



Data Mining in Action

Лекция 2. Простые методы и математика

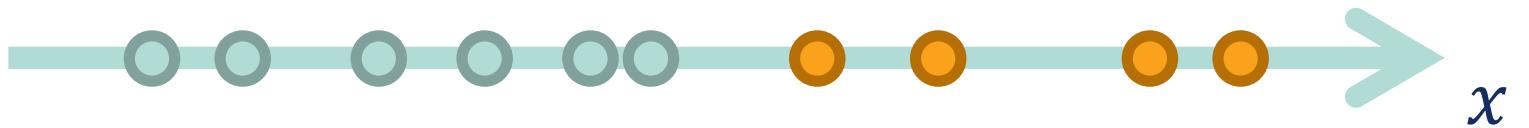
План лекции

1. Порог по одному признаку
2. От пней к деревьям
3. Сложные границы и соседи
4. Плотность и наивный байес

1. Порог по одному признаку

Выборка

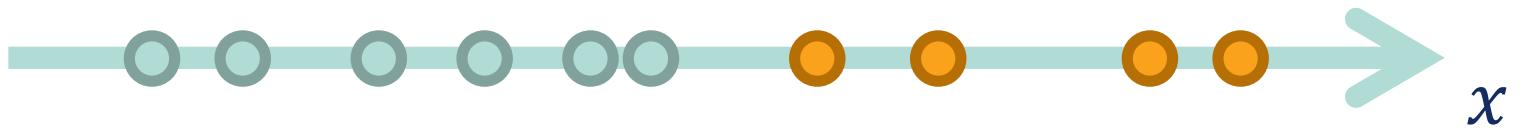
Рассмотрим выборку объектов с одним признаком x :



Как подобрать порог по признаку в задаче бинарной классификации?

Выборка

Рассмотрим выборку объектов с одним признаком x :

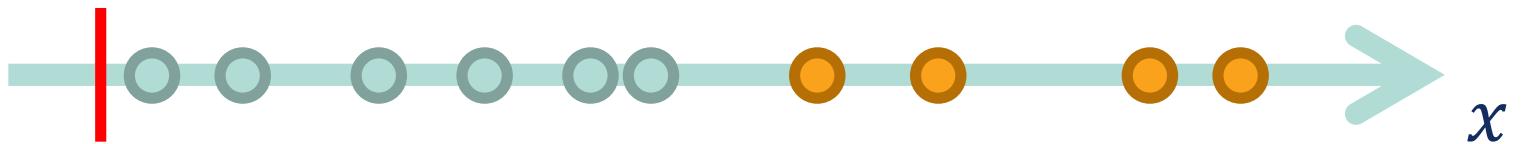


Как подобрать порог по признаку в задаче бинарной классификации?

Можно попробовать двигать перебрать пороги.
Например, сдвигая на один пример.

Выборка

Рассмотрим выборку объектов с одним признаком x :

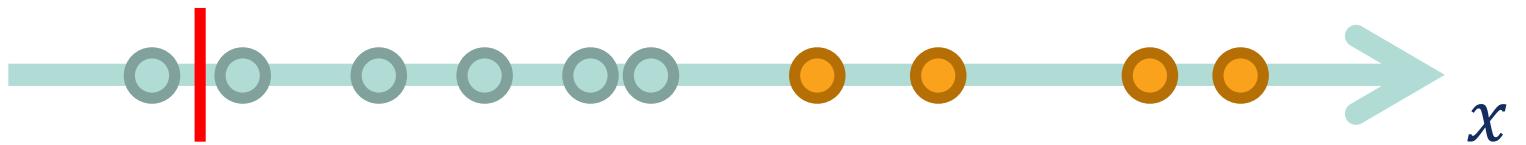


Как подобрать порог по признаку в задаче бинарной классификации?

Можно попробовать двигать перебрать пороги.
Например, сдвигая на один пример.

Выборка

Рассмотрим выборку объектов с одним признаком x :

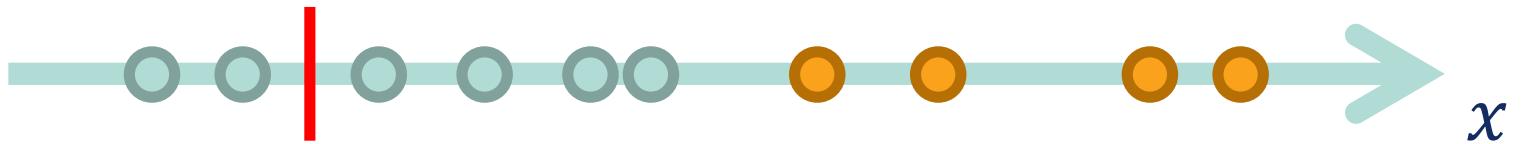


Как подобрать порог по признаку в задаче бинарной классификации?

Можно попробовать двигать перебрать пороги.
Например, сдвигая на один пример.

Выборка

Рассмотрим выборку объектов с одним признаком x :

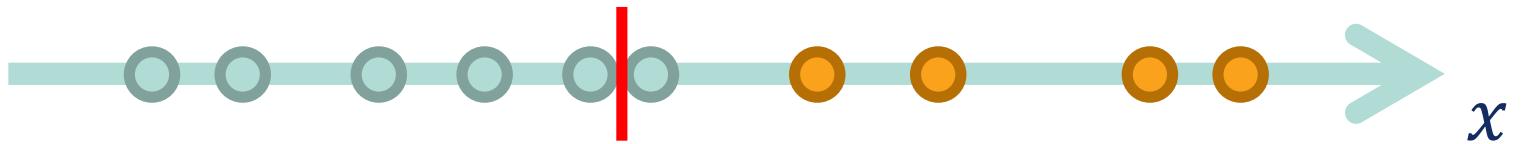


Как подобрать порог по признаку в задаче бинарной классификации?

Можно попробовать двигать перебрать пороги.
Например, сдвигая на один пример.

Выборка

Рассмотрим выборку объектов с одним признаком x :

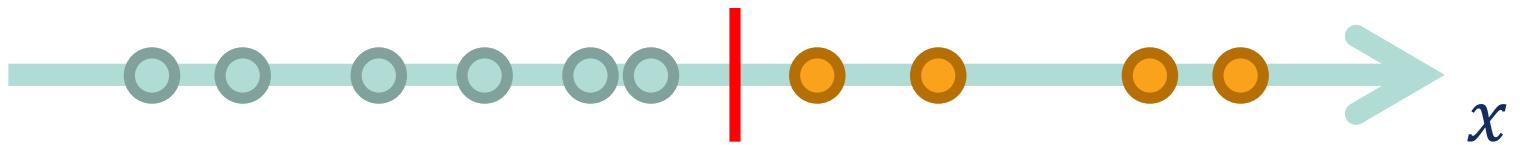


Как подобрать порог по признаку в задаче бинарной классификации?

Можно попробовать двигать перебрать пороги.
Например, сдвигая на один пример.

Выборка

Рассмотрим выборку объектов с одним признаком x :

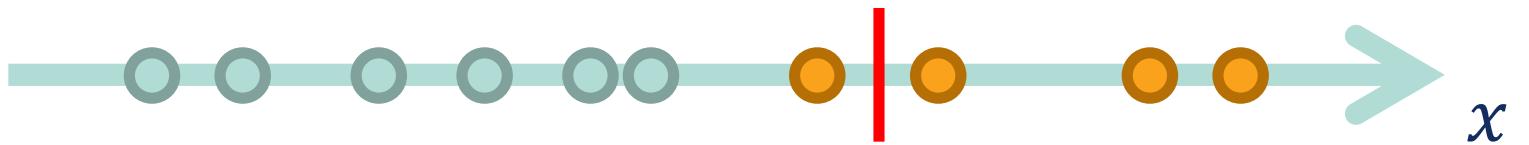


Как подобрать порог по признаку в задаче бинарной классификации?

Можно попробовать двигать перебрать пороги.
Например, сдвигая на один пример.

Выборка

Рассмотрим выборку объектов с одним признаком x :

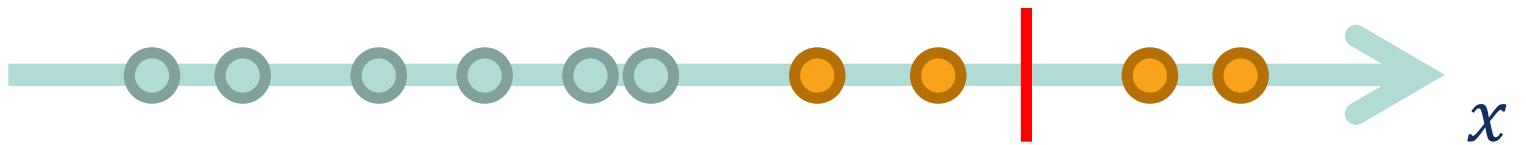


Как подобрать порог по признаку в задаче бинарной классификации?

Можно попробовать двигать перебрать пороги.
Например, сдвигая на один пример.

Выборка

Рассмотрим выборку объектов с одним признаком x :

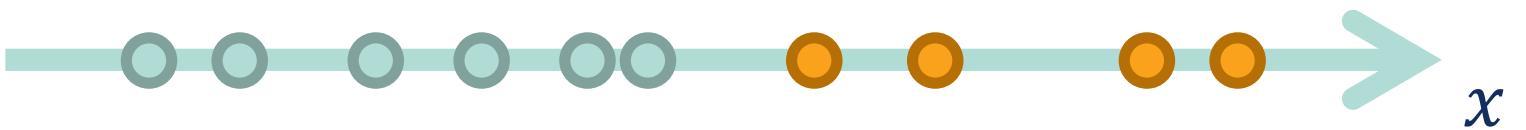


Как подобрать порог по признаку в задаче бинарной классификации?

Можно попробовать двигать перебрать пороги.
Например, сдвигая на один пример.

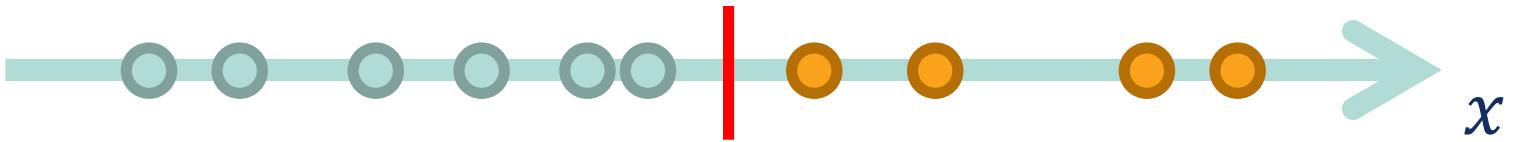
Выборка

Рассмотрим выборку объектов с одним признаком x :



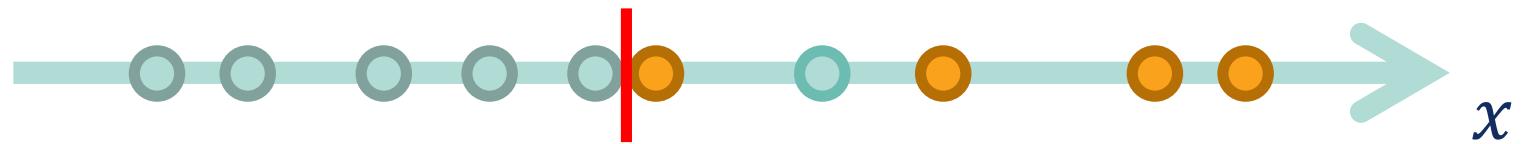
Как подобрать порог по признаку в задаче бинарной классификации?

Можно провести между последним объектом одного класса и первым объектом другого (если выборка разделима):



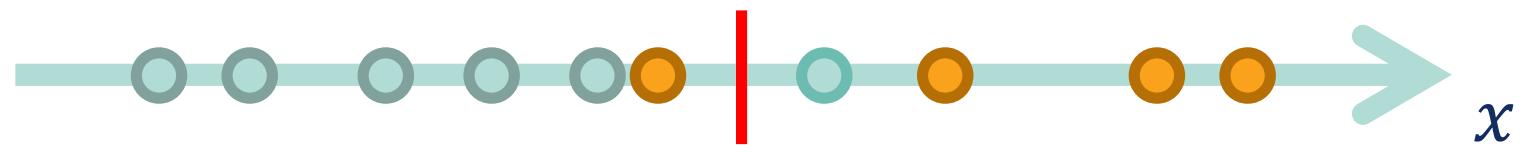
Проблема выбора

Часто есть несколько неплохих порогов:



Проблема выбора

Часто есть несколько неплохих порогов:



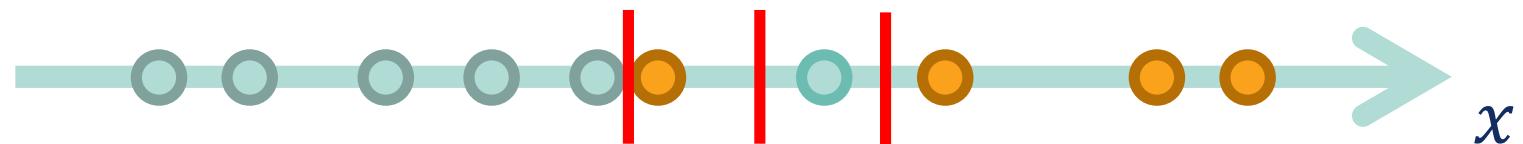
Проблема выбора

Часто есть несколько вариантов деления:



Усложнение модели

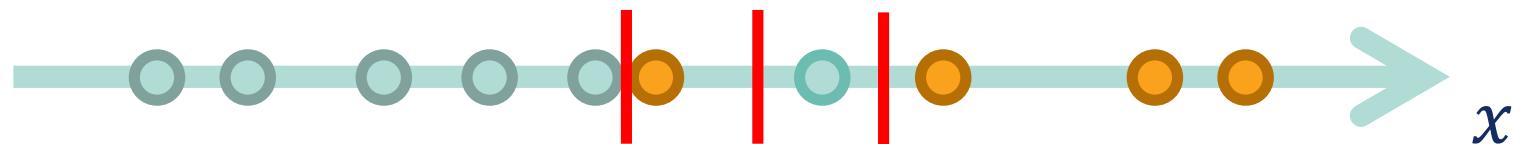
Если потребовать от модели максимальной точности



и не ограничивать количество порогов, можно было бы разделить выборку идеально

Усложнение модели

Если потребовать от модели максимальной точности

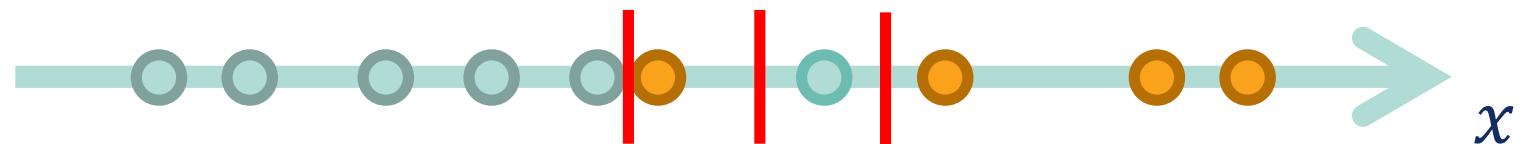


и не ограничивать количество порогов, можно было бы разделить выборку идеально

Но это просто запоминание выборки – такой классификатор легко может оказаться слишком переобученным.

Усложнение модели

Если потребовать от модели максимальной точности



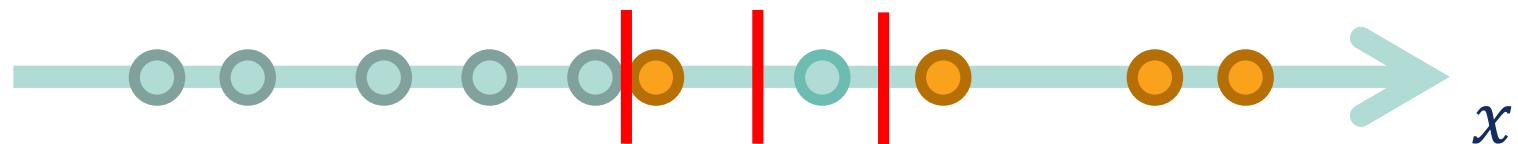
и не ограничивать количество порогов, можно было бы разделить выборку идеально

Но это просто запоминание выборки – такой классификатор легко может оказаться слишком переобученным.

Вопрос тем, кто уже знает, что такое kNN: подумайте, как он связан с обсуждаемым сейчас классификатором

Случай множества порогов

Если потребовать от модели максимальной точности



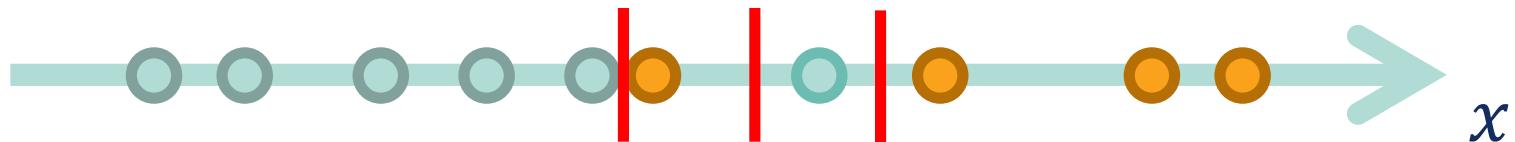
и не ограничивать количество порогов, можно было бы разделить выборку идеально

Но это просто запоминание выборки – такой классификатор легко может оказаться слишком переобученным:



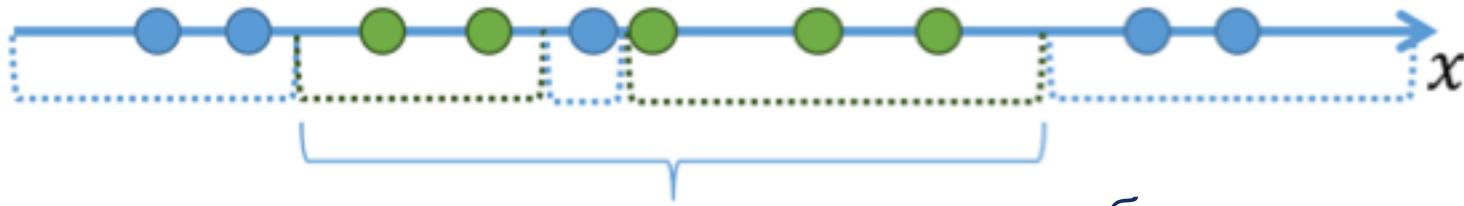
Случай множества порогов

Если потребовать от модели максимальной точности



и не ограничивать количество порогов, можно было бы разделить выборку идеально

Но это просто запоминание выборки – такой классификатор легко может оказаться слишком переобученным:



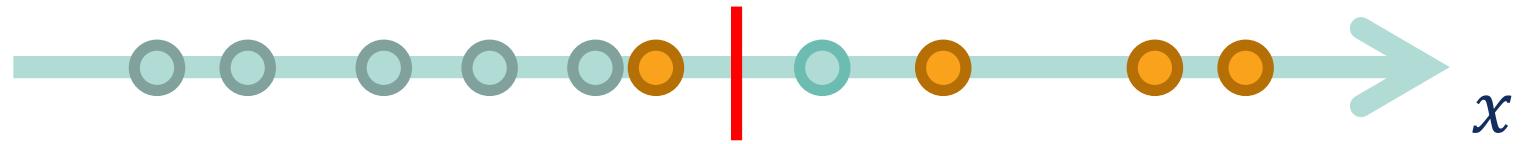
Возможно какие-то интервалы лучше объединить

Разделение неразделимо- й выборки

Предположим, выборка не разделима идеально:

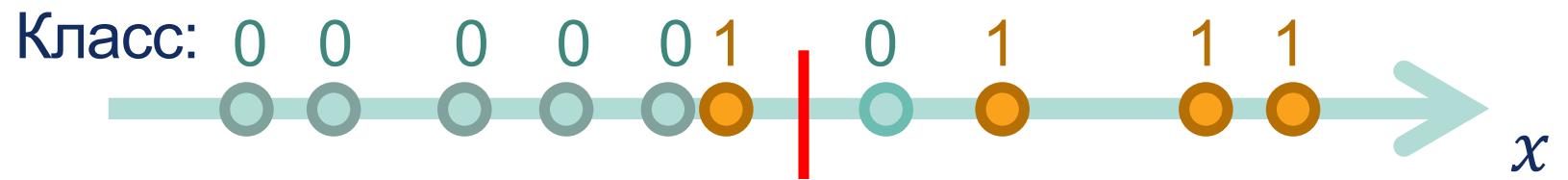


Но нужно по-прежнему адекватно ее разделить:

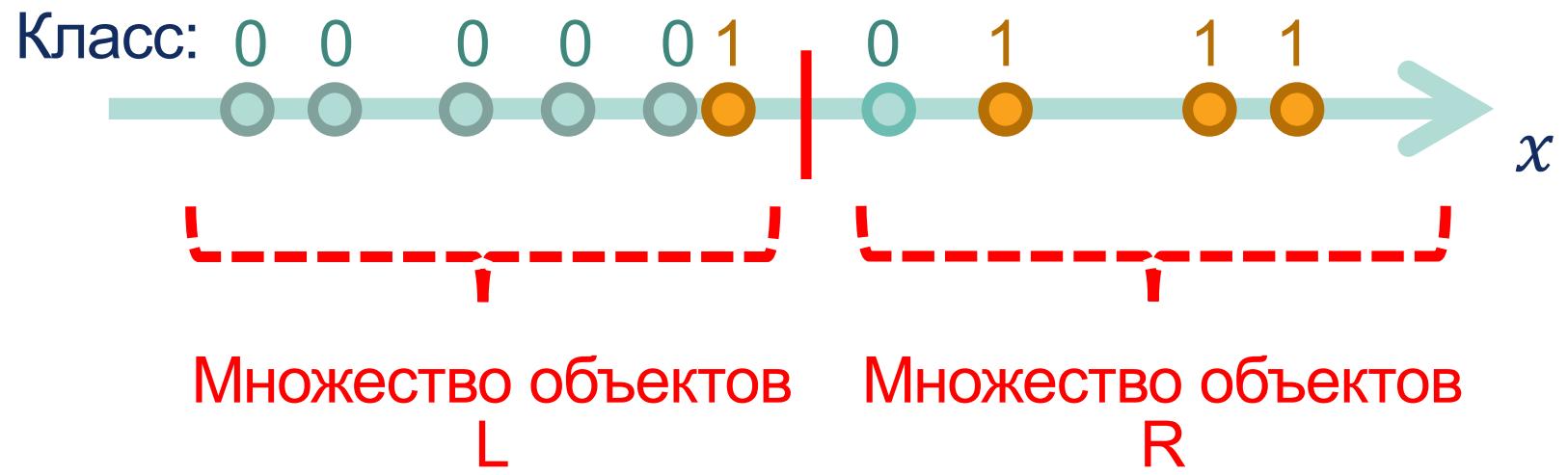


Как это записать?

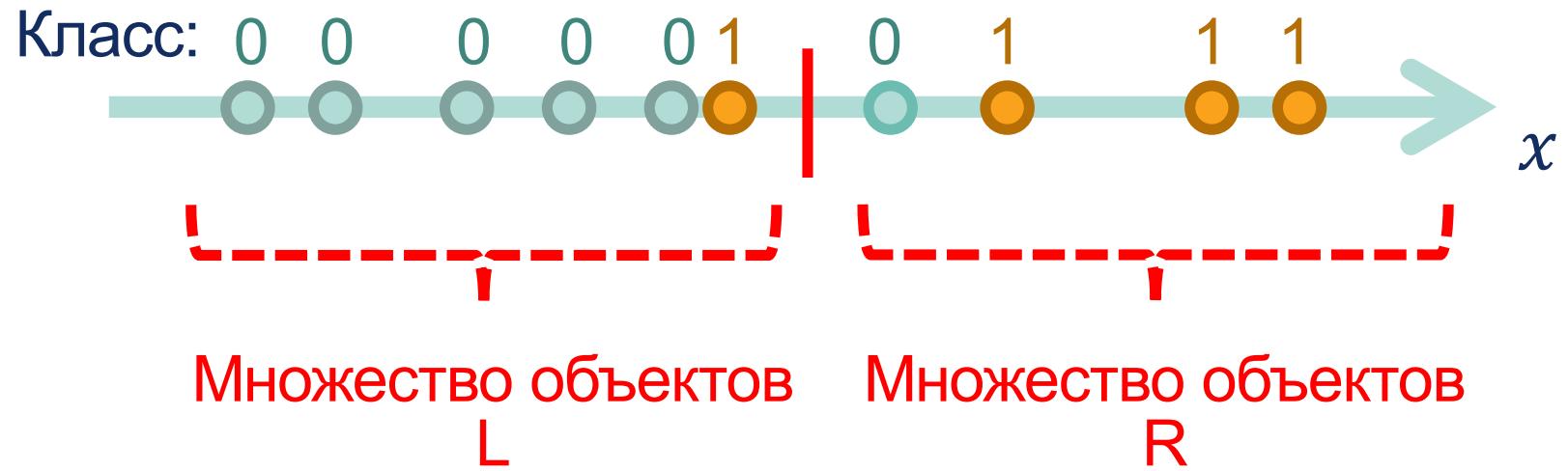
Задача оптимизации



Задача оптимизации

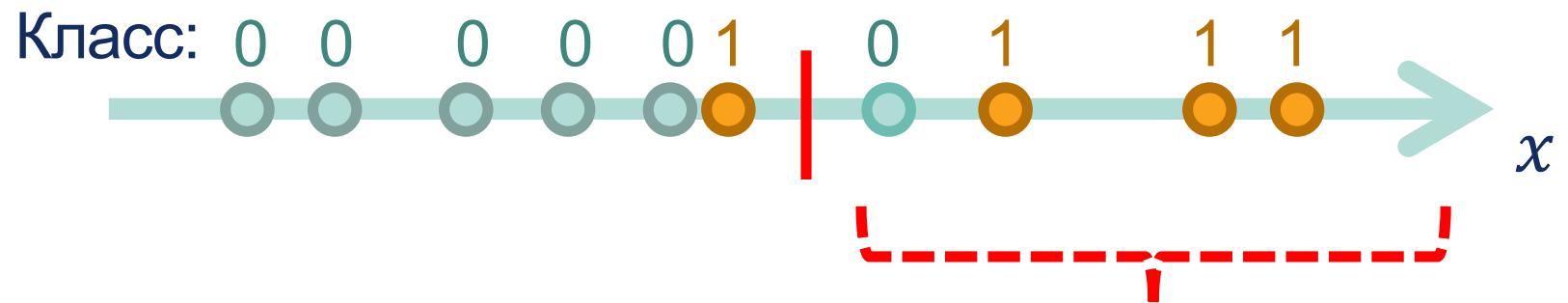


Задача оптимизации



Чтобы разделить классы хорошо – нужно, чтобы и в L и в R преобладал только один класс

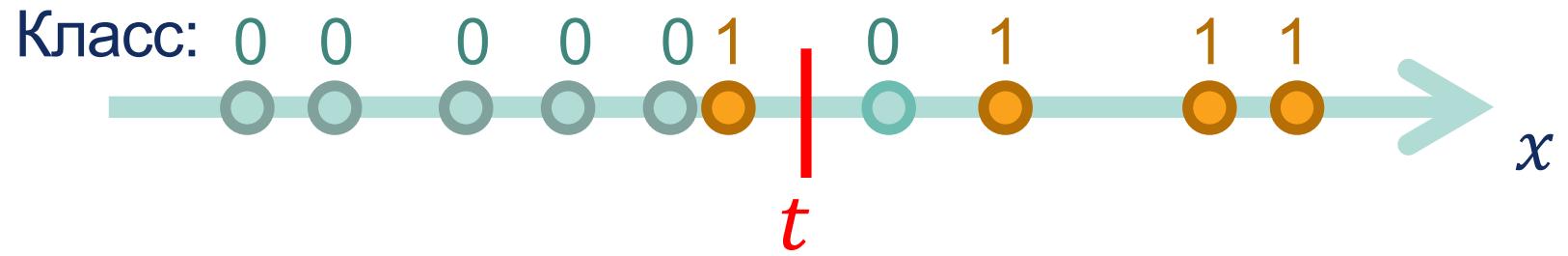
Задача оптимизации



Пусть p_0 – доля класса 0 в R , а p_1 – доля класса 1 в R .
В нашем примере $p_0 = \frac{1}{4}$, а $p_1 = \frac{3}{4}$.

Как записать, что один из классов преобладает?

Задача оптимизации

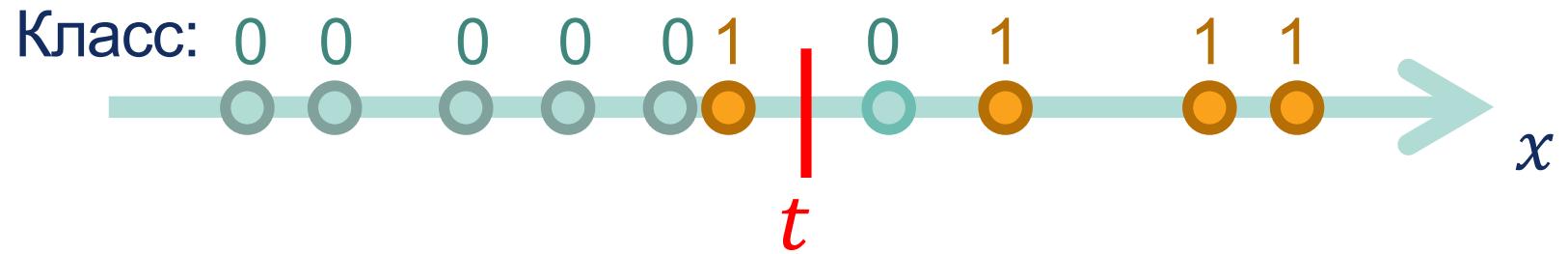


Как записать, что один из классов должен преобладать в R?

Например, так:

$$p_{max} = \max\{p_0, p_1\} \rightarrow \max_t$$

Задача оптимизации



Как записать, что один из классов должен преобладать в R?

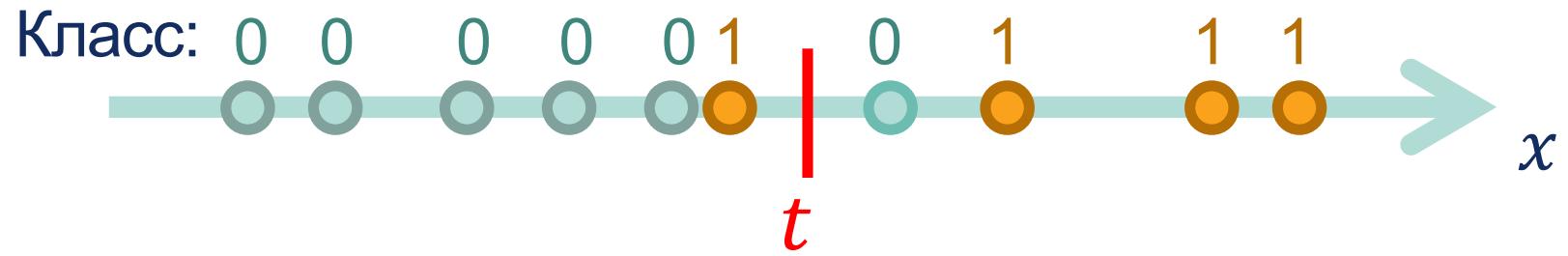
Например, так:

$$p_{max} = \max\{p_0, p_1\} \rightarrow \max_t$$

Или так:

$$1 - p_{max} \rightarrow \min_t$$

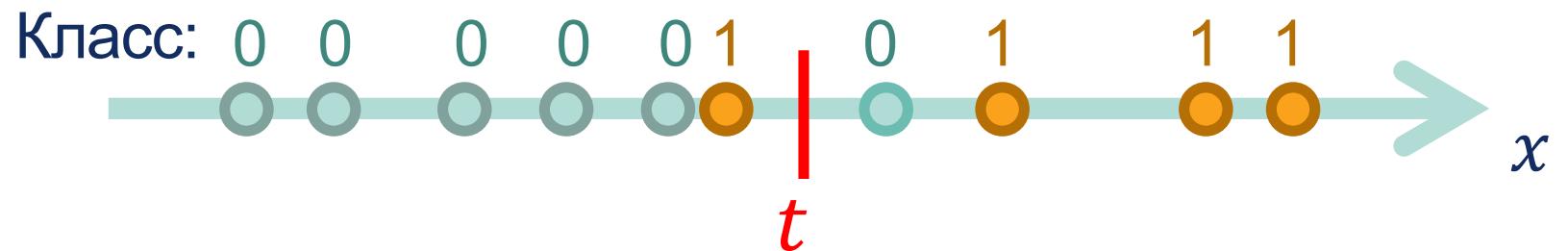
Задача оптимизации



Другой вариант:

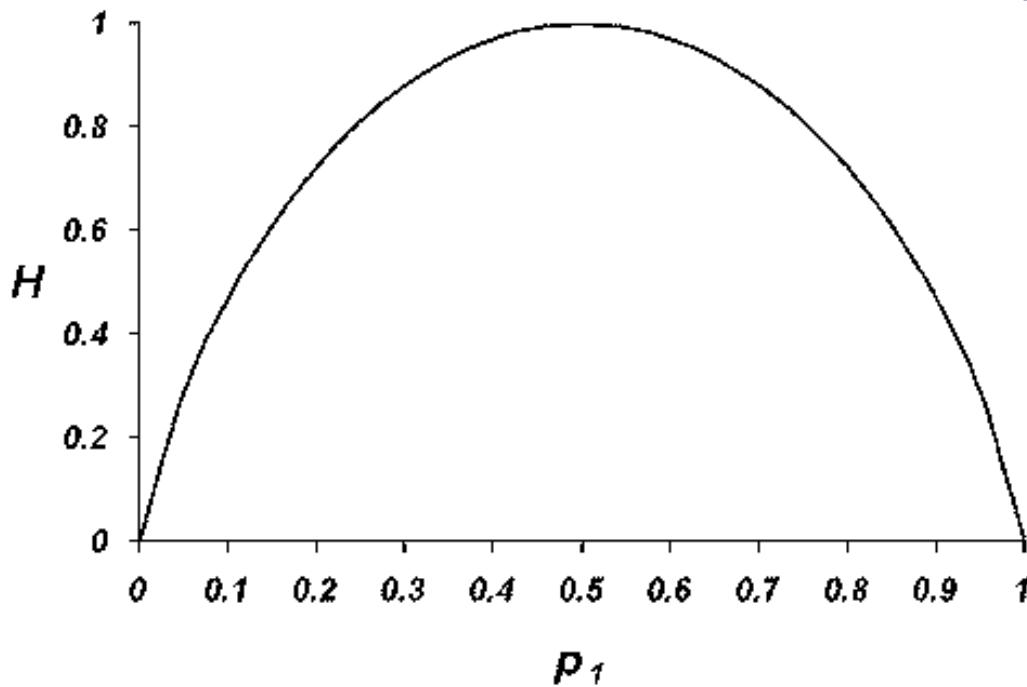
$$H(R) = -p_0 \ln p_0 - p_1 \ln p_1 \rightarrow \min_t$$

Задача оптимизации

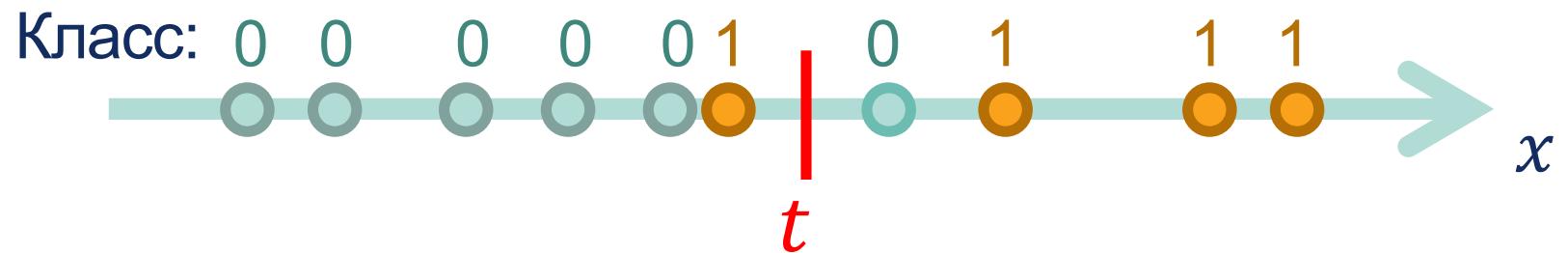


Другой вариант:

$$H(R) = -p_0 \ln p_0 - p_1 \ln p_1 \rightarrow \min_t$$



Задача оптимизации

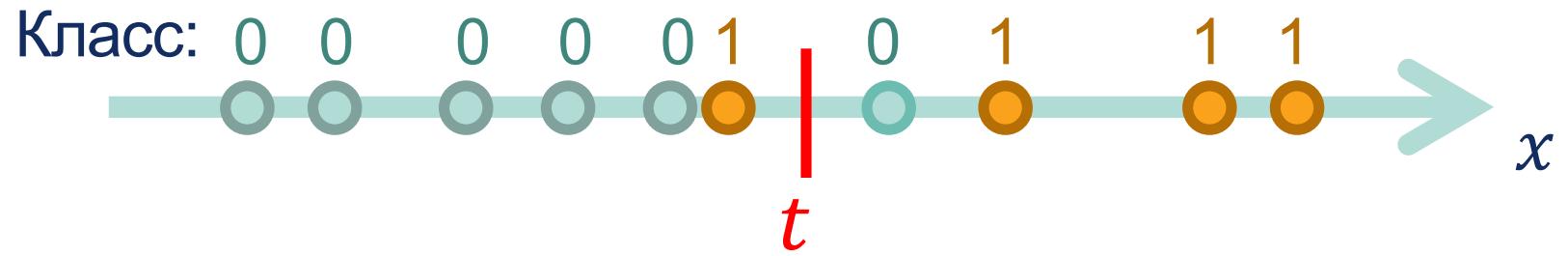


Другой вариант:

$$H(R) = -p_0 \ln p_0 - p_1 \ln p_1 \rightarrow \min_t$$



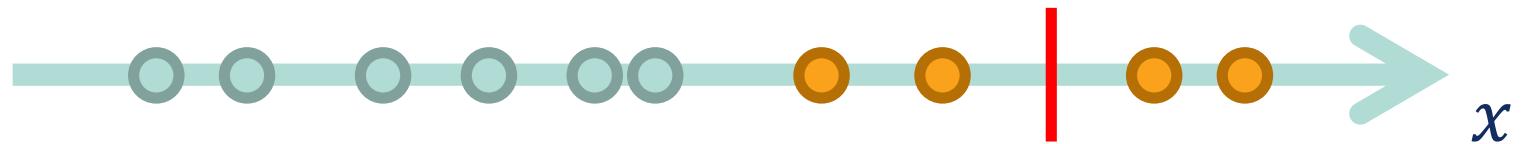
Задача оптимизации



Все это разные способы задать оптимизационную задачу, которую мы можем решить перебирая порог t

Уточнение

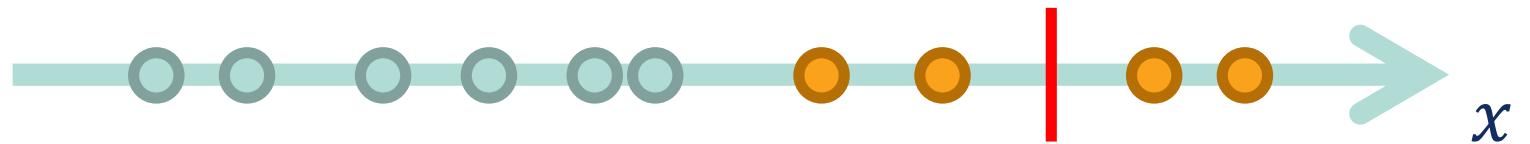
Но если смотреть только на R, можем нечаянно разделить выборку так:



Здесь проблемы только в левой части, в правой все хорошо с преобладанием одного класса

Уточнение

Но если смотреть только на R, можем нечаянно разделить выборку так:

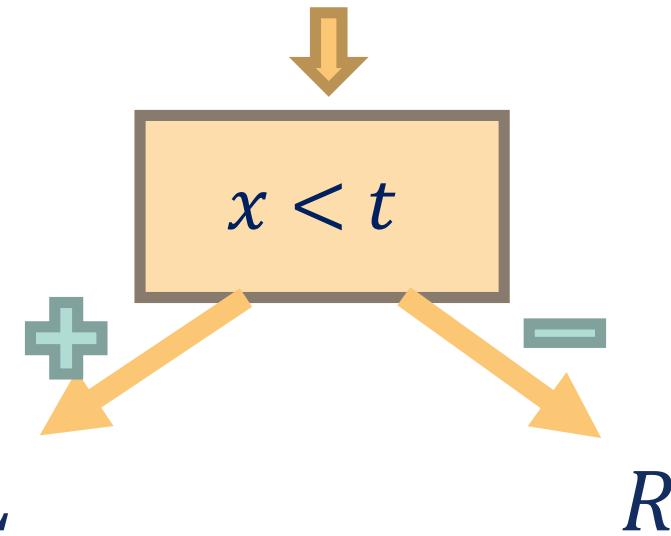


Здесь проблемы только в левой части, в правой все хорошо с преобладанием одного класса

Значит надо учитывать обе части: R и L

Выбор разбиения

Вся выборка (n объектов)

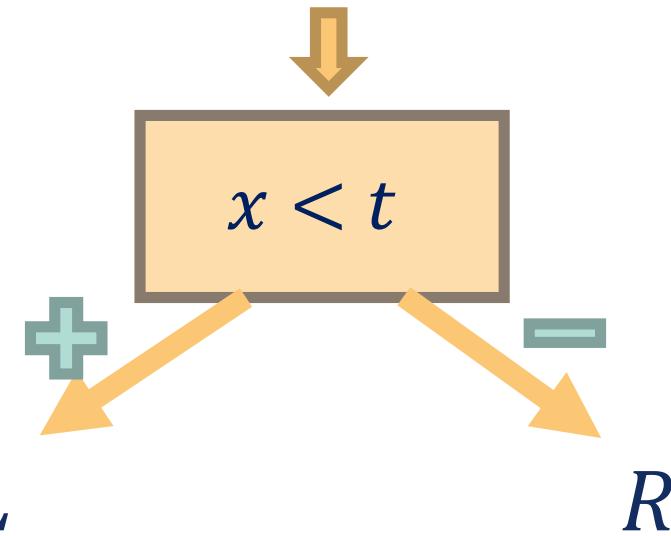


$$G(t) = H(L) + H(R) \rightarrow \min_t$$

$H(R)$ - мера «неоднородности»
(impurity) множества R

Выбор разбиения

Вся выборка (n объектов)

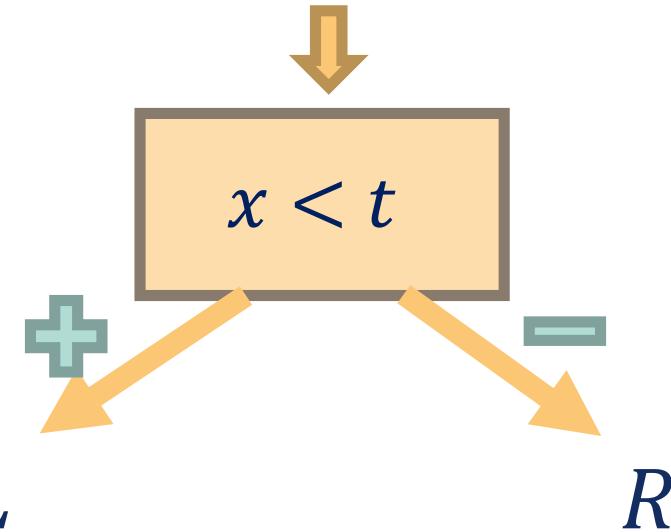


$$G(t) = H(L) + H(R) \rightarrow \min_t$$

Но что если L и R сильно разного размера?
Учтем это.

Выбор разбиения

Вся выборка (n объектов)



$$G(t) = \frac{|L|}{n} H(L) + \frac{|R|}{n} H(R) \rightarrow \min_t$$

Критерии построения разбиений

$H(R)$ – мера «неоднородности» множества R

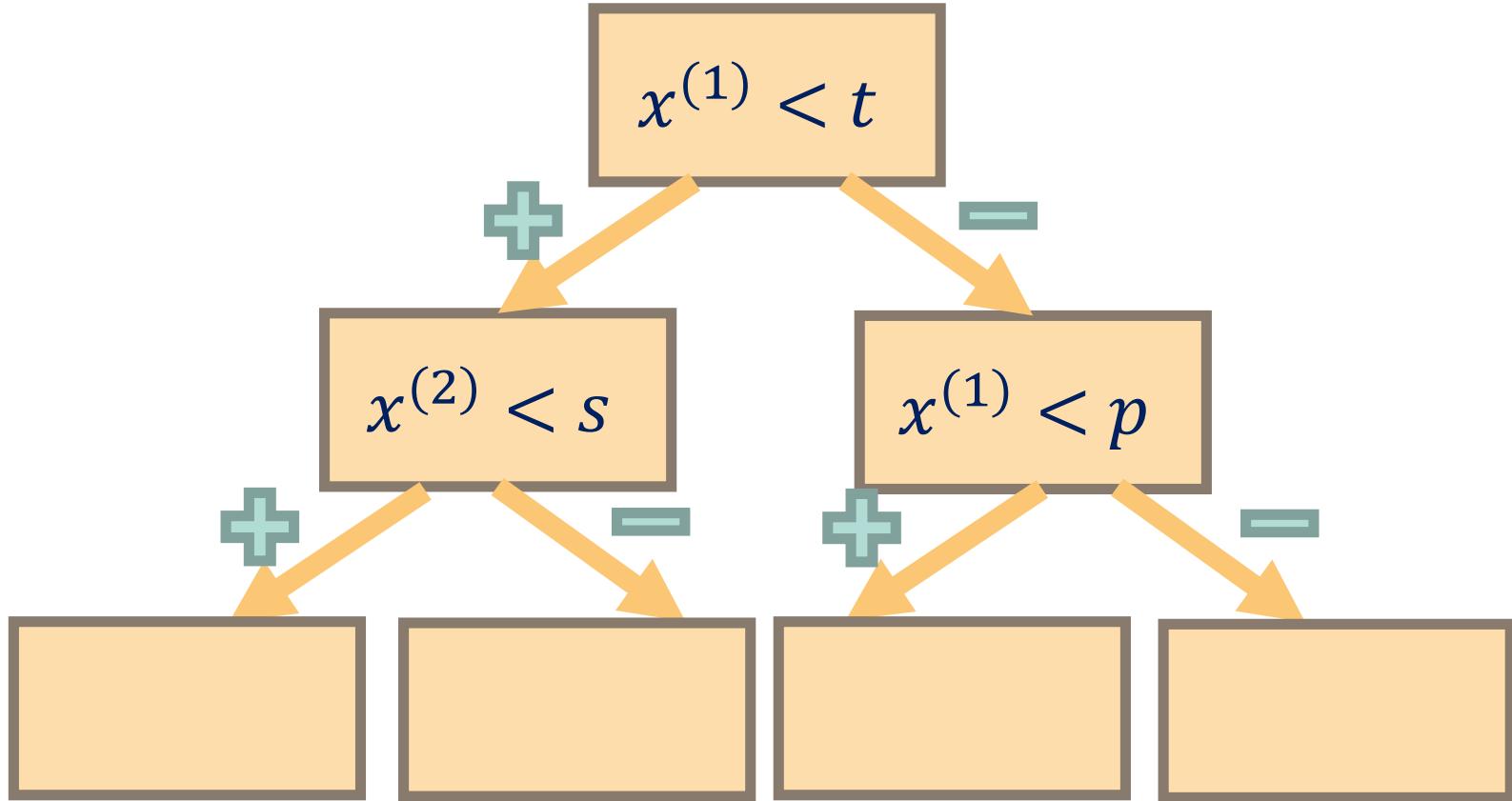
Критерии построения разбиений

$H(R)$ – мера «неоднородности» множества R

Варианты этой функции:

- 1) Misclassification criteria: $H(R) = 1 - \max\{p_0, p_1\}$
- 2) Entropy criteria: $H(R) = -p_0 \ln p_0 - p_1 \ln p_1$
- 3) Gini criteria: $H(R) = 1 - p_0^2 - p_1^2 = 2p_0p_1$

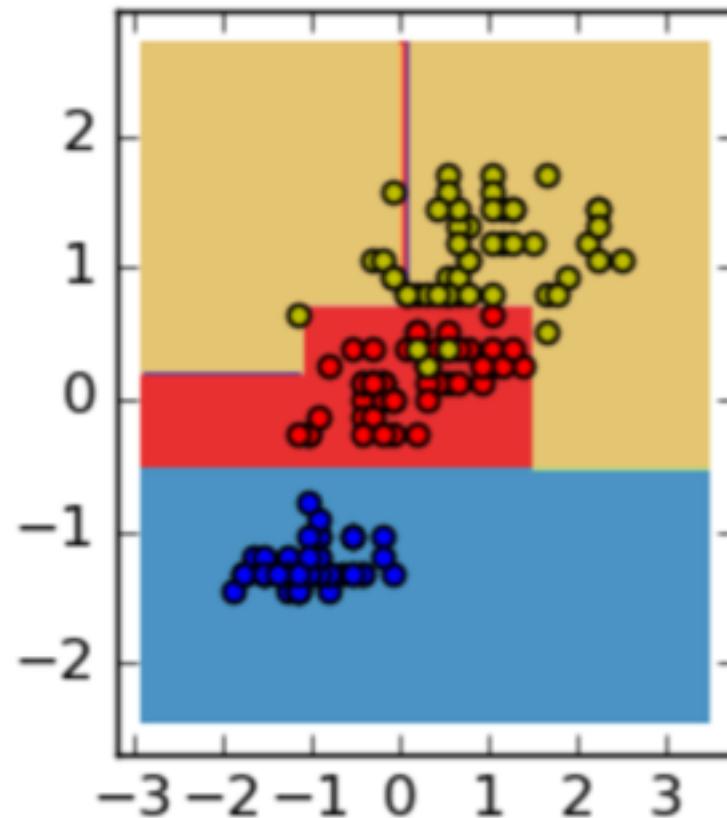
Если
признаков и
порогов
больше



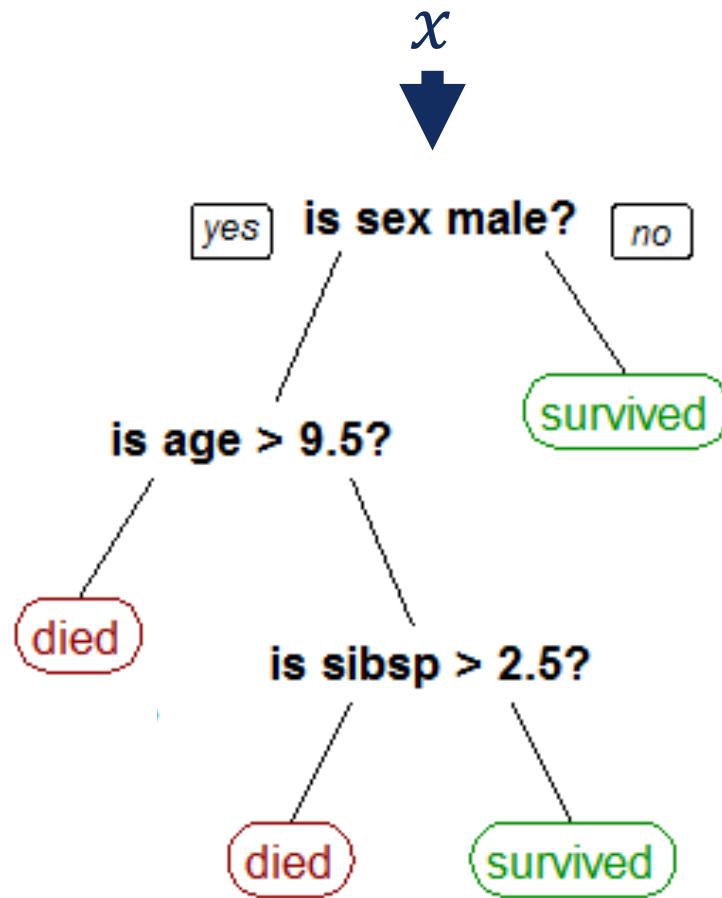
2. От пней к деревьям

Границы классов

Пример границ для 3 классов при 2 признаках:



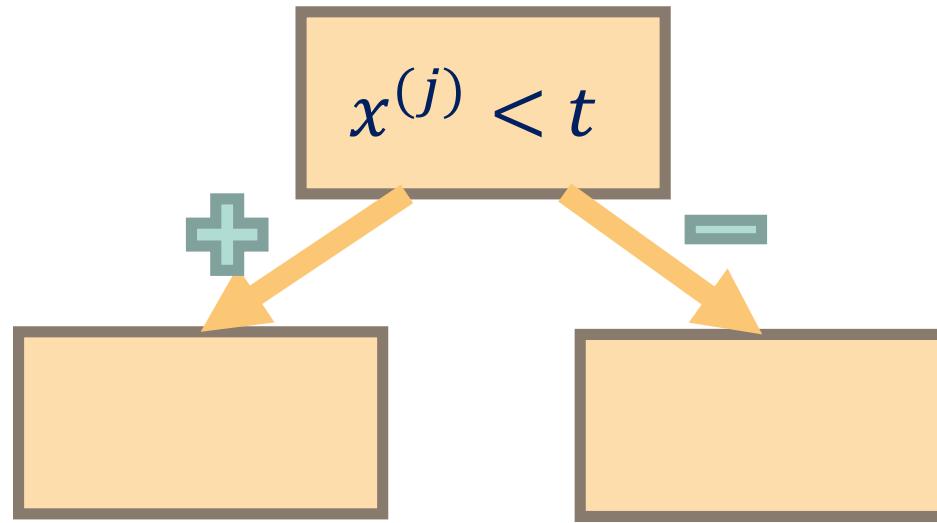
Решающее дерево



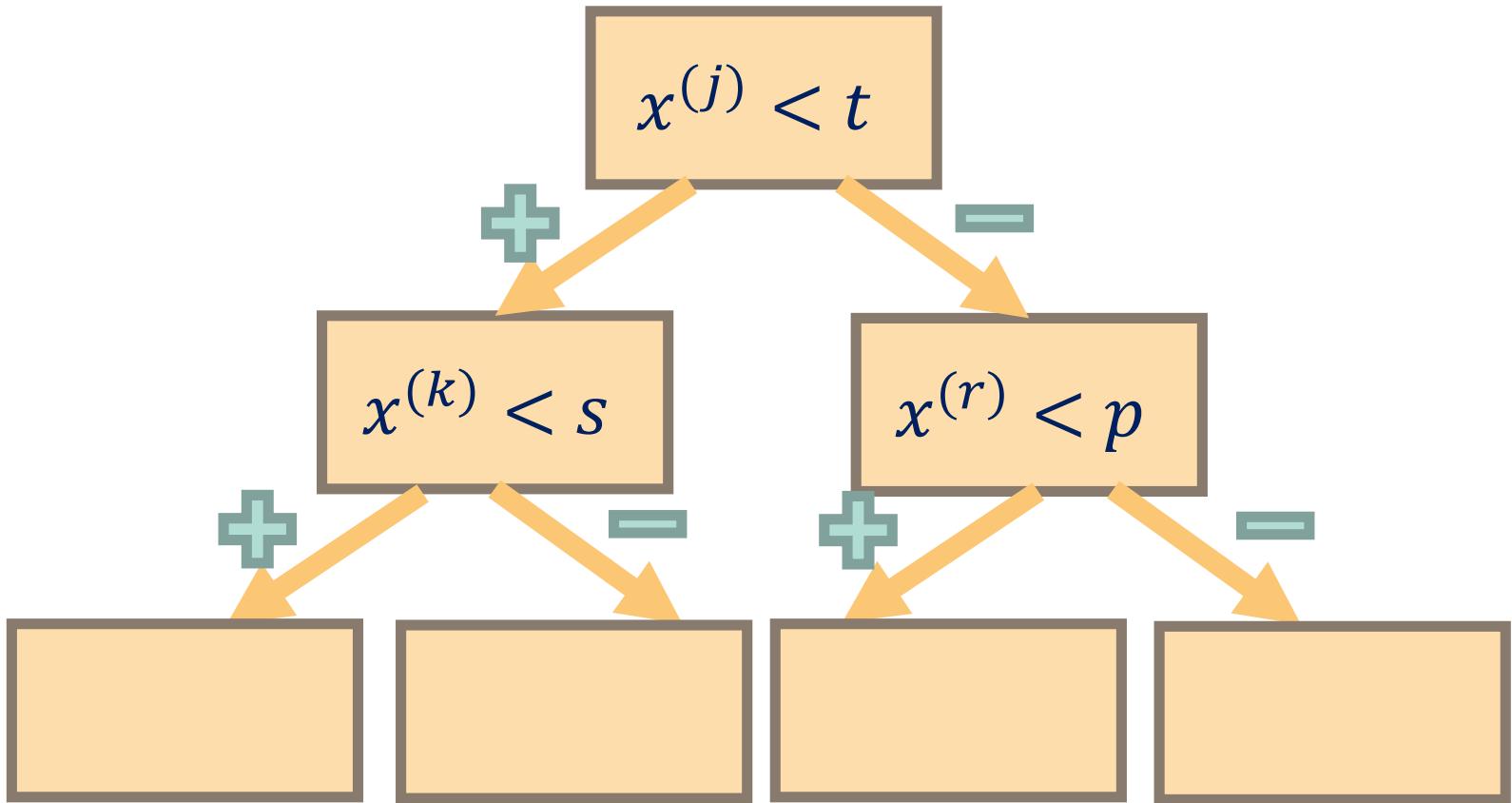
Рекурсивное построение

$$x^{(j)} < t$$

Рекурсивное построение

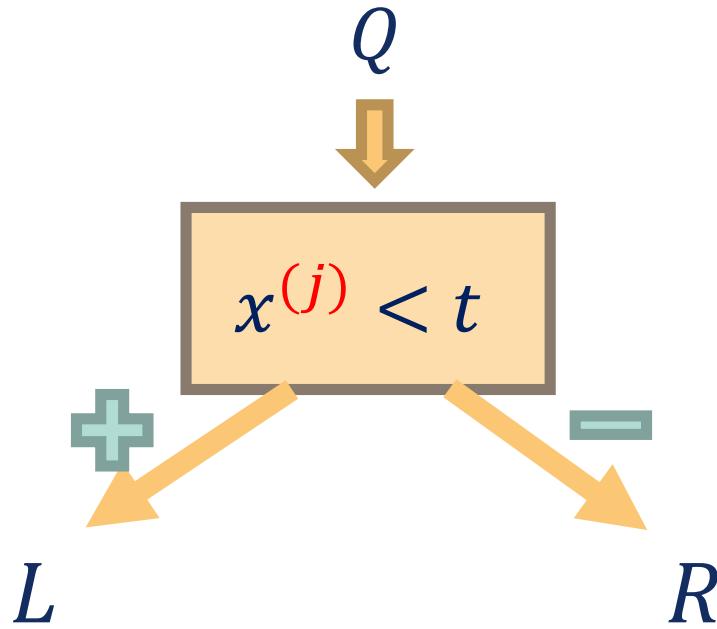


Рекурсивное построение



Процесс можно продолжать в тех узлах, в
которые попадает достаточно много объектов

Выбор разбиения



$$G(j, t) = \frac{|L|}{|Q|} H(L) + \frac{|R|}{|Q|} H(R) \rightarrow \min_{j,t}$$

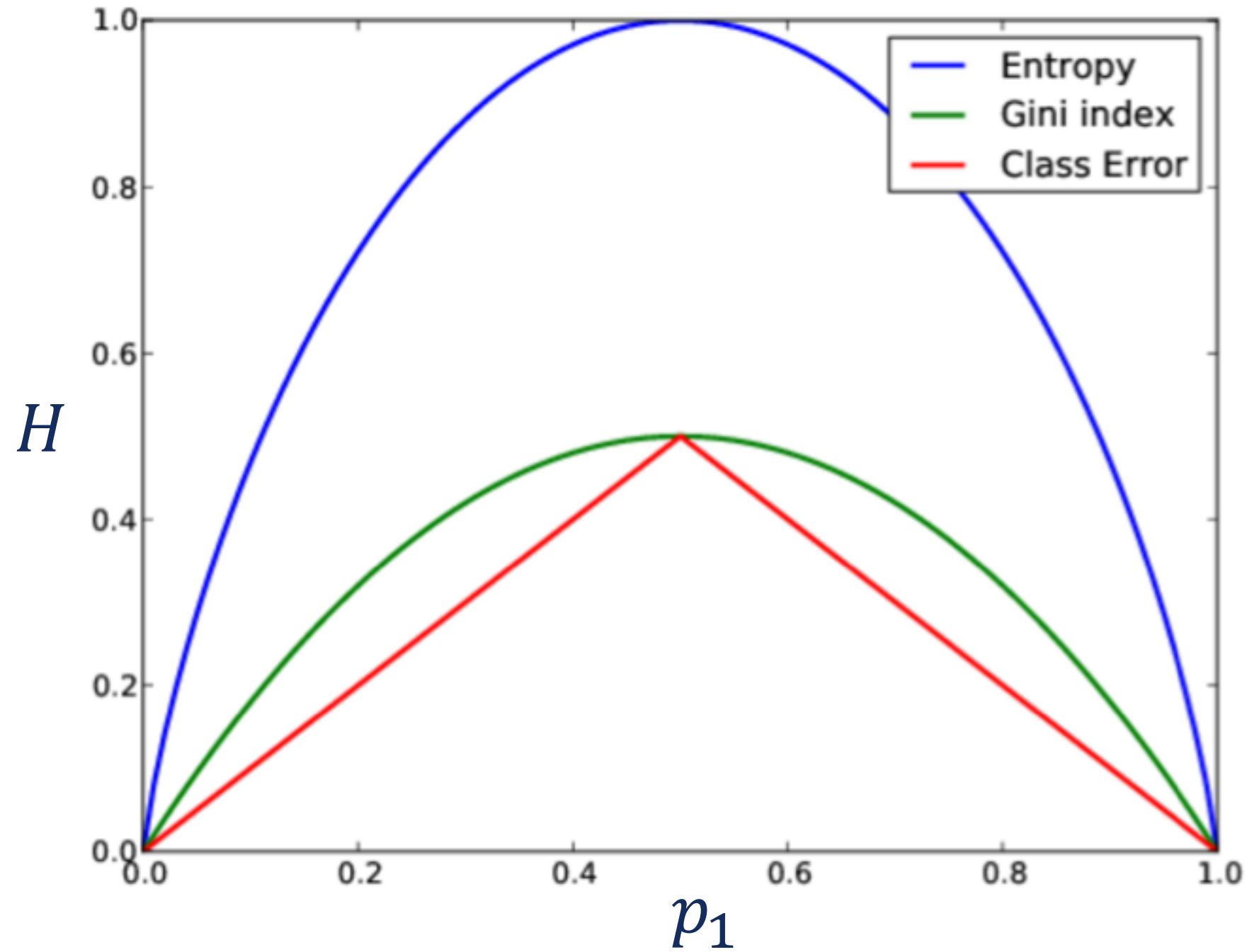
Критерии построения разбиений

$H(R)$ – мера «неоднородности» множества R

Варианты этой функции:

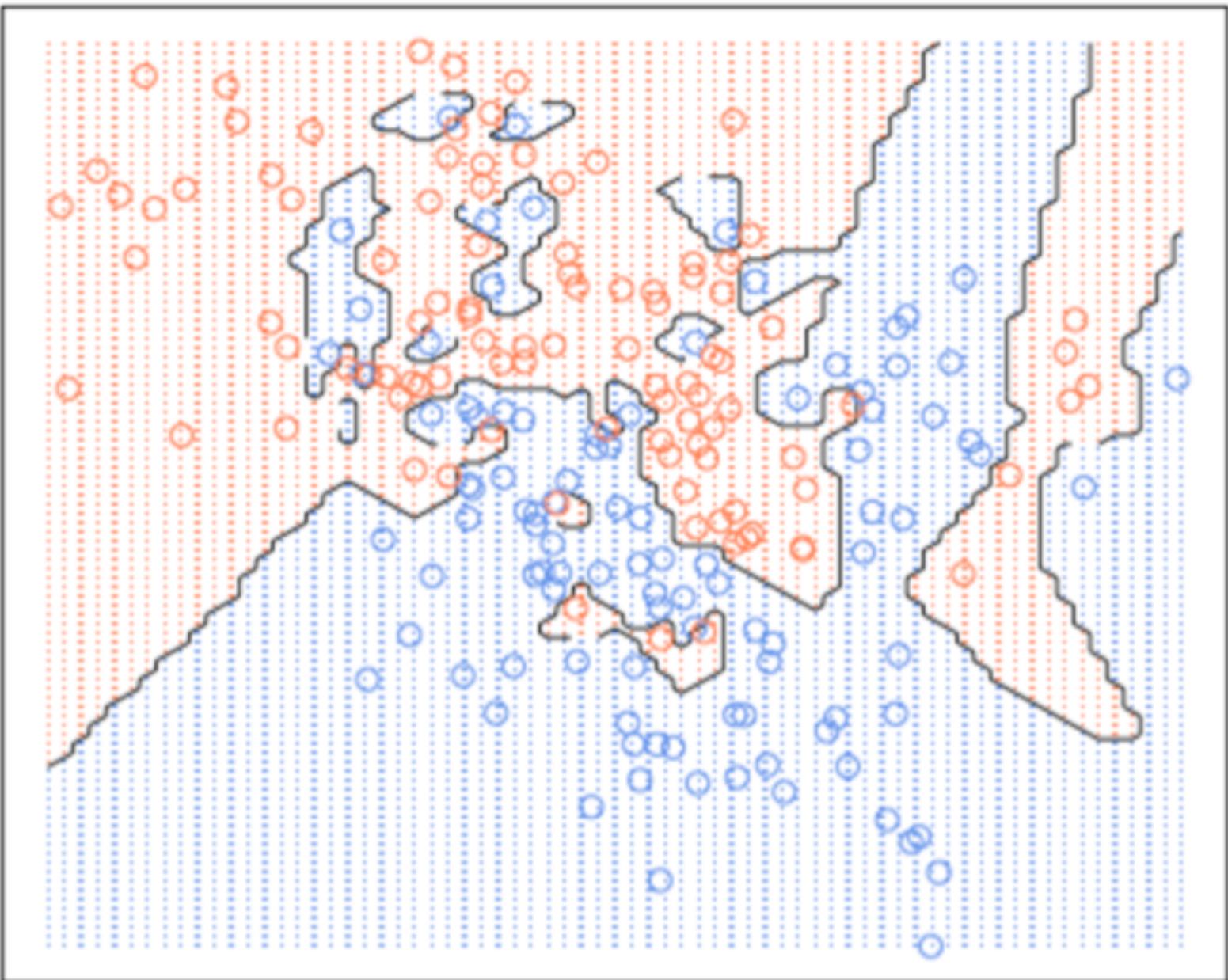
- 1) Misclassification criteria: $H(R) = 1 - \max\{p_0, p_1\}$
- 2) Entropy criteria: $H(R) = -p_0 \ln p_0 - p_1 \ln p_1$
- 3) Gini criteria: $H(R) = 1 - p_0^2 - p_1^2 = 2p_0p_1$

Критерии построения разбиений

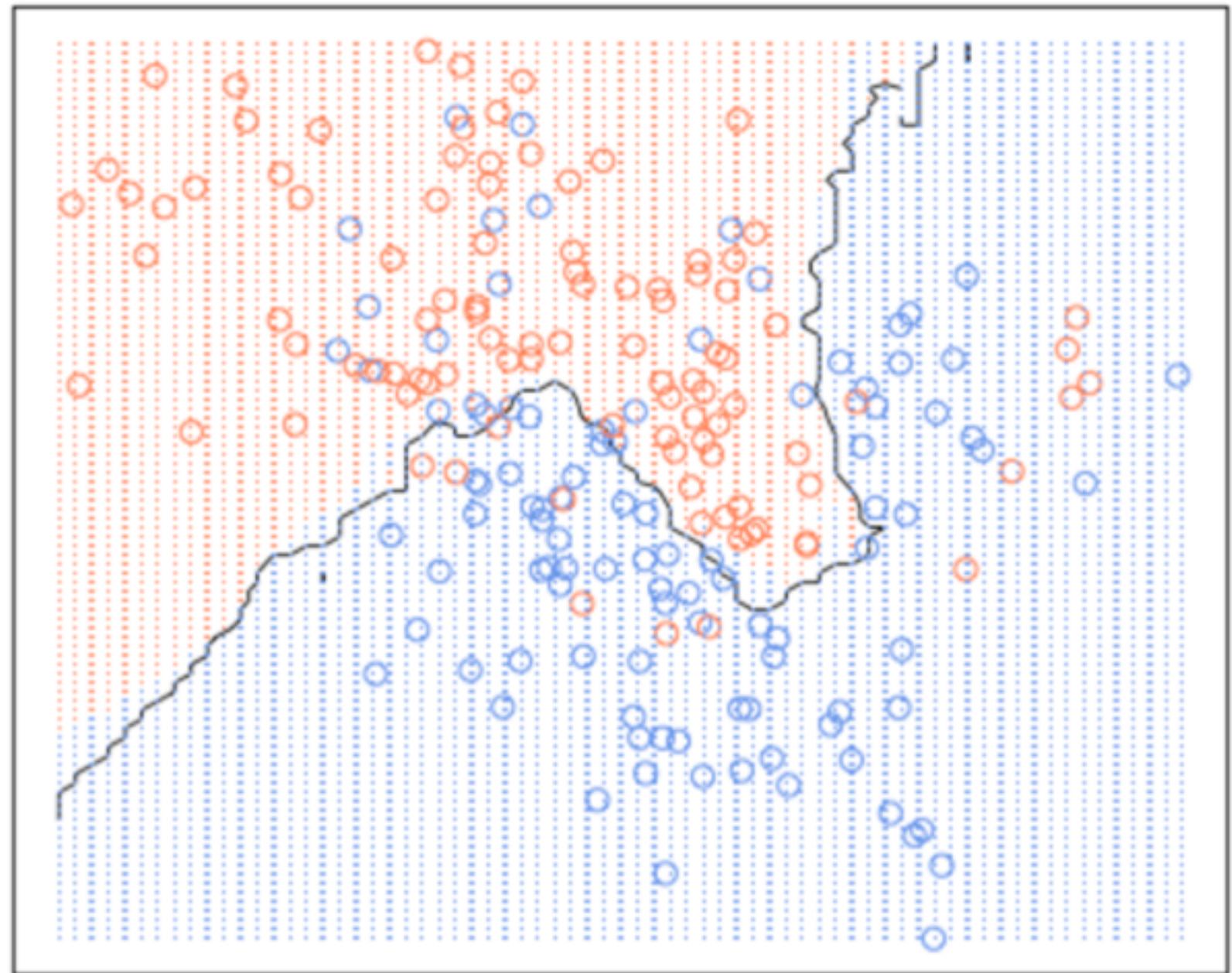


3. Сложные границы и соседи

Сложные границы



Сложные границы

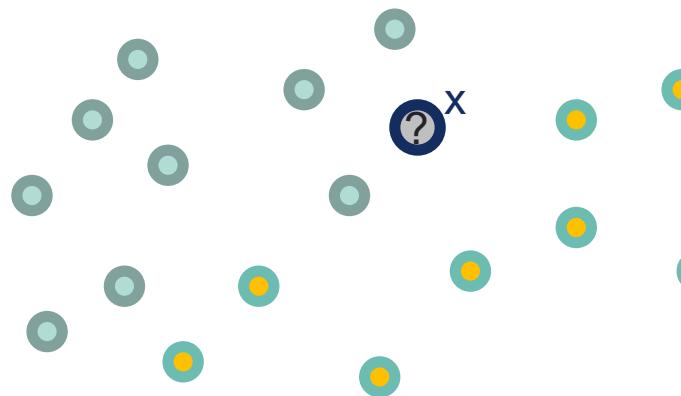


Сложные границы



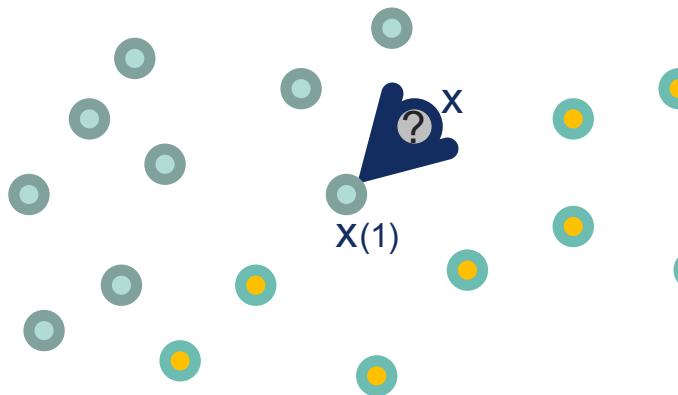
Метод ближайшего соседа

Пример классификации:



Метод ближайшего соседа (1NN)

Пример классификации:



Что такое расстояние

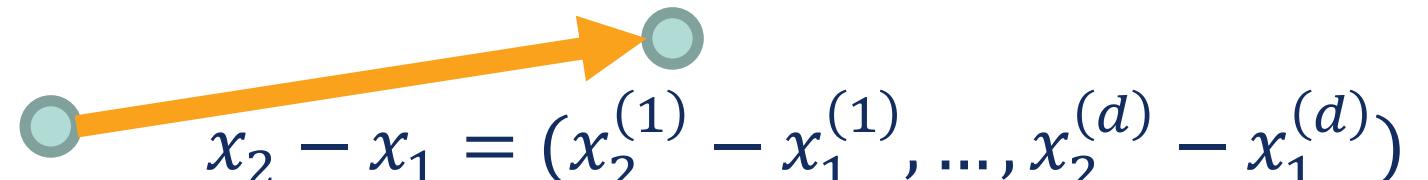
Есть две точки в многомерном пространстве: x_1 и x_2
Как ввести расстояние между ними?


$$x_2 = (x_2^{(1)}, \dots, x_2^{(d)})$$


$$x_1 = (x_1^{(1)}, \dots, x_1^{(d)})$$

Что такое расстояние

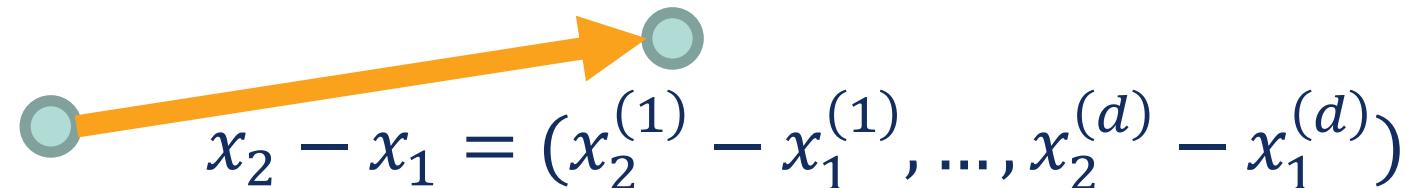
Есть две точки в многомерном пространстве: x_1 и x_2
Как ввести расстояние между ними?



Частая практика: $d(x_1, x_2) = d(x_2, x_1) = \|x_2 - x_1\|$

Что такое расстояние

Есть две точки в многомерном пространстве: x_1 и x_2
Как ввести расстояние между ними?



Частая практика: $d(x_1, x_2) = d(x_2, x_1) = \|x_2 - x_1\|$

Евклидово расстояние (как в жизни, но в многомерном пространстве):

$$d(x_1, x_2) = \sqrt{\left(x_2^{(1)} - x_1^{(1)}\right)^2 + \dots + \left(x_2^{(d)} - x_1^{(d)}\right)^2}$$

Примеры норм

В зависимости от выбора способа вычислять норму (длину) вектора получаем разные метрики.

Примеры норм:

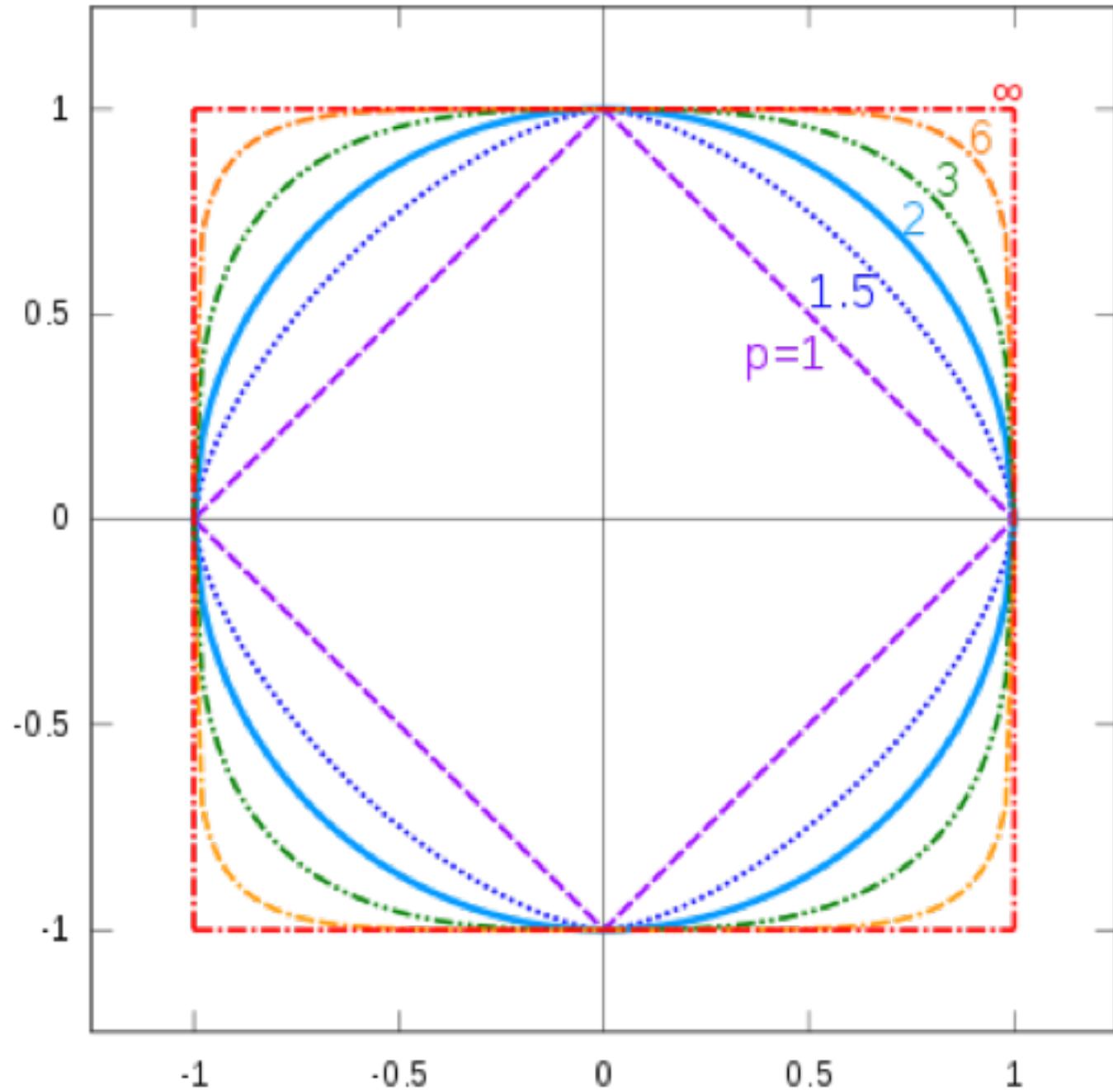
$$\|x\|_{\ell_2} = \sqrt{\left(x^{(1)}\right)^2 + \dots + \left(x^{(d)}\right)^2}$$

$$\|x\|_{\ell_1} = |x^{(1)}| + \dots + |x^{(d)}|$$

$$\|x\|_{\ell_\infty} = \max \{|x^{(1)}|, \dots, |x^{(d)}|\}$$

$$\|x\|_{\ell_p} = \sqrt[p]{\left|x^{(1)}\right|^p + \dots + \left|x^{(d)}\right|^p}$$

Примеры норм



Функция близости

Но можно ввести расстояние каким-то своим особым способом или вообще ввести не расстояние, а **функцию близости**

Пример: Косинусная мера близости (cosine similarity)

Функция близости

Но можно ввести расстояние каким-то своим особым способом или вообще ввести не расстояние, а **функцию близости**

Пример: Косинусная мера близости (cosine similarity)

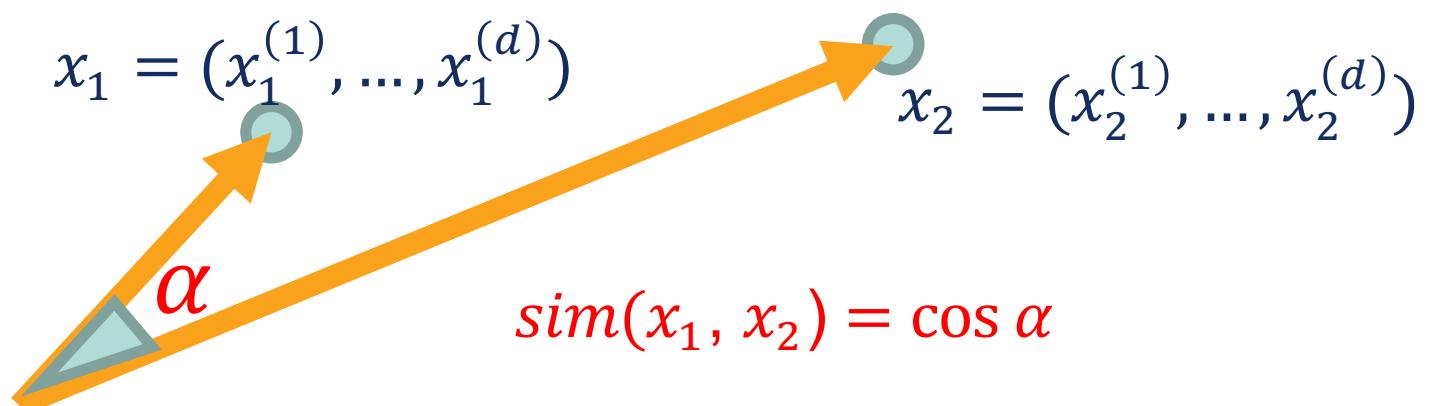
$$sim(x_1, x_2) = \frac{\langle x_1, x_2 \rangle}{\|x_1\| \cdot \|x_2\|} = \frac{x_1^{(1)} \cdot x_2^{(1)} + \dots + x_1^{(d)} \cdot x_2^{(d)}}{\|x_1\| \cdot \|x_2\|}$$

Функция близости

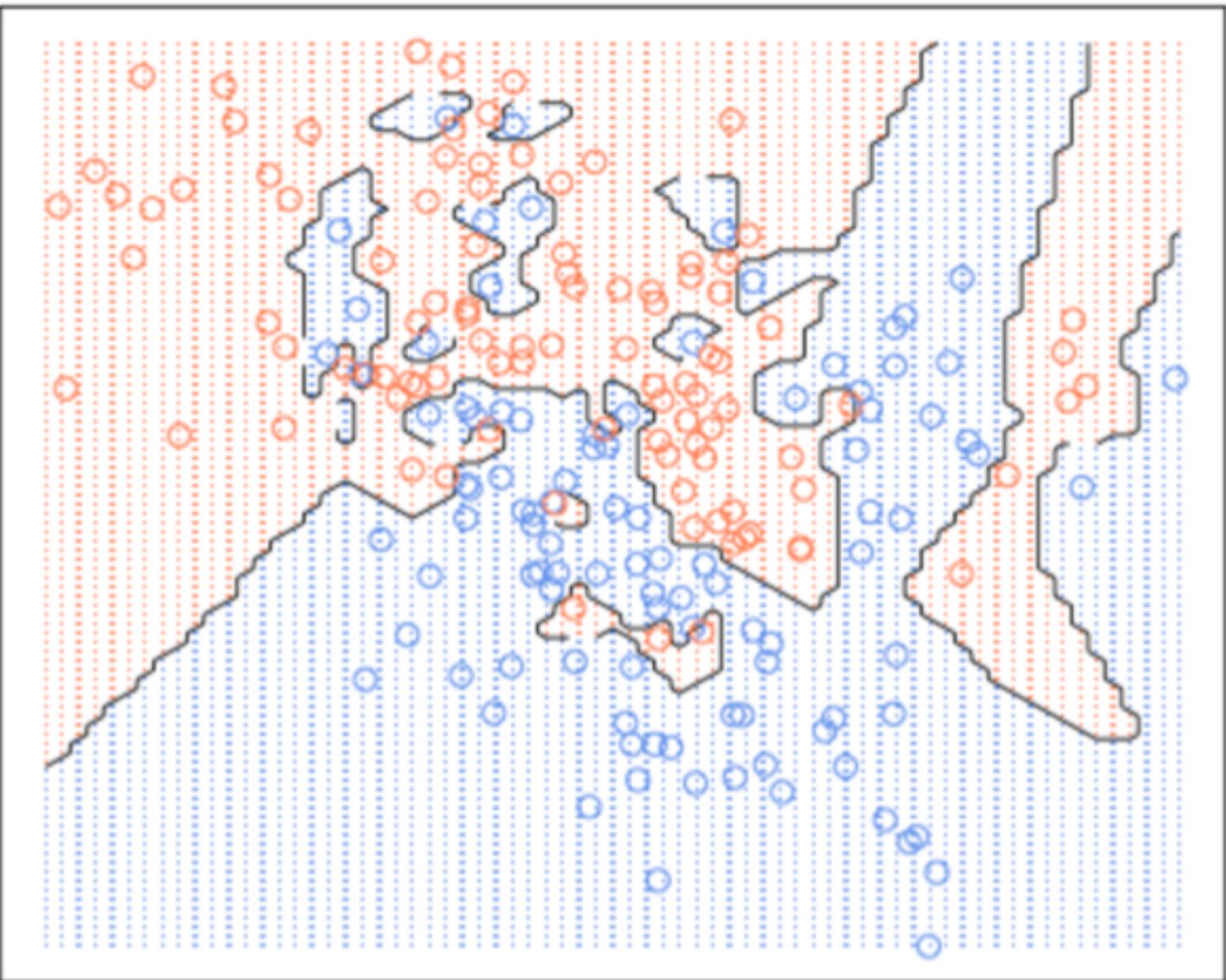
Но можно ввести расстояние каким-то своим особым способом или вообще ввести не расстояние, а **функцию близости**

Пример: Косинусная мера близости (cosine similarity)

$$sim(x_1, x_2) = \frac{\langle x_1, x_2 \rangle}{\|x_1\| \cdot \|x_2\|} = \frac{x_1^{(1)} \cdot x_2^{(1)} + \dots + x_1^{(d)} \cdot x_2^{(d)}}{\|x_1\| \cdot \|x_2\|}$$

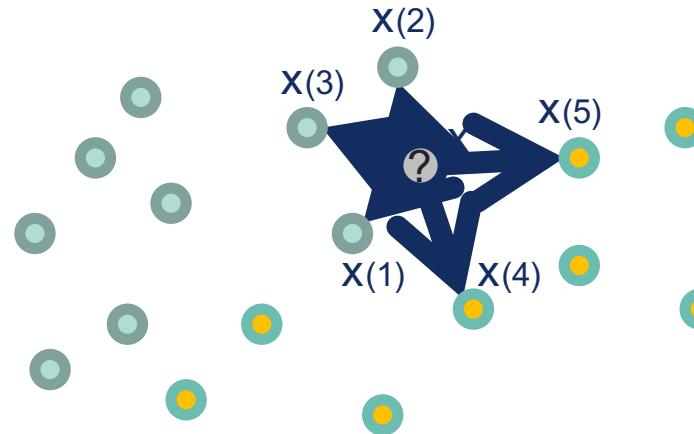


Границы классов в 1NN



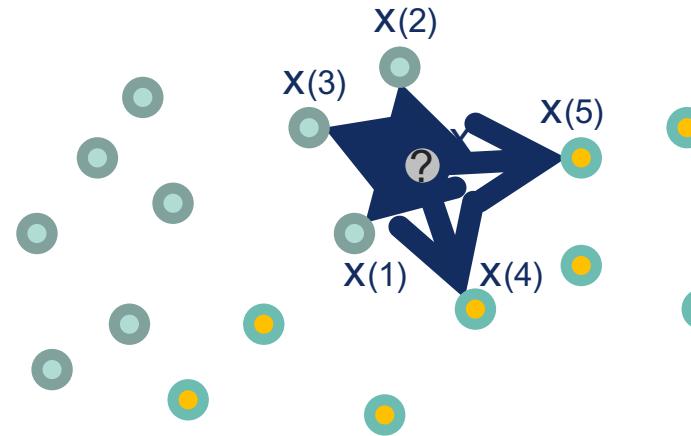
Метод k ближайших соседей (kNN)

Пример классификации ($k = 5$):



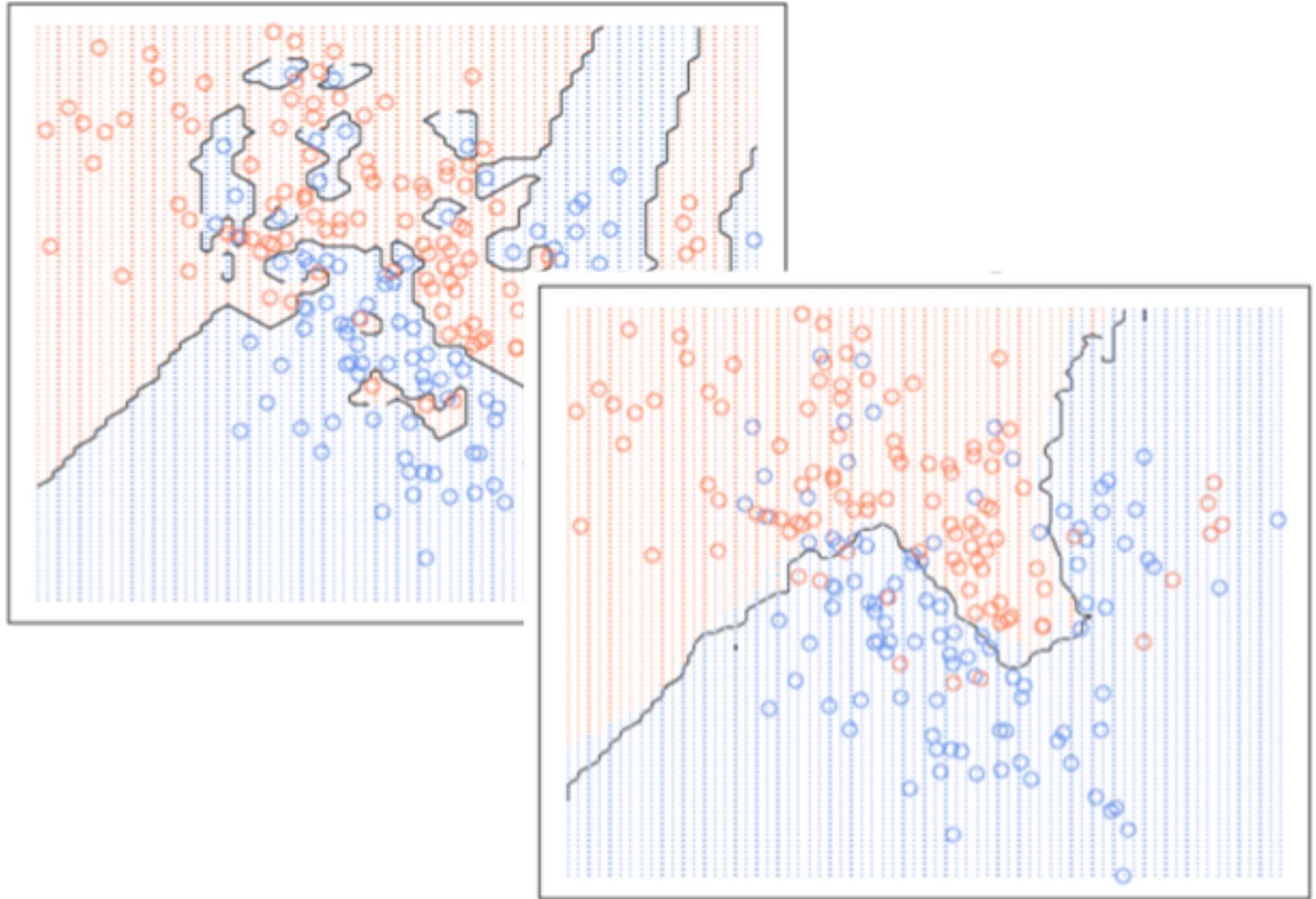
Метод k ближайших соседей (kNN)

Пример классификации ($k = 5$):



Выбираем класс, который преобладает

Сглаживание границ



Вопрос про настройку параметров

Какое количество соседей оптимально брать с точки зрения **качества работы на обучающей выборке?**

Вопрос про настройку параметров

Какое количество соседей оптимально брать с точки зрения **качества работы на обучающей выборке?**

Правильно, $k=1$ – для каждого объекта обучающей выборки смотрим на ближайшего соседа (этот же объект)

Вопрос про настройку параметров

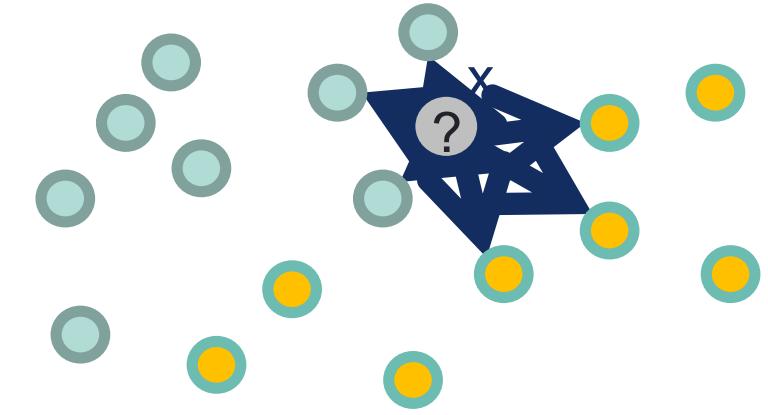
Какое количество соседей оптимально брать с точки зрения **качества работы на обучающей выборке?**

Правильно, $k=1$ – для каждого объекта обучающей выборки смотрим на ближайшего соседа (этот же объект)

Вывод: некоторые параметры алгоритмов (например, количество соседей k) нужно подбирать на отложенной выборке или кросс-валидации

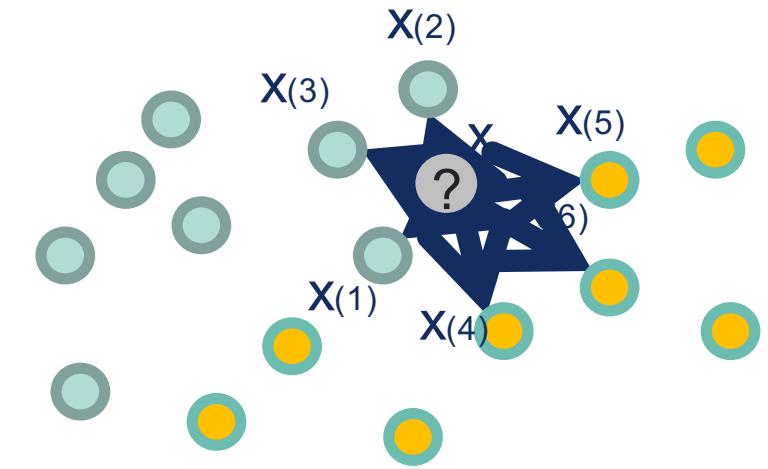
kNN с весами

Пример классификации ($k = 6$):



kNN с весами

Пример классификации ($k = 6$):



kNN с весами

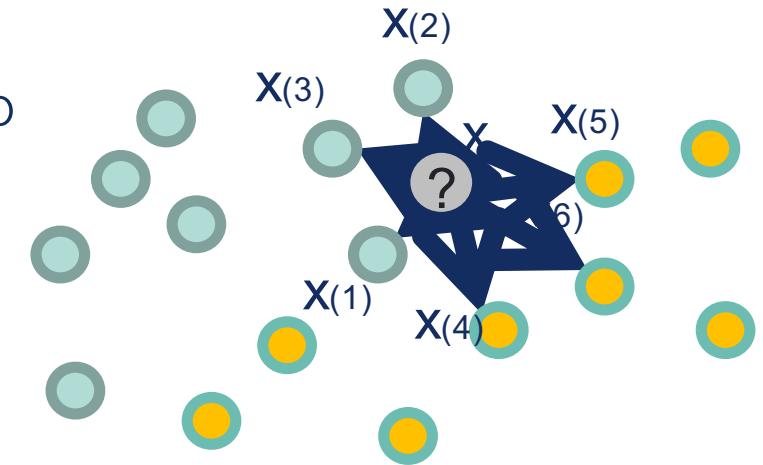
Пример классификации ($k = 6$):

Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x(i)) = w(d(x, x_{(i)}))$$



kNN с весами

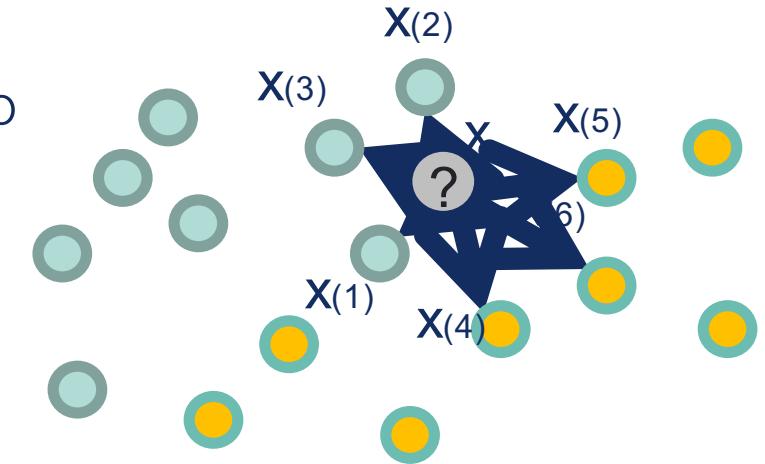
Пример классификации ($k = 6$):

Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x(i)) = w(d(x, x_{(i)}))$$



$$Z_{\text{teal}} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$Z_{\text{yellow}} = \frac{w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

kNN с весами

Пример классификации ($k = 6$):

Веса можно определить как функцию от соседа или его номера:

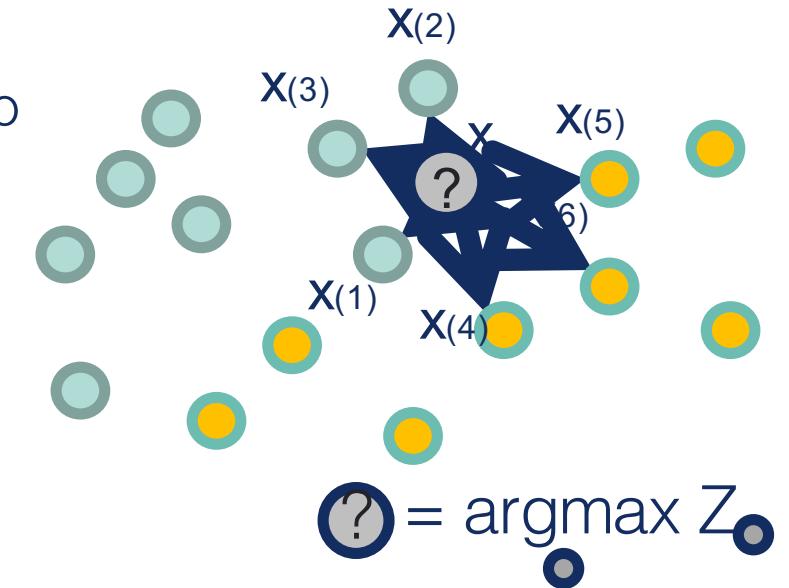
$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x(i)) = w(d(x, x_{(i)}))$$

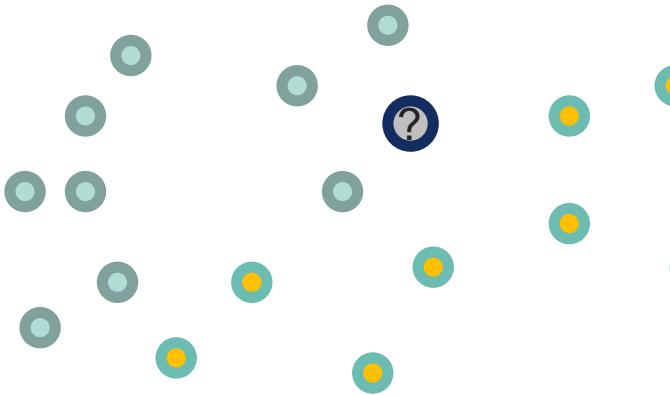
$$Z_{\text{●}} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$Z_{\text{○}} = \frac{w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$



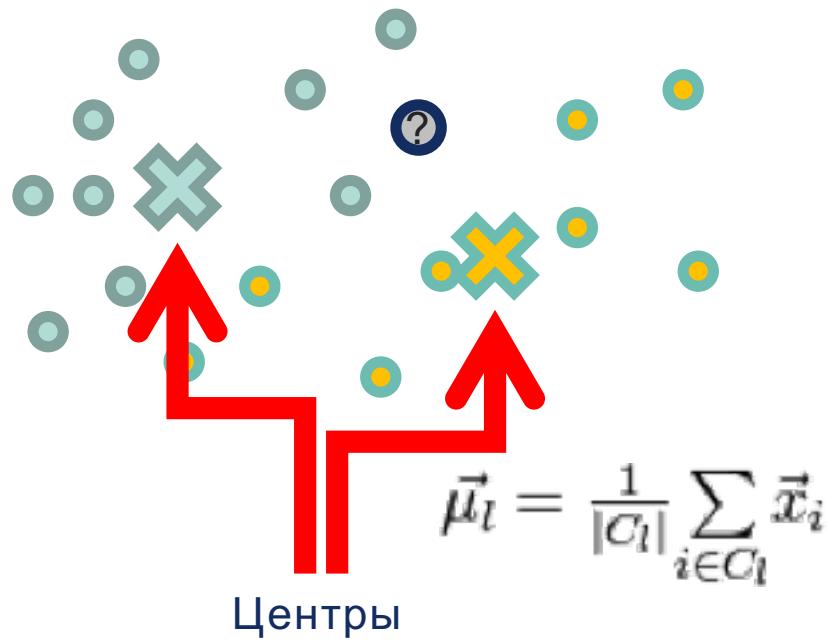
**Другой
похожий
алгоритм**

Центроидный классификатор



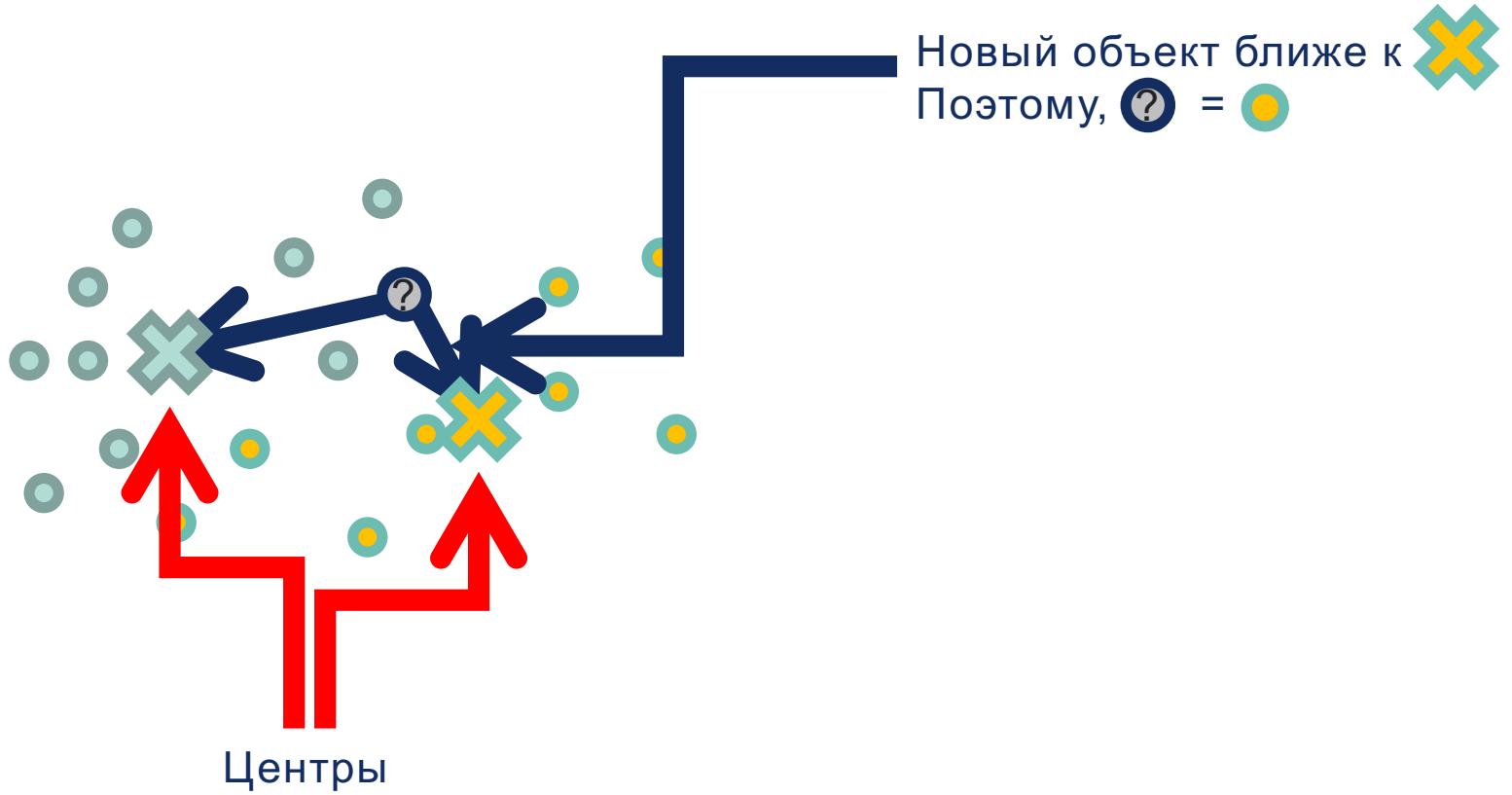
Другой похожий алгоритм

Центроидный классификатор



Другой похожий алгоритм

Центроидный классификатор



Метрические алгоритмы

Мы формируем понятие близости объекта к классу, как правило используя расстояния в пространстве признаков.

Общая идея

Метрические алгоритмы

Мы формируем понятие близости объекта к классу, как правило используя расстояния в пространстве признаков.

Общая идея

Можно брать расстояния до самих объектов класса, можно до центров класса.

Метрические алгоритмы

Мы формируем понятие близости объекта к классу, как правило используя расстояния в пространстве признаков.

Общая идея

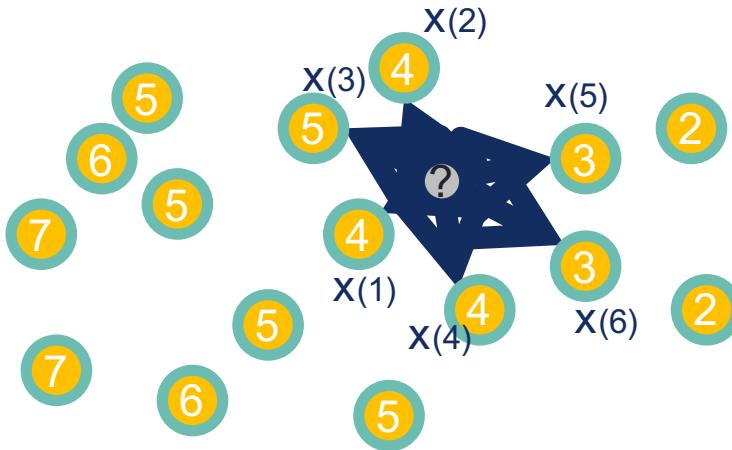
Можно брать расстояния до самих объектов класса, можно до центров класса.

Каждый класс «голосует» таким образом за себя и мы выбираем класс, набирающий больше всего «голосов»

Обобщение для регрессии

Все это применимо и в регрессии

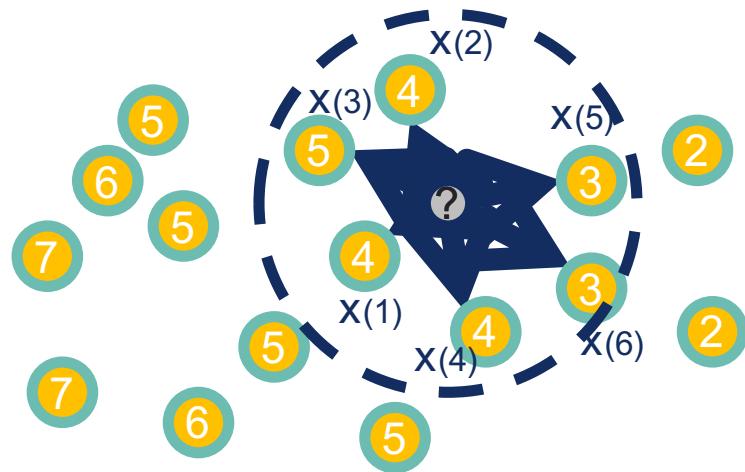
Пример взвешенного kNN ($k = 6$) в задаче регрессии:



Обобщение для регрессии

Все это применимо и в регрессии

Пример взвешенного kNN ($k = 6$) в задаче регрессии:



Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

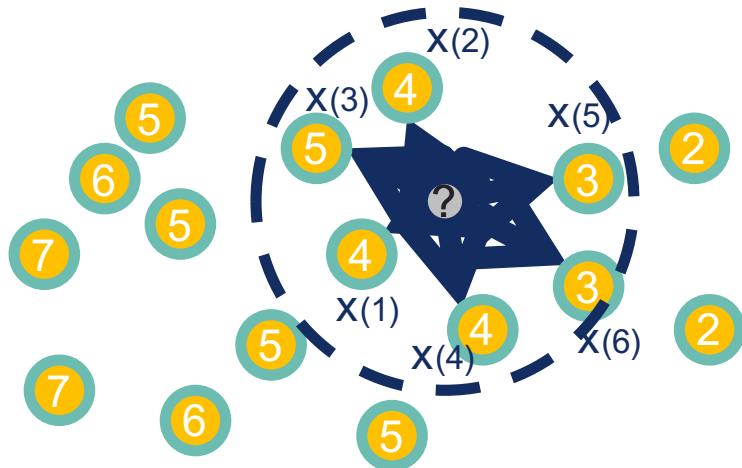
или как функцию расстояния:

$$w(x(i)) = w(d(x, x_{(i)}))$$

Обобщение для регрессии

Все это применимо и в регрессии

Пример взвешенного kNN ($k = 6$) в задаче регрессии:



Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

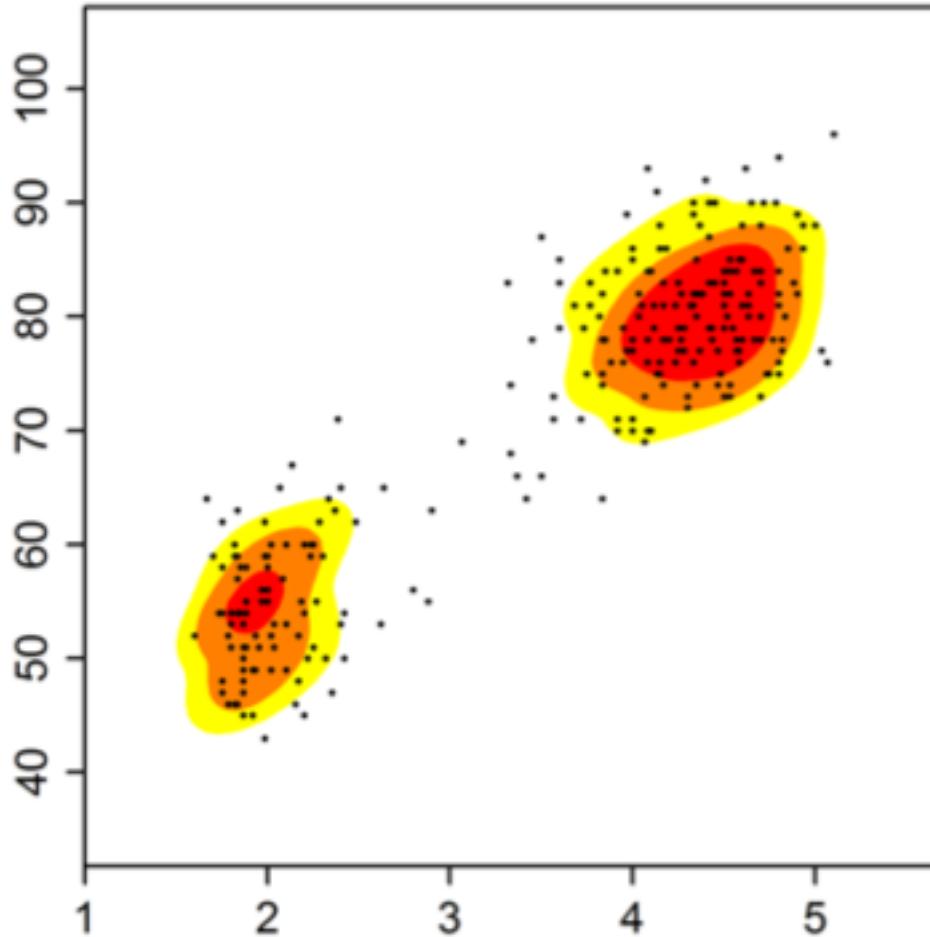
$$w(x(i)) = w(d(x, x_{(i)}))$$

$$\text{?} = \frac{4 \cdot w(x_{(1)}) + 4 \cdot w(x_{(2)}) + 5 \cdot w(x_{(3)}) + 4 \cdot w(x_{(4)}) + 3 \cdot w(x_{(5)}) + 3 \cdot w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

4. Плотность и наивный байес

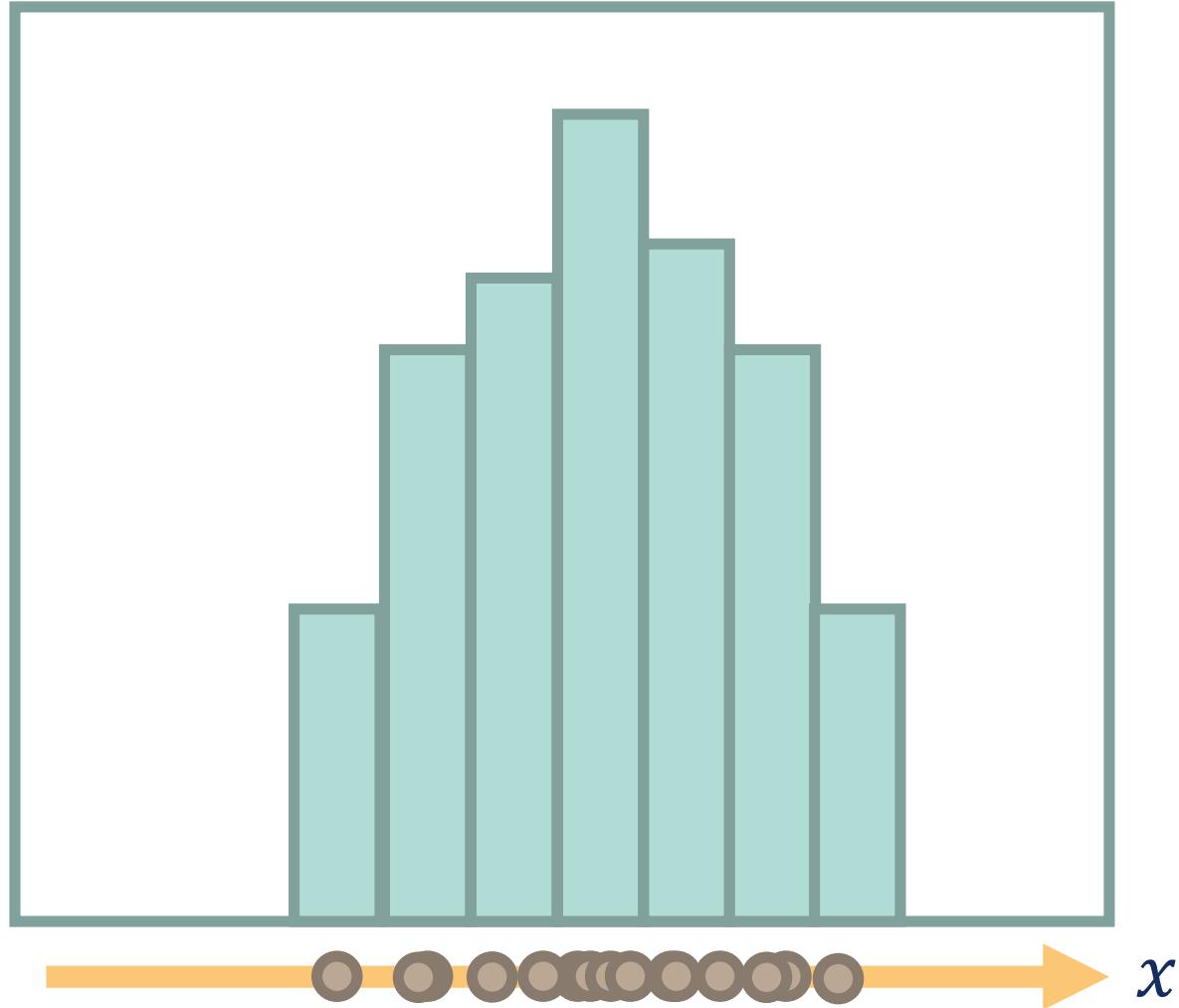
Плотность

Пример для двух признаков



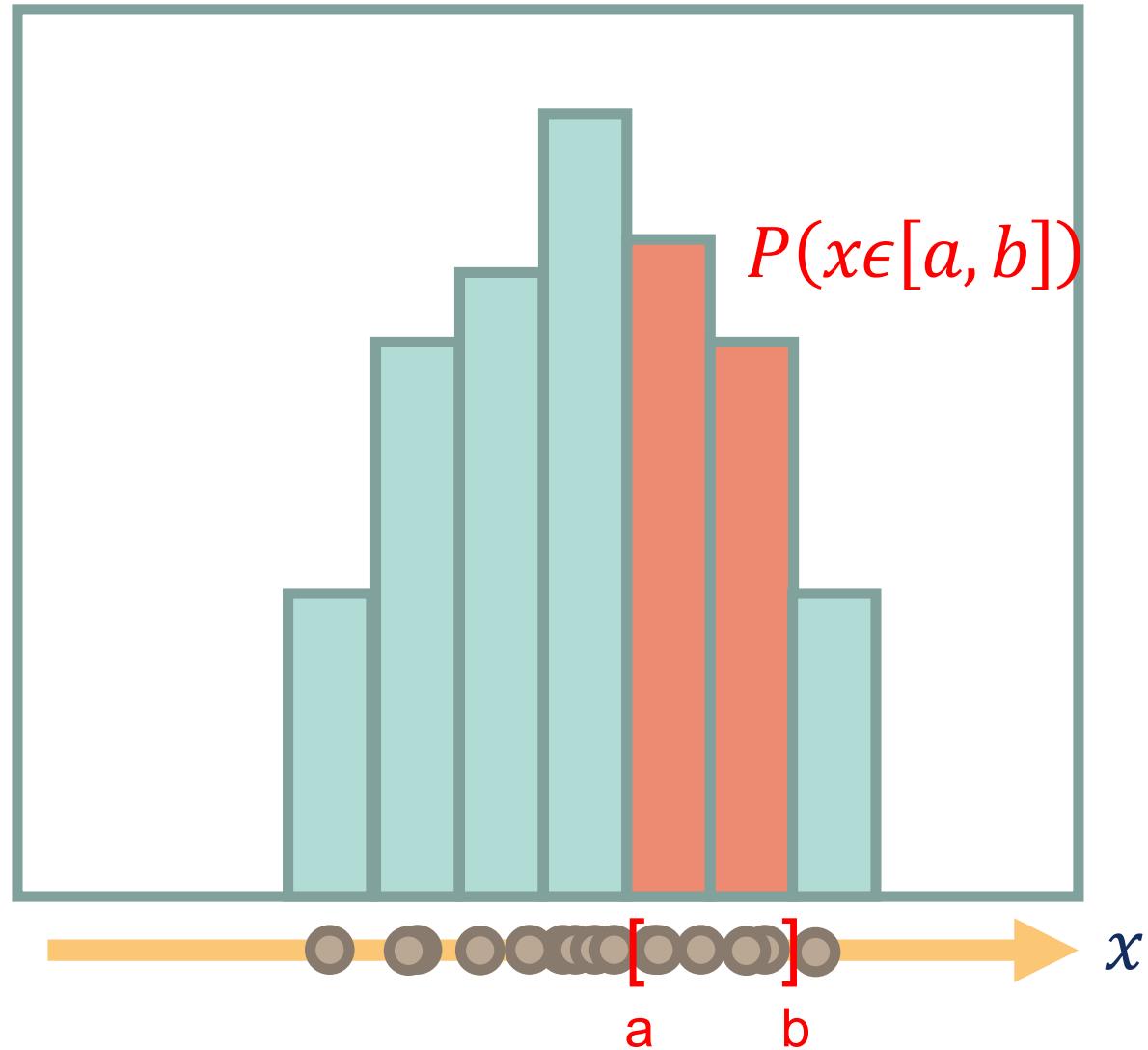
Плотность

Одномерный случай



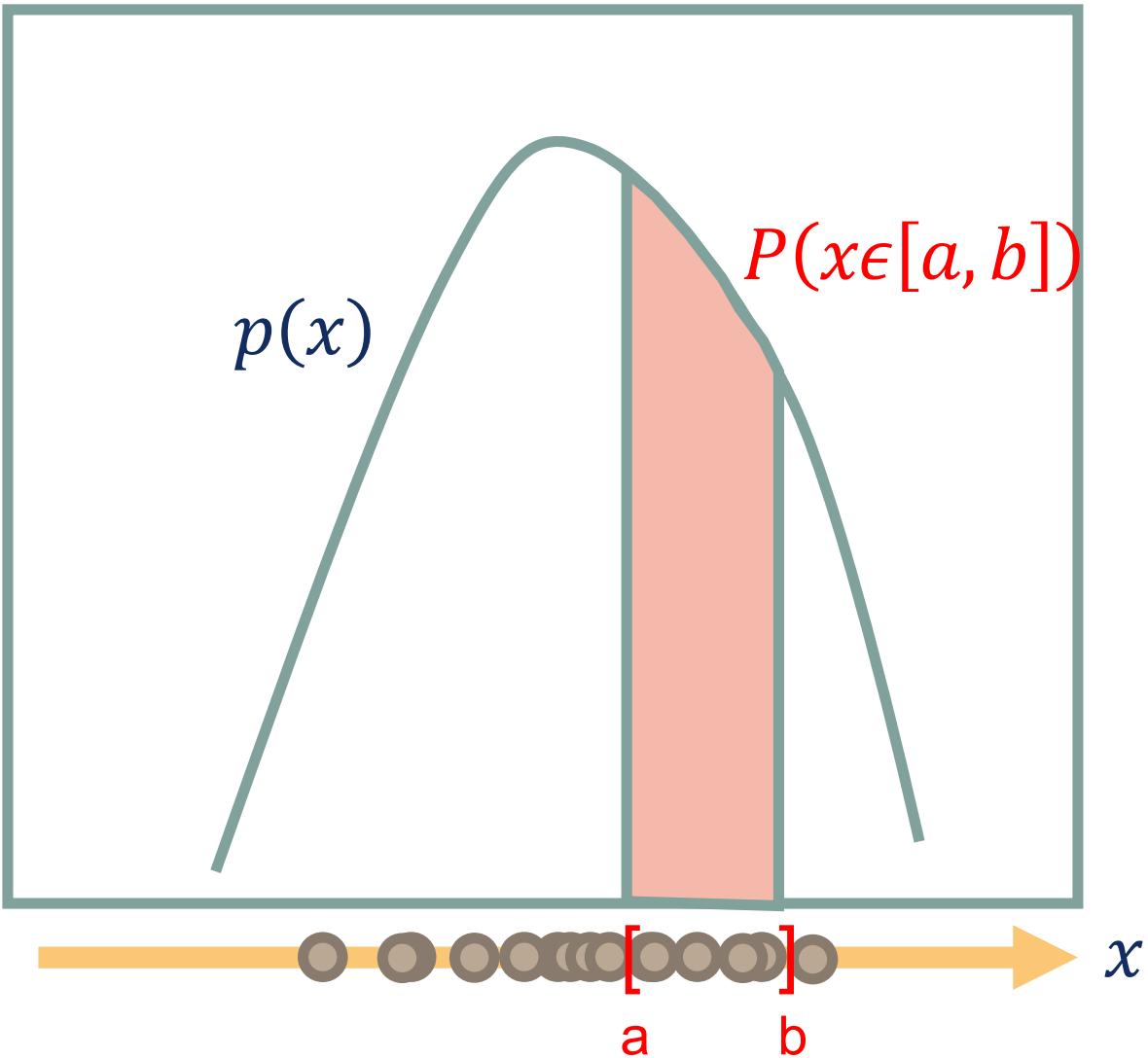
Плотность

Одномерный случай



Плотность

Плотность распределения

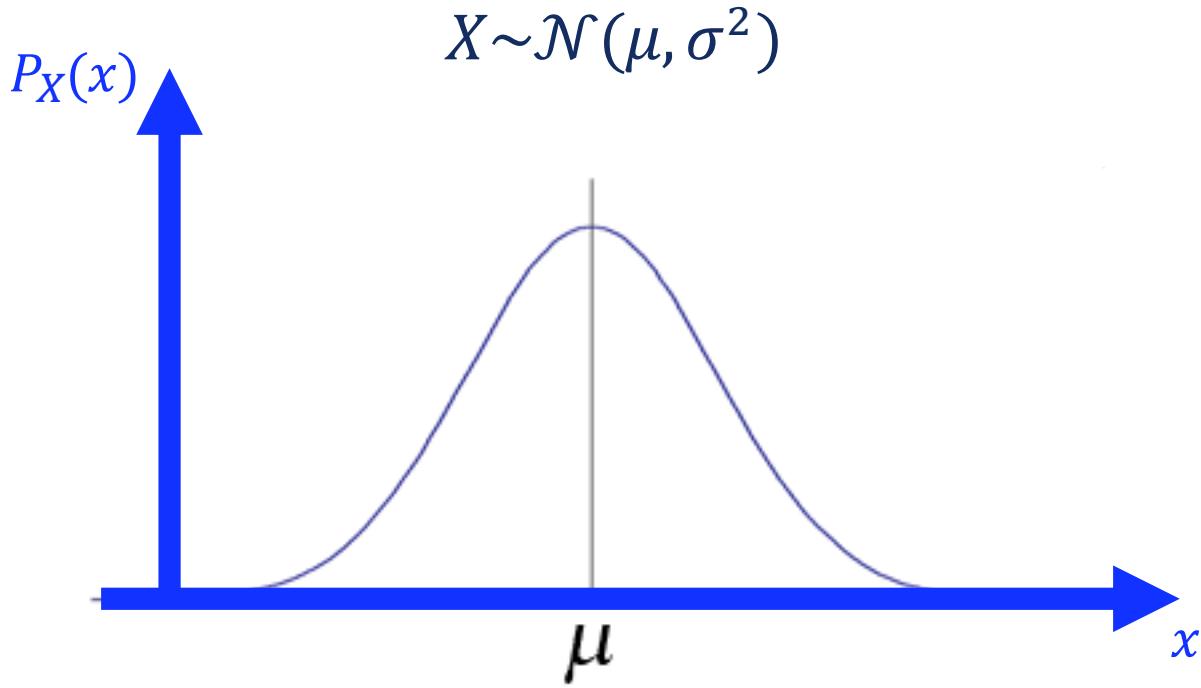


Подходы к оценке плотности

1. Непараметрическая оценка плотности
2. Параметрическая оценка плотности
 - a) Оценка параметров некоторого стандартного распределения (нормальное, мультиномиальное, бернулли)
 - b) Восстановление смеси распределений

Нормальное распределение

Пример
оценки
параметров



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Нормальное распределение

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

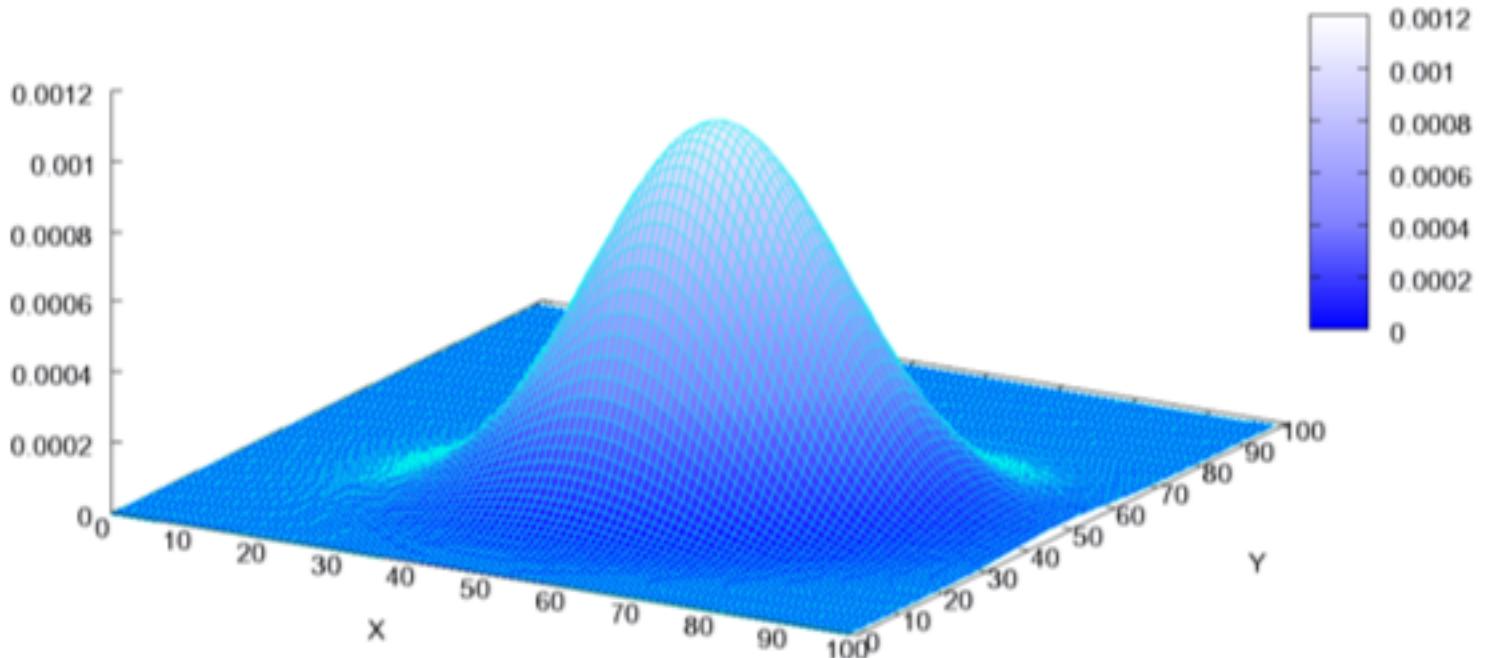
Пример
оценки
параметров

другой вариант оценки для σ^2 :

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

Многомерный пример

Многомерное нормальное распределение

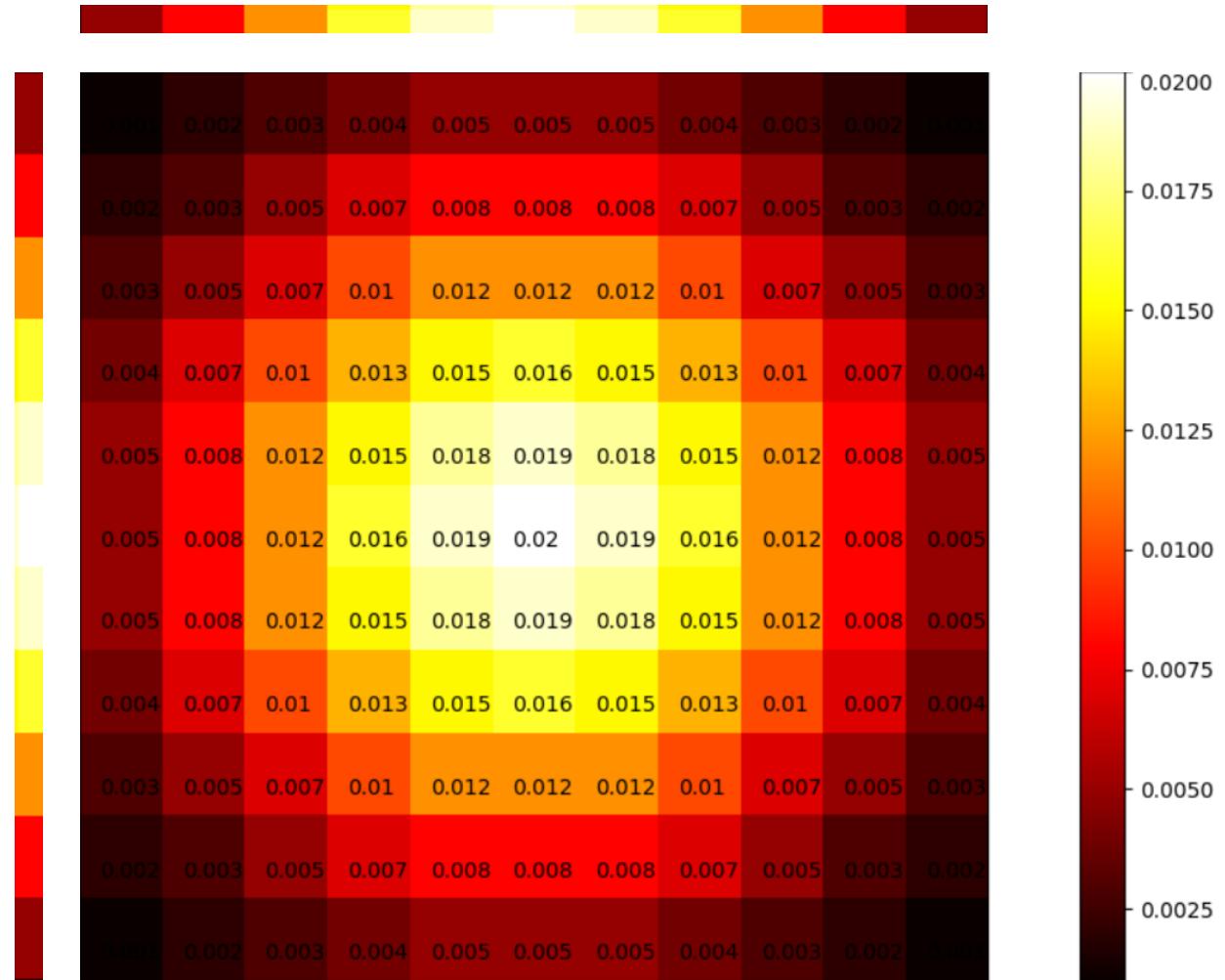


$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Очень много параметров: вектор средних μ и матрица ковариаций Σ

Можно представить $p(x) = p(x^{(1)})p(x^{(2)})$

Произведение одномерных плотностей



Наивная гипотеза

На самом деле так можно не всегда

Если признаки $x^{(1)}, \dots, x^{(d)}$ распределены независимо:

$$p(x) = p(x^{(1)}) \dots p(x^{(d)})$$

В общем же случае это не так. Но даже если признаки не независимы, мы можем сказать «давайте с какой-то степенью точности считать, что это равенство выполнено».

Гипотеза о независимости признаков и дает **наивному байесовскому классификатору** название «наивный»

Умножение вероятностей

Почему вероятности умножаются?

Простой пример:

По данным некоторого опроса выяснилось, что 7/10 опрошенных любят кофе и эта доля одинаковая как среди мужчин, так и среди женщин (не зависит от этого признака). Половина опрошенных были мужчинами, половина – женщинами.

Какую долю среди всех опрошенных составляют мужчины, которые любят кофе?

Умножение вероятностей

Почему вероятности умножаются?

Простой пример:

По данным некоторого опроса выяснилось, что $7/10$ опрошенных любят кофе и эта доля одинаковая как среди мужчин, так и среди женщин (не зависит от этого признака). Половина опрошенных были мужчинами, половина – женщинами.

Какую долю среди всех опрошенных составляют мужчины, которые любят кофе?

Ответ: $\frac{1}{2} \cdot \frac{7}{10}$

Умножение вероятностей

Почему вероятности умножаются?

Простой пример:

По данным некоторого опроса выяснилось, что $7/10$ опрошенных любят кофе и эта доля одинаковая как среди мужчин, так и среди женщин (не зависит от этого признака). Половина опрошенных были мужчинами, половина – женщинами.

Какую долю среди всех опрошенных составляют мужчины, которые любят кофе?

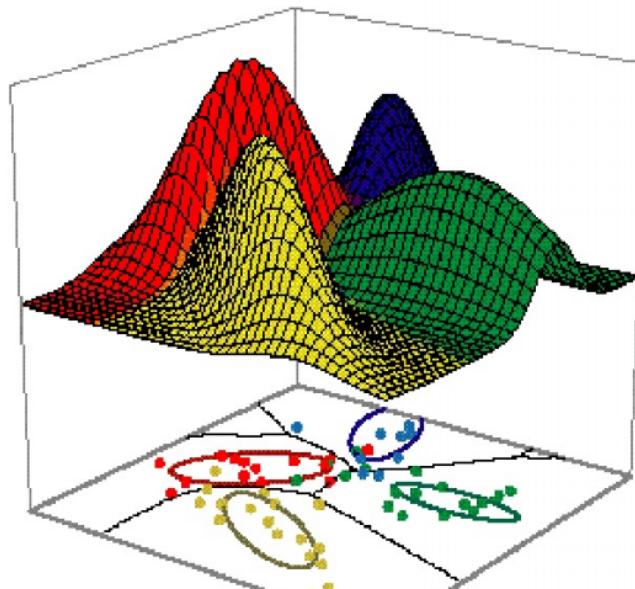
Ответ: $\frac{1}{2} \cdot \frac{7}{10}$

А вот если бы пол влиял на любовь к кофе, вместо $7/10$ в ответе было бы другое число

Классификация

Как можно определять класс

Если мы знаем плотности классов, то можем относить объект выборки к тому классу, плотность которого в этой точке признакового пространства больше:



Как решить задачу классификации

1. Считаем, что $p(x) = p(x^{(1)}) \dots p(x^{(d)})$
2. Оцениваем **для каждого класса каждую из одномерных плотностей** по выборке (например, считаем нормальными и вычисляем параметры по формуле)
3. Классифицируя объект x выбираем класс с максимальной плотностью в точке x

Первая идея

Как решить задачу классификации

1. Считаем, что $p(x) = p(x^{(1)}) \dots p(x^{(d)})$
2. Оцениваем **для каждого класса каждую из одномерных плотностей** по выборке (например, считаем нормальными и вычисляем параметры по формуле)
3. Классифицируя объект x выбираем класс с максимальной плотностью в точке x

Первая идея

Проблема: как сделать поправку на то, что какой-то класс в принципе редко встречается?

Наивный Байес

Наивный байесовский классификатор

$p(x|y)$ - Плотность класса y

Считаем, что $p(x|y) = p(x^{(1)}|y) \dots p(x^{(d)}|y)$

Обучение модели:

1. Оцениваем **для каждого класса y каждую из одномерных плотностей $p(x^{(k)}|y)$** по выборке (например, считаем нормальными и вычисляем параметры по формуле)
2. Оцениваем для **для каждого класса y его априорную вероятность $P(y)$**

Применение модели:

$$a(x) = \operatorname{argmax}_y \left(P(y) p(x^{(1)}|y) \dots p(x^{(d)}|y) \right)$$

План лекции

1. Порог по одному признаку
2. От пней к деревьям
3. Сложные границы и соседи
4. Плотность и наивный байес

Data Mining in Action

Лекция 2

Группа курса в Telegram:



<https://t.me/joinchat/B1OITk74nRV56Dp1TDJGNA>