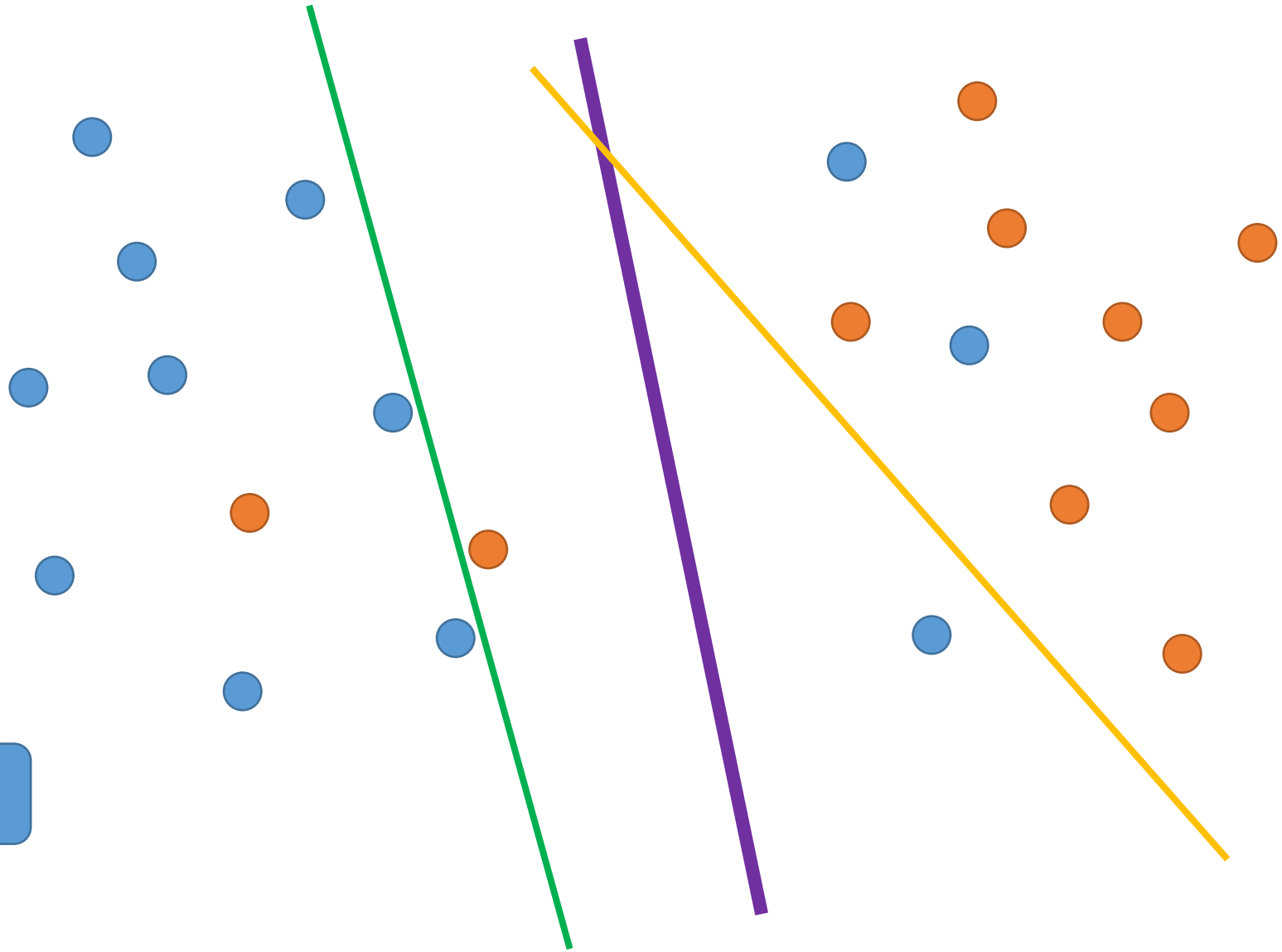
Отступ классификатора



Метод опорных векторов

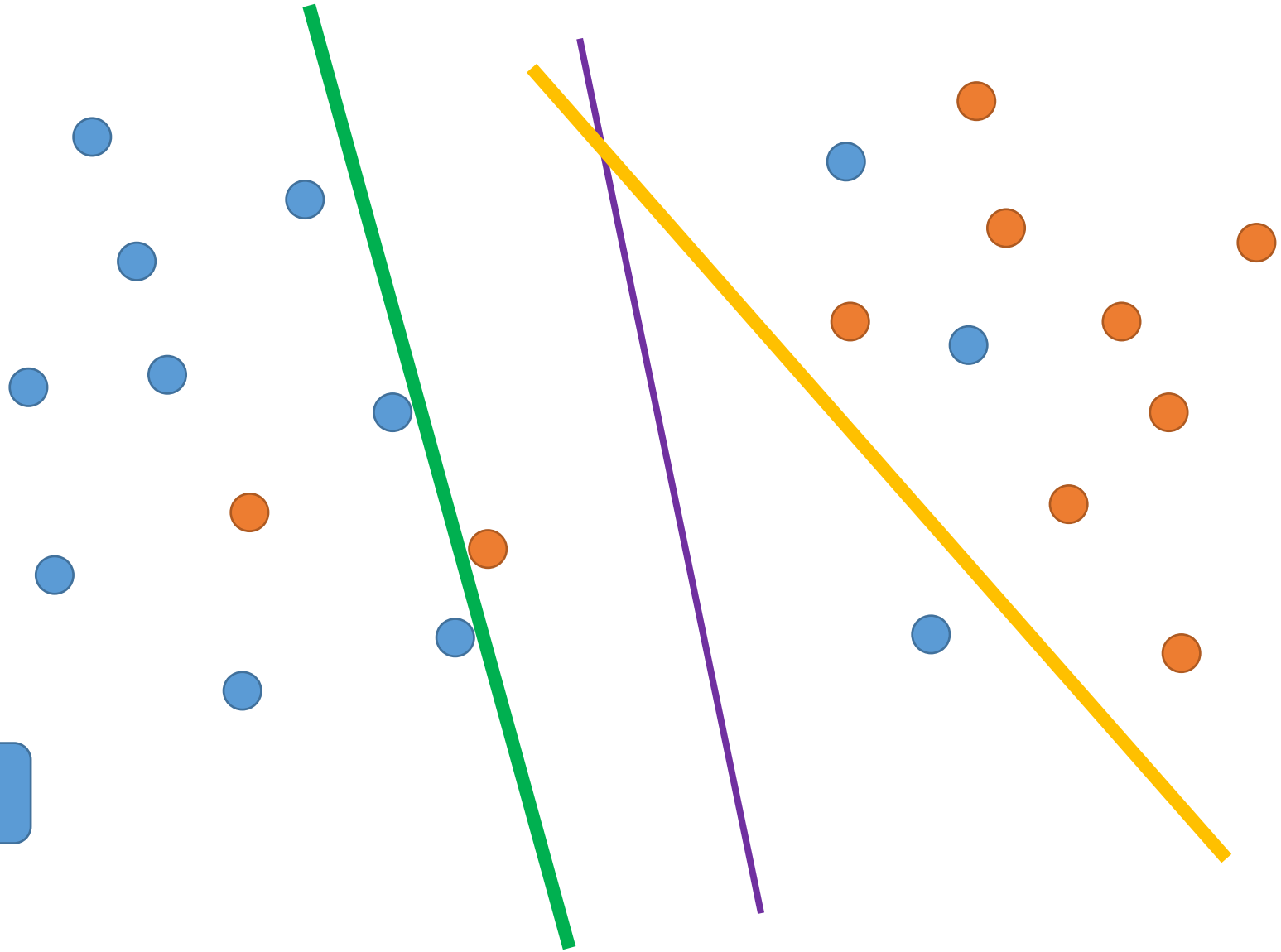
$$\begin{cases} \|w\|^2 + C\sum_{i=1}^{\ell} \xi_i \to \min\limits_{w, w_0, \xi_i} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

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:

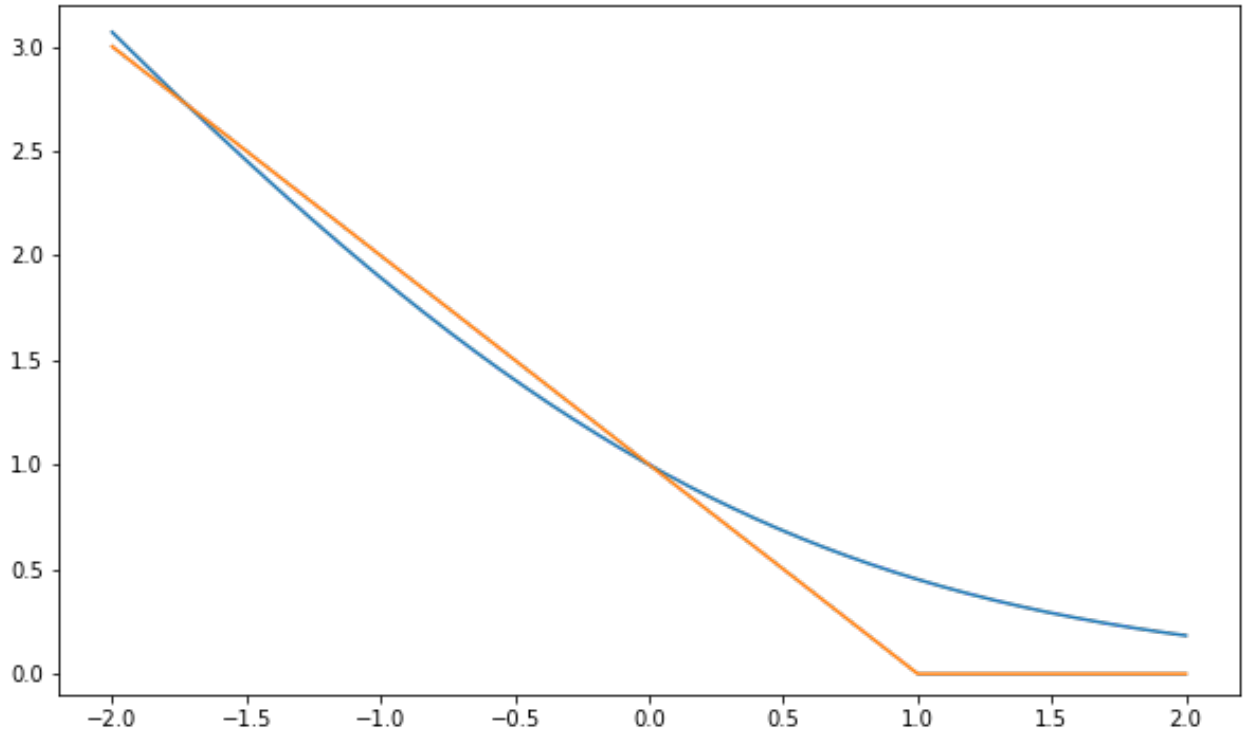
$$\xi_i \geq \max\big(0, 1 - y_i(\langle w, x_i \rangle + w_0)\big)$$

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



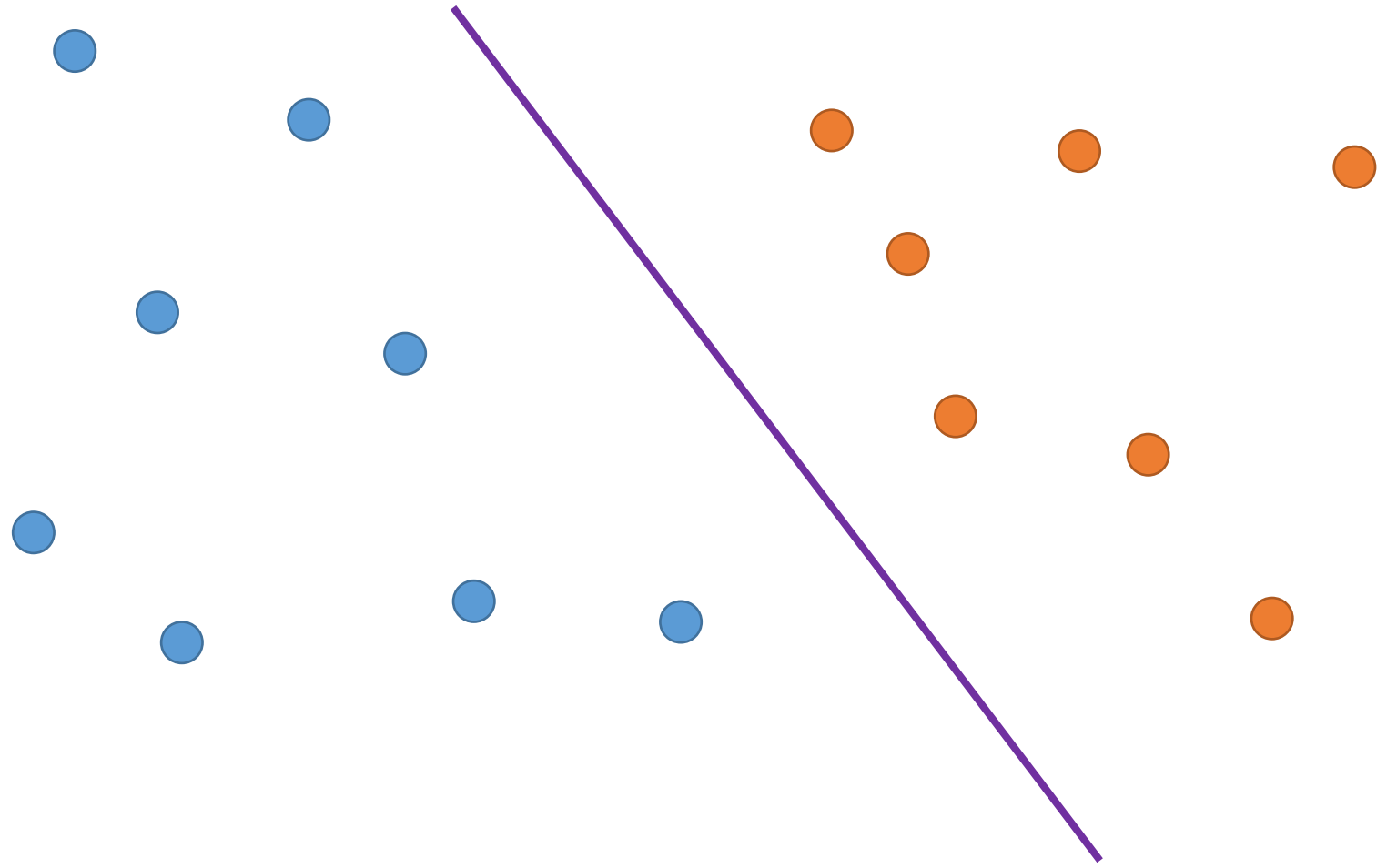
Резюме

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

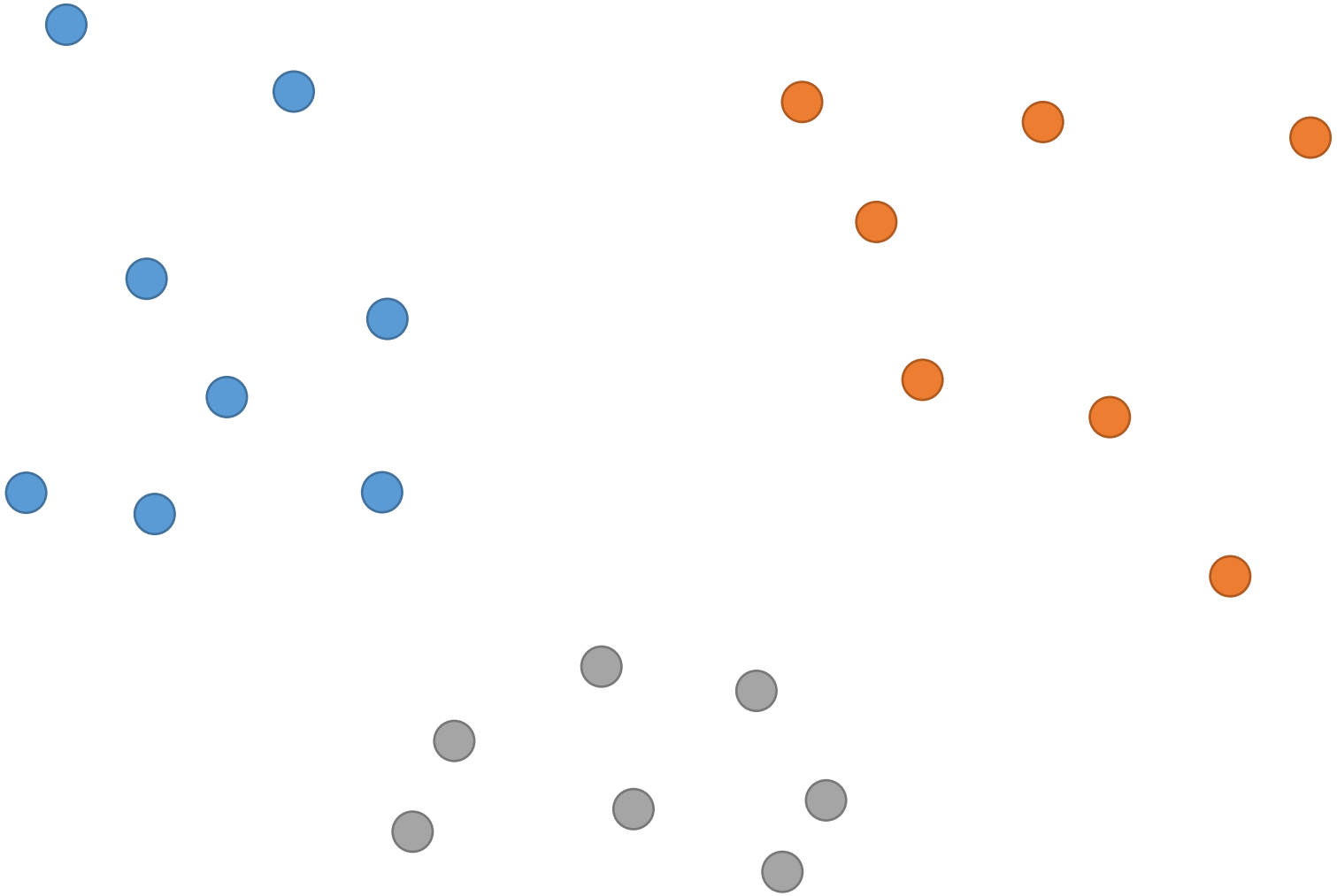
- Метод опорных векторов основан на идее максимизации отступа классификатора

Многоклассовая классификация

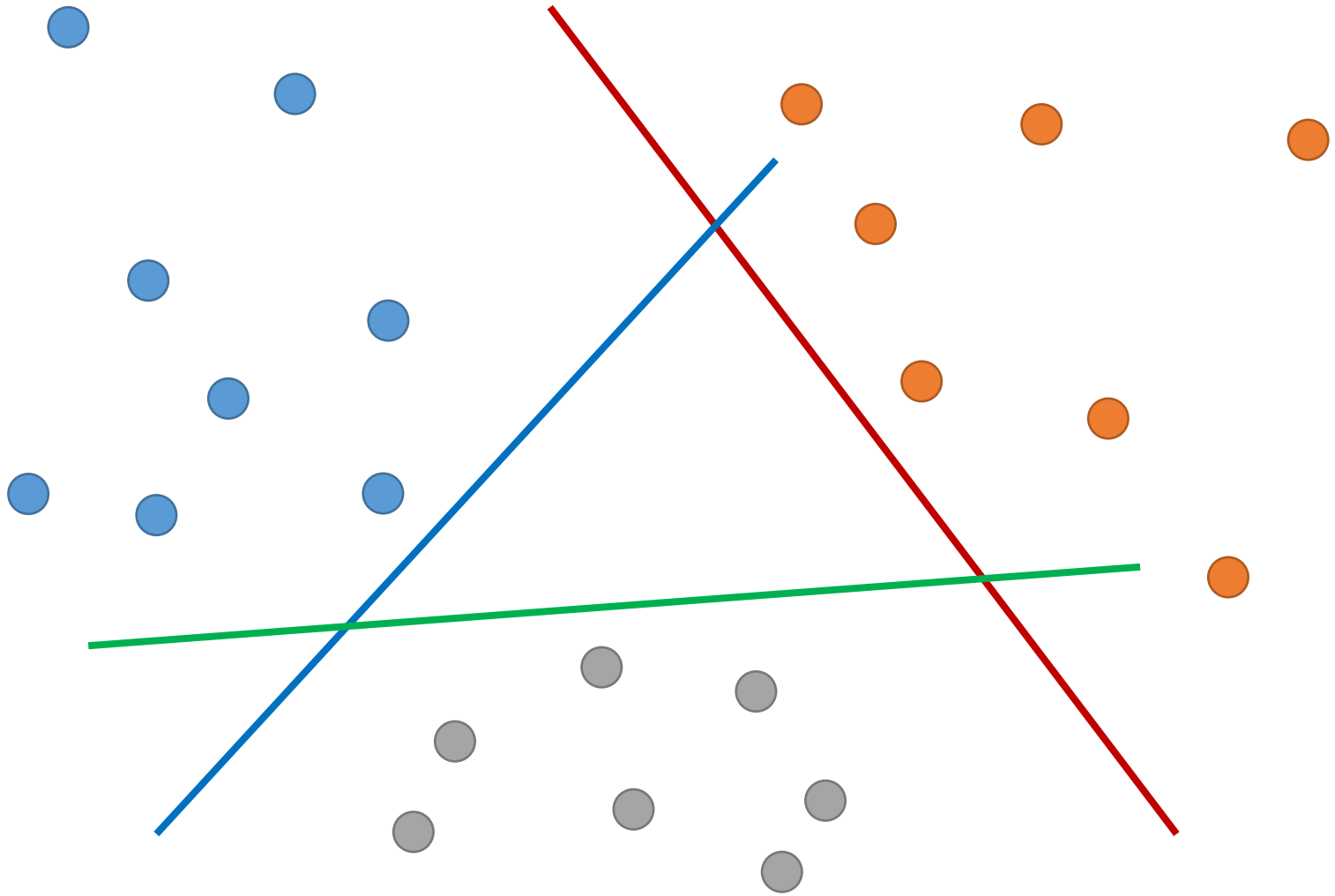
Бинарная классификация



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

Как делать нелинейные модели

Предсказание стоимости квартиры

- Признаки: площадь, этаж, расстояние до метро и т.д.
- Целевая переменная: рыночная стоимость квартиры

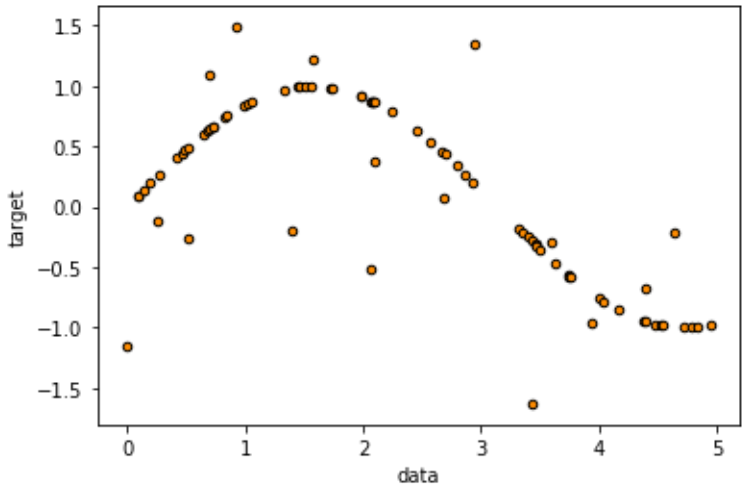
Предсказание стоимости квартиры

- Линейная модель:

$$a(x) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{этаж})$$

$$+w_3 * (\text{расстояние до метро}) + \cdots$$

- Вряд ли признаки линейно связаны с целевой переменной



Предсказание стоимости квартиры

- Линейная модель:

$$a(x) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{этаж})$$
$$+w_3 * (\text{расстояние до метро}) + \cdots$$

- Вряд ли признаки не связаны между собой

Предсказание стоимости квартиры

- Линейная модель с полиномиальными признаками:

$$a(x) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{этаж})$$
$$+w_3 * (\text{расстояние до метро}) + w_4 * (\text{площадь})^2$$
$$+w_5 * (\text{этаж})^2 + w_6 * (\text{расстояние до метро})^2$$
$$+w_7 * (\text{площадь}) * (\text{этаж}) + \cdots$$

Предсказание стоимости квартиры

- Линейная модель с полиномиальными признаками:

$$a(x) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{этаж})$$
$$+w_3 * (\text{расстояние до метро}) + w_4 * (\text{площадь})^2$$
$$+w_5 * (\text{этаж})^2 + w_6 * (\text{расстояние до метро})^2$$
$$+w_7 * (\text{площадь}) * (\text{этаж}) + \cdots$$

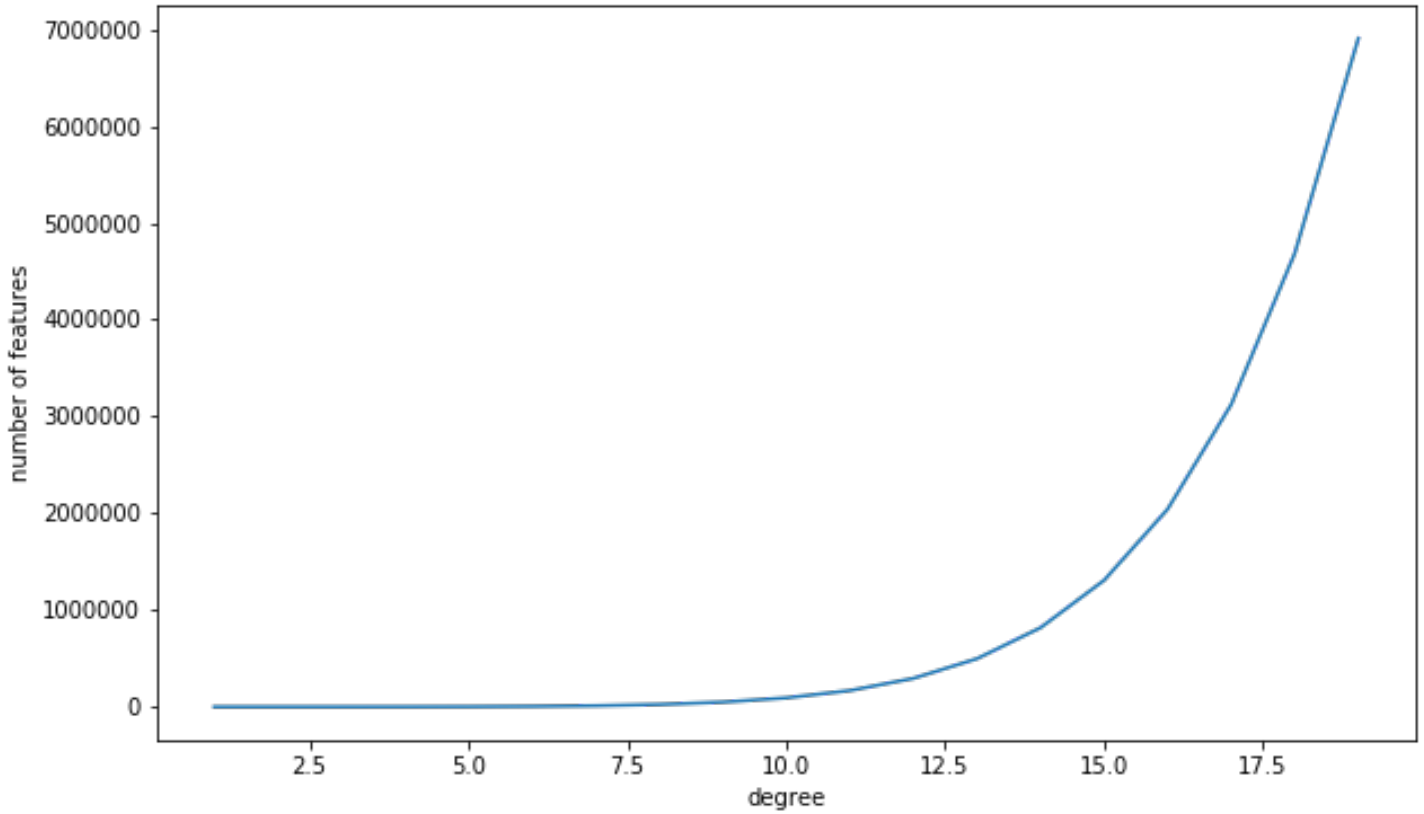
- Может быть сложно интерпретировать модель
- Что такое $(\text{расстояние до метро}) * (\text{этаж})^2$?

Предсказание стоимости квартиры

- Допустим, изначально имеем 10 признаков
- Полиномиальных степени 2: 55
- Полиномиальных степени 3: 220
- Полиномиальных степени 4: 715

Предсказание стоимости квартиры

- Линейная модель с полиномиальными признаками:



Предсказание стоимости квартиры

- Линейная модель с полиномиальными признаками:

$$a(x) = w_0 + w_1 * [30 < \text{площадь} < 50]$$

$$+w_2 * [50 < \text{площадь} < 80] + \cdots$$

$$+w_{20} * [2 < \text{этаж} < 5] + \cdots$$

$$+w_{100} * [30 < \text{площадь} < 50][2 < \text{этаж} < 5] + \cdots$$

- Признаки интерпретируются куда лучше: $[30 < \text{площадь} < 50][2 < \text{этаж} < 5][100 < \text{расстояние до метро} < 500]$
- Но их станет ещё больше!