Data Mining in Action

Лекция 1. Простые методы и математика

Базовый модуль

Занятия

1. Простые методы и математика
2. Решающие деревья и ансамбли
3. Линейные модели
4. Валидация качества
5. Введение в обучение без учителя

Задания

1. Простые методы: теория + код
2. Ансамбли: контест + код
3. Линейные модели: контест + код
4. Валидация качества: теория + квест

Переход на другие модули и отсев:

1. Индустрия: топ 15-20 рейтинга
2. Deep Learning: топ 40 рейтинга
3. Отсев после майских праздников: те, кто не прислал ни одного задания

Обозначения

	Средний заказ	Жалоб за неделю	Дней не активен	...	Лет с нами	Ушел в отток?
x_1	400 р	5	10	$\cdots$	0.1	**Да**
x_2	600 р	1	8	$\cdots$	1.5	**Нет**
x_3	200 р	0	6	$\cdots$	0.2	**Да**
...						

Обозначения

	Средний заказ	Жалоб за неделю	Дней не активен	…	Лет с нами	Ушел в отток?
x_1	x_{11} 400 р	x_{12} 5	x_{13} 10	…	0.1	Да
x_2	600 р	1	8	…	1.5	Нет
x_3	200 р	0	6	…	0.2	Да
…						

Обозначения

	Средний заказ	Жалоб за неделю	Дней не активен	…	Лет с нами	Ушел в отток?
x_1	x_{11}	x_{12}	x_{13}	…	0.1	Да
x_2	x_{21}	x_{22}	8	…	1.5	Нет
x_3	200 р	0	6	…	0.2	Да
…						

Обозначения

	Средний заказ	Жалоб за неделю	Дней не активен	…	Лет с нами	Ушел в отток?
x_1	x_{11}	x_{12}	x_{13}	…	0.1	y_1
x_2	x_{21}	x_{22}	8	…	1.5	y_2
x_3	200 р	0	6	…	0.2	y_3
…						…

Терминология

Выборка: $\{x_1, \dots, x_n\}$

Объекты выборки:

$$x_1 = (x_{11}, \dots, x_{1d})$$
$$\dots$$
$$x_n = (x_{n1}, \dots, x_{nd})$$

$x_{i1}, \dots, x_{id}$ - признаки i-того объекта

Терминология

Матрица признаков:

$$\begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix}$$

$y_1, \dots, y_n$ - значения **целевой переменной** (target) на объектах выборки

X – пространство признаков, Y – множество ответов

$$\forall i = 1 \dots n \quad x_i \in X \quad y_i \in Y$$

Выборки

Обучающая выборка:

$$(X^\ell, Y^\ell) = (\{x_1, \dots, x_\ell\}, \{y_1, \dots, y_\ell\})$$

Тестовая выборка:

$$(X^t, Y^t) = (\{x_{\ell+1}, \dots, x_{\ell+t}\}, \{y_{\ell+1}, \dots, y_{\ell+t}\})$$

Задача: восстановить отображение $X \to Y$, построив на обучающей выборке **модель** (алгоритм) a, такую, что для объектов x, похожих на объекты обучающей выборки, $a(x)$ в среднем хорошо предсказывает ответ y

Выборки

Обучающая выборка:

$$(X^\ell, Y^\ell) = (\{x_1, \dots, x_\ell\}, \{y_1, \dots, y_\ell\})$$

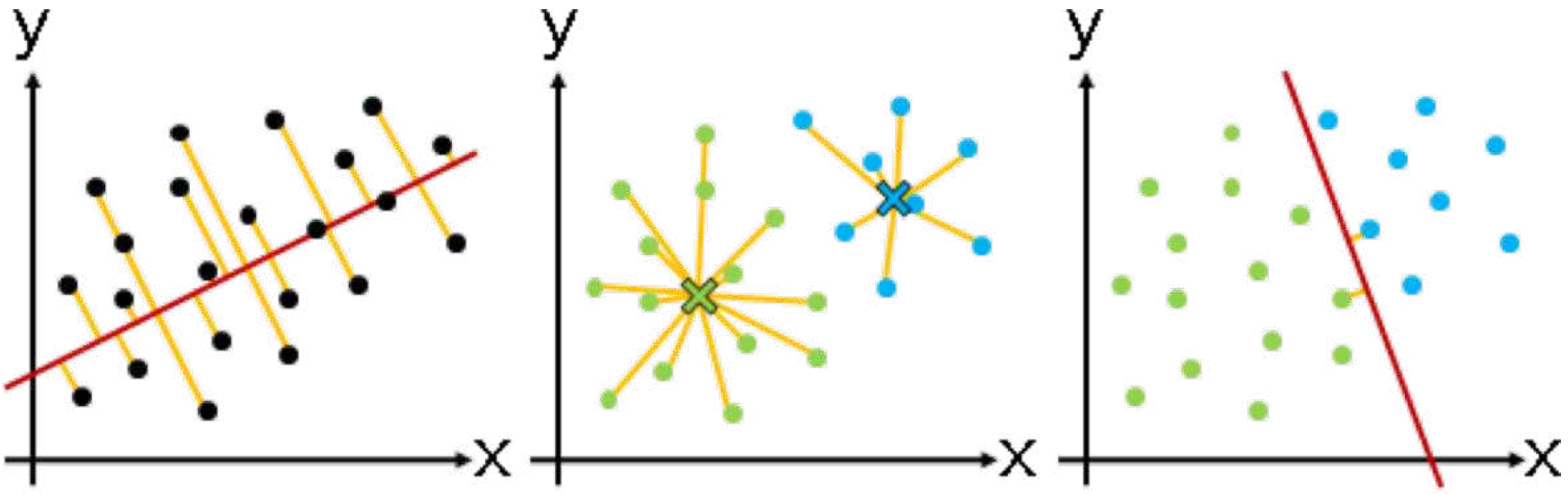
Тестовая выборка:

$$(X^t, Y^t) = (\{x_{\ell+1}, \dots, x_{\ell+t}\}, \{y_{\ell+1}, \dots, y_{\ell+t}\})$$

На обучающей выборке настраиваем алгоритм

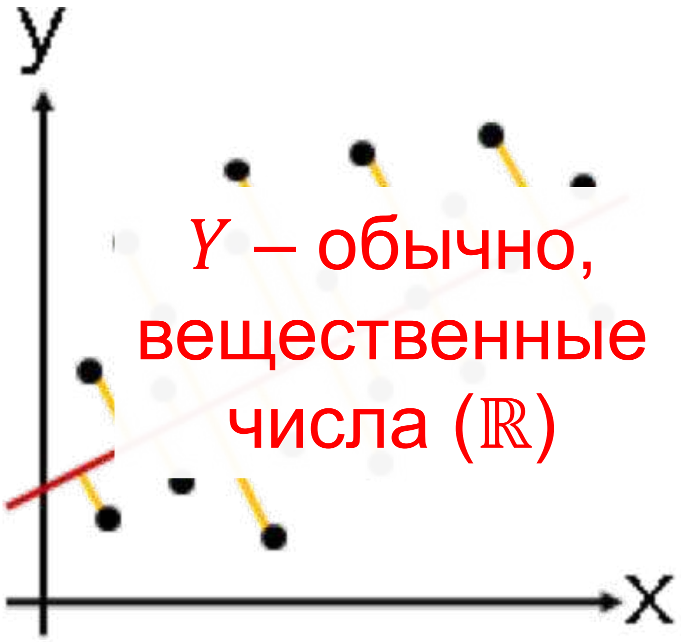
На тестовой – оцениваем качество прогнозирования

Стандартные задачи



Регрессия Кластеризация Классификация

Стандартные задачи



Регрессия

Кластеризация

Классификация

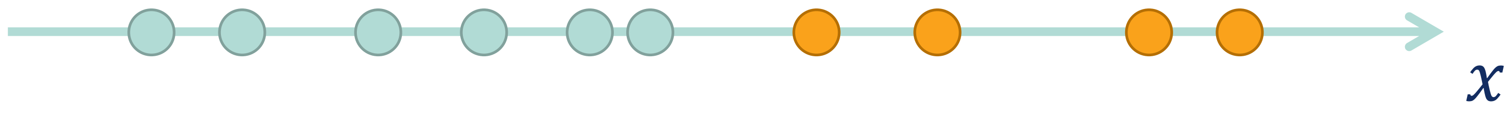
План лекции

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

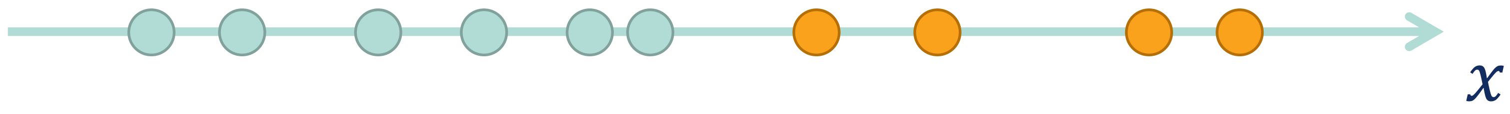
Рассмотрим выборку объектов с одним признаком x:



Как подобрать порог по признаку в задаче бинарной классификации?

Выборка

Рассмотрим выборку объектов с одним признаком x:

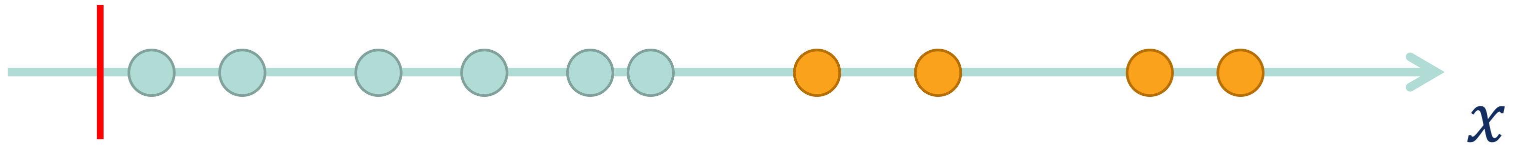


Как подобрать порог по признаку в задаче бинарной классификации?

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

Выборка

Рассмотрим выборку объектов с одним признаком x:

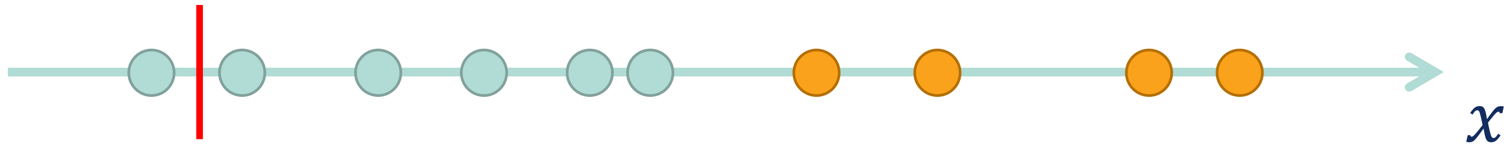


Как подобрать порог по признаку в задаче бинарной классификации?

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

Выборка

Рассмотрим выборку объектов с одним признаком x:

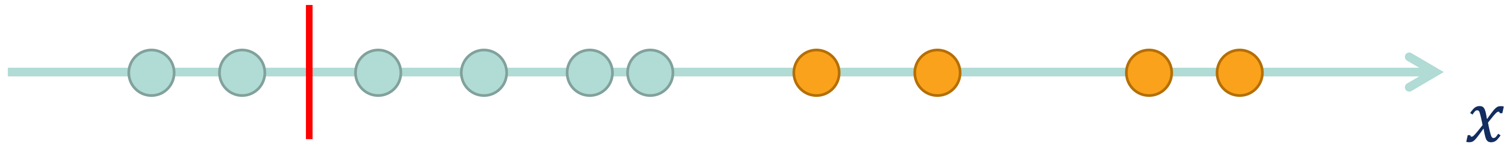


Как подобрать порог по признаку в задаче бинарной классификации?

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

Выборка

Рассмотрим выборку объектов с одним признаком x:

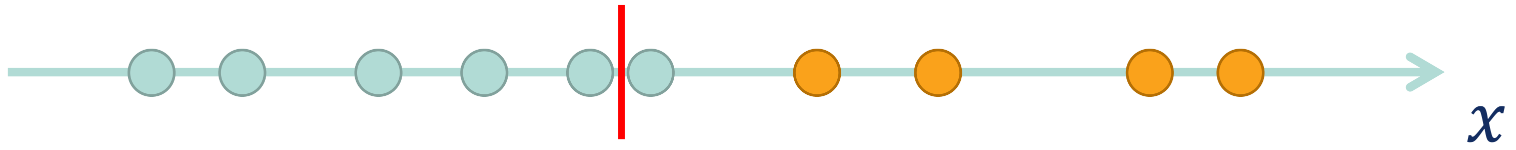


Как подобрать порог по признаку в задаче бинарной классификации?

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

Выборка

Рассмотрим выборку объектов с одним признаком x:

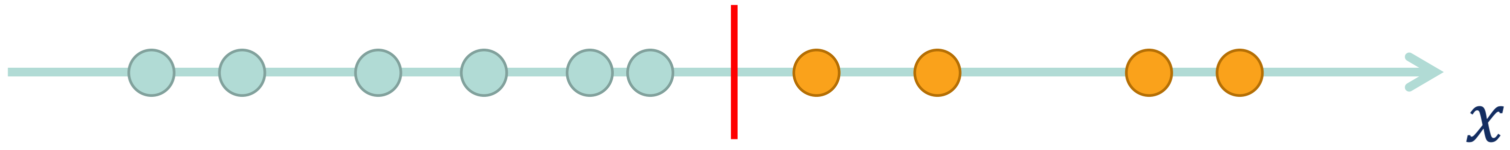


Как подобрать порог по признаку в задаче бинарной классификации?

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

Выборка

Рассмотрим выборку объектов с одним признаком x:

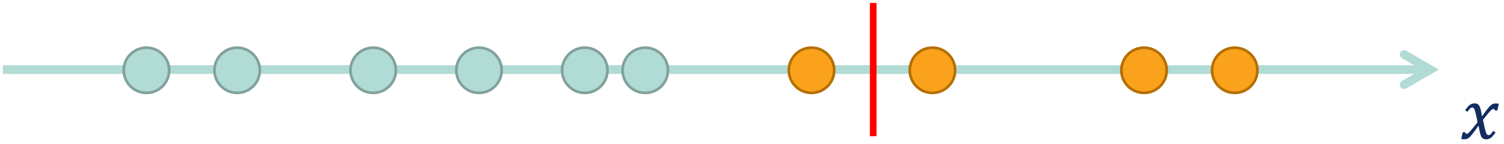


Как подобрать порог по признаку в задаче бинарной классификации?

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

Выборка

Рассмотрим выборку объектов с одним признаком x:

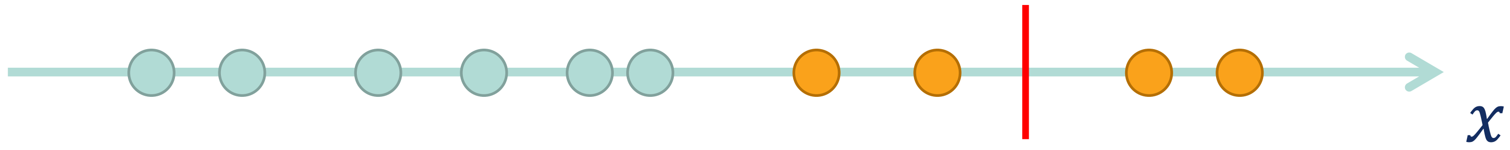


Как подобрать порог по признаку в задаче бинарной классификации?

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

Выборка

Рассмотрим выборку объектов с одним признаком x:

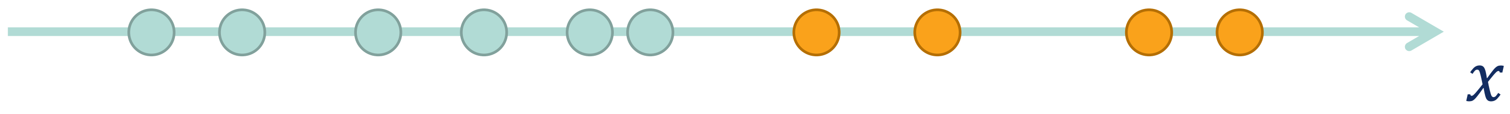


Как подобрать порог по признаку в задаче бинарной классификации?

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

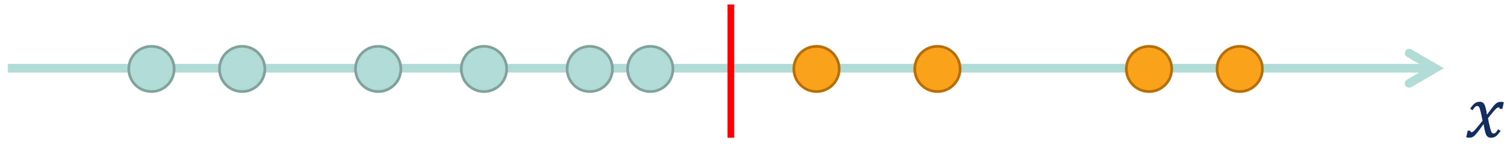
Выборка

Рассмотрим выборку объектов с одним признаком x:



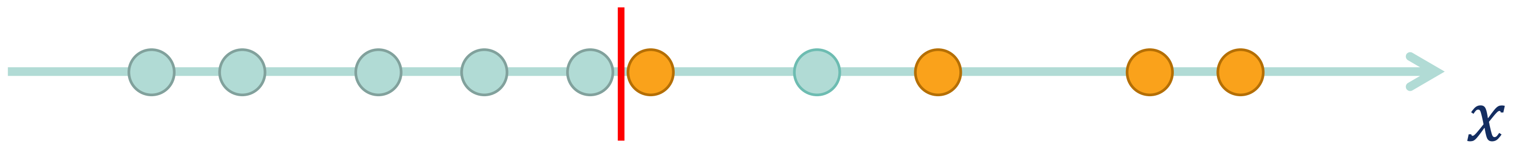
Как подобрать порог по признаку в задаче бинарной классификации?

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



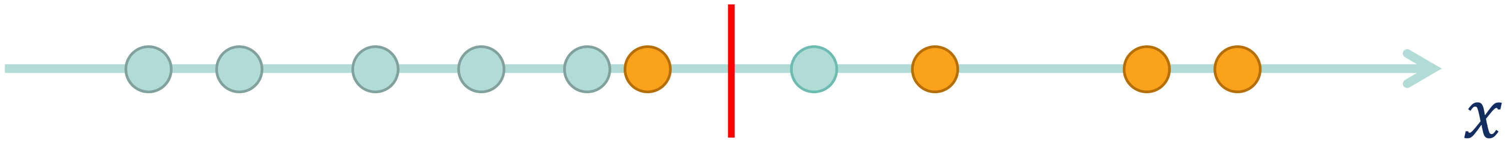
Проблема выбора

Часто есть несколько неплохих порогов:



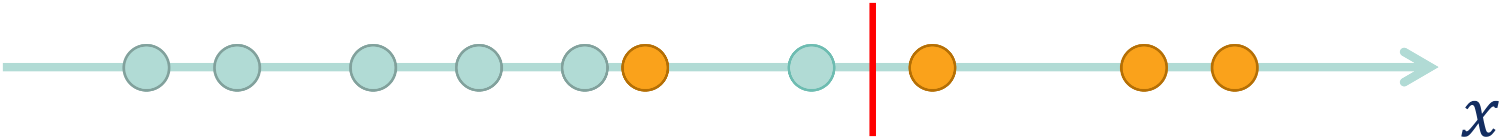
Проблема выбора

Часто есть несколько неплохих порогов:



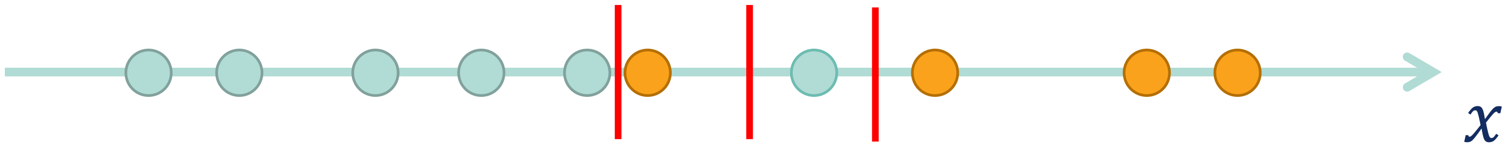
Проблема выбора

Часто есть несколько вариантов деления:



Усложнение модели

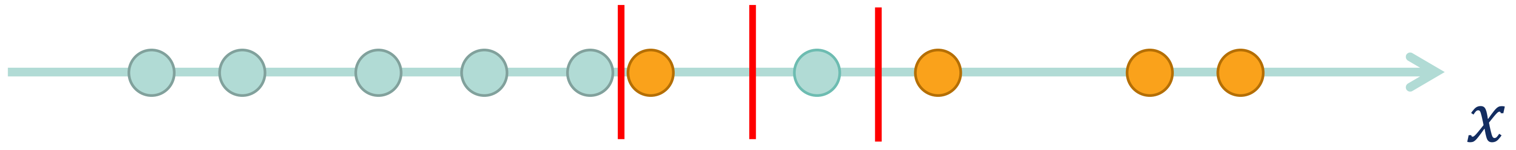
Если потребовать от модели максимальной точности



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

Усложнение модели

Если потребовать от модели максимальной точности

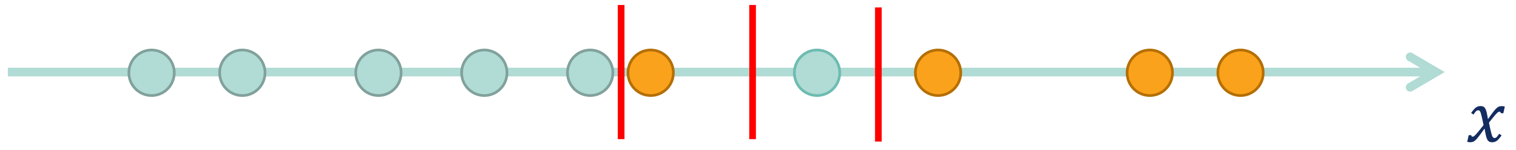


и не ограничивать количество порогов, можно было бы разделить выборку идеально

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

Усложнение модели

Если потребовать от модели максимальной точности



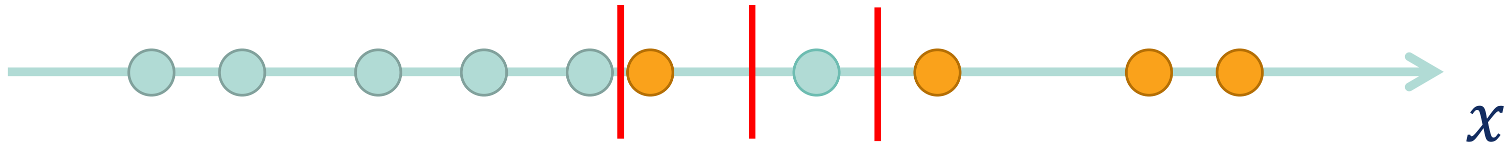
и не ограничивать количество порогов, можно было бы разделить выборку идеально

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

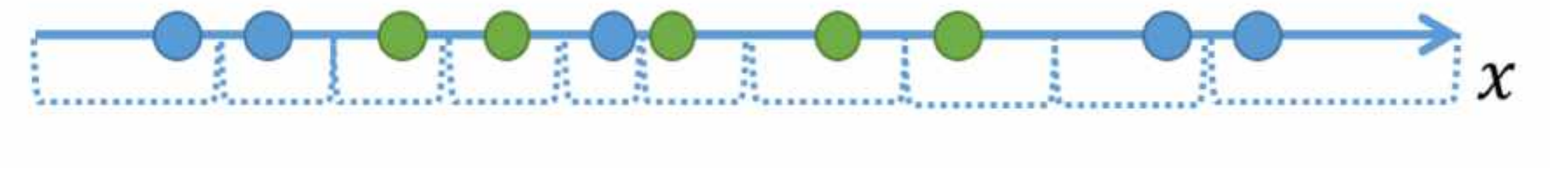
Случай множества порогов

Если потребовать от модели максимальной точности



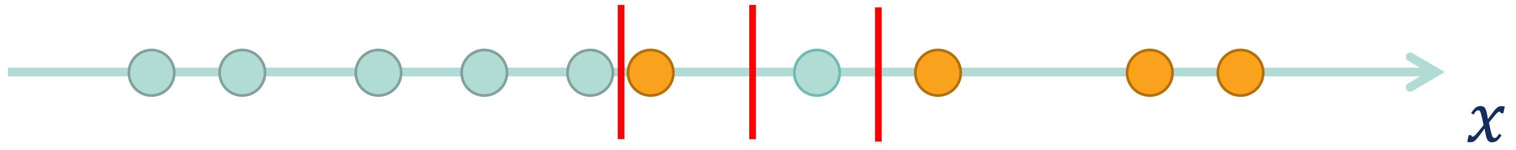
и не ограничивать количество порогов, можно было бы разделить выборку идеально

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



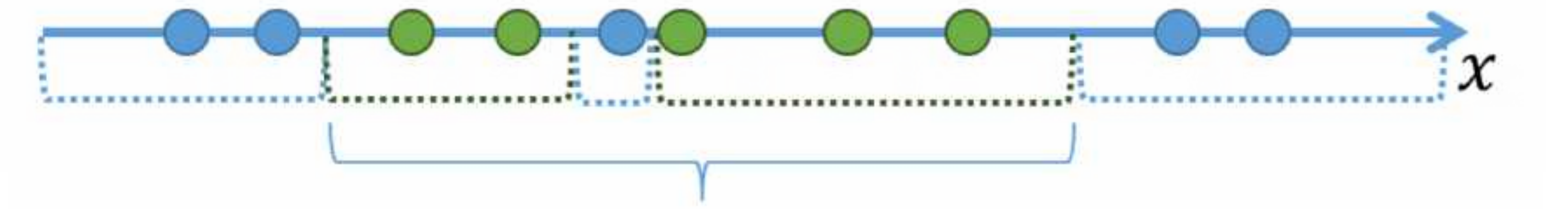
Случай множества порогов

Если потребовать от модели максимальной точности



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

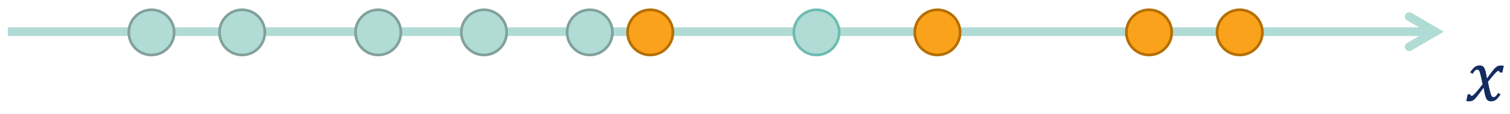
Но это просто запоминание выборки – такой классификатор легко может оказаться слишком переобученным:



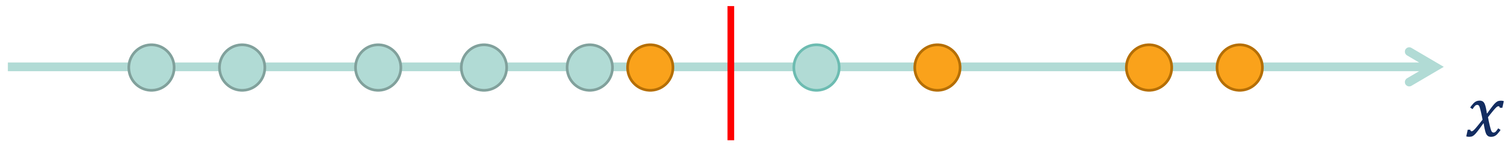
Возможно какие-то интервалы лучше объединить

Разделение неразделимой выборки

Предположим, выборка не разделима идеально:

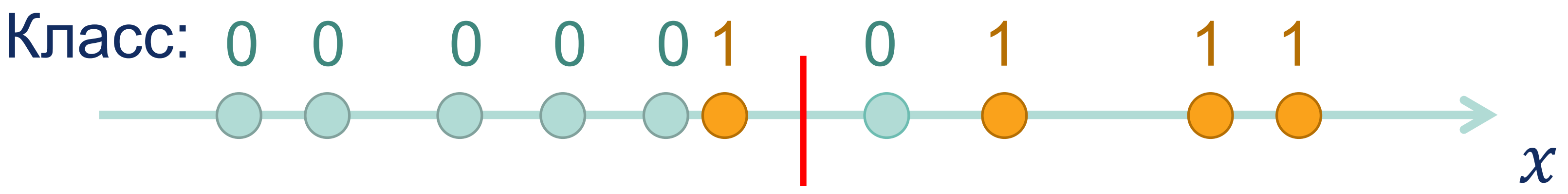


Но нужно по-прежнему адекватно ее разделить:

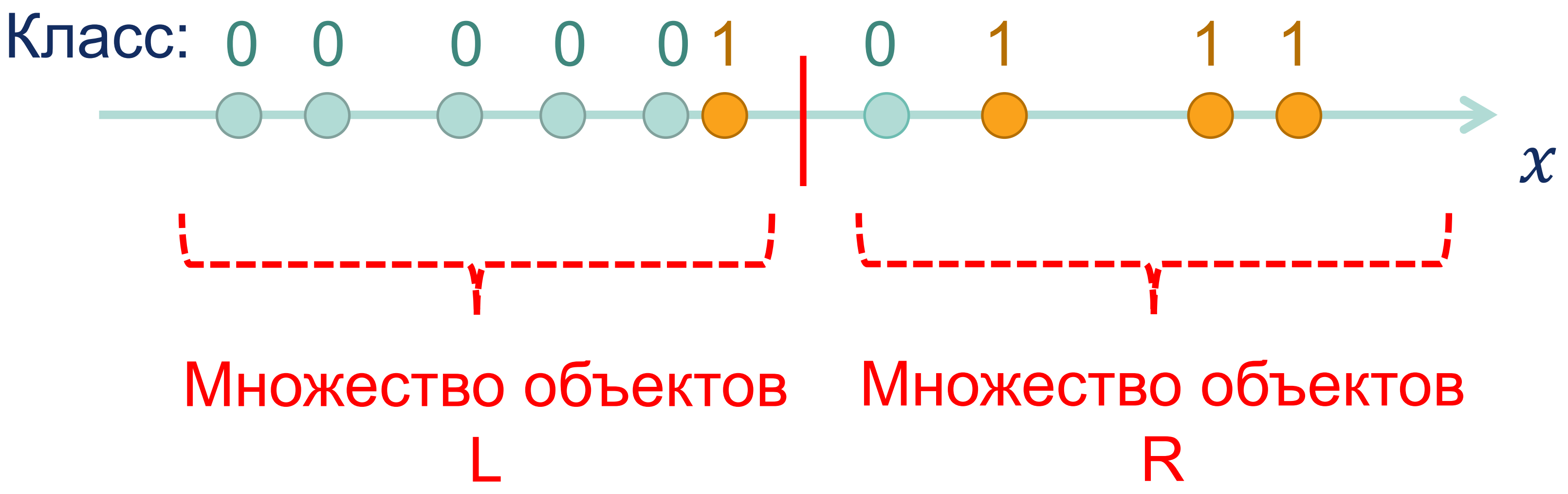


Как это записать?

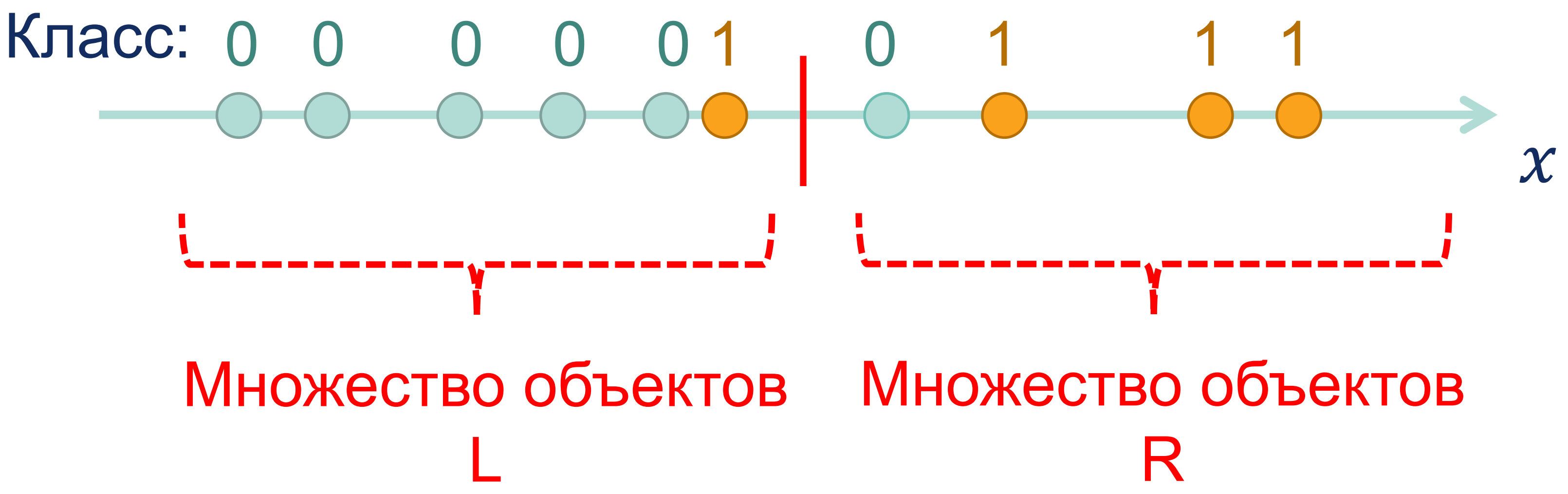
Задача оптимизации



Задача оптимизации

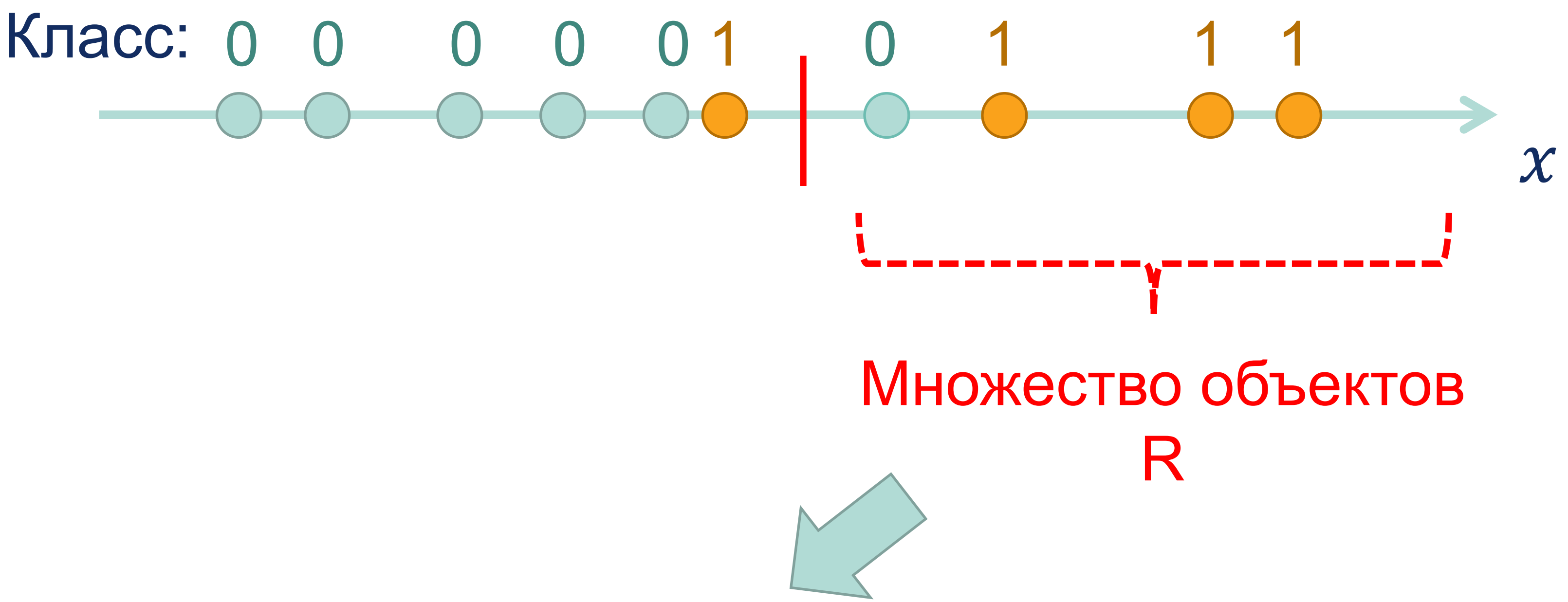


Задача оптимизации



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

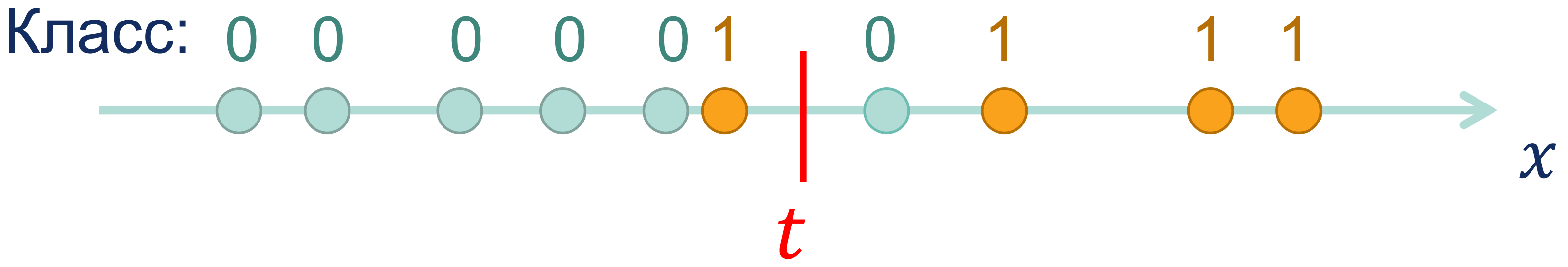
Задача оптимизации



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

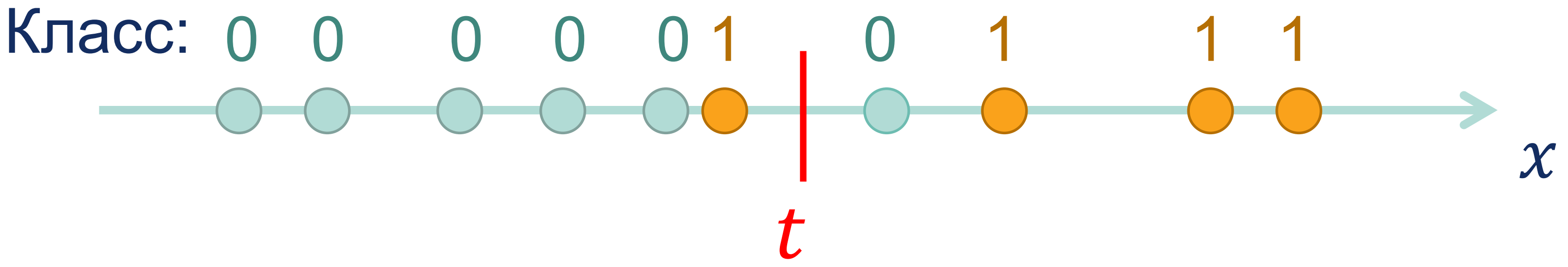


Как записать, что один из классов должен преобладать в R?

Например, так:

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

Задача оптимизации



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

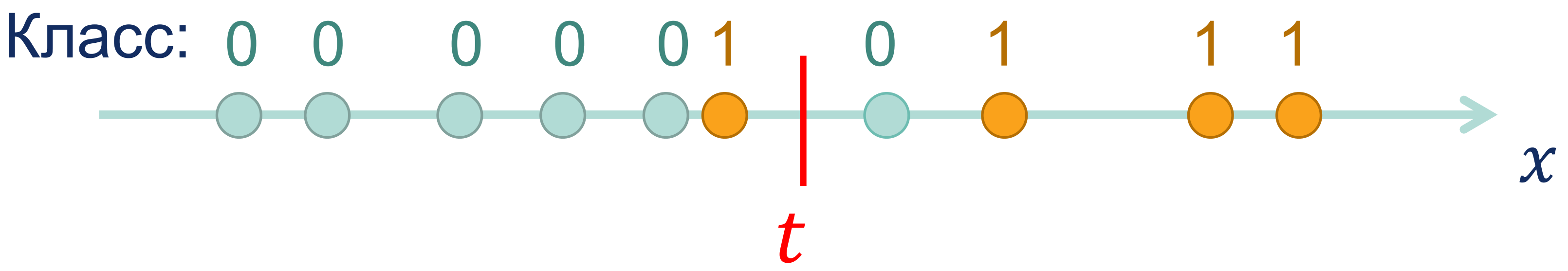
Например, так:

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

Или так:

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

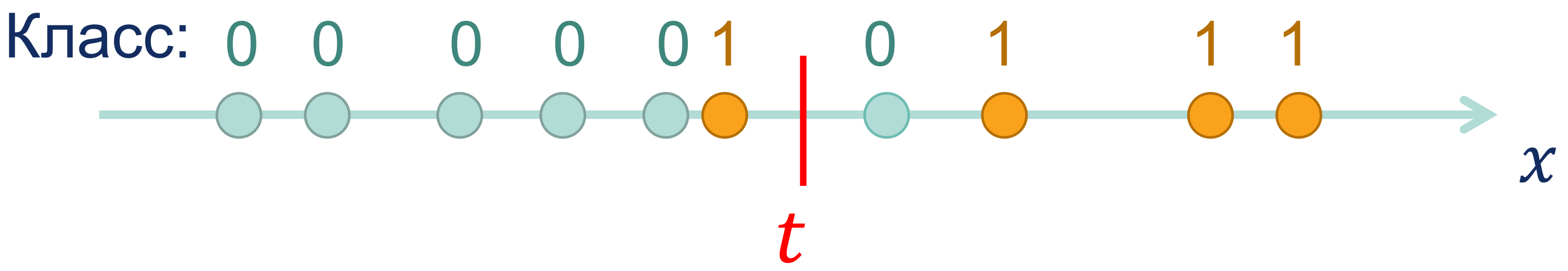
Задача оптимизации



Другой вариант:

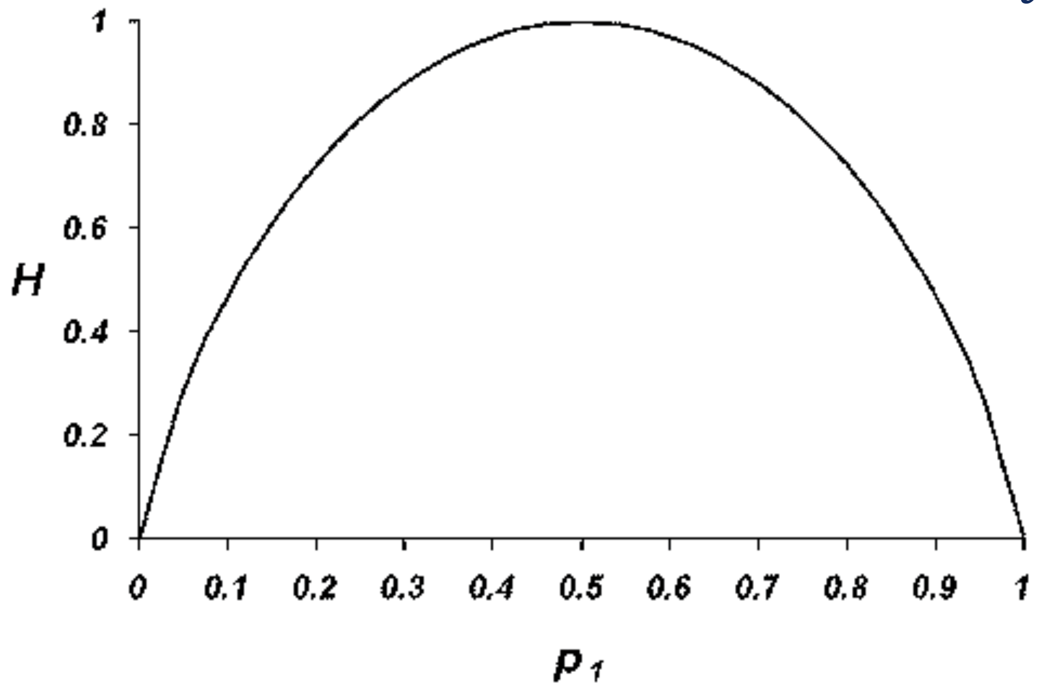
$$H(R) = -p_0 \ln p_0 - p_1 \ln p_1 \to \min_t$$

Задача оптимизации

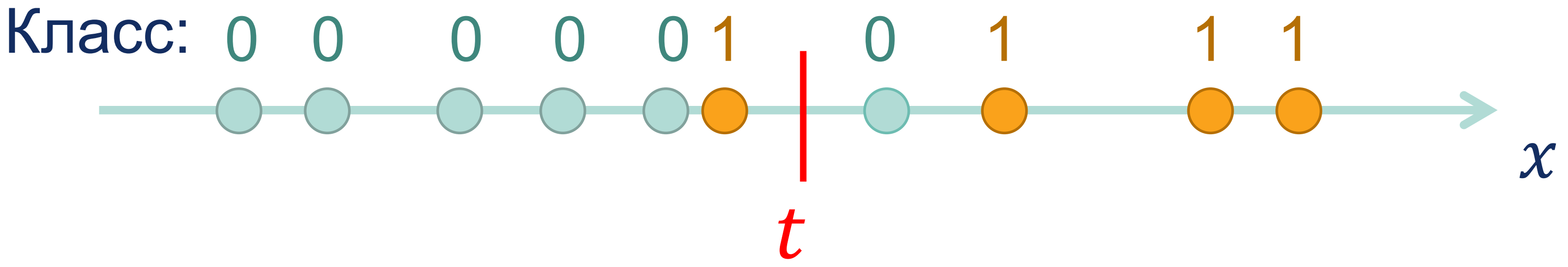


Другой вариант:

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

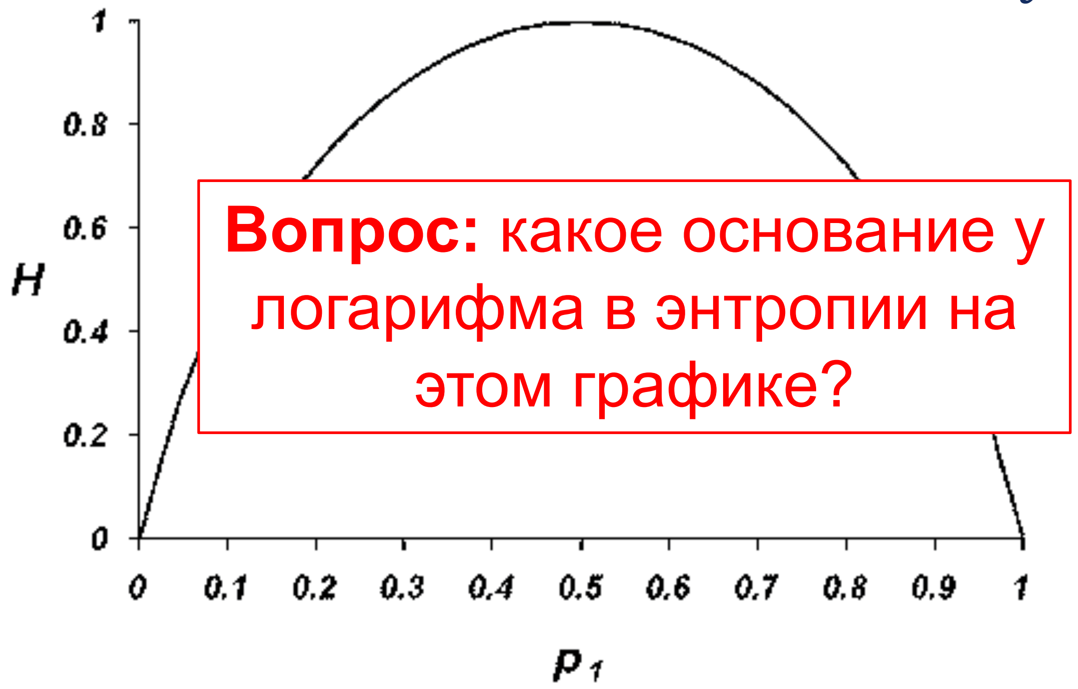


Задача оптимизации



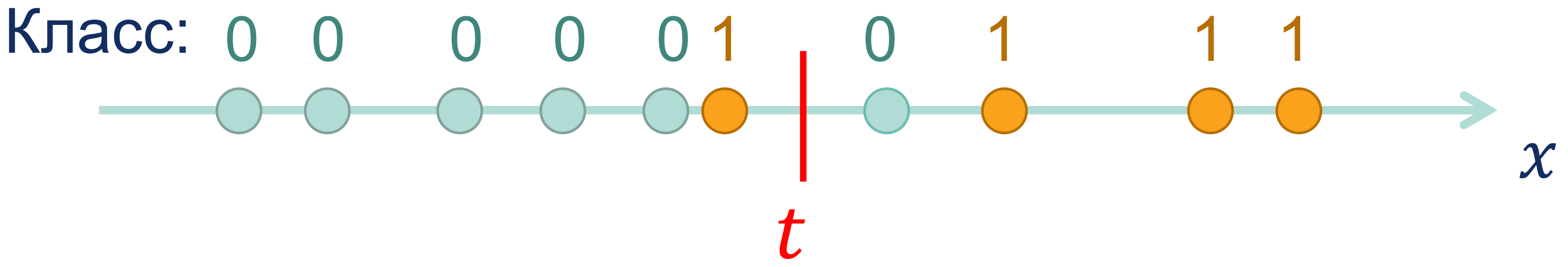
Другой вариант:

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



Вопрос: какое основание у логарифма в энтропии на этом графике?

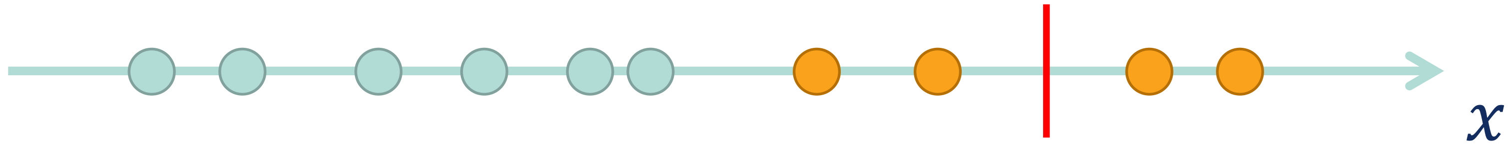
Задача оптимизации



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

Уточнение

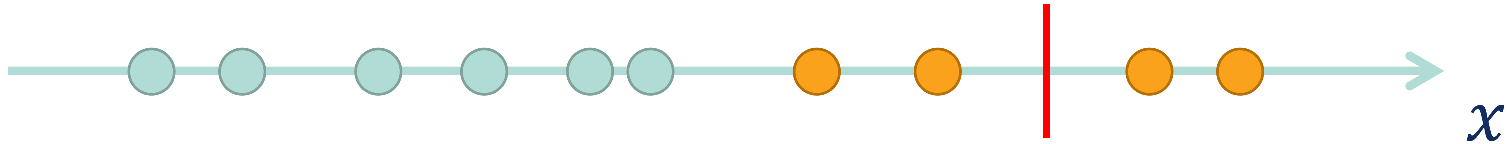
Но если смотреть только на R, можем нечаянно разделить выборку так:



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

Уточнение

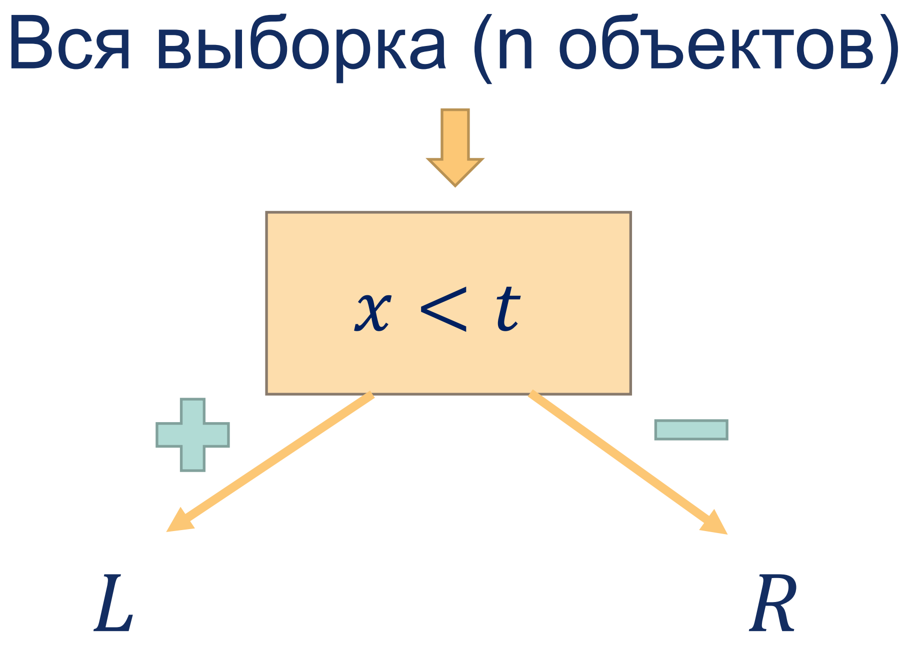
Но если смотреть только на R, можем нечаянно разделить выборку так:



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

Значит надо учитывать обе части: R и L

Выбор разбиения

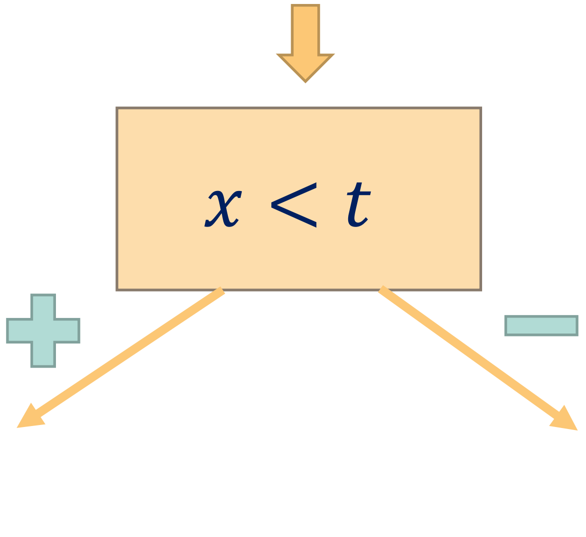


$$G(t) = H(L) + H(R) \to \min_t$$

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

Выбор разбиения

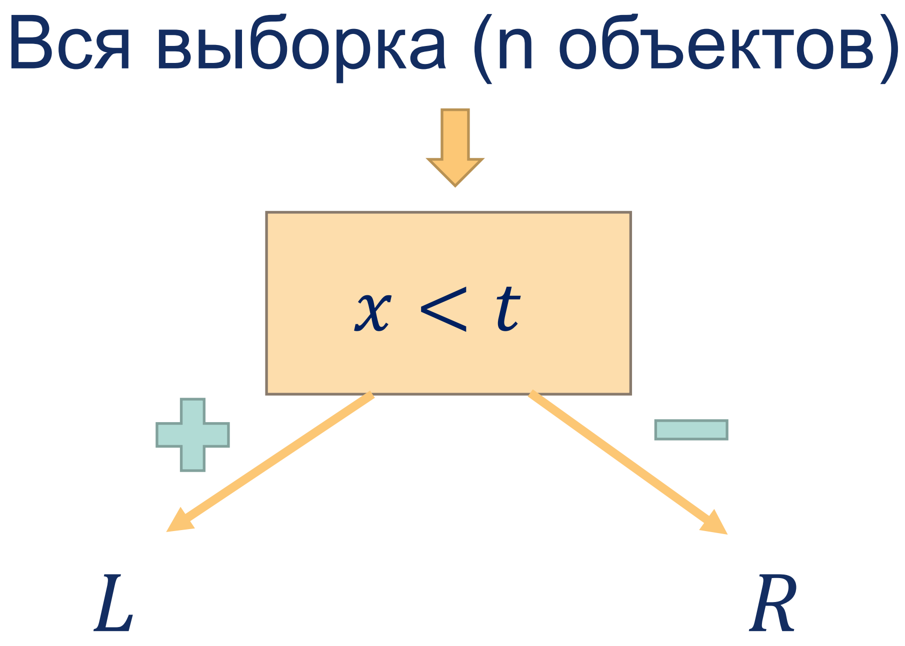
Вся выборка (n объектов)



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

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

Выбор разбиения



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

Критерии построения разбиений

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

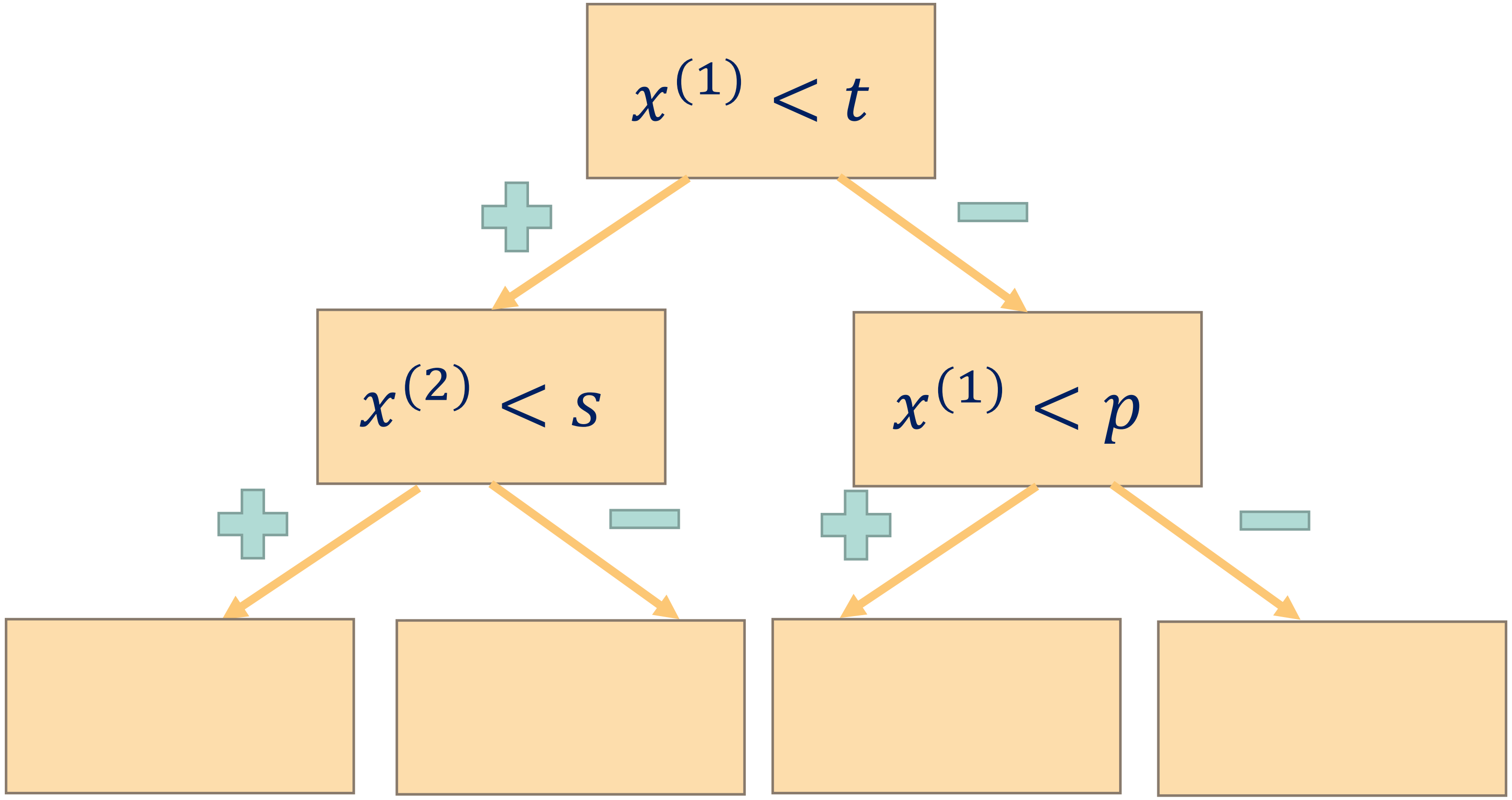
Критерии построения разбиений

$H(R)$ – 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: $H(R) = 1 - {p_0}^2 - {p_1}^2 = 2p_0p_1$

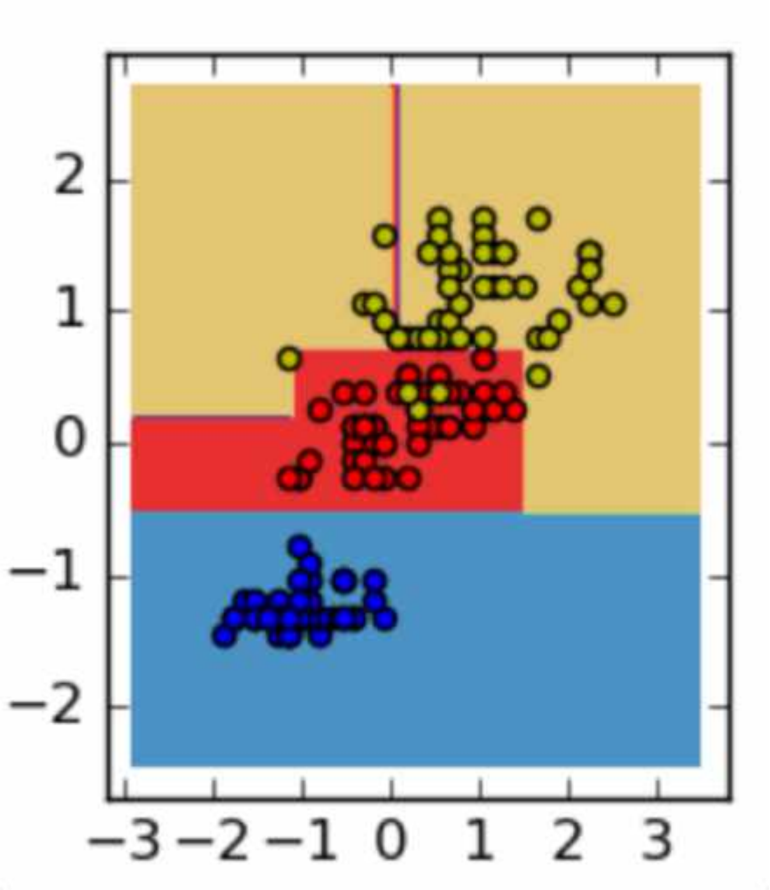
Если признаков и порогов больше



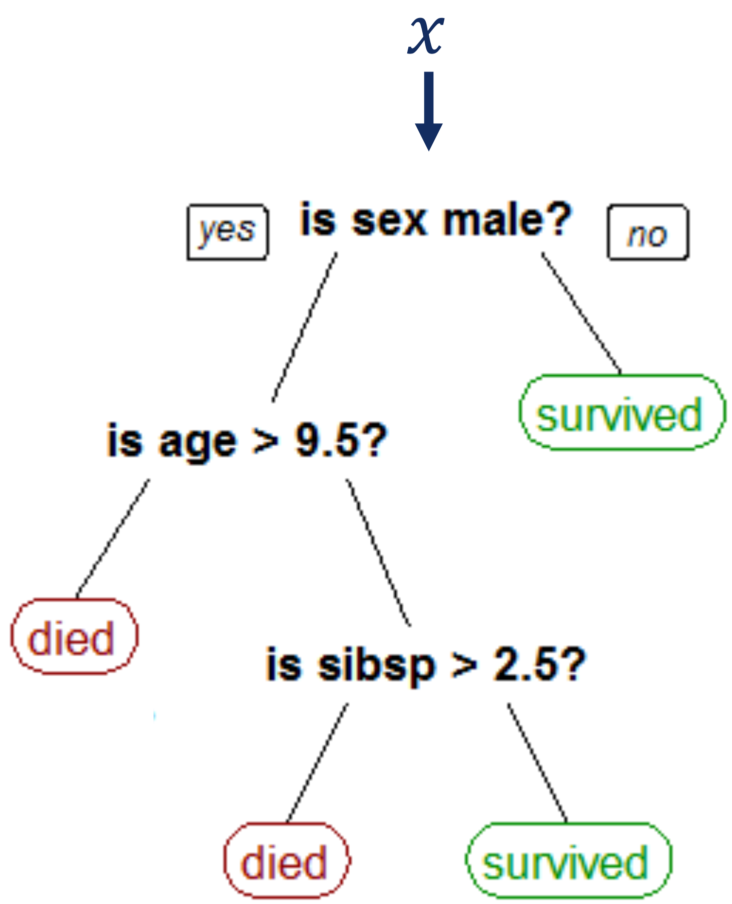
2. От пней к деревьям

Границы классов

Пример границ для 3 классов при 2 признаках:



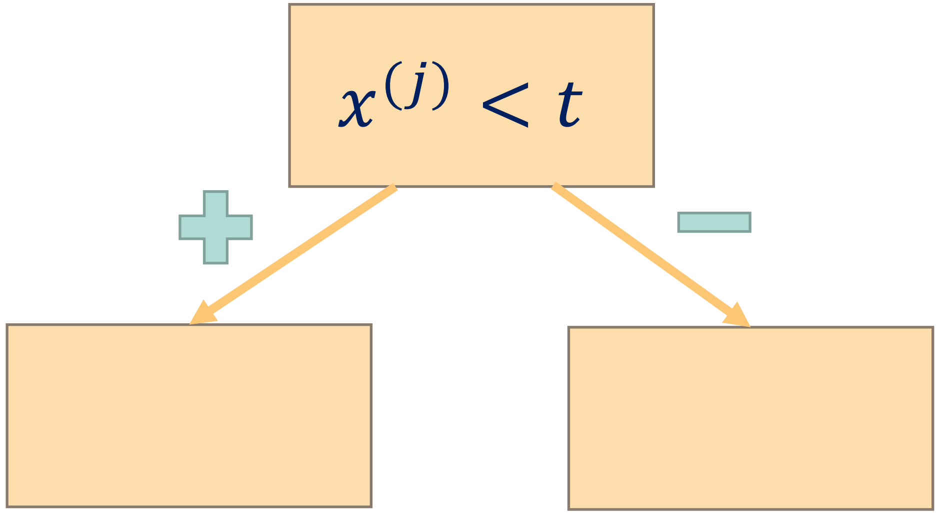
Решающее дерево



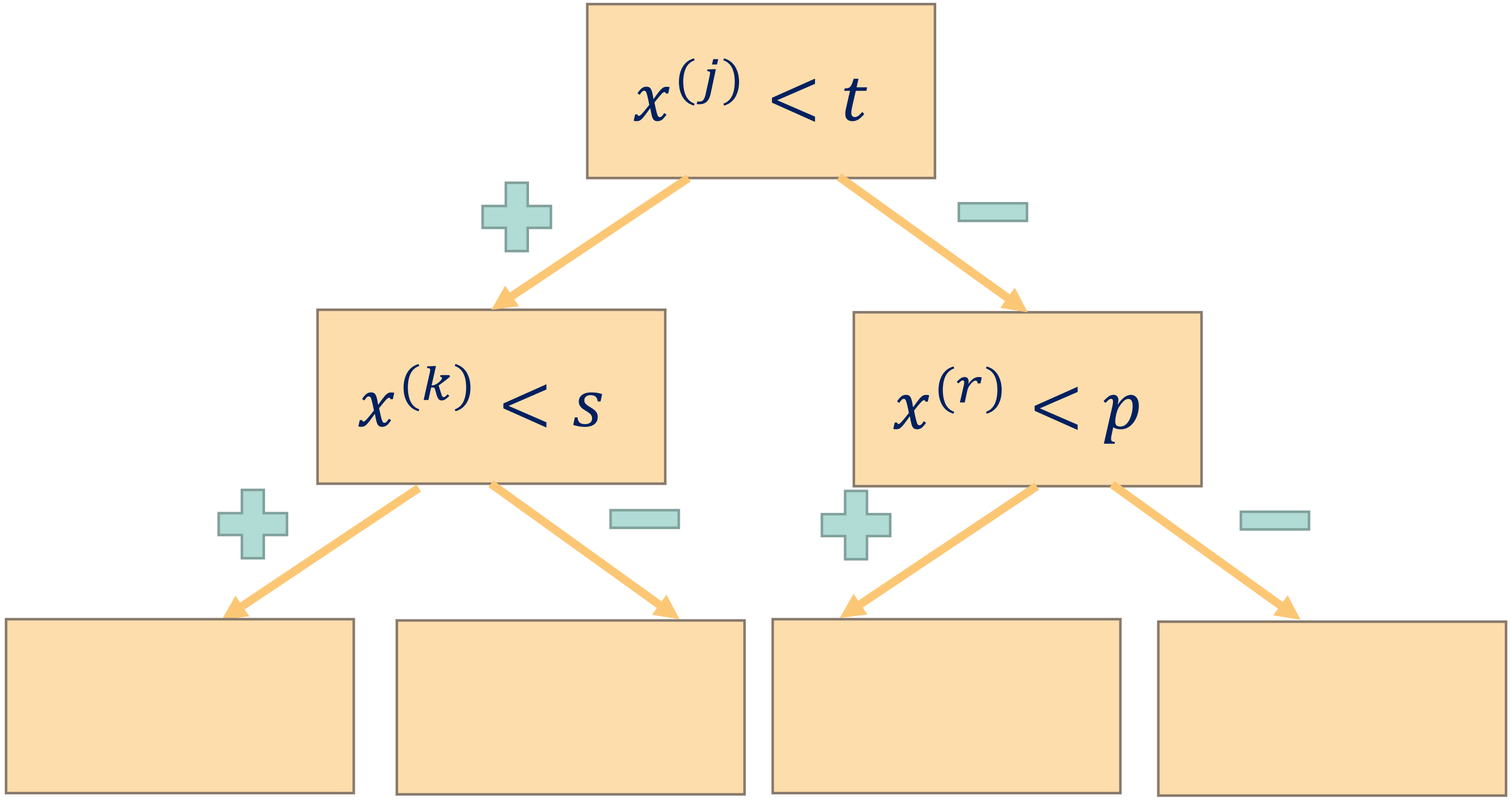
Рекурсивное построение

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

Рекурсивное построение

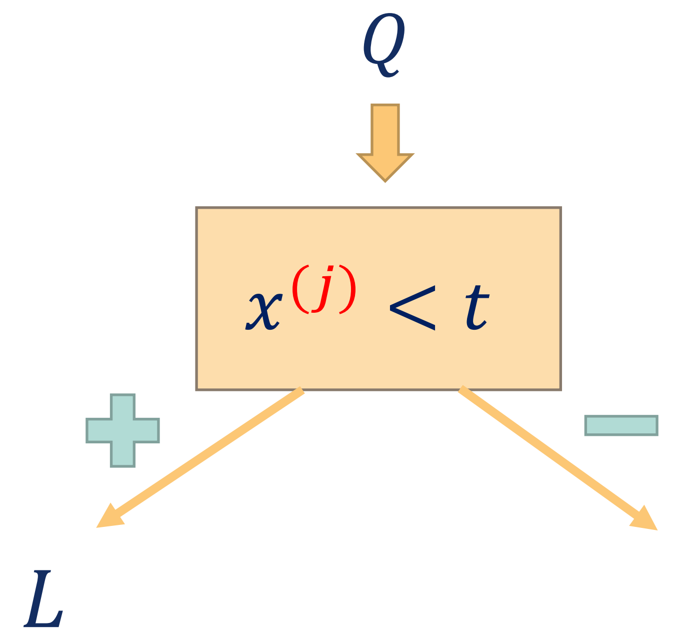


Рекурсивное построение



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

Выбор разбиения



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

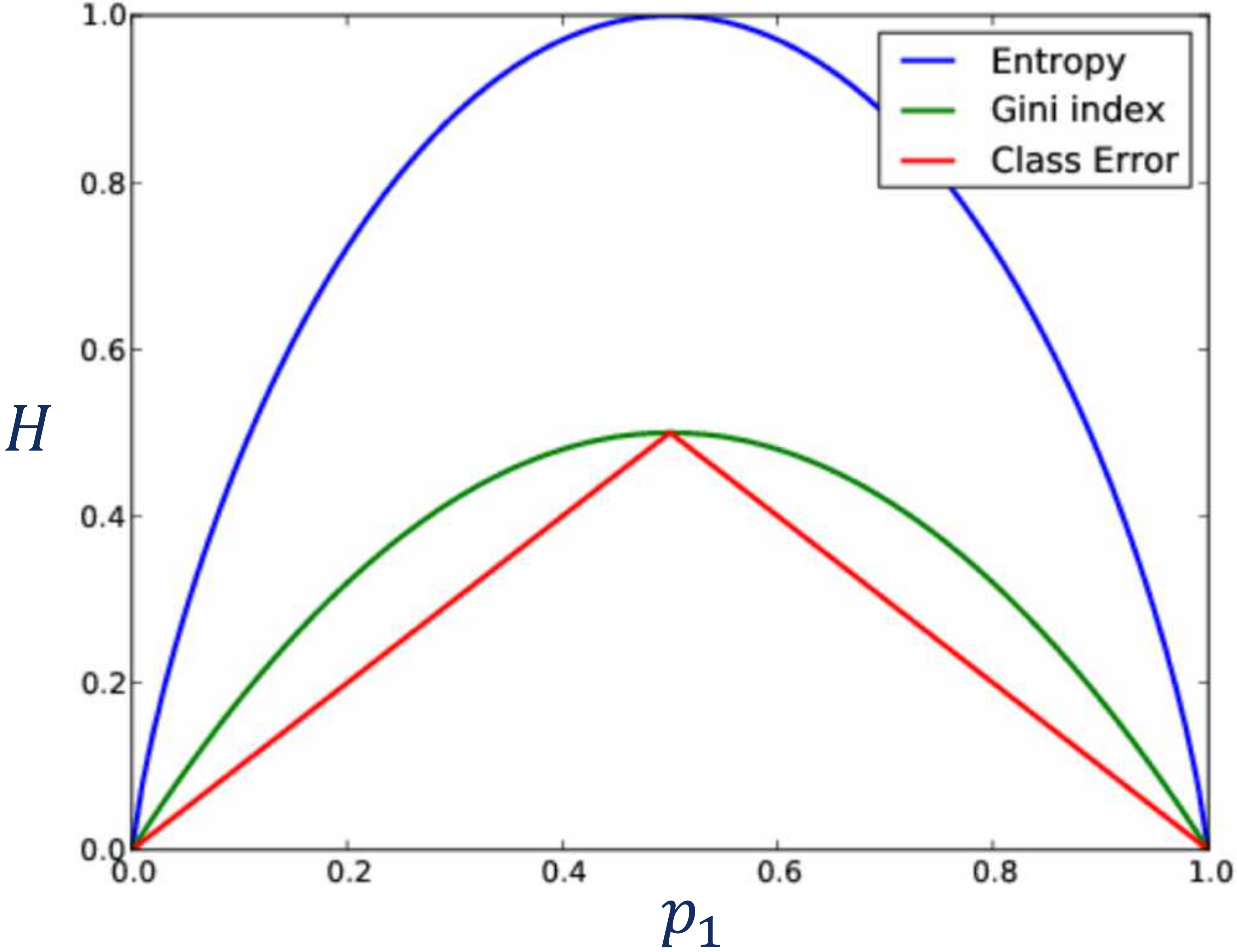
Критерии построения разбиений

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

Варианты этой функции:

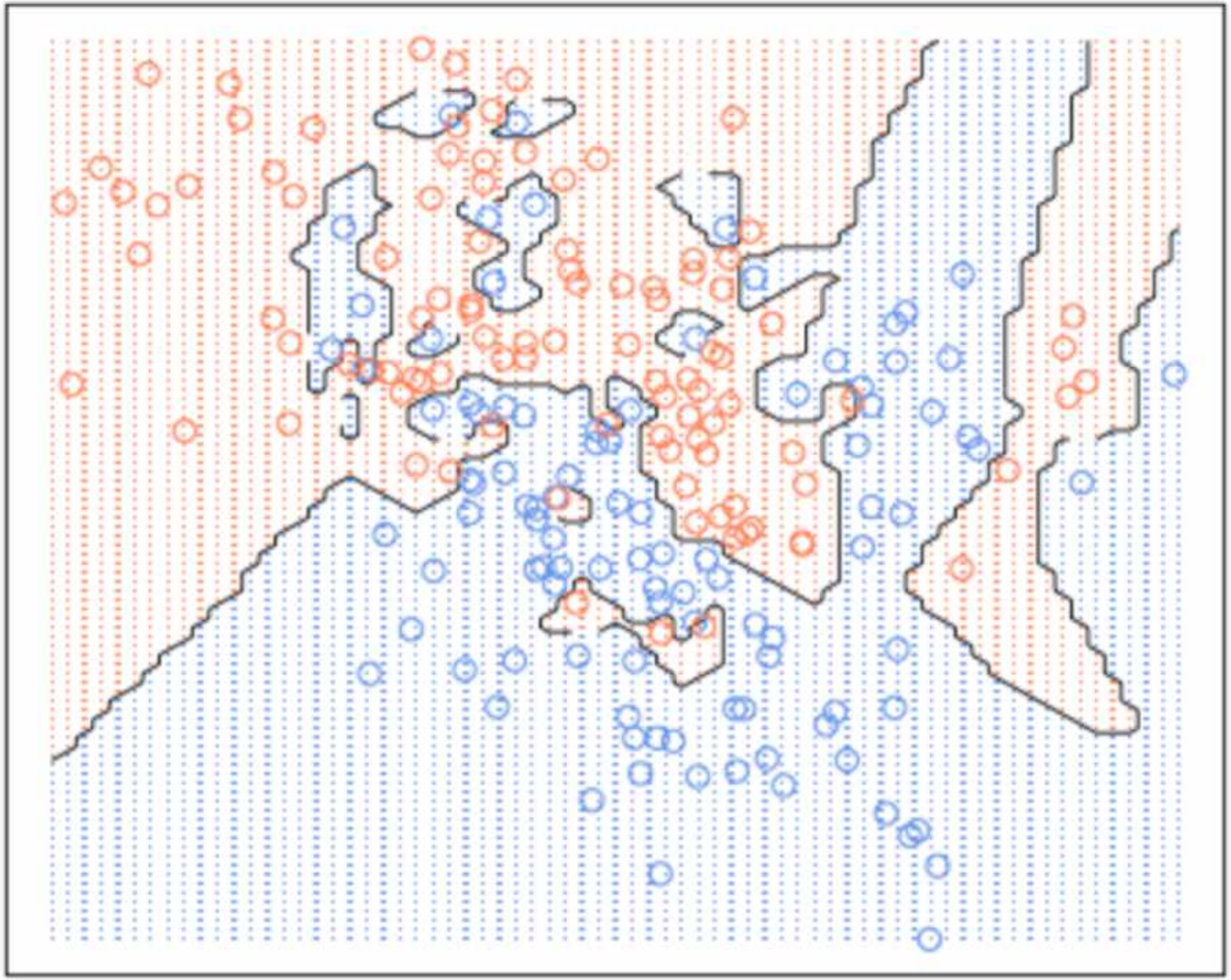
1) Misclassification criteria: $H(R) = 1 - \max\{p_0, p_1\}$
2) Entropy criteria: $H(R) = -p_0 \ln p_0 - p_1 \ln p_1$
3) Gini criteria: $H(R) = 1 - {p_0}^2 - {p_1}^2 = 2p_0 p_1$

Критерии построения разбиений

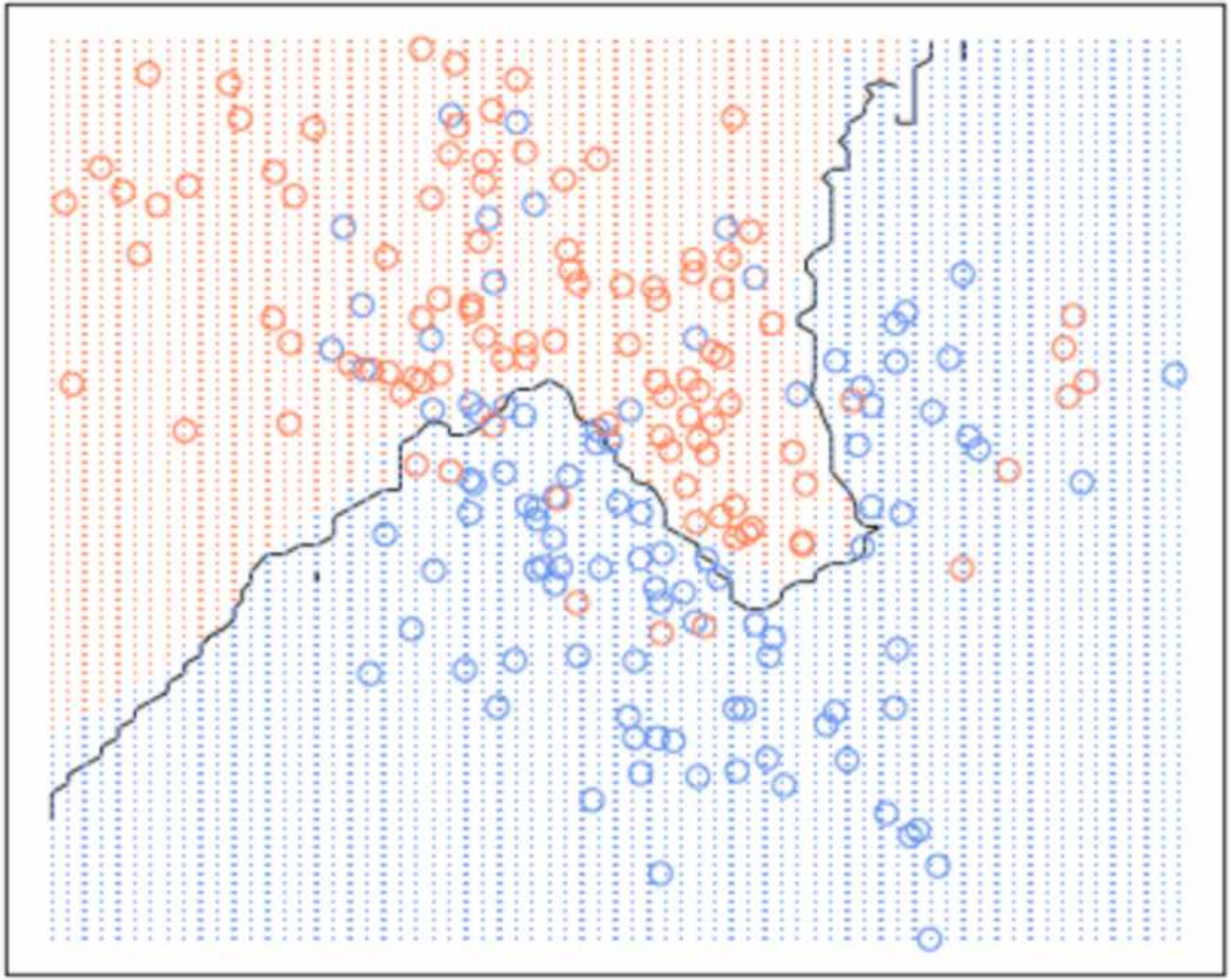


3. Сложные границы и соседи

Сложные границы



Сложные границы

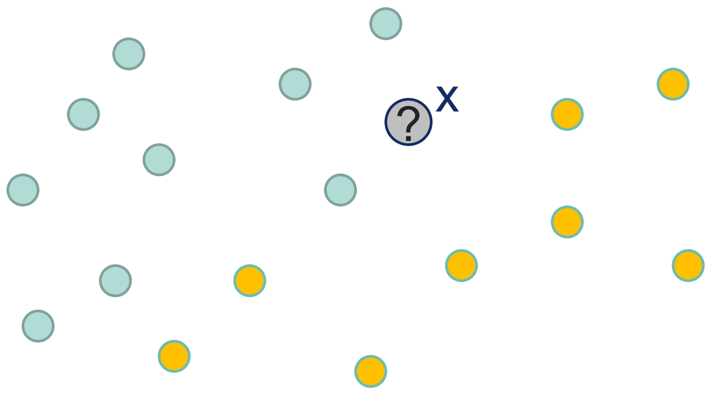


Сложные границы

Дереву такие границы строить тяжело: понадобится много разбиений

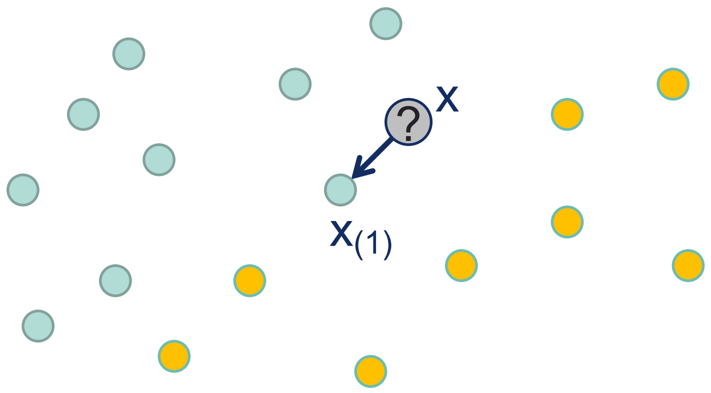
Метод ближайшего соседа

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?

$$x_2 - x_1 = (x_2^{(1)} - x_1^{(1)}, \dots, x_2^{(d)} - x_1^{(d)})$$

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

Что такое расстояние

Есть две точки в многомерном пространстве: x_1 и x_2

Как ввести расстояние между ними?

$$x_2 - x_1 = (x_2^{(1)} - x_1^{(1)}, \dots, x_2^{(d)} - x_1^{(d)})$$

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

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

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

Примеры норм

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

Примеры норм:

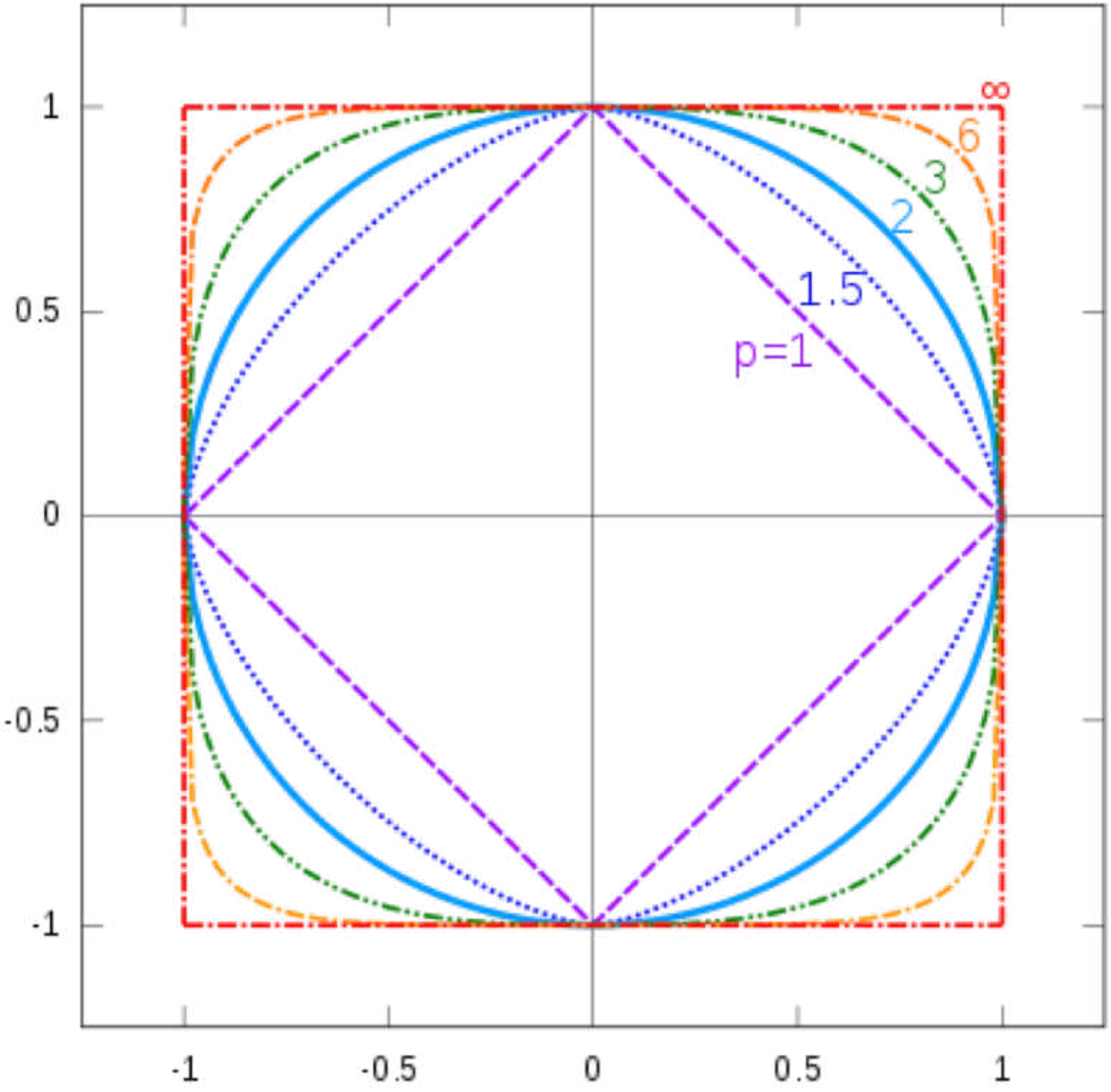
$$\|x\|_{\ell 2} = \sqrt{\left(x^{(1)}\right)^2 + \cdots + \left(x^{(d)}\right)^2}$$

$$\|x\|_{\ell 1} = \left|x^{(1)}\right| + \cdots + \left|x^{(d)}\right|$$

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

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

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)

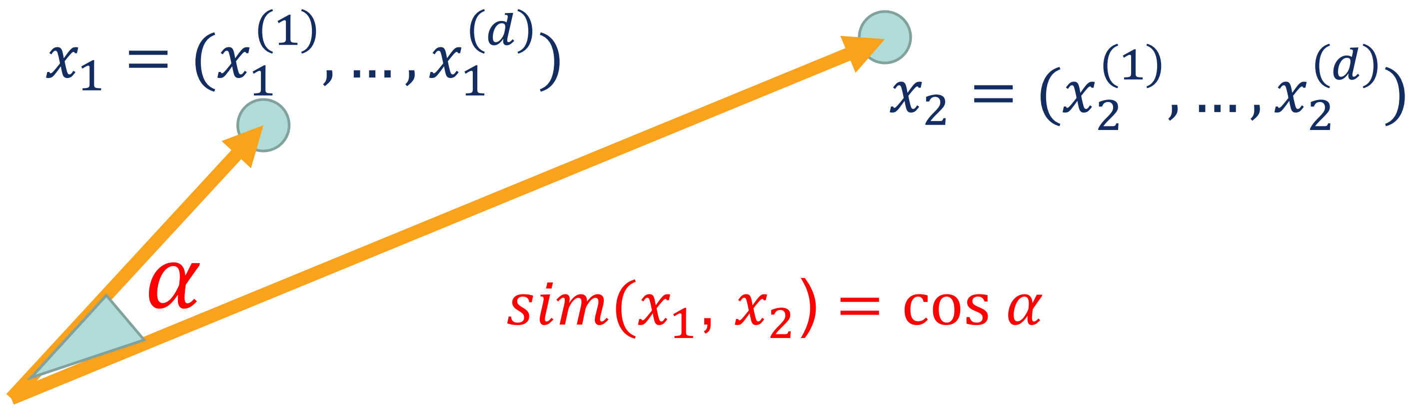
$$sim(x_1, x_2) = \frac{\langle x_1, x_2 \rangle}{\|x_1\| \cdot \|x_2\|} = \frac{x_1^{(1)} \cdot x_2^{(1)} + \ldots + x_1^{(d)} \cdot x_2^{(d)}}{\|x_1\| \cdot \|x_2\|}$$

Функция близости

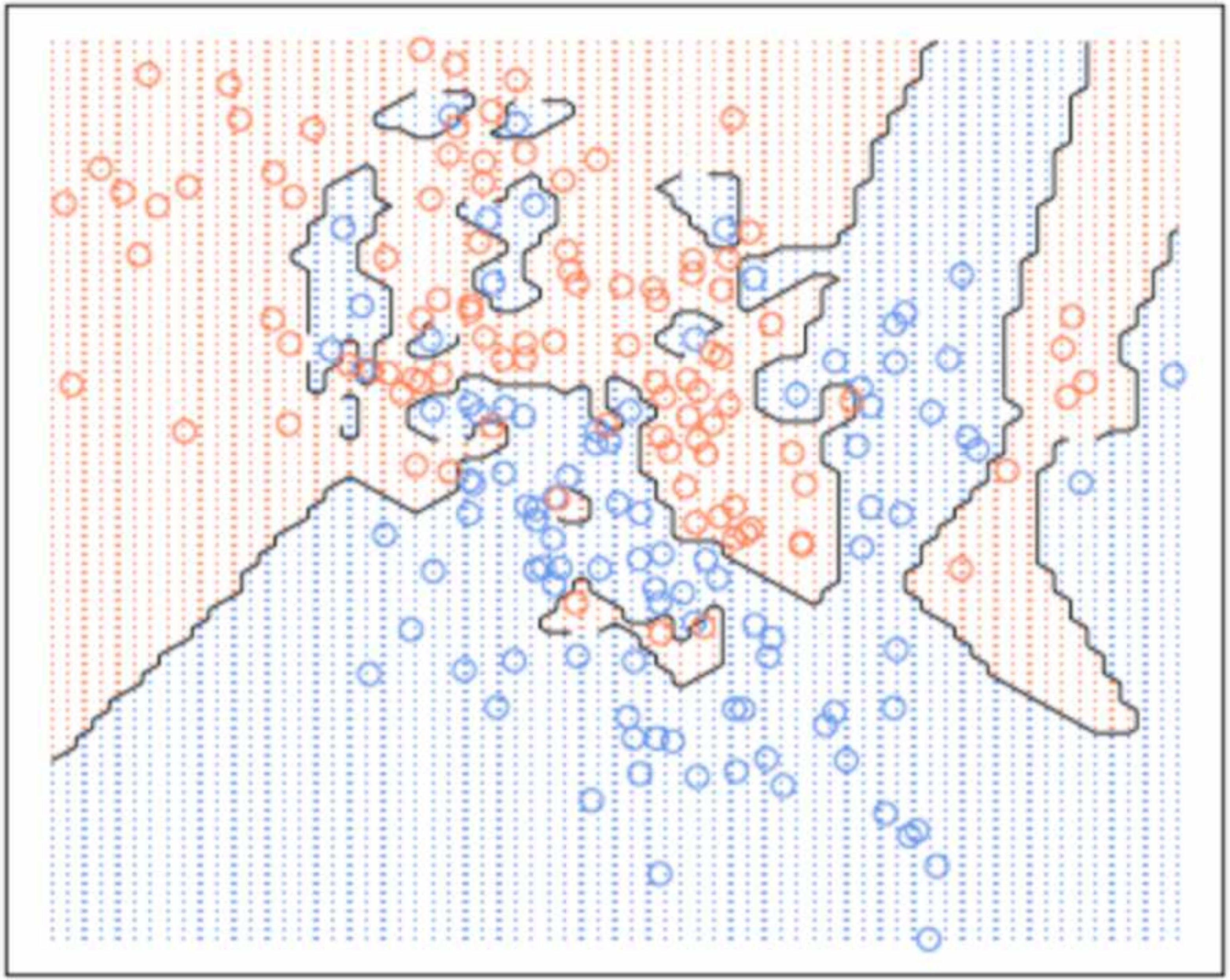
Но можно ввести расстояние каким-то своим особым способом или вообще ввести не расстояние, а **функцию близости**

Пример: Косинусная мера близости (cosine similarity)

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

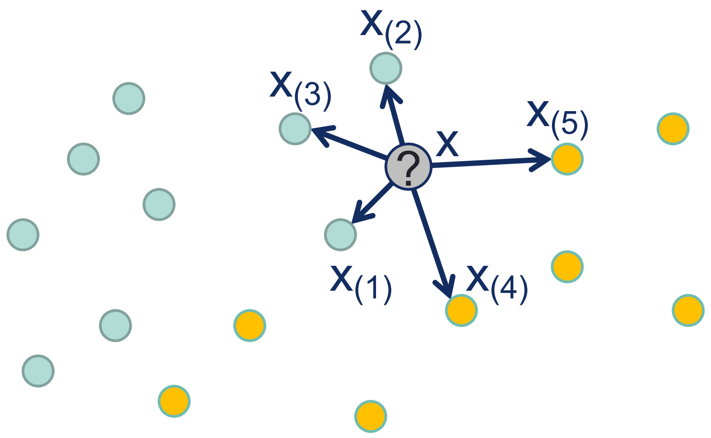


Границы классов в 1NN



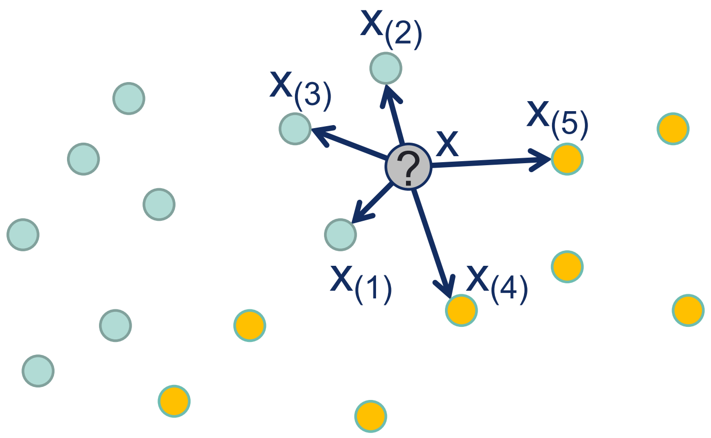
Метод k ближайших соседей (kNN)

Пример классификации (k = 5):



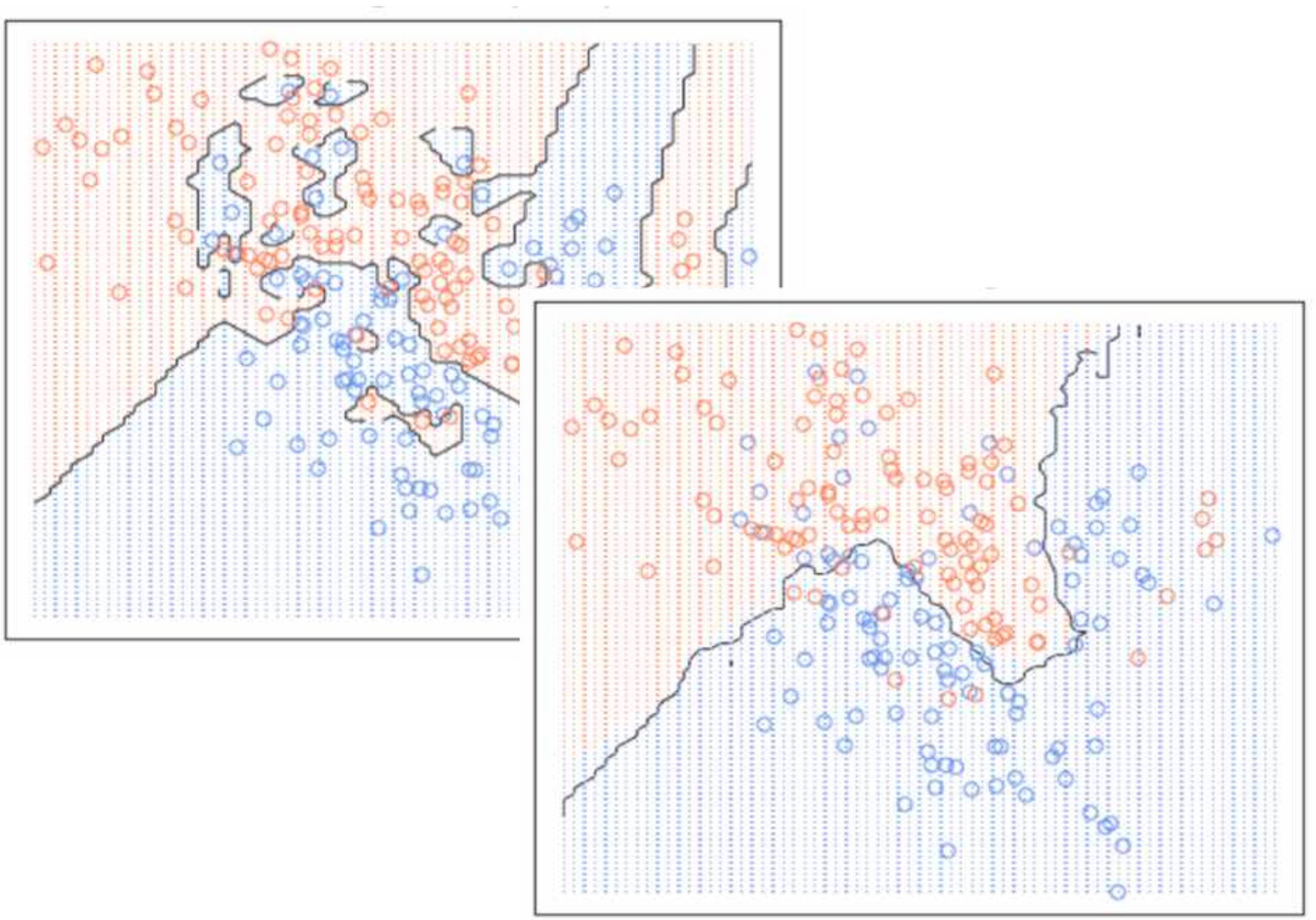
Метод k ближайших соседей (kNN)

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

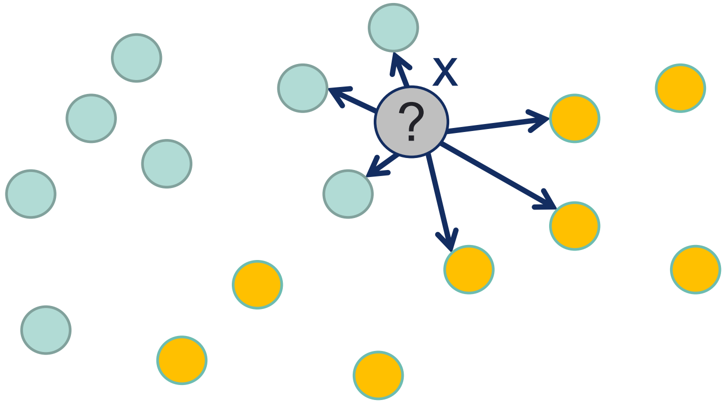
Какое количество соседей оптимально брать с точки зрения **качества работы на обучающей выборке?**

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

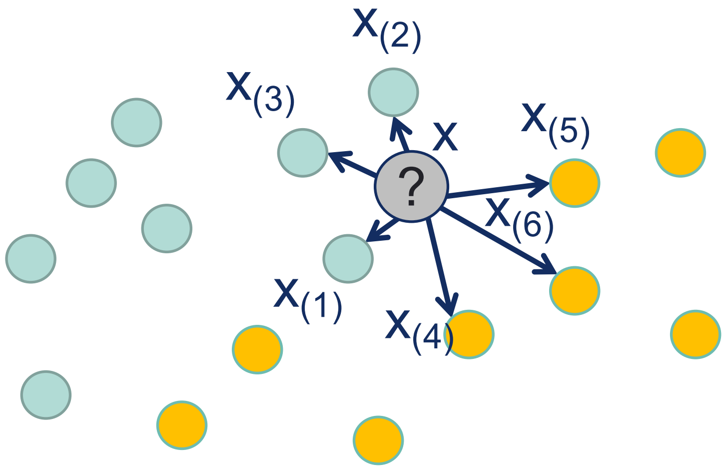
kNN с весами

Пример классификации (k = 6):



kNN с весами

Пример классификации (k = 6):



kNN с весами

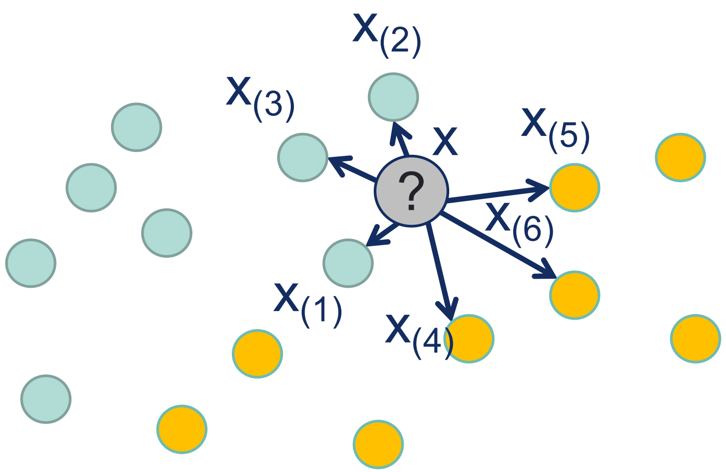
Пример классификации ($k = 6$):

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

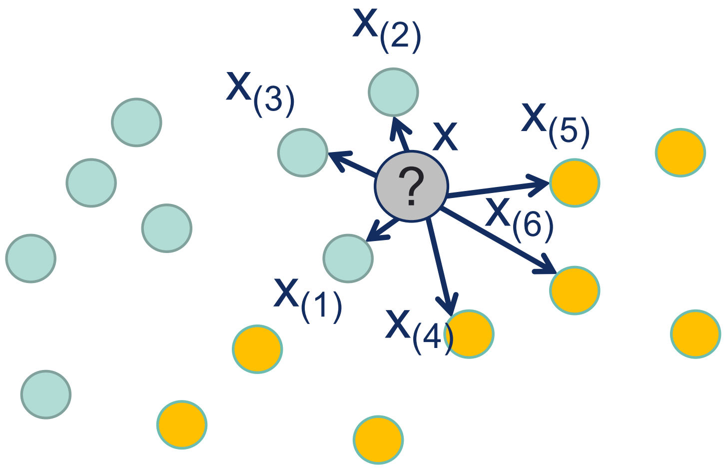
Пример классификации (k = 6):

Веса можно определить как функцию от соседа или его номера:

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

или как функцию расстояния:

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



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

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

kNN с весами

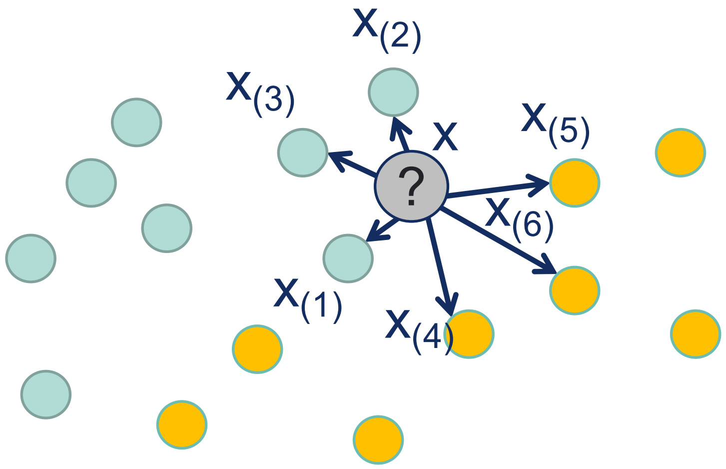
Пример классификации (k = 6):

Веса можно определить как функцию от соседа или его номера:

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

или как функцию расстояния:

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



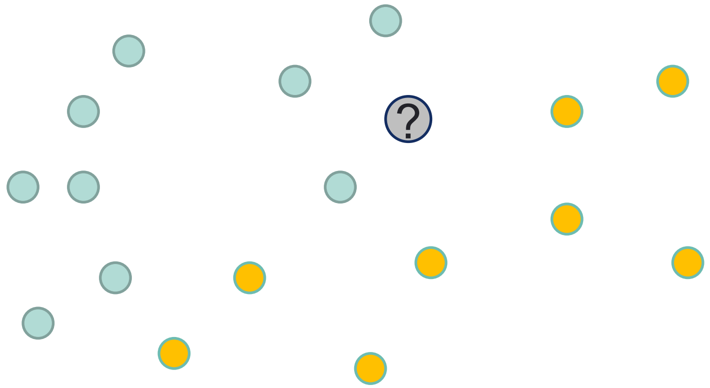
? = argmax Z

$$Z = \frac{\boxed{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$Z = \frac{\boxed{w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

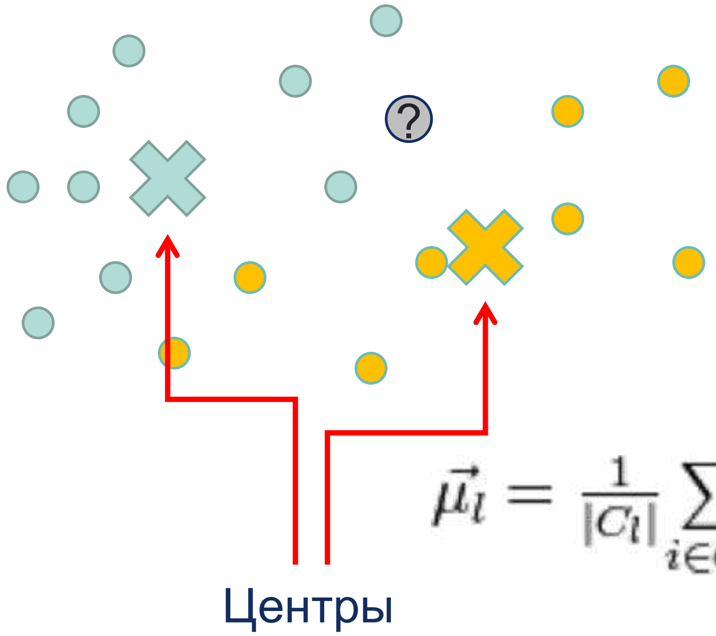
Другой похожий алгоритм

Центроидный классификатор



Центроидный классификатор

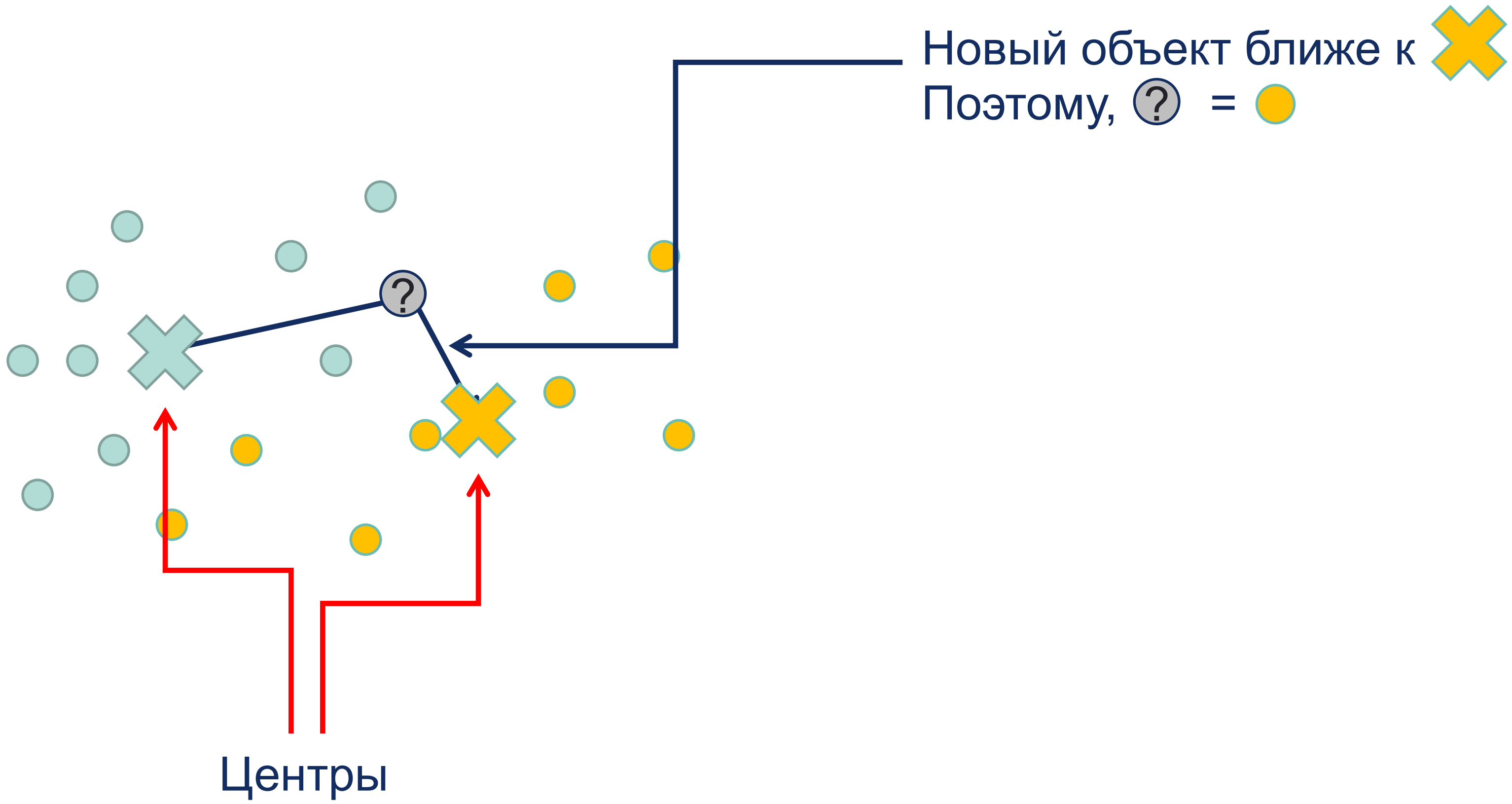
Другой похожий алгоритм



$$\vec{\mu_l} = \frac{1}{|C_l|}\sum_{i \in C_l} \vec{x}_i$$

Другой похожий алгоритм

Центроидный классификатор



Общая идея

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.

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

Общая идея

Метрические алгоритмы

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

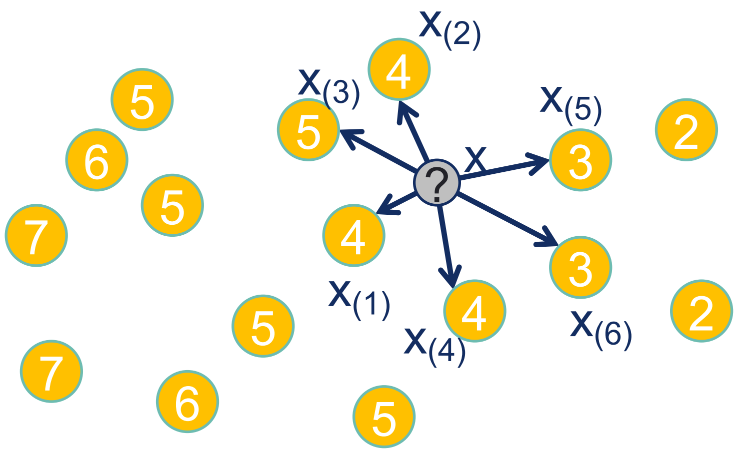
Можно брать расстояния до самих объектов класса, можно до центров класса.

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

Обобщение для регрессии

Все это применимо и в регрессии

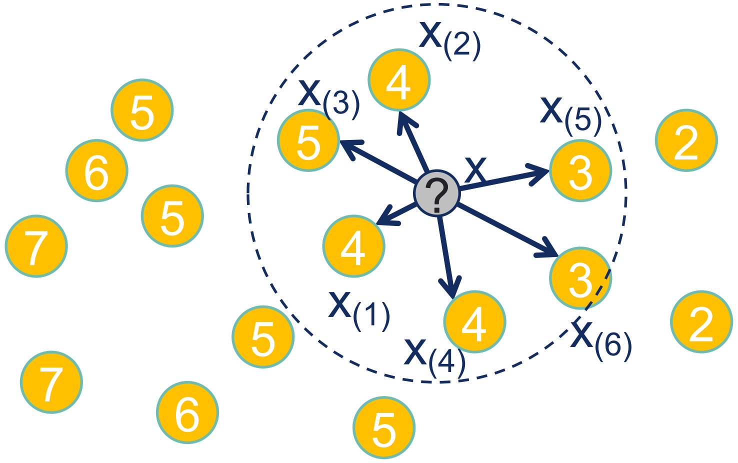
Пример взвешенного kNN ($k = 6$) в задаче регрессии:



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:

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

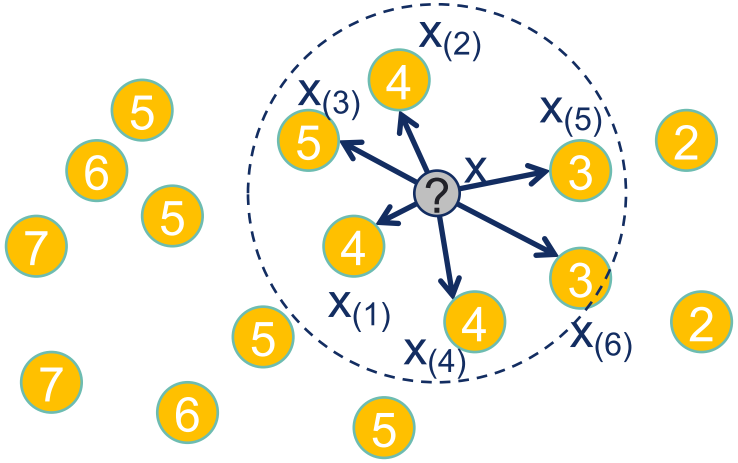
или как функцию расстояния:

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

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:

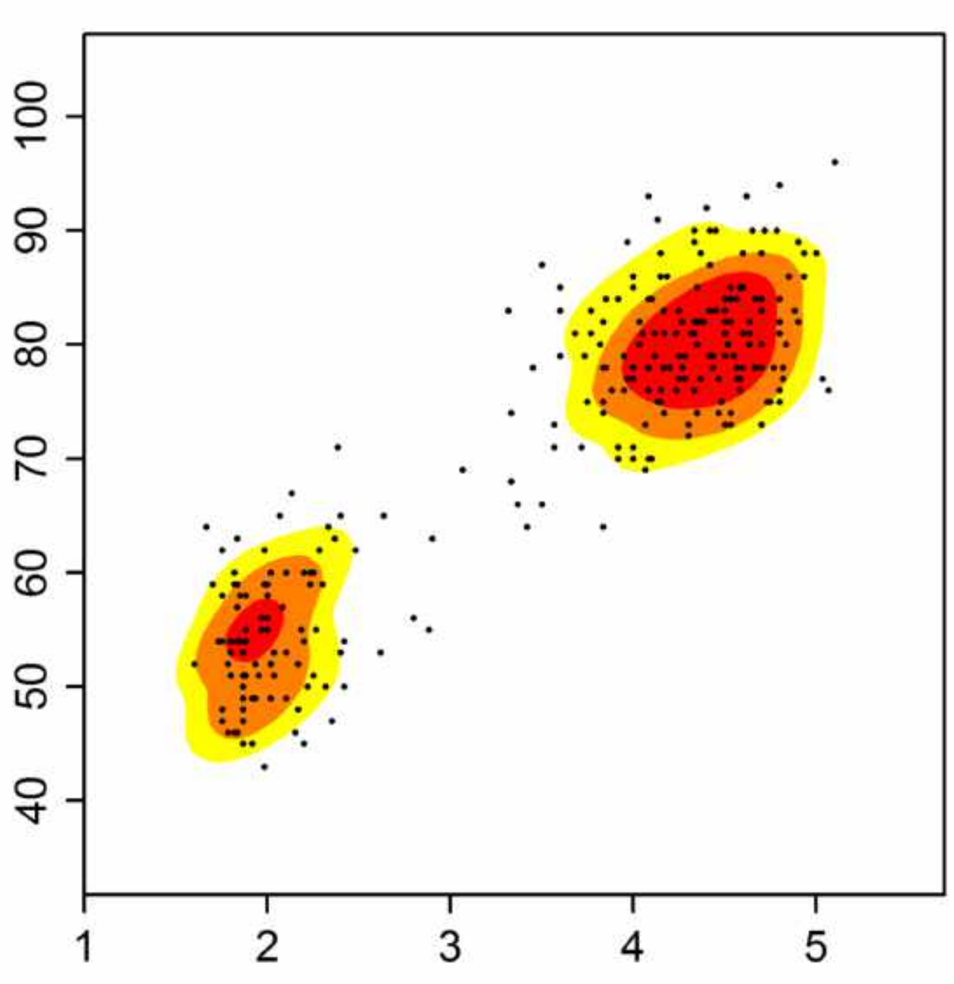
$$w(x(i)) = w(d(x, x_{(i)}))$$

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

4. Плотность и наивный байес

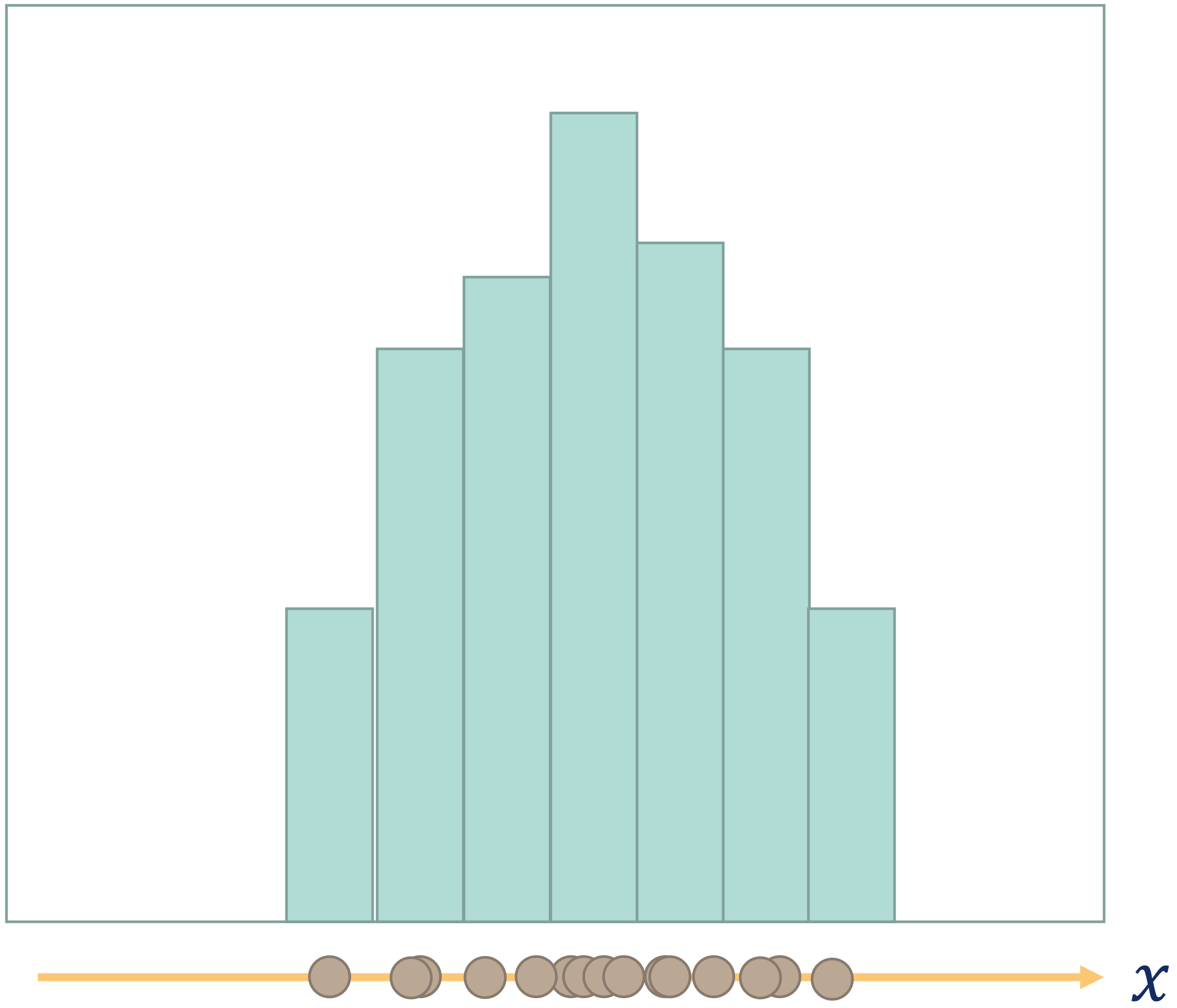
Плотность

Пример для двух признаков



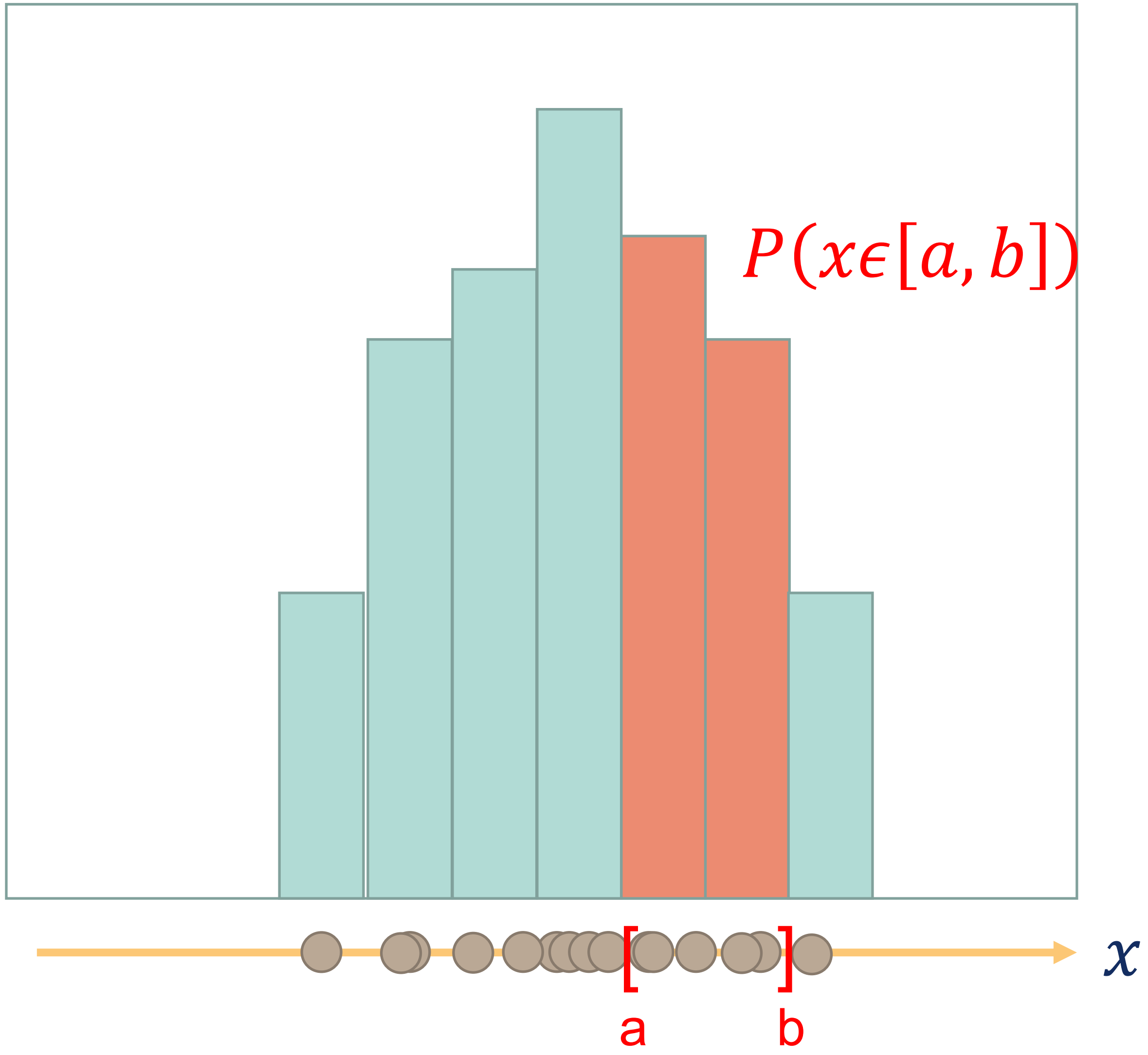
Плотность

Одномерный случай



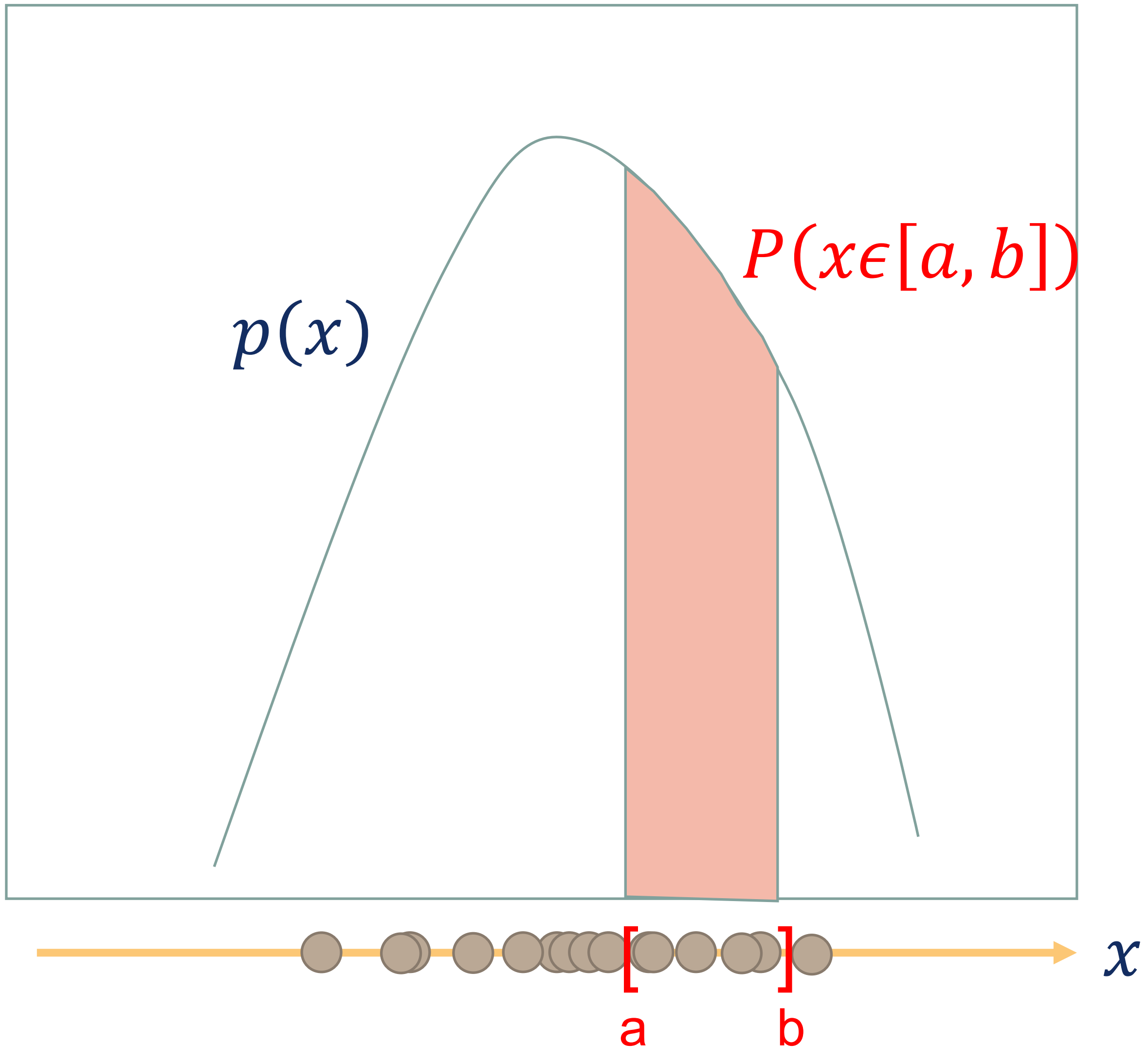
Одномерный случай

Плотность



Плотность

Плотность распределения



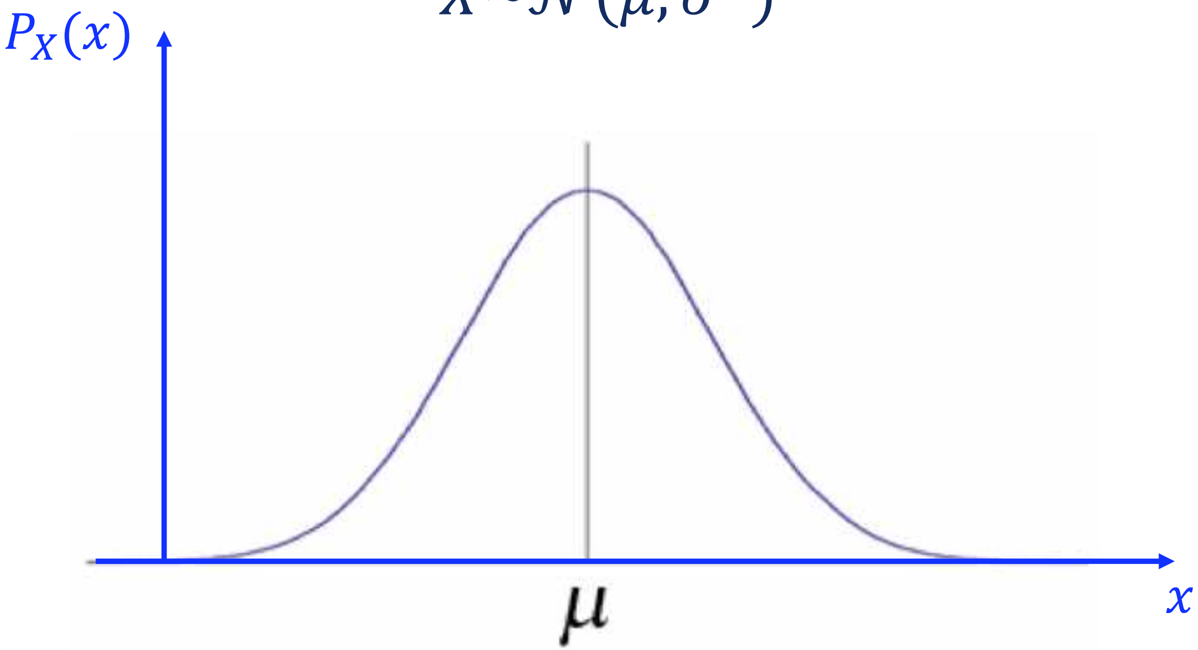
Подходы к оценке плотности

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

Пример оценки параметров

Нормальное распределение

$$X \sim \mathcal{N}(\mu, \sigma^2)$$



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

Пример оценки параметров

Нормальное распределение

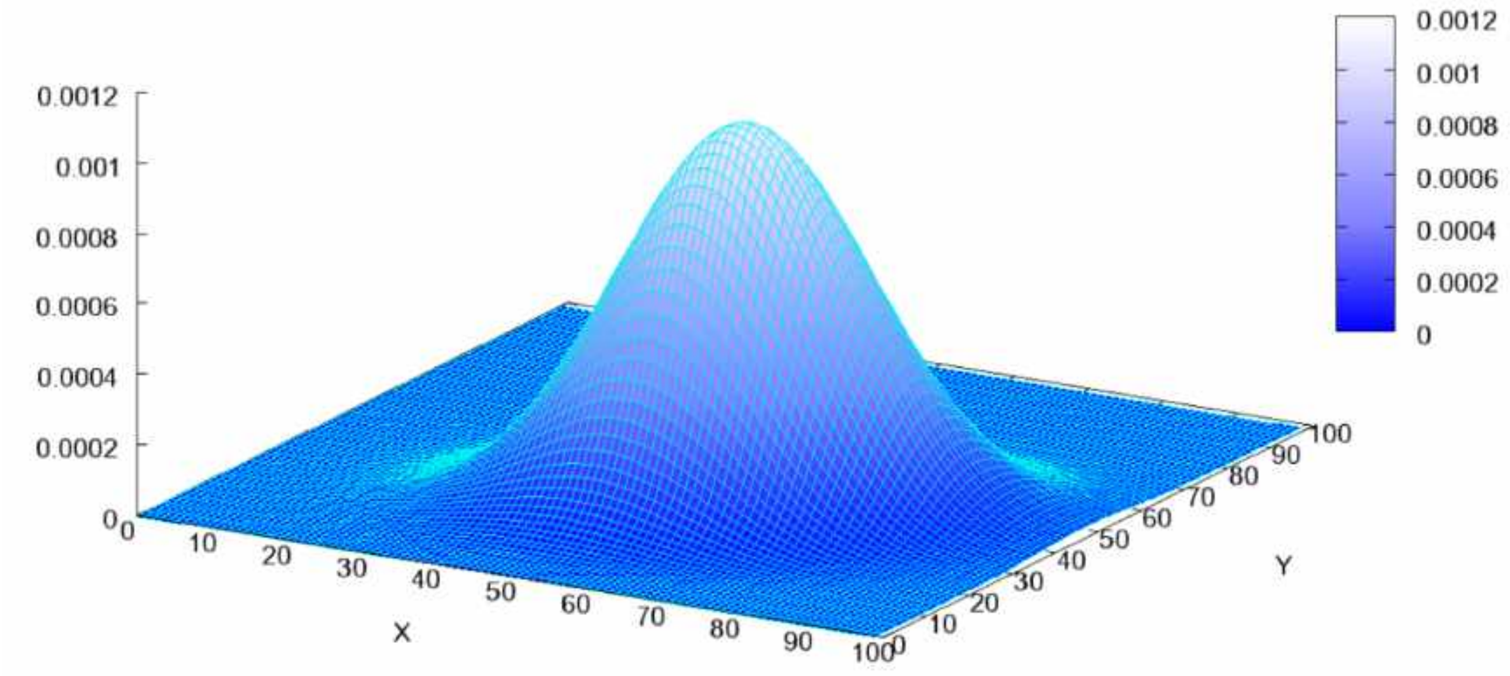
$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

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

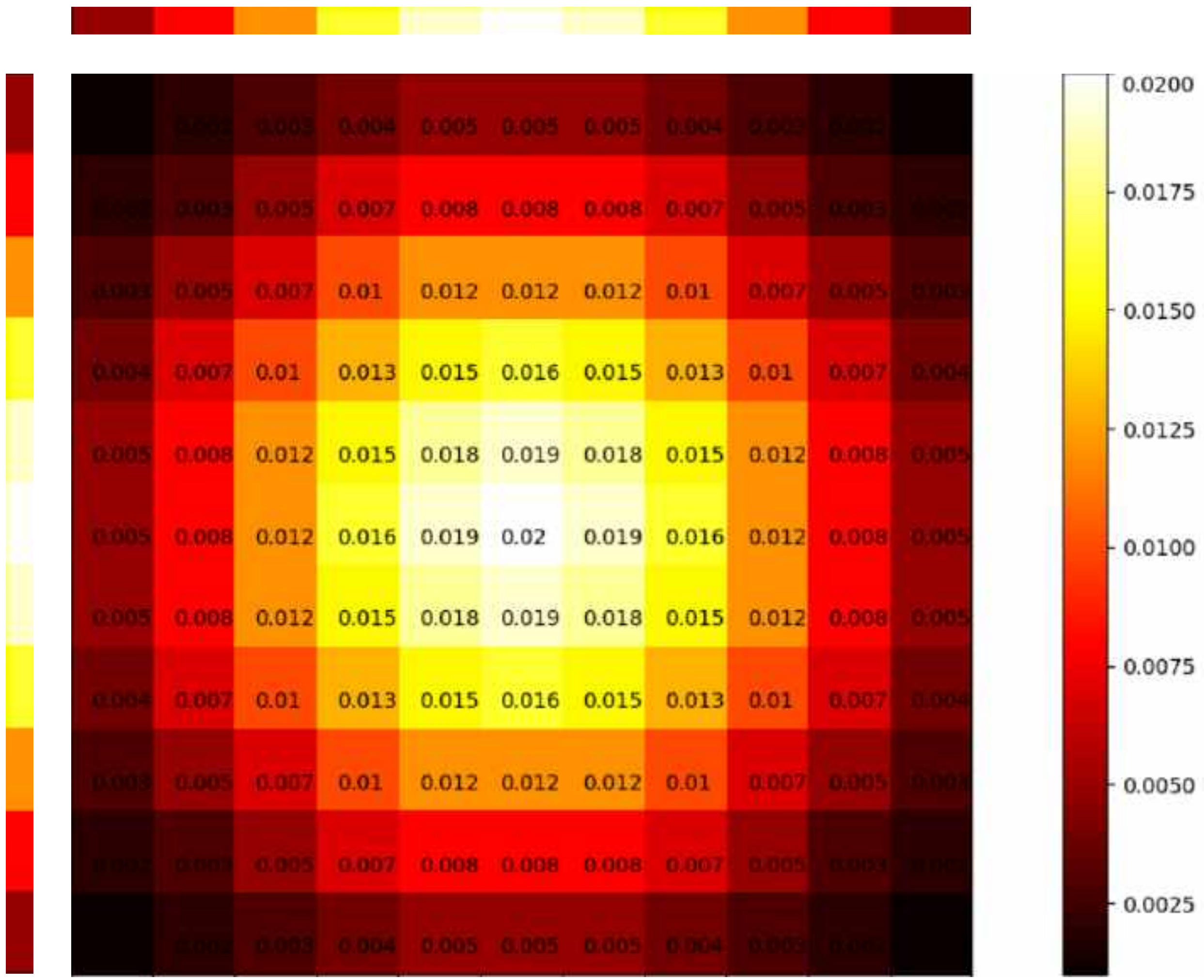


$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det\Sigma}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

Очень много параметров: вектор средних μ и матрица ковариаций Σ

Произведение одномерных плотностей

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



Наивная гипотеза

На самом деле так можно не всегда

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

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

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?

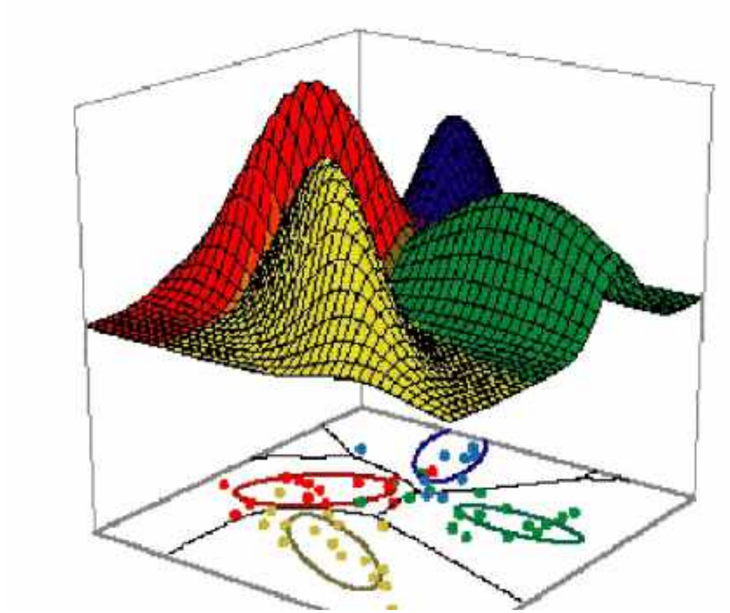
Ответ: $\frac{1}{2} \cdot \frac{7}{10}$

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:



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

Как решить задачу классификации

1. Считаем, что $p(x) = p\left(x^{(1)}\right) \dots p\left(x^{(d)}\right)$
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

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 $p(x|y) = p\left(x^{(1)}|y\right) \dots p\left(x^{(d)}|y\right)$

Обучение модели:

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

Применение модели:

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

План лекции

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

Приложение

Минимизация риска и байесовская классификация

Байесовский классификатор

По известному вектору признаков x алгоритм относит объект к классу a(x) по правилу:

$$a(x) = \mathrm{argmax}_y\, P(y|x)$$

Байесовский классификатор

$$a(x) = \mathrm{argmax}_y \, P(y|x)$$

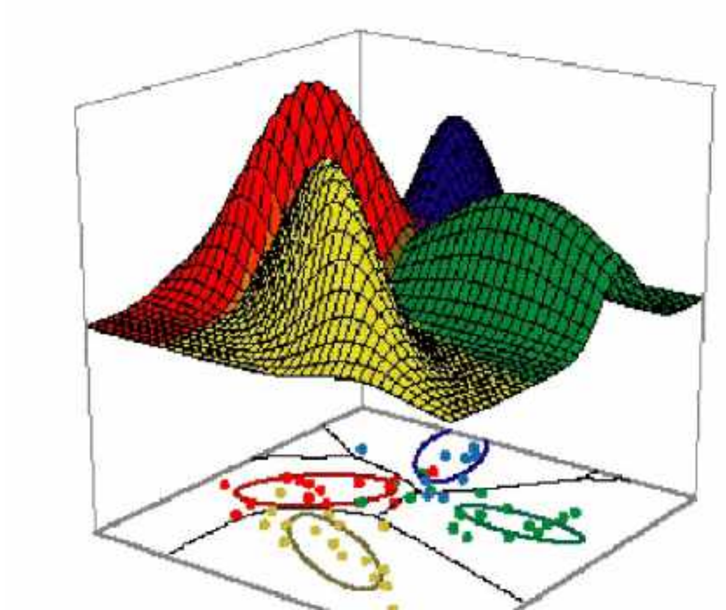
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$a(x) = \mathrm{argmax}_y \, P(x|y) \, P(y)$$

Байесовский классификатор

$$a(x) = \mathrm{argmax}_y\ P(x|y)\ P(y)$$

Если P(y) одинаковы для всех классов – мы просто выбираем класс, плотность которого больше в точке x



Зачем нам понадобилась теорема Байеса

- $P(y|x)$ – вероятность класса y при признаках x
- X часто из вещественных чисел и признаков часто очень много
- Всевозможных значений признаков так много, что скорее всего каждый вектор x встретится только один или несколько раз
- Этого недостаточно для оценки $P(y|x)$

Что оценивается по обучающей выборке

- $P(x|y)$ – вероятность увидеть набор признаков x в классе y, если x дискретный
- Если координаты вектора x – вещественные, $P(x|y)$ – плотность распределения x
- Именно эту величину и можно оценивать по обучающей выборке
- А затем подставлять в классификатор:

$$a(x) = \mathrm{argmax}_y\, P(x|y)\, P(y)$$

Проблема нехватки данных

- Пример: в обучающей выборке 100 000 объектов с 10 000 признаков
- 100 000 точек в пространстве размерности 10 000 – очень мало
- Например, если x – бинарный, то у него может быть 2^{10000} значений, что сильно больше 100 000
- Поэтому восстановить P(x|y) как функцию от признаков x довольно трудно

Байесовская регрессия

$a(x) = \operatorname{argmax}_y P(y)P(x|y)$?

x – признаковое описание
y – прогнозируемая величина

Байесовская регрессия

$a(\cancel{x) = \arg\max_y} P(y)P(x|y)$?

x – признаковое описание
y – прогнозируемая величина

Вряд ли получится восстановить P(x|y)

Байесовская регрессия

$\cancel{a(x) = \mathrm{argmax}_y P(y)P(x|y)}$?

x – признаковое описание
y – прогнозируемая величина

Вряд ли получится восстановить $P(x|y)$

$a(x) = \mathrm{argmax}_y P(y|x)$ тоже сомнительный вариант для регрессии

Штрафы за ошибки

- В классификации
  - Разные ошибки классификации могут быть в разной степени критичны
  - Пример: классификация мест, в которых может быть обнаружено месторождение нефти на классы «есть нефть» и «нет нефти».
  - Можем захотеть назначить разные штрафы за разные ошибки

Штрафы за ошибки

- В классификации
  - Разные ошибки классификации могут быть в разной степени критичны
  - Пример: классификация мест, в которых может быть обнаружено месторождение нефти на классы «есть нефть» и «нет нефти».
  - Можем захотеть назначить разные штрафы за разные ошибки
- В регрессии
  - Зависимость в любом случае восстанавливается неточно
  - Квадратичные потери:

$$MSE = \frac{1}{l}\sum_{i=1}^{l}\big(y_i - a(x_i)\big)^2$$

  - Сумма модулей отклонений:

$$MAE = \frac{1}{l}\sum_{i=1}^{l}|y_i - a(x_i)|$$

Более общий подход

- Для объекта x мы делаем прогноз a(x)
- Правильный ответ на этом объекте - y
- Величину ошибки алгоритма оцениваем как L(y, a(x)) (функцию выбираем сами)
- Пример функции L для задачи классификации:
$$L(y, a(x)) = [y \neq a(x)]$$
- Пример функции L для задачи регрессии:
$$L(y, a(x)) = (y - a(x))^2$$

Функционал риска

$R(a(x), x) = \mathbb{E}(L(y, a(x)) \mid x)$

Можно давать на объекте x ответ, который минимизирует ожидаемую ошибку:

$a(x) = \mathrm{argmin}_s R(s, x)$

Оптимальный байесовский классификатор

Для классификации:

$$R(a(x),x) = \mathbb{E}\big(L\big(y,a(x)\big)\big|x\big) = \sum_{y\in Y} L\big(y,a(x)\big)P(y|x)$$

$$\boxed{a(x)} = arg\min_s R(s,x) = arg\min_s \sum_{y\in Y} L(y,s)P(y|x) =$$

$$= \boxed{arg\min_s \sum_{y\in Y} L(y,s)P(y)P(x|y)}$$

Реальный классификатор в точности оптимальным не будет из-за погрешности в восстановлении плотностей

Оптимальный байесовский регрессор

Для регрессии:

$$R(a(x), x) = \mathbb{E}\big(L\big(y, a(x)\big)\big|x\big) = \int\limits_{y \in Y} L\big(y, a(x)\big) p(y|x) dy$$

$$\boxed{a(x)} = arg \min_s R(s, x) = \boxed{arg \min_s \int\limits_{y \in Y} L(y, s) p(y|x) dy}$$

Функционал среднего риска

- Можно рассмотреть $R(a) = \mathbb{E}_x R(a(x), x)$ (по всем x из X)
- Для определенности рассмотрим случай классификации объектов с дискретными признаками:

$$R(a) = \sum_{x \in X} R(a(x), x)\, P(x)$$

- Можно оценить R(a) снизу:

$$\sum_{x \in X} R(a(x), x)\, P(x) \geq \sum_{x \in X} P(x) \min_{s} R(s, x)$$

Функционал среднего риска

$$R(a) = \sum_{x \in X} R(a(x), x)\, P(x)$$

- Можно оценить R(a) снизу:

$$\sum_{x \in X} R(a(x), x)\, P(x) \geq \sum_{x \in X} P(x) \min_{s} R(s, x)$$

- Если a(x) – оптимальный байесовский, он минимизирует R(a(x), x)
- Значит оценка достигается и R(a) он тоже минимизирует

Оптимальный байесовский классификатор

$$a(x) = \underset{s}{argmin} \sum_{y \in Y} L(s, y) P(y) P(x|y)$$

Частный случай

Если L(s, y) = [y != s]:

$$\sum_{y \in Y} L(s, y)\, P(y|x) \to min$$

Частный случай

Если L(s, y) = [y != s]:

$$\sum_{y \in Y} L(s, y)\, P(y|x) \to min$$

$$\sum_{y \in Y \setminus \{s\}} P(y|x) = \left(\sum_{y \in Y} P(y|x) \right) - P(s|x) \to min$$

Частный случай

Если L(s, y) = [y != s]:

$$\sum_{y \in Y} L(s, y)\, P(y|x) \to \min_{s}$$

$$\sum_{y \in Y \setminus \{s\}} P(y|x) = \left(\sum_{y \in Y} P(y|x) \right) - P(s|x) \to min$$

$$P(s|x) \to max$$

Частный случай

$$a(x) = arg\min_{y} P(y|x) = arg\min_{y} P(y)P(x|y)$$

Квадратичная функция потерь в регрессии

$$\int_Y (t-y)^2 p(y|x)\,dy \to \min_t$$

Квадратичная функция потерь в регрессии

$$\int_Y (t-y)^2 p(y|x)dy \to \min_t$$

$$\frac{\partial}{\partial t}\int_Y (t-y)^2 p(y|x)dy = 2\int_Y (t-y)p(y|x)dy =$$

Квадратичная функция потерь в регрессии

$$\int_Y (t-y)^2 p(y|x)dy \to \min_t$$

$$\frac{\partial}{\partial t}\int_Y (t-y)^2 p(y|x)dy = 2\int_Y (t-y)p(y|x)dy =$$

$$= 2\left(\int_Y tp(y|x)dy - \int_Y yp(y|x)dy\right) = 0$$

Квадратичная функция потерь в регрессии

$$\int_Y (t-y)^2 p(y|x)dy \to \min_t$$

$$\frac{\partial}{\partial t}\int_Y (t-y)^2 p(y|x)dy = 2\int_Y (t-y)p(y|x)dy =$$

$$= 2\left(\int_Y tp(y|x)dy - \int_Y yp(y|x)dy\right) = 0$$

$$a(x) = t = \int_Y yp(y|x)dy = \mathsf{E}(y|x)$$

Абсолютное отклонение

$$\int_Y |t - y| p(y|x) dy \to \min_t$$

Абсолютное отклонение

$$\int_Y |t - y| p(y|x) dy \to \min_t$$

$$\frac{\partial}{\partial t} \int_Y |t - y| p(y|x) dy = \frac{\partial}{\partial t} \int_{Y \setminus \{t\}} |t - y| p(y|x) dy =$$

Абсолютное отклонение

$$\int_Y |t - y| p(y|x) dy \to \min_t$$

$$\frac{\partial}{\partial t} \int_Y |t - y| p(y|x) dy = \frac{\partial}{\partial t} \int_{Y \setminus \{t\}} |t - y| p(y|x) dy =$$

$$= \int_{Y \setminus \{t\}} \operatorname{sign}(t - y) p(y|x) dy = \int_{\{t>y\}} p(y|x) dy - \int_{\{t<y\}} p(y|x) dy =$$

Абсолютное отклонение

$$\int_Y |t - y| p(y|x) dy \to \min_t$$

$$\frac{\partial}{\partial t} \int_Y |t - y| p(y|x) dy = \frac{\partial}{\partial t} \int_{Y \setminus \{t\}} |t - y| p(y|x) dy =$$

$$= \int_{Y \setminus \{t\}} \operatorname{sign}(t - y) p(y|x) dy = \int_{\{t>y\}} p(y|x) dy - \int_{\{t<y\}} p(y|x) dy =$$

$$= \mathsf{P}(\{t > y\}|x) - \mathsf{P}(\{t < y\}|x) = 0.$$

Абсолютное отклонение

$$\int_Y |a(x) - y| p(y|x) dy$$

$$\frac{\partial}{\partial t} \int_Y |t - y| p(y|x) dy = \frac{\partial}{\partial t} \int_{Y \setminus \{t\}} |t - y| p(y|x) dy =$$

$$= \int_{Y \setminus \{t\}} \operatorname{sign}(t - y) p(y|x) dy = \int_{\{t>y\}} p(y|x) dy - \int_{\{t<y\}} p(y|x) dy =$$

$$= \mathsf{P}(\{t > y\}|x) - \mathsf{P}(\{t < y\}|x) = 0.$$

$$\mathsf{P}(\{t = y\}|x) = 0$$

Абсолютное отклонение

$$\int_Y |a(x) - y| p(y|x) dy$$

$$\frac{\partial}{\partial t} \int_Y |t - y| p(y|x) dy = \frac{\partial}{\partial t} \int_{Y \setminus \{t\}} |t - y| p(y|x) dy =$$

$$= \int_{Y \setminus \{t\}} \operatorname{sign}(t - y) p(y|x) dy = \int_{\{t>y\}} p(y|x) dy - \int_{\{t<y\}} p(y|x) dy =$$

$$= \mathsf{P}(\{t > y\}|x) - \mathsf{P}(\{t < y\}|x) = 0.$$

$$\mathsf{P}(\{t = y\}|x) = 0 \implies P(\{t \leq y\}|x) = P(\{t > y\}|x) = \frac{1}{2}$$

Оценка вероятностей

$Y = \{0; 1\}$ $\qquad$ $L\big(y, a(x)\big) = -y \ln a(x) - (1 - y) \ln(1 - a(x))$

Оценка вероятностей

$Y = \{0; 1\}$

$$L\big(y, a(x)\big) = -y \ln a(x) - (1 - y) \ln(1 - a(x))$$

$$\sum_{y \in Y} (-y \ln t - (1 - y) \ln(1 - t))\, P(y|x) \to \min_t$$

Оценка вероятностей

$Y = \{0; 1\}$

$$L\big(y, a(x)\big) = -y \ln a(x) - (1 - y) \ln(1 - a(x))$$

$$\sum_{y \in Y} (-y \ln t - (1 - y) \ln(1 - t)) \, P(y|x) \to \min_t$$

$\mathrm{P}(1|x) = p$

Оценка вероятностей

$Y = \{0; 1\}$

$$L\big(y, a(x)\big) = -y \ln a(x) - (1 - y) \ln(1 - a(x))$$

$$\sum_{y \in Y} (-y \ln t - (1 - y) \ln(1 - t))\, P(y|x) \to \min_t$$

$$\mathrm{P}(1|x) = p$$

$$-(1 - p) \ln(1 - t) - p \ln t \to \min_t$$

Оценка вероятностей

$Y = \{0; 1\}$

$$L\big(y, a(x)\big) = -y \ln a(x) - (1-y)\ln(1-a(x))$$

$$\sum_{y \in Y} (-y \ln t - (1-y)\ln(1-t))\, P(y|x) \to \min_t$$

$\mathsf{P}(1|x) = p$

$$-(1-p)\ln(1-t) - p\ln t \to \min_t$$

$$\frac{\partial}{\partial t}(-(1-p)\ln(1-t) - p\ln t) = \frac{1-p}{1-t} - \frac{p}{t} =$$

Оценка вероятностей

$Y = \{0; 1\}$

$$L(y, a(x)) = -y \ln a(x) - (1 - y) \ln(1 - a(x))$$

$$\sum_{y \in Y} (-y \ln t - (1 - y) \ln(1 - t))\, P(y|x) \to \min_t$$

$\mathrm{P}(1|x) = p$

$$-(1 - p) \ln(1 - t) - p \ln t \to \min_t$$

$$\frac{\partial}{\partial t}(-(1 - p) \ln(1 - t) - p \ln t) = \frac{1 - p}{1 - t} - \frac{p}{t} =$$

$$= \frac{(1 - p)t - p(1 - t)}{(1 - t)t} = \frac{t - p}{(1 - t)t} = 0$$

Оценка вероятностей

$Y = \{0; 1\}$

$$L\big(y, a(x)\big) = -y \ln a(x) - (1-y)\ln(1 - a(x))$$

$$\sum_{y \in Y} (-y \ln t - (1-y)\ln(1-t))\, P(y|x) \to \min_t$$

$\mathrm{P}(1|x) = p$

$$-(1-p)\ln(1-t) - p \ln t \to \min_t$$

$$\frac{\partial}{\partial t}(-(1-p)\ln(1-t) - p\ln t) = \frac{1-p}{1-t} - \frac{p}{t} =$$

$$= \frac{(1-p)t - p(1-t)}{(1-t)t} = \frac{t-p}{(1-t)t} = 0 \quad \Longrightarrow t = p$$

Почему это все работает

- Средний риск:

$$R(a) = \mathbb{E}_{x,y} L(y, a(x))$$

Почему это все работает не только в байесовском классификаторе

- Средний риск:

$$R(a) = \mathbb{E}_{x,y} L(y, a(x))$$

- Ошибка на обучающей выборке:

$$Q = \frac{1}{l} \sum_{i=1}^{l} L(y_i, a(x_i)) \approx \mathbb{E}_{x,y} L(y, a(x))$$

Резюме

- Принцип минимизации функционала среднего риска
- Анализ функций потерь
- Квадратичная функция потерь – для оценки матожидания
- Абсолютные отклонения – для оценки ½ квантили
- Log loss – для оценки вероятностей
- Понимание неудачного выбора функции потерь

Data Mining in Action

Лекция 1

Группа курса в Telegram:

https://t.me/joinchat/B1OITk74nRV56Dp1TDJGNA