

Машинное обучение: простые алгоритмы обучения с учителем

MADE academy

Эмели Драль

Базовые концепции машинного обучения

1. Виды обучения, виды задач, базовые концепции
2. **Простые алгоритмы: логика построения и связь с математикой**
3. Оценка качества в машинном обучении

Простые методы обучения с учителем

1. Логический подход
2. Метрический подход
3. Вероятностный подход

Логический подход

Простейшая выборка

Рассмотрим выборку объектов с одним признаком x :



Как подобрать порог по признаку в задаче бинарной классификации?

Логический ПОДХОД

Простейшая выборка

Рассмотрим выборку объектов с одним признаком x :

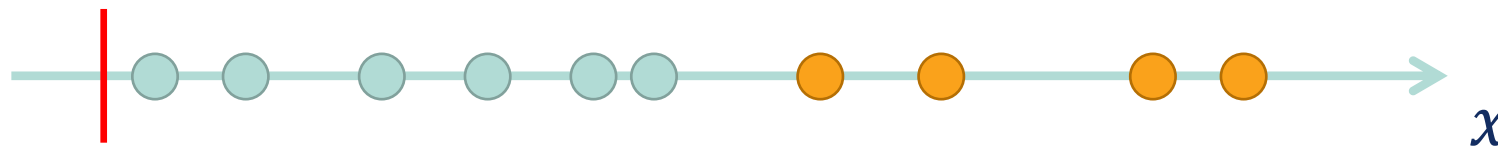


Как подобрать порог по признаку в задаче бинарной классификации?

Можно перебрать пороги, например, сдвигая на один пример.

Простейшая выборка

Рассмотрим выборку объектов с одним признаком x :



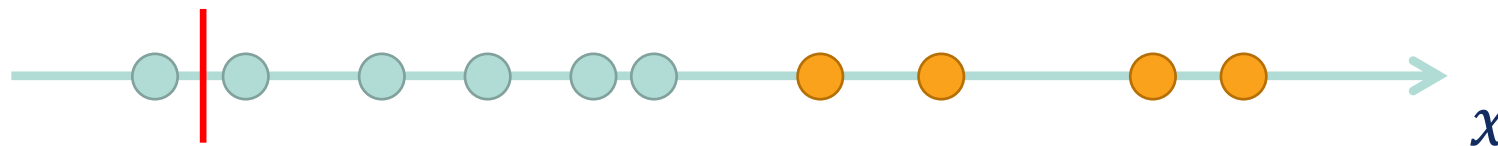
Логический
ПОДХОД

Как подобрать порог по признаку в задаче бинарной классификации?

Можно перебрать пороги, например, сдвигая на один пример.

Простейшая выборка

Рассмотрим выборку объектов с одним признаком x :



Как подобрать порог по признаку в задаче бинарной классификации?

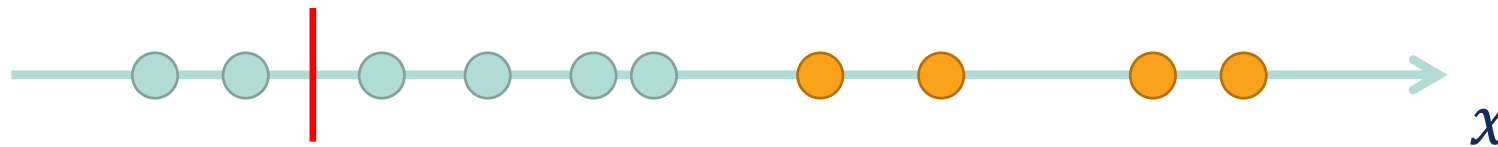
Можно перебрать пороги, например, сдвигая на один пример.

Логический
ПОДХОД

Логический ПОДХОД

Простейшая выборка

Рассмотрим выборку объектов с одним признаком x :



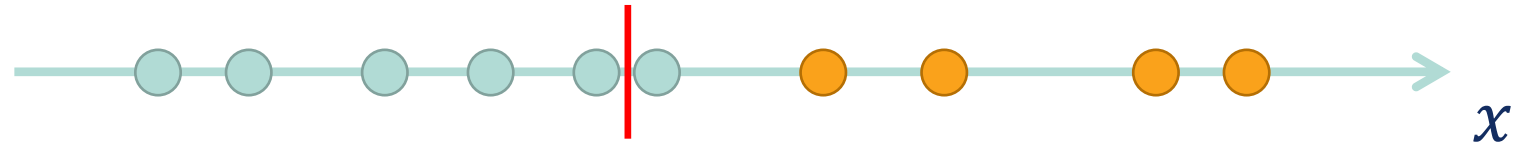
Как подобрать порог по признаку в задаче бинарной классификации?

Можно перебрать пороги, например, сдвигая на один пример.

Логический ПОДХОД

Простейшая выборка

Рассмотрим выборку объектов с одним признаком x :



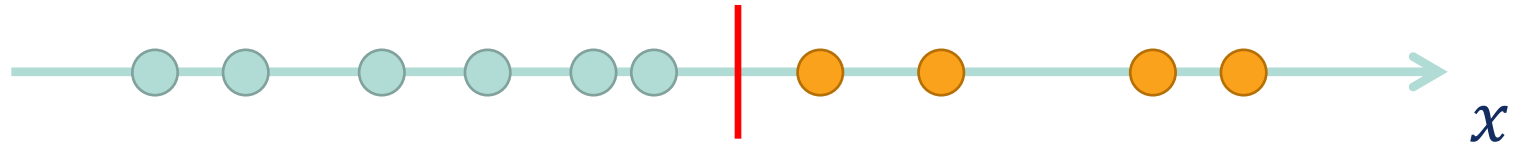
Как подобрать порог по признаку в задаче бинарной классификации?

Можно перебрать пороги, например, сдвигая на один пример.

Логический ПОДХОД

Простейшая выборка

Рассмотрим выборку объектов с одним признаком x :



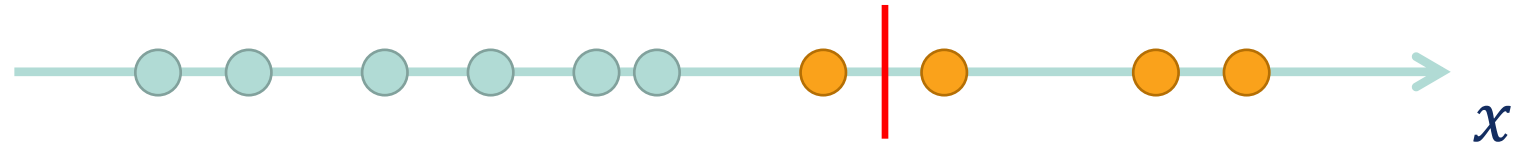
Как подобрать порог по признаку в задаче бинарной классификации?

Можно перебрать пороги, например, сдвигая на один пример.

Логический ПОДХОД

Простейшая выборка

Рассмотрим выборку объектов с одним признаком x :

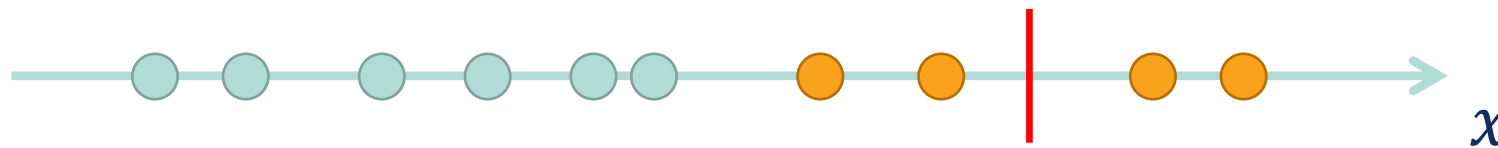


Как подобрать порог по признаку в задаче бинарной классификации?

Можно перебрать пороги, например, сдвигая на один пример.

Простейшая выборка

Рассмотрим выборку объектов с одним признаком x :



Как подобрать порог по признаку в задаче бинарной классификации?

Можно перебрать пороги, например, сдвигая на один пример.

Логический
ПОДХОД

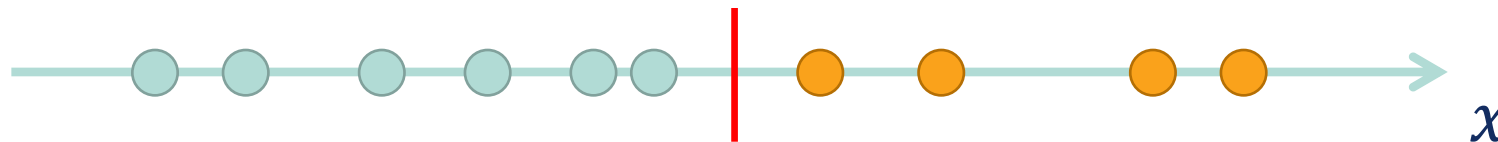
Простейшая выборка

Рассмотрим выборку объектов с одним признаком x :



Как подобрать порог по признаку в задаче бинарной классификации?

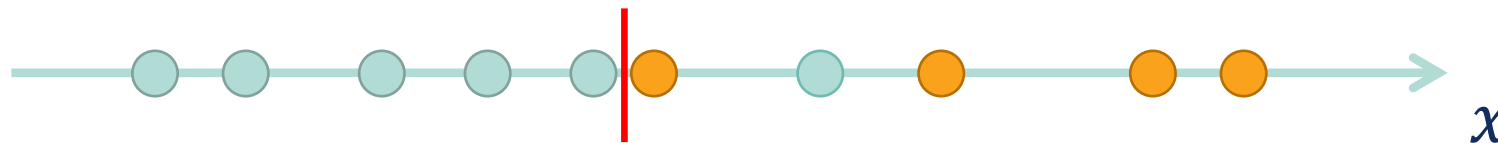
Если выборка разделима, оптимальный порог - между последним объектом одного класса и первым объектом:



Логический
ПОДХОД

Простейшая выборка

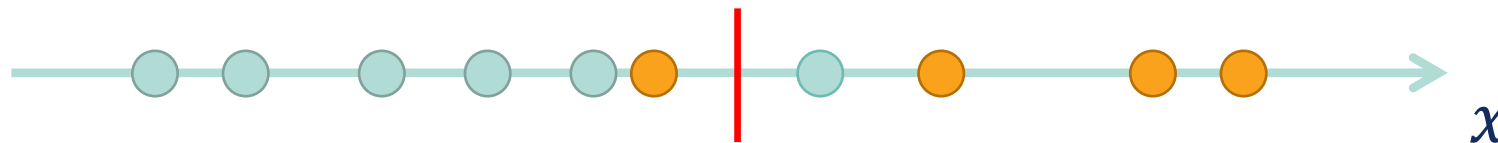
Часто выборка не разделима и есть несколько неплохих порогов:



Логический
подход

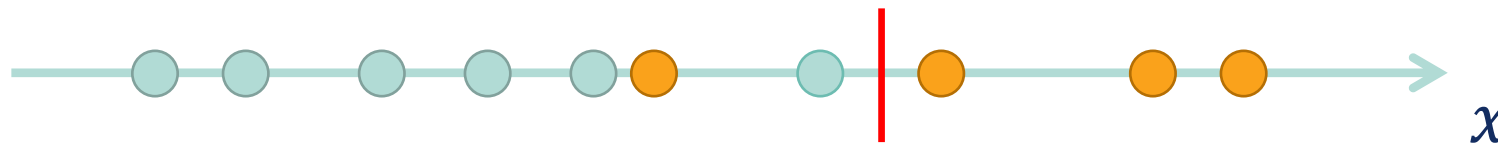
Простейшая выборка

Часто выборка не разделима и есть несколько неплохих порогов:



Простейшая выборка

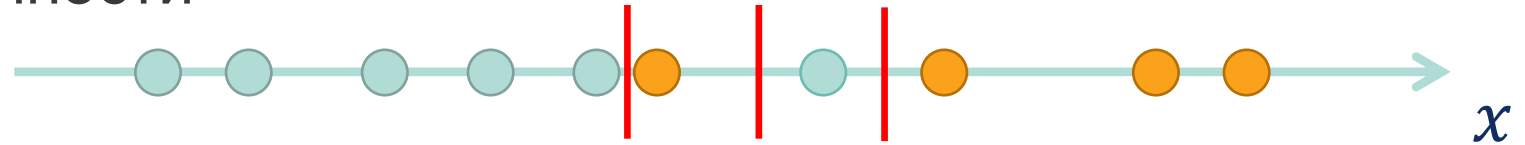
Часто выборка не разделима и есть несколько неплохих порогов:



Логический подход

Простейшая выборка

Вариант 1: потребовать от модели максимальной точности

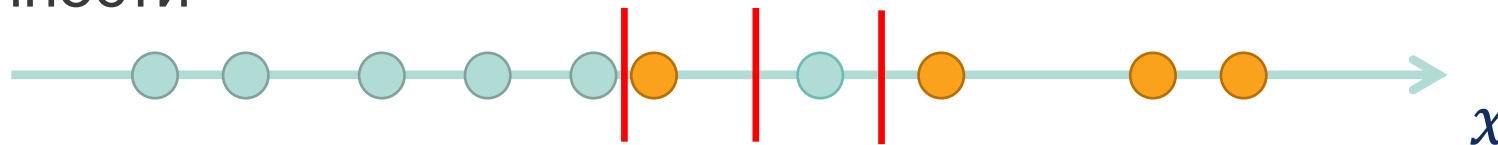


и не ограничивать количество порогов, чтобы разделить выборку идеально

Логический подход

Простейшая выборка

Вариант 1: потребовать от модели максимальной точности



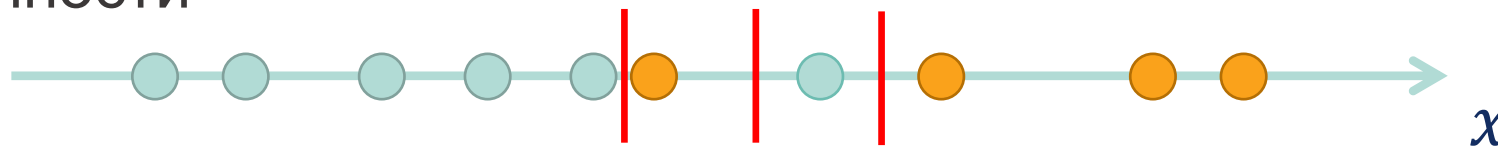
и не ограничивать количество порогов, чтобы разделить выборку идеально

Проблема: запоминание выборки вместо обучения

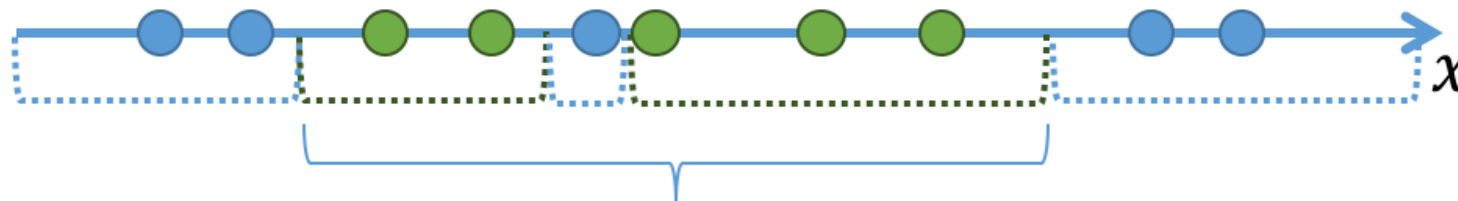
Логический подход

Простейшая выборка

Вариант 1: потребовать от модели максимальной точности



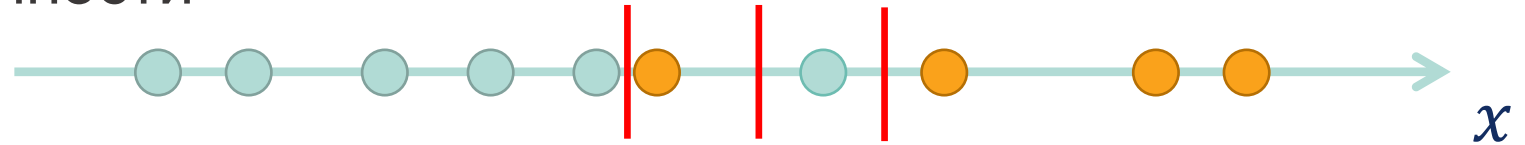
Вариант 2: разрешить объединение интервалов



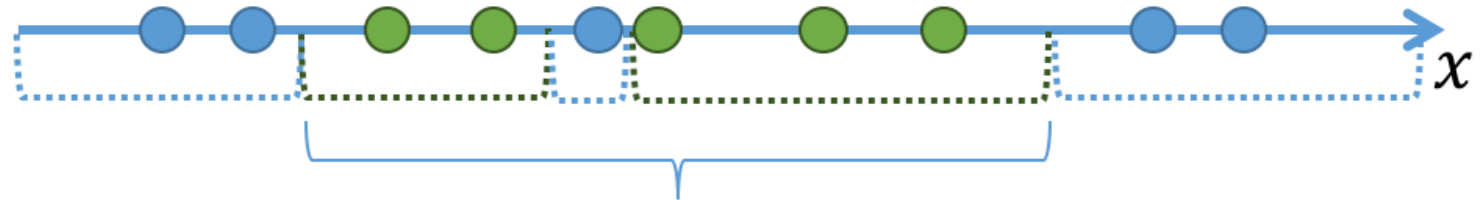
Логический подход

Простейшая выборка

Вариант 1: потребовать от модели максимальной точности



Вариант 2: разрешить объединение интервалов



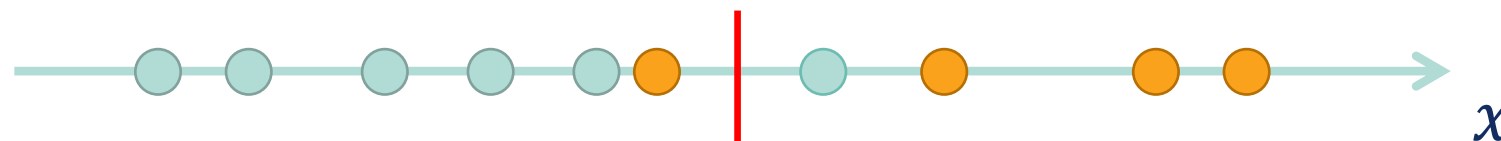
Такие интервалы можно строить последовательно

Простая выборка

Итак, выборка линейно не разделима

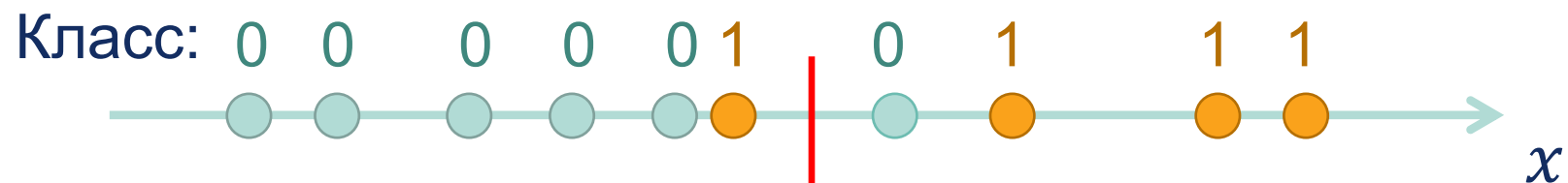


Требуется выбрать оптимальный порог:



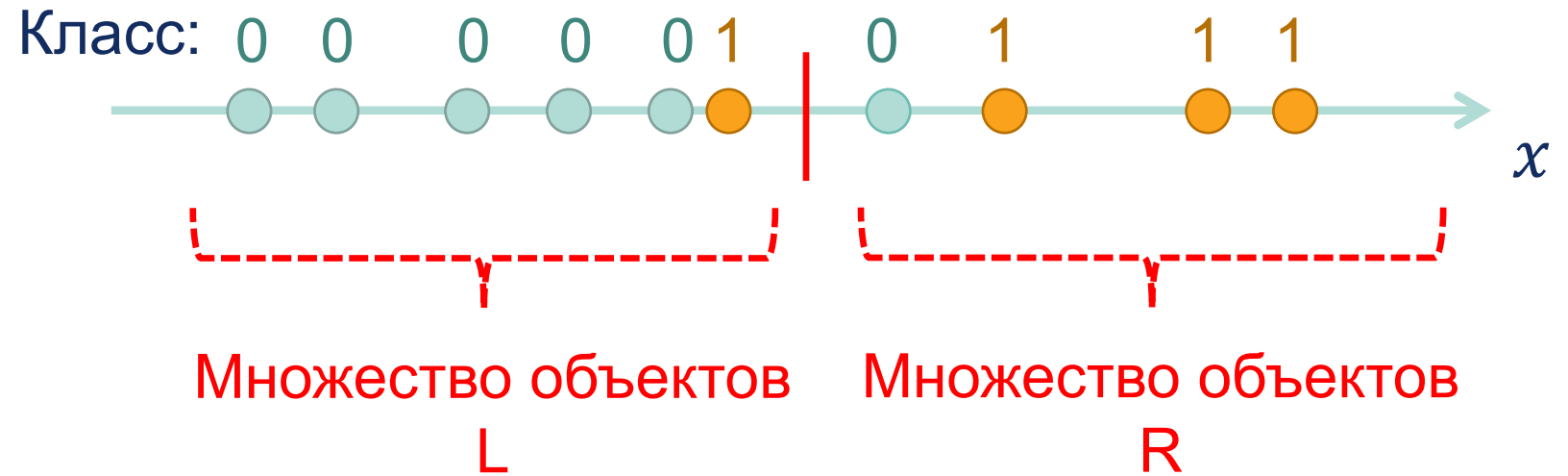
Как поставить задачу?

Задача оптимизации



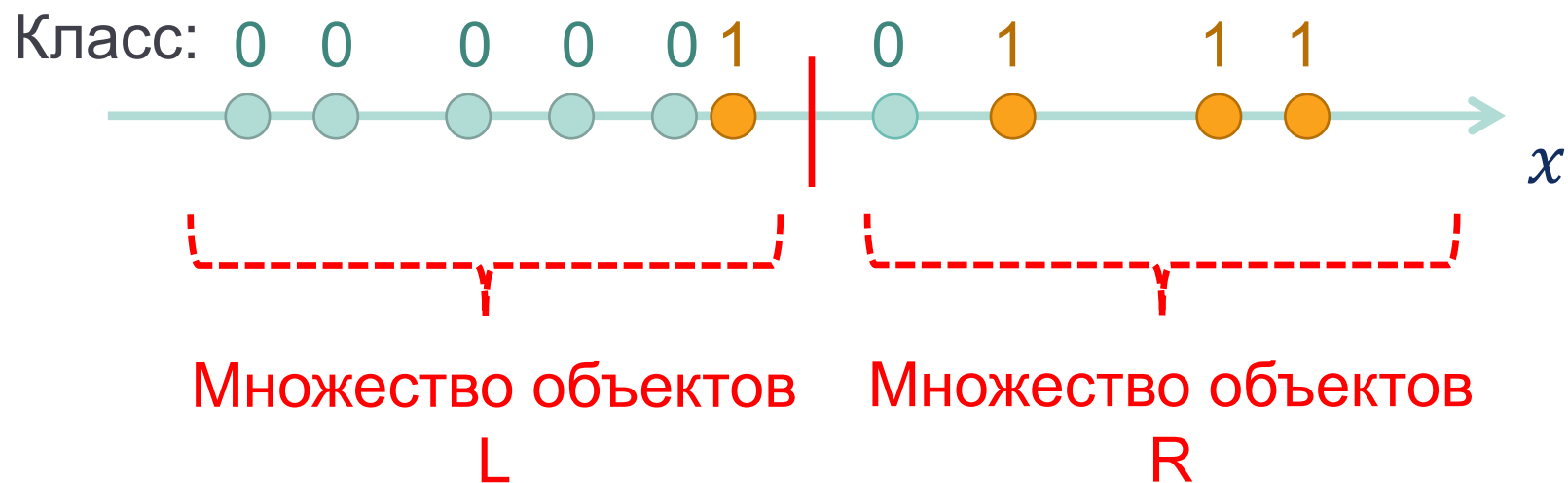
Логический
подход

Задача оптимизации



Логический
подход

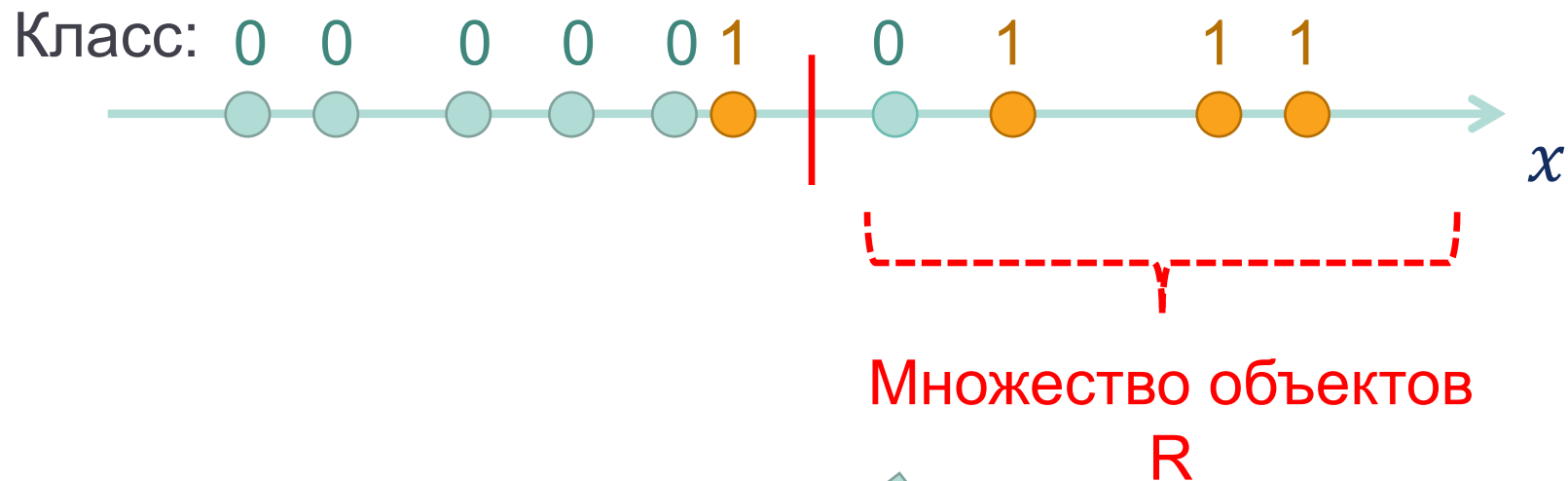
Задача оптимизации



Логический
подход

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

Задача оптимизации



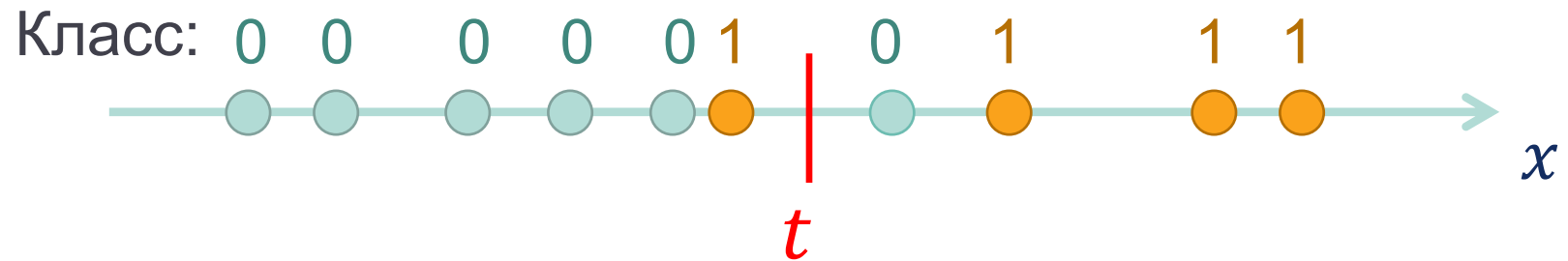
Логический
подход

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?

Логический подход

Задача оптимизации



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

Например, так:

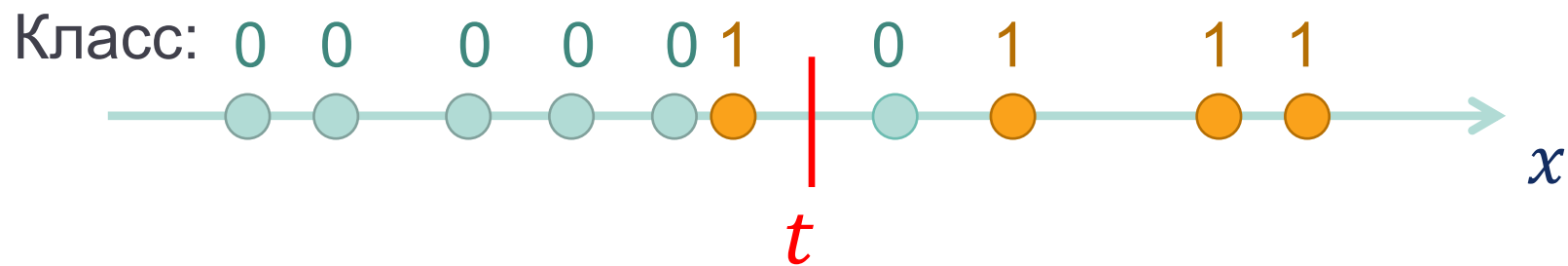
$$p_{max} = \max\{p_0, p_1\} \rightarrow \max_t$$

Или так:

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

Логический подход

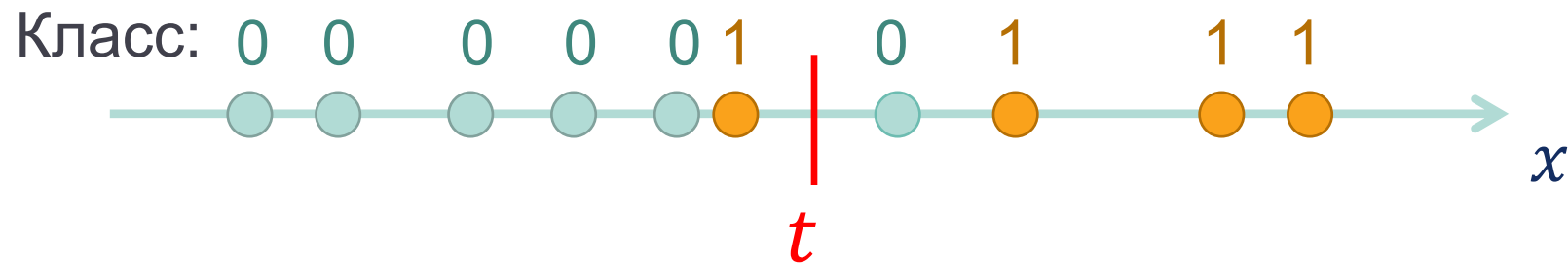
Задача оптимизации



Другой вариант:

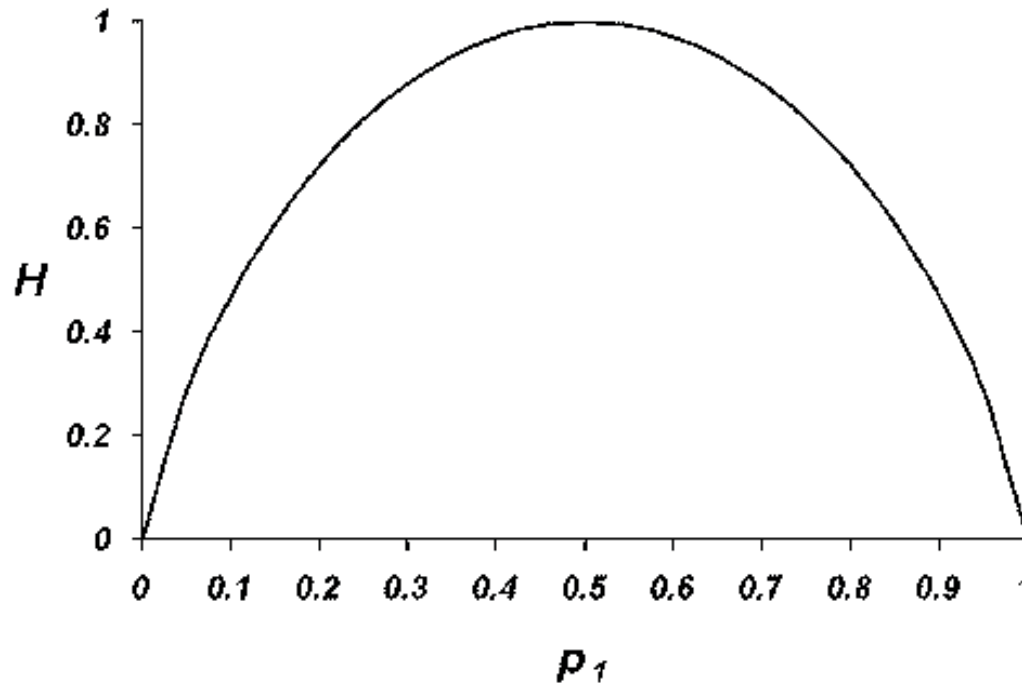
$$H(R) = -p_0 \ln p_0 - p_1 \ln p_1 \rightarrow \min_t$$

Задача оптимизации



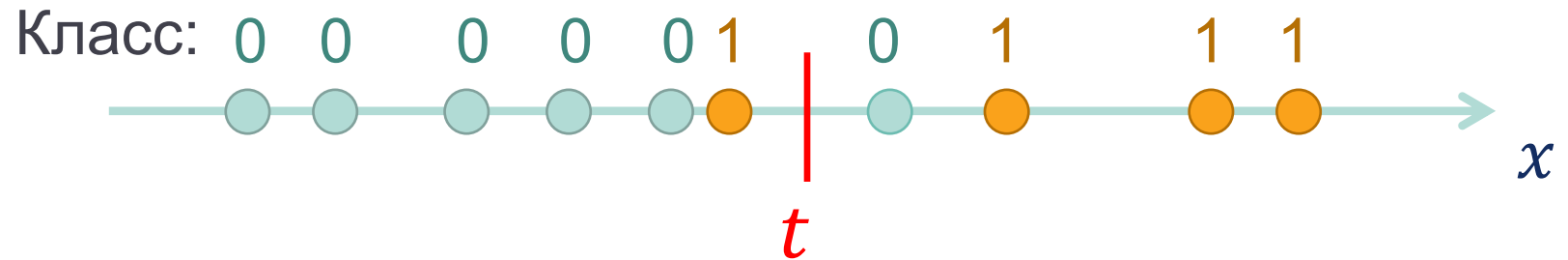
Другой вариант:

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



Логический
подход

Задача оптимизации

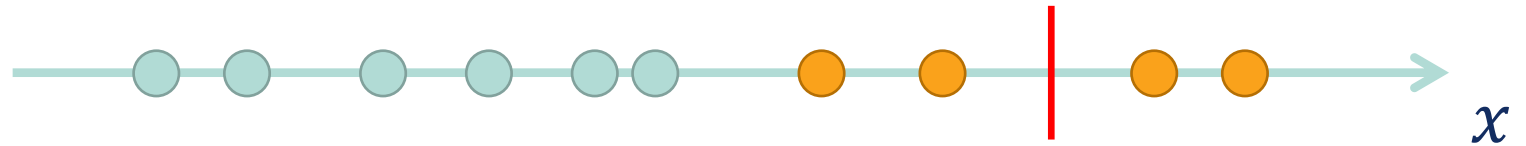


Логический
подход

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

Логический ПОДХОД

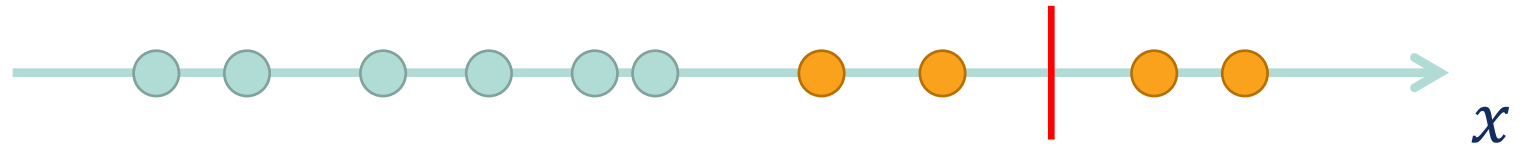
Но если смотреть только на R , можем разделить выборку так:



Здесь проблема возникает только в левой части, в правой части преобладает один класс

Логический ПОДХОД

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

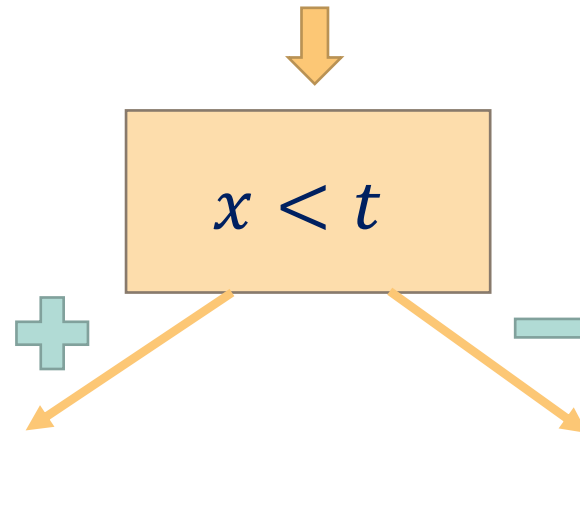


Здесь проблема возникает только в левой части, в правой части преобладает один класс

Значит надо учитывать обе части: R и L

Оптимизация разбиения

Вся выборка (n объектов)



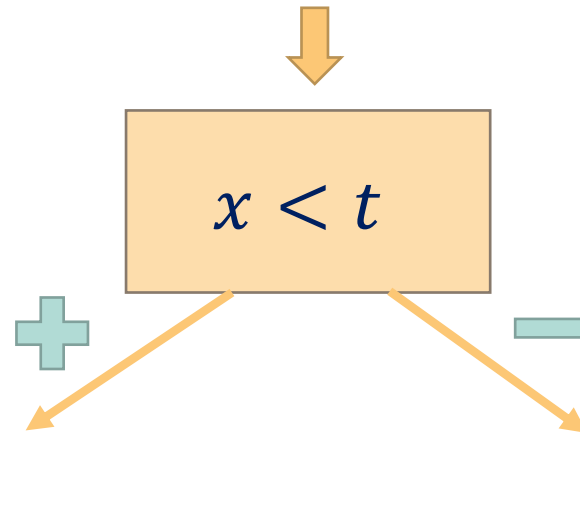
Логический
подход

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

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

Оптимизация разбиения

Вся выборка (n объектов)



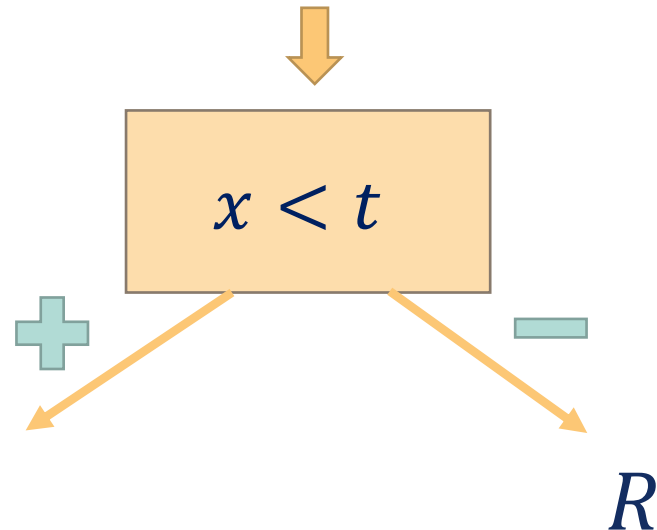
Логический
подход

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

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

Оптимизация разбиения

Вся выборка (n объектов)



Логический
подход

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

Оптимизация разбиения

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

Логический
подход

Оптимизация разбиения

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

Варианты этой функции:

1) Misclassification criteria: $H(R) = 1 - \max\{p_0, p_1\}$

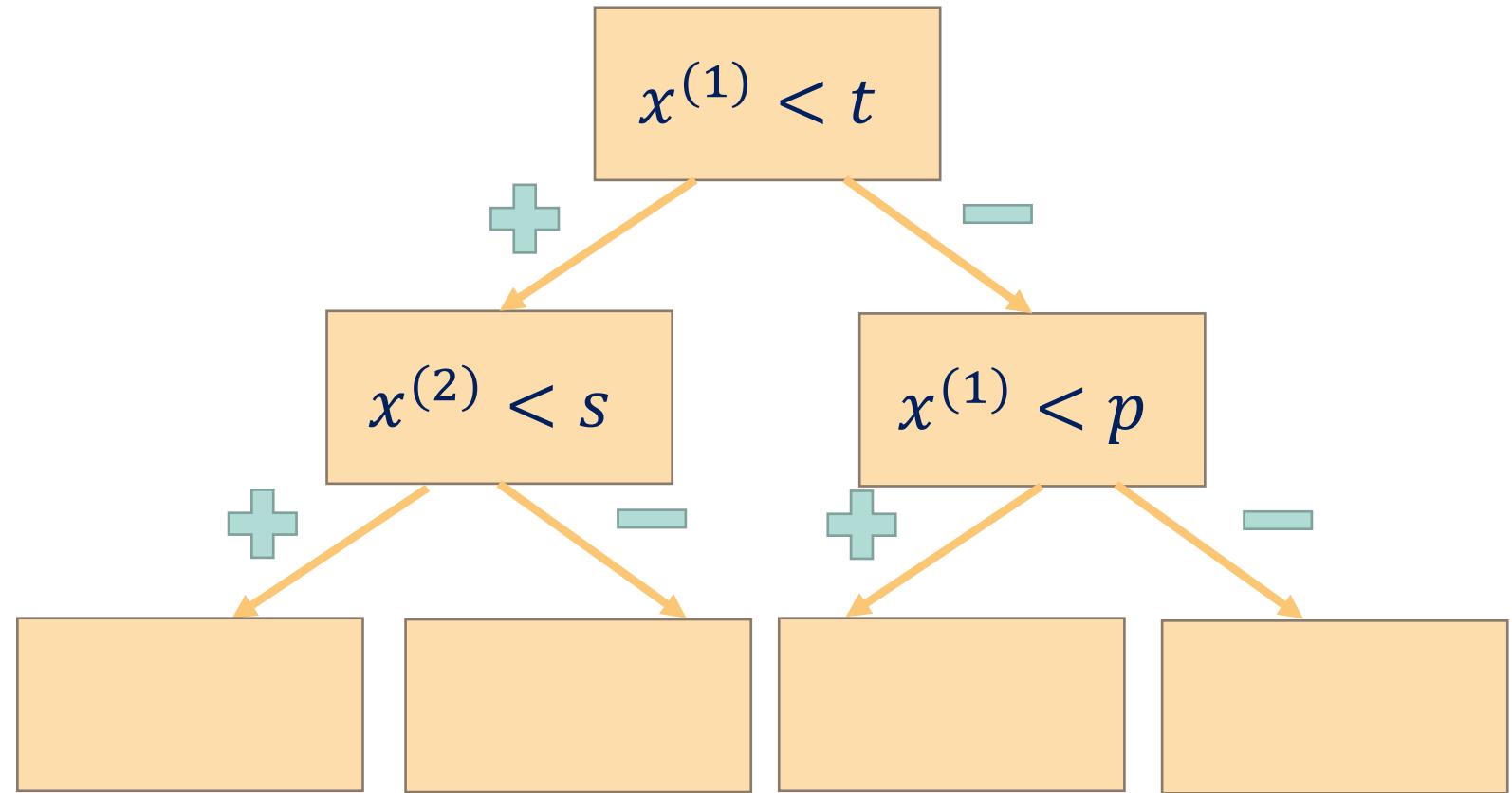
2) Entropy criteria: $H(R) = -p_0 \ln p_0 - p_1 \ln p_1$

3) Gini criteria: $H(R) = 1 - p_0^2 - p_1^2 = 2p_0p_1$

Логический
ПОДХОД

Обобщение для N признаков

Логический
подход



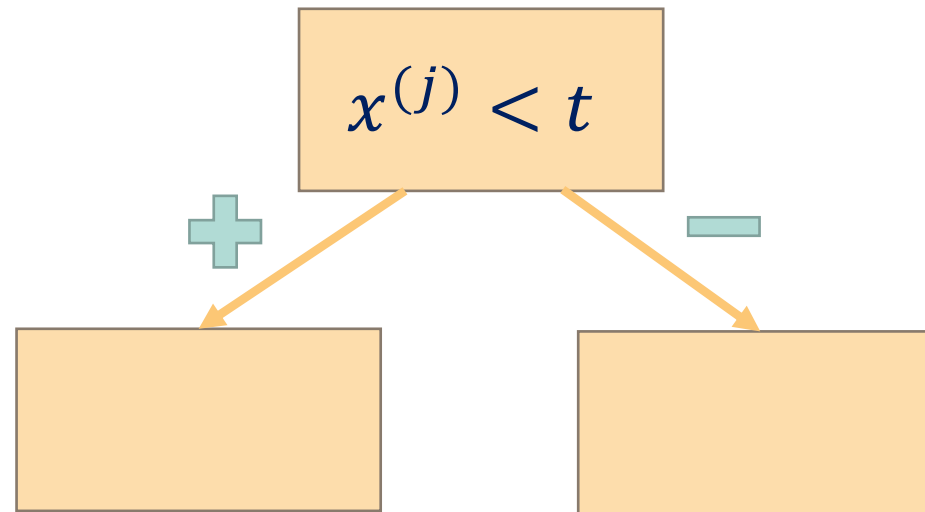
Рекурсивное построение

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

Логический
подход

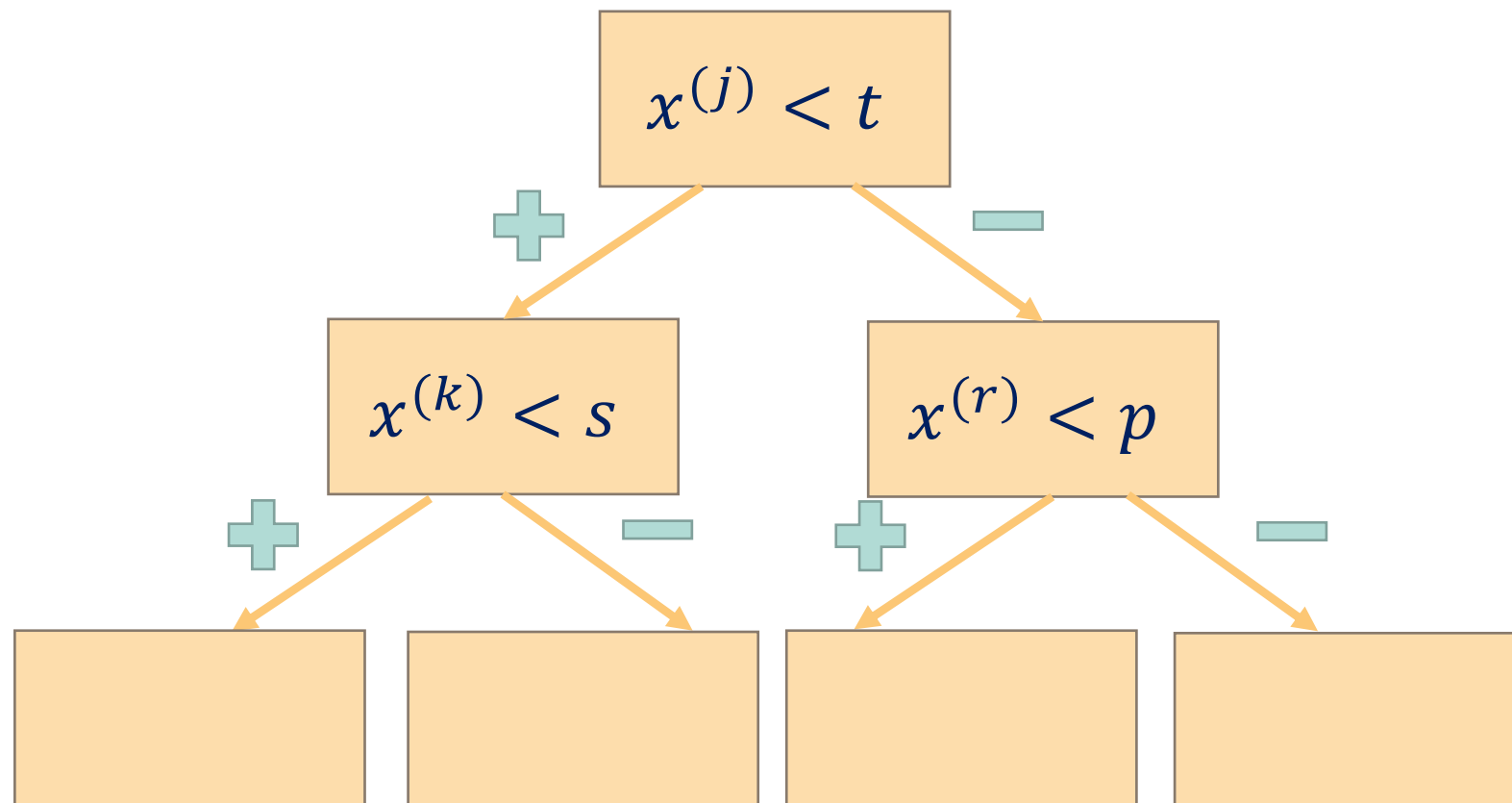
Рекурсивное построение

Логический
подход



Рекурсивное построение

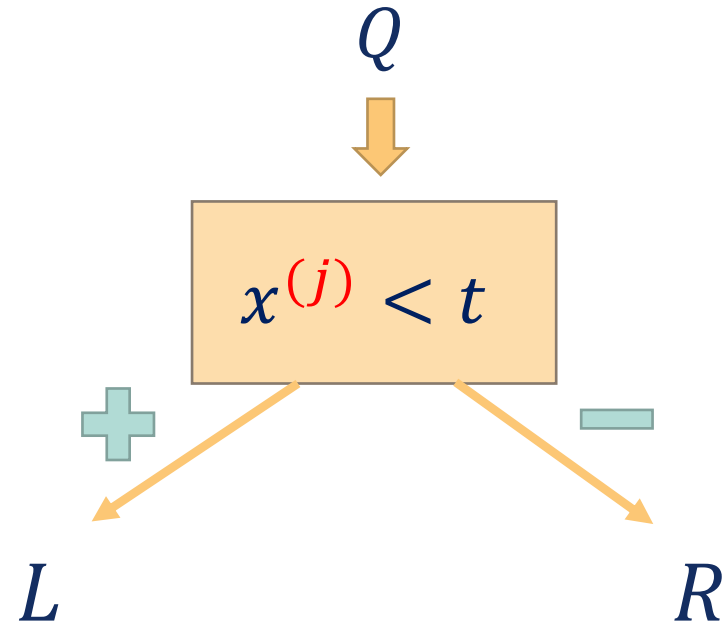
Логический
подход



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

Логический
ПОДХОД

Рекурсивное построение



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

Рекурсивное построение

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

Варианты этой функции:

1) Misclassification criteria: $H(R) = 1 - \max\{p_0, p_1\}$

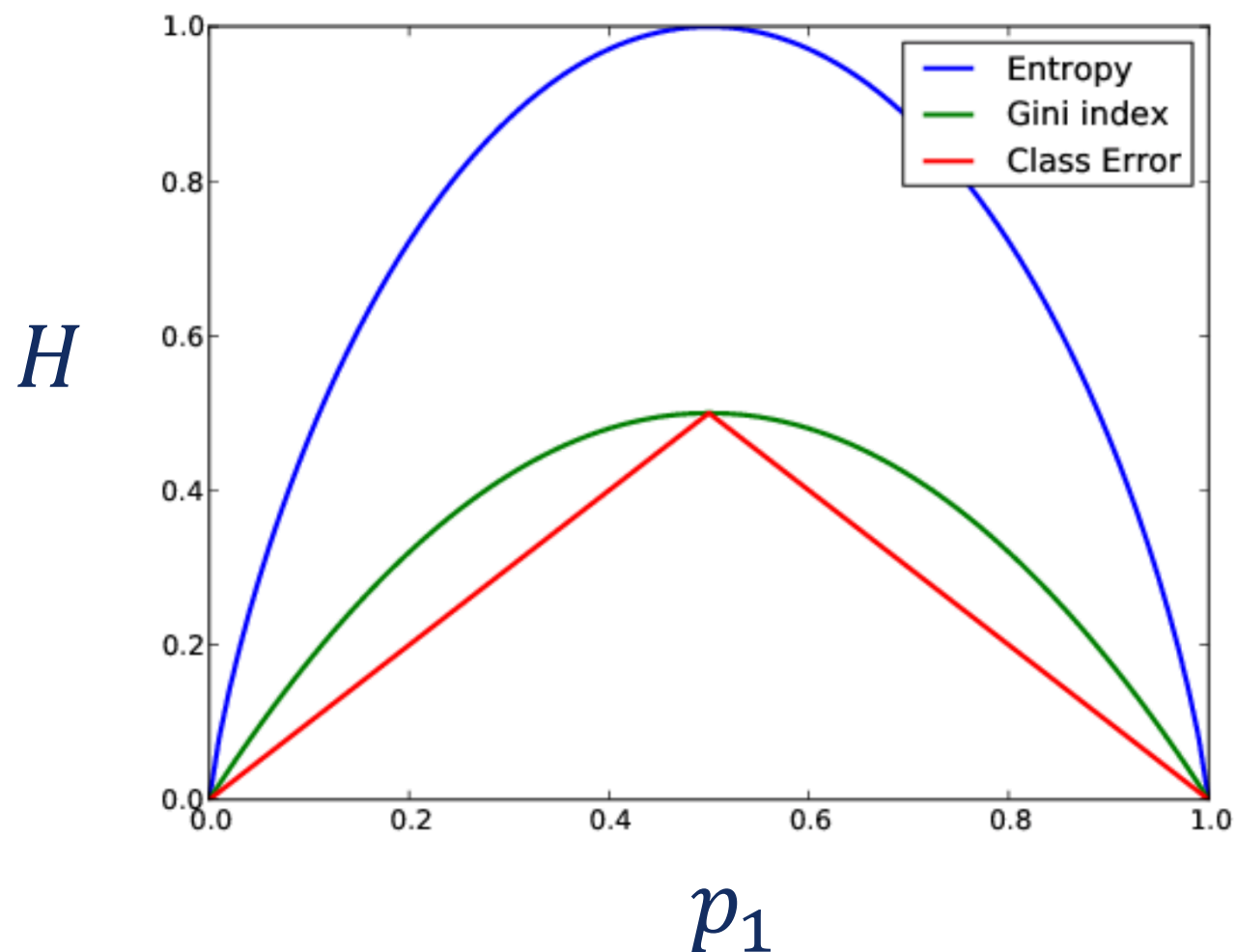
2) Entropy criteria: $H(R) = -p_0 \ln p_0 - p_1 \ln p_1$

3) Gini criteria: $H(R) = 1 - p_0^2 - p_1^2 = 2p_0p_1$

Логический
ПОДХОД

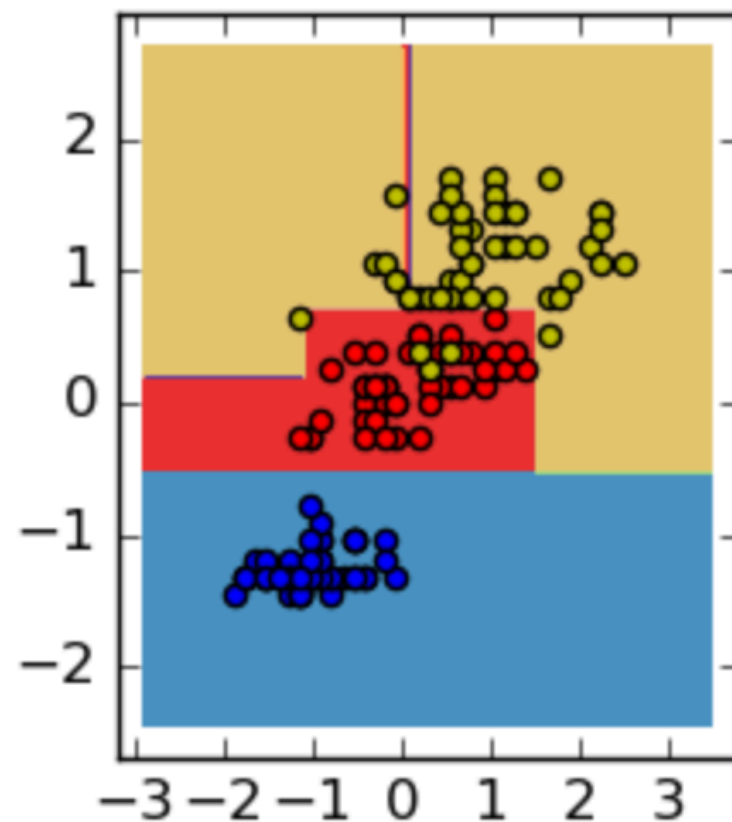
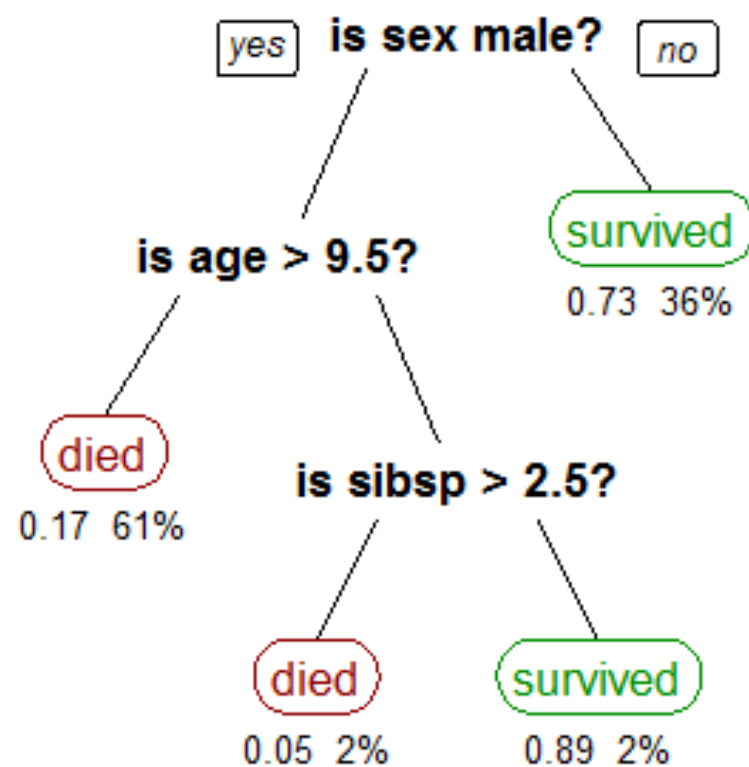
Критерии разбиений

Логический
подход



Дерево решений

Логический
ПОДХОД



Логический ПОДХОД

Деревья решений

Область применения:

- базовый алгоритм в ансамбле
- очень небольшие выборки
- алгоритм для интерпретации сложной модели

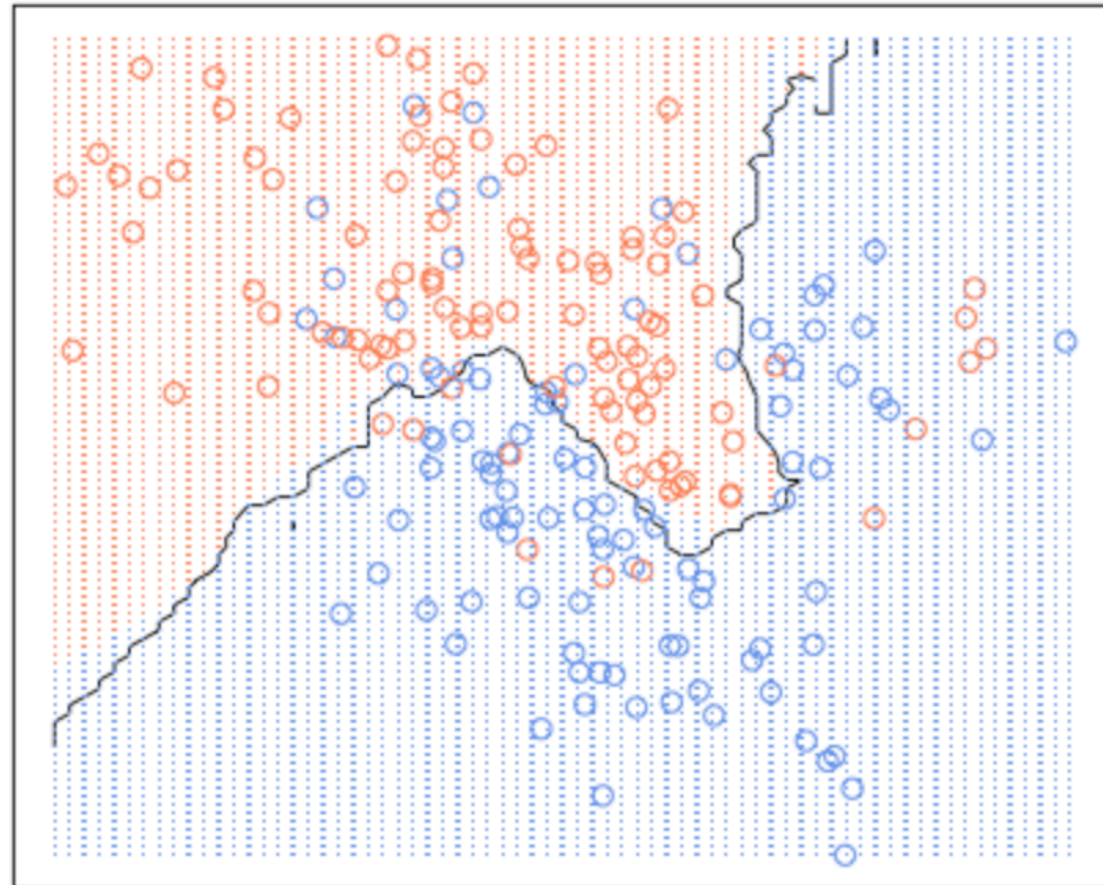
Ограничения:

- сильнейшее переобучение

Метрический подход

Границы сложной формы

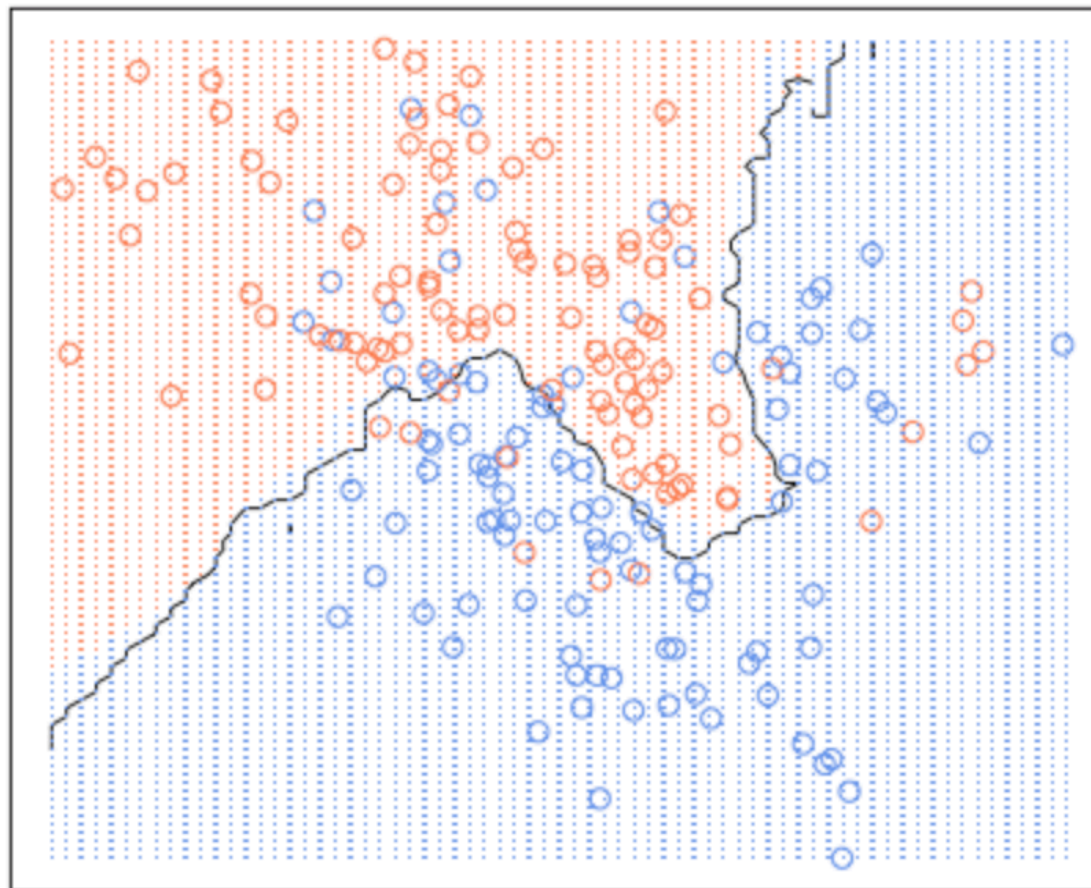
Метрический
подход



Метрический подход

Границы сложной формы

С помощью дерева решений такие границы строить неудобно



Границы сложной формы

Гипотеза о “компактности”:

- объекты одного класса похожи на представителей своего класса, а значит расположены в пространстве рядом друг с другом

Сложные границы

Границы сложной формы

Гипотеза о “компактности”:

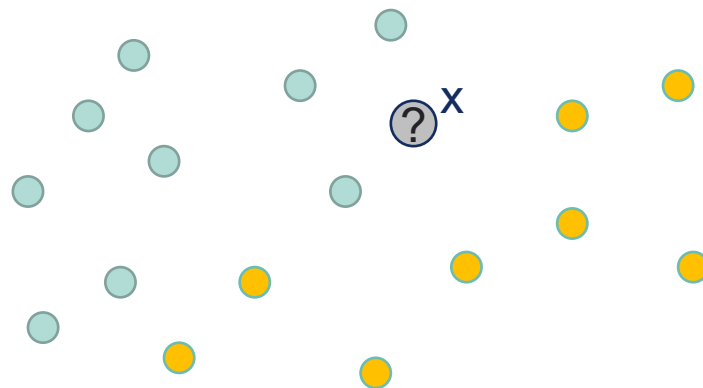
- объекты одного класса похожи на представителей своего класса, а значит расположены в пространстве рядом друг с другом

Идея:

- давайте классифицировать объекты на основе близости

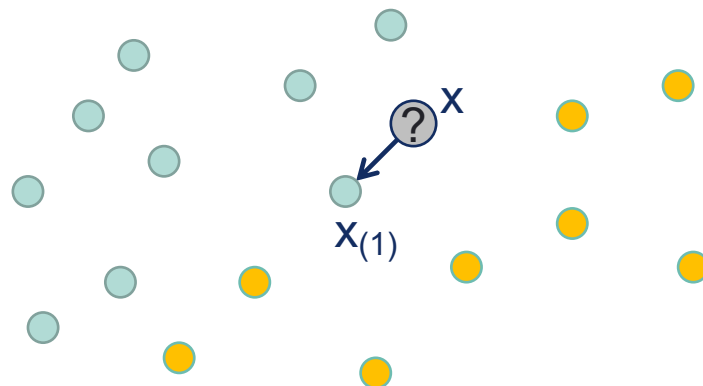
Бинарная классификация

Метрический
подход



Бинарная классификация

Метрический
подход





Бинарная классификация

Возьмём две точки в многомерном пространстве:

x_1 и x_2

Как ввести расстояние между ними?

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


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

Метрический подход

Расстояние между объектами

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:


$$(1) d(x_1, x_2) = d(x_2, x_1) = \|x_2 - x_1\|$$

Метрический подход

Расстояние между объектами

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:

$$(1) d(x_1, x_2) = d(x_2, x_1) = \|x_2 - x_1\|$$

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

Расстояние между объектами

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

Примеры норм:

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

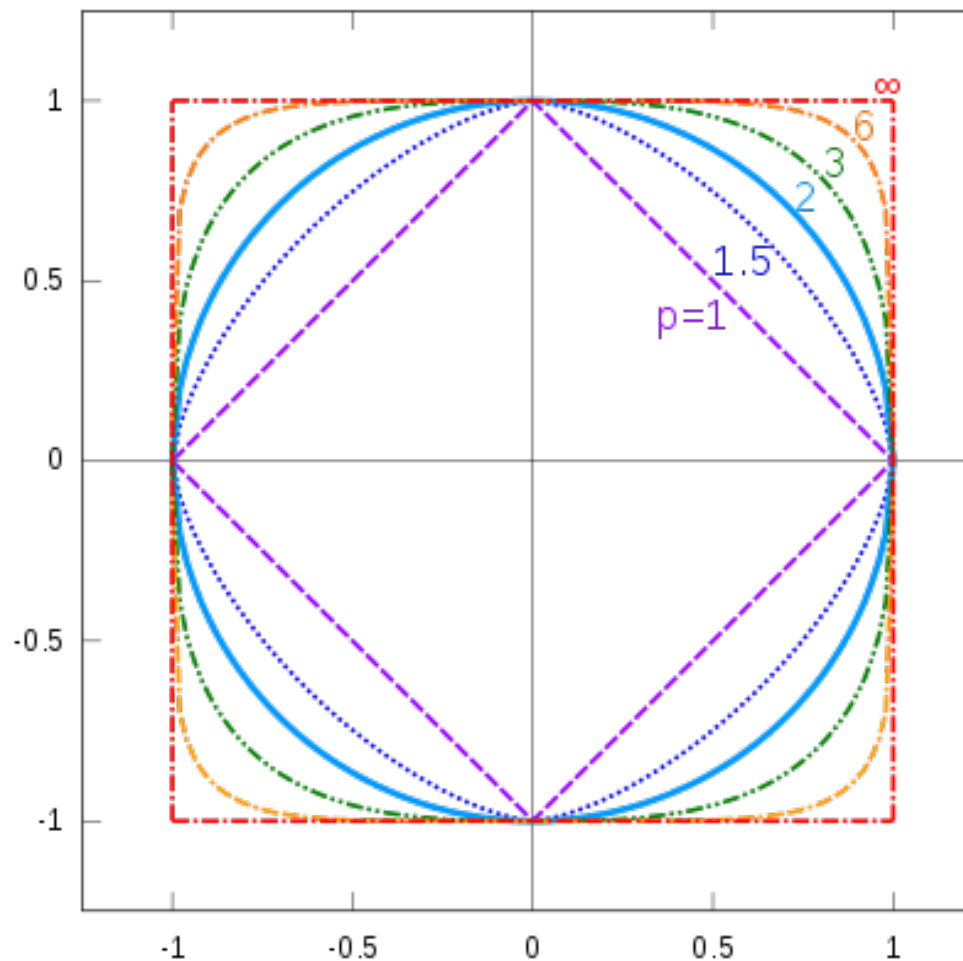
$$\|x\|_{\ell_1} = \left|x^{(1)}\right| + \dots + \left|x^{(d)}\right|$$

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

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

Метрический
подход

Варианты норм



Метрический подход

Расстояние между объектами

Расстояние должно иметь смысл для решаемой задачи:

- Как оценить расстояние между клиентами?
- Как оценить расстояние между фильмами?

Метрический подход

Расстояние между объектами

Расстояние должно иметь смысл для решаемой задачи:

- Как оценить расстояние между клиентами?
- Как оценить расстояние между фильмами?

Идеи:

- Можно ввести кастомизированную метрику (кажется, можно даже не метрику, а функцию)
- Можно вместо расстояния ввести меру близости

Расстояние между объектами

Пример:

Косинусная мера близости (cosine similarity)

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

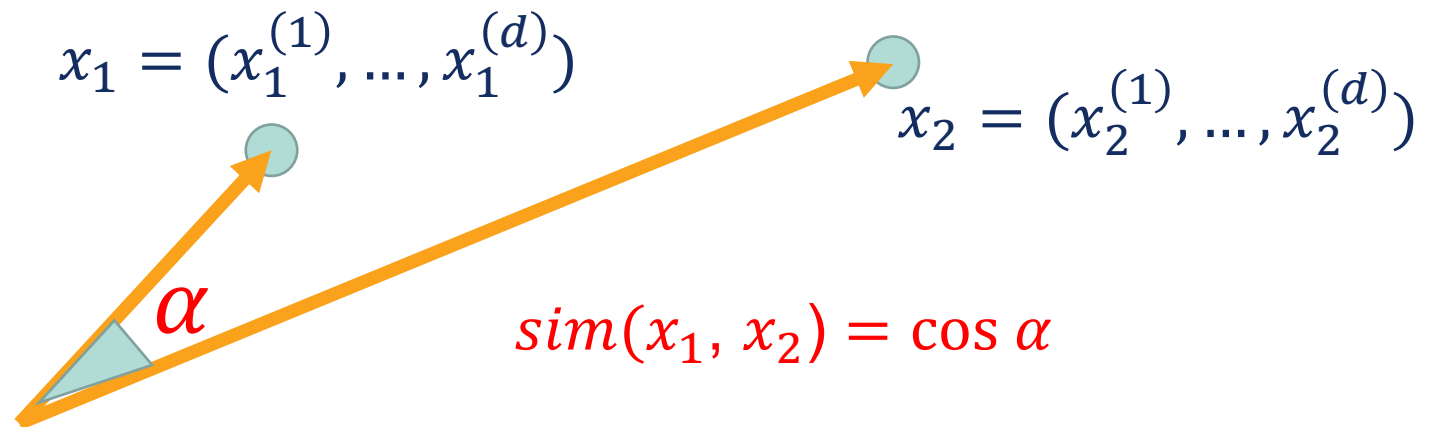
Метрический подход

Расстояние между объектами

Пример:

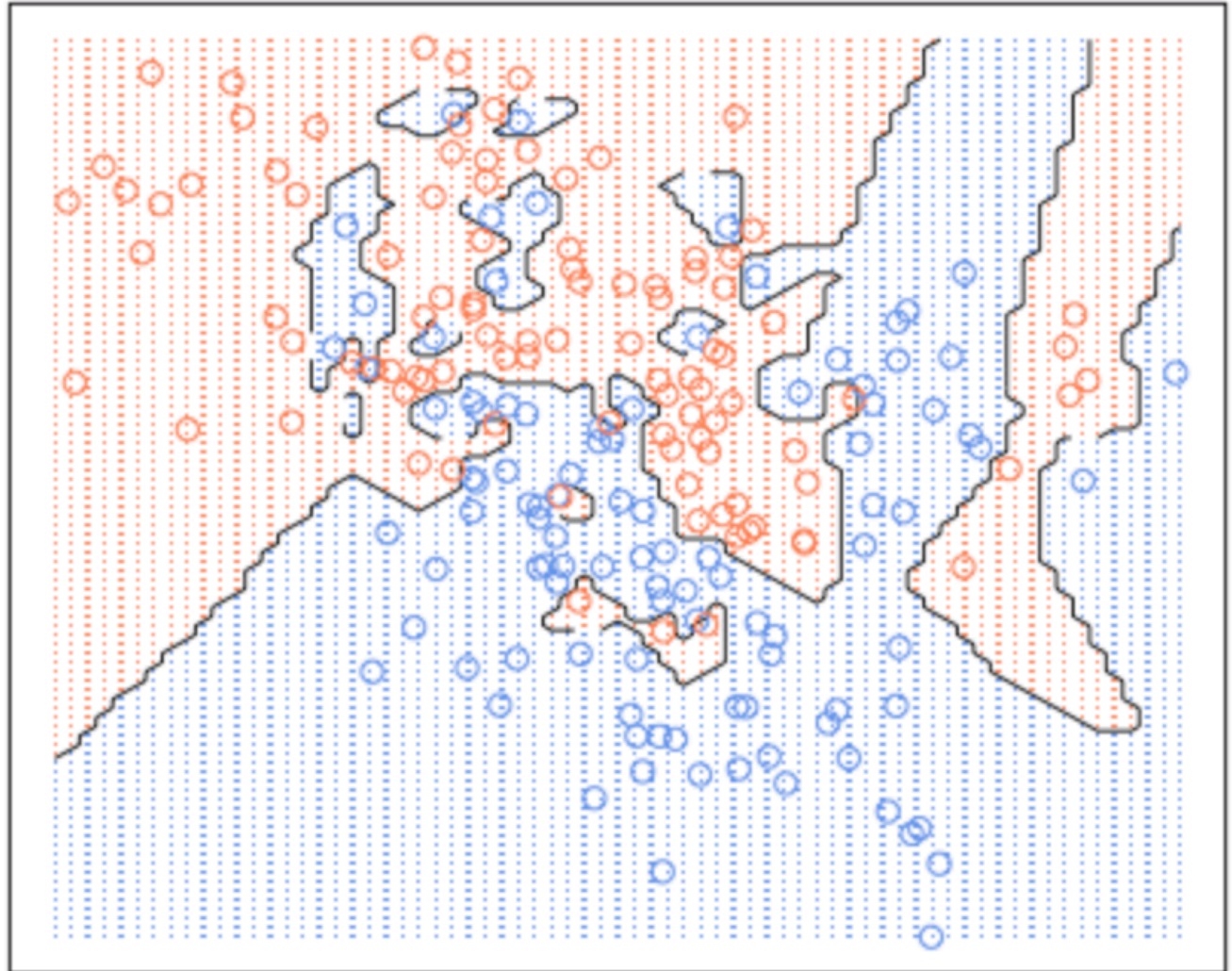
Косинусная мера близости (cosine similarity)

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



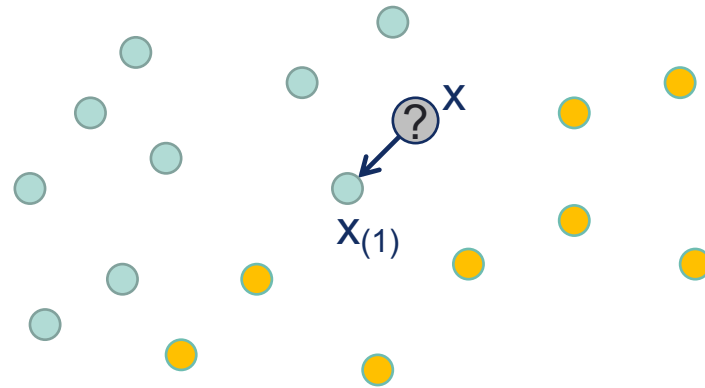
Метод 1NN

Метрический
подход



Метрический
подход

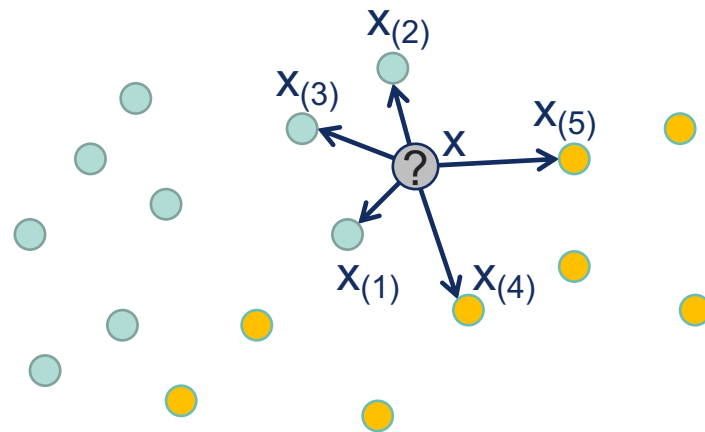
Метод 1NN



Метрический подход

Метод kNN

Пример классификации для $k = 5$:

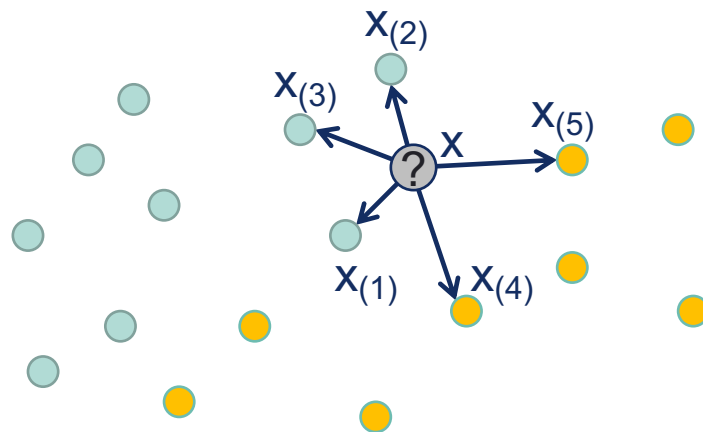


Метрический подход

Метод kNN

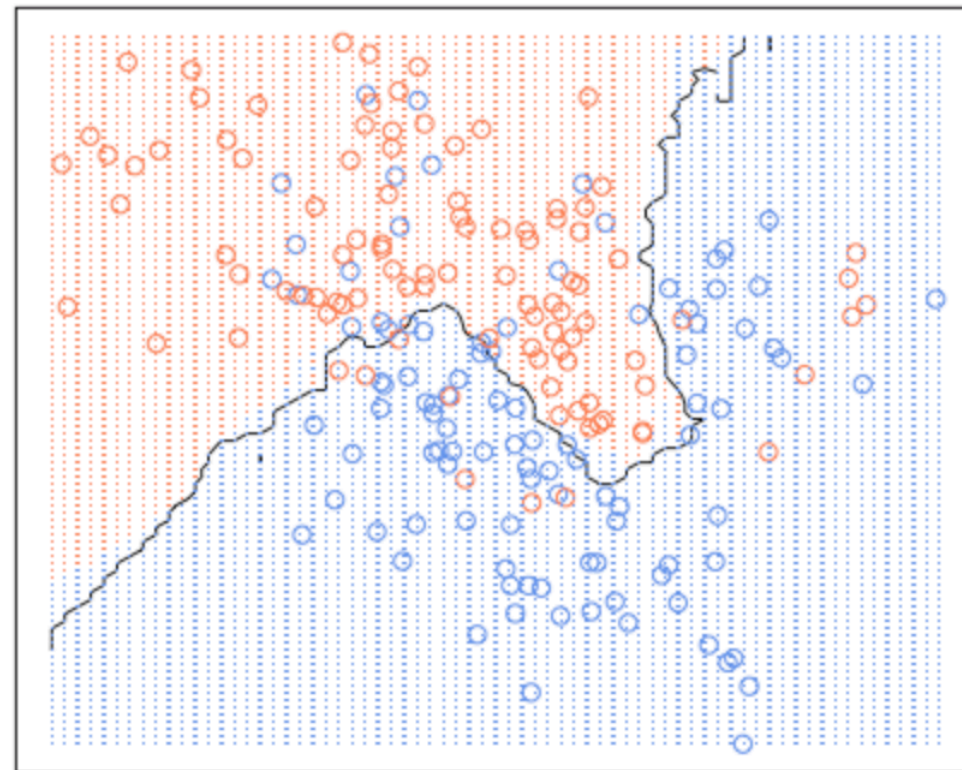
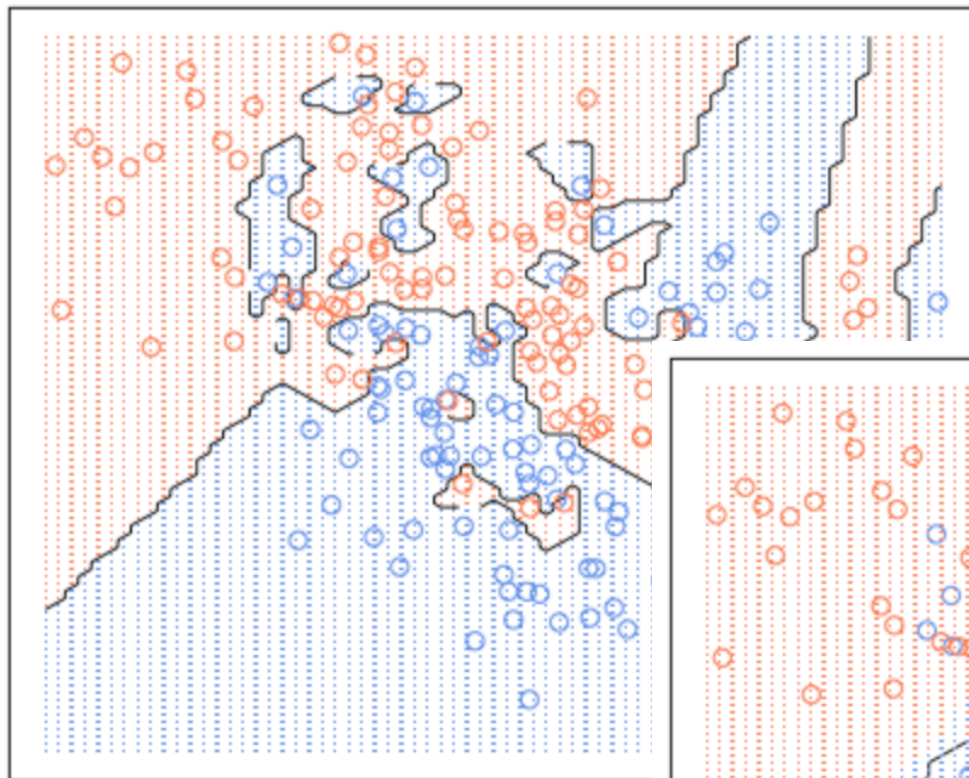
Пример классификации для $k = 5$:

Выбираем класс, который преобладает



Метод kNN

Метрический
подход



Метрический подход

Метод kNN

Как подобрать оптимальное значение k ?

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

Метрический подход

Метод kNN

Как подобрать оптимальное значение k ?

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

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

Метрический подход

Метод kNN

Как подобрать оптимальное значение k ?

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

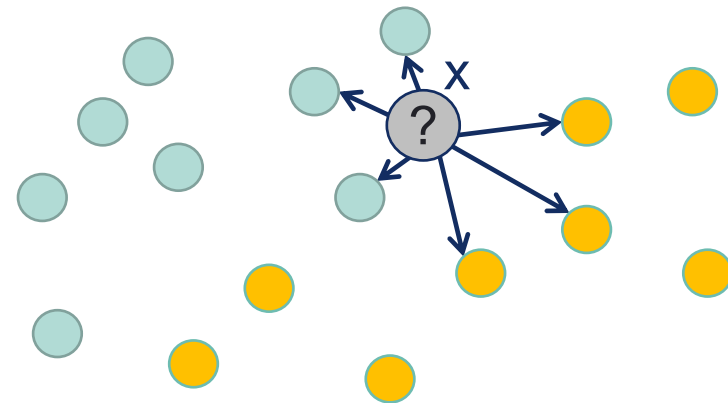
Правильно, $k=1$ – для каждого объекта обучающей выборки смотрим на ближайшего соседа (этот же объект)

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

Метрический подход

Метод kNN

Пример классификации ($k = 6$):

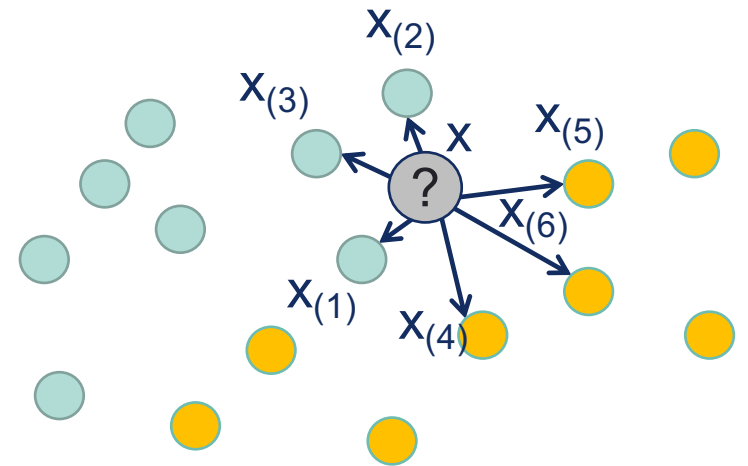


Как принять решение, если за каждый класс голосует одинаковое количество объектов?

Метрический подход

Метод kNN

Пример классификации ($k = 6$):



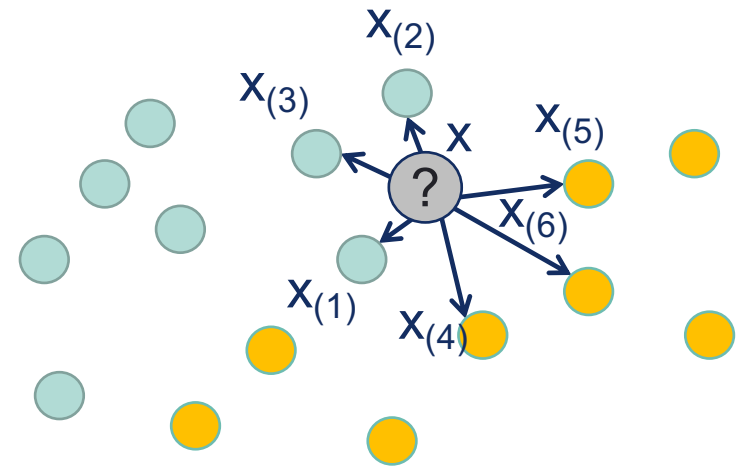
Как принять решение, если за каждый класс голосует одинаковое количество объектов?

Идея: давайте взвесим вклад от соседей

Метрический подход

Метод kNN

Пример классификации ($k = 6$):



Веса:

- функция от номера объекта $w(x_{(i)}) = w(i)$
- функция от расстояния $w(x_{(i)}) = w(d(x, x_{(i)}))$

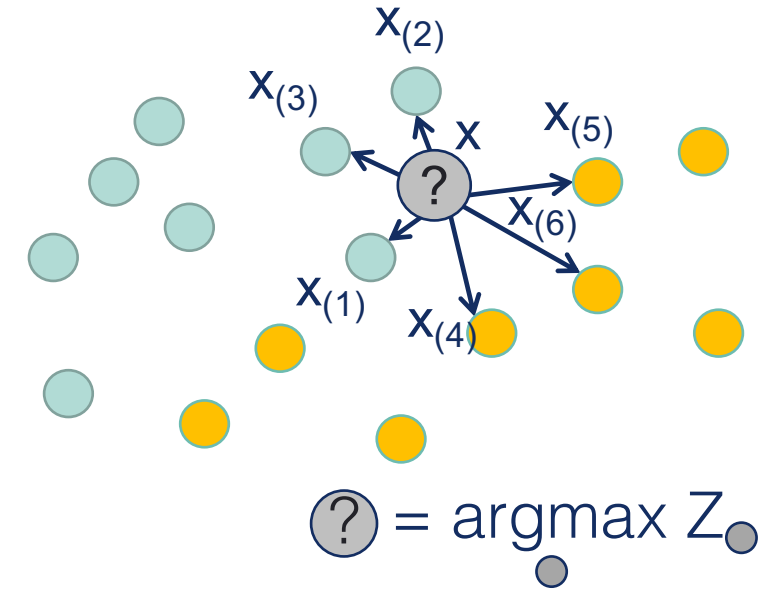
Метрический подход

Метод kNN

Пример классификации ($k = 6$):

Веса:

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



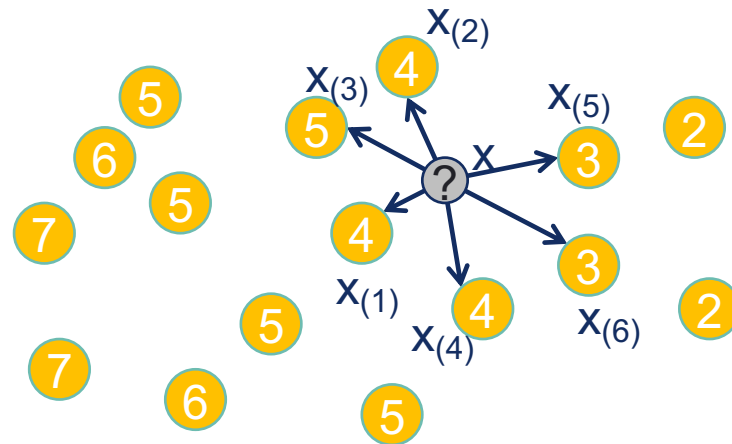
$$Z_{\text{blue}} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

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

Метрический подход

Метод kNN в задаче регрессии

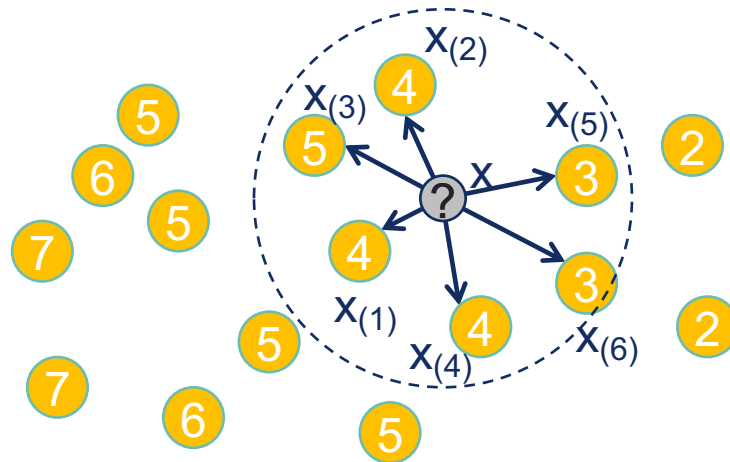
Пример взвешенного kNN ($k = 6$) в задаче регрессии:



Метрический подход

Метод kNN в задаче регрессии

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



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

Обобщенный метрический классификатор

Метрический ПОДХОД

$$a(x; X^l) = \underset{y \in Y}{\operatorname{argmax}} \underbrace{\sum_{i=1}^l [y^{(i)} = y] w(i; x)}_{\Gamma_y(x)}$$

$w(i; x)$ – вес i -го соседа объекта x

$\Gamma_y(x)$ - оценка близости объекта x к классу y

Метрический ПОДХОД

Метод Парзенковского окна

$$w(i; x) = K\left(\frac{\rho(x, x^{(i)})}{h}\right), \text{ где } h \text{ — ширина окна}$$

$K(r)$ — ядро, не возрастает и положительно на $[0, 1]$

Метод Парзенковского окна фиксированной ширины:

$$a(x; X^l, h, K) = \operatorname{argmax}_{y \in Y} \sum_{i=1}^l [y^{(i)} = y] K\left(\frac{\rho(x, x^{(i)})}{h}\right)$$

Метод Парзенковского окна переменной ширины:

$$a(x; X^l, k, K) = \operatorname{argmax}_{y \in Y} \sum_{i=1}^l [y^{(i)} = y] K\left(\frac{\rho(x, x^{(i)})}{\rho(x, x^{(k+1)})}\right)$$

Метрический ПОДХОД

Метод потенциальных функций

$$w(i; x) = \gamma^{(i)} K \left(\frac{\rho(x, x^{(i)})}{h} \right), \text{ где } \gamma^{(i)} \text{ вес объекта } i$$

$$a(x; Xl, h, K) = \operatorname{argmax}_{y \in Y} \sum_{i=1}^l [y^{(i)} = y] \gamma^{(i)} K \left(\frac{\rho(x, x^{(i)})}{h_i} \right)$$

Аналогия из физики:

- $\gamma^{(i)}$ - величина заряда в точке x_i
- h_i — радиус действия потенциала с центром в точке x_i
- y_i — знак заряда
- $K(r) = 1/r$

Метрический подход

Метод Парзенковского окна

Область применения:

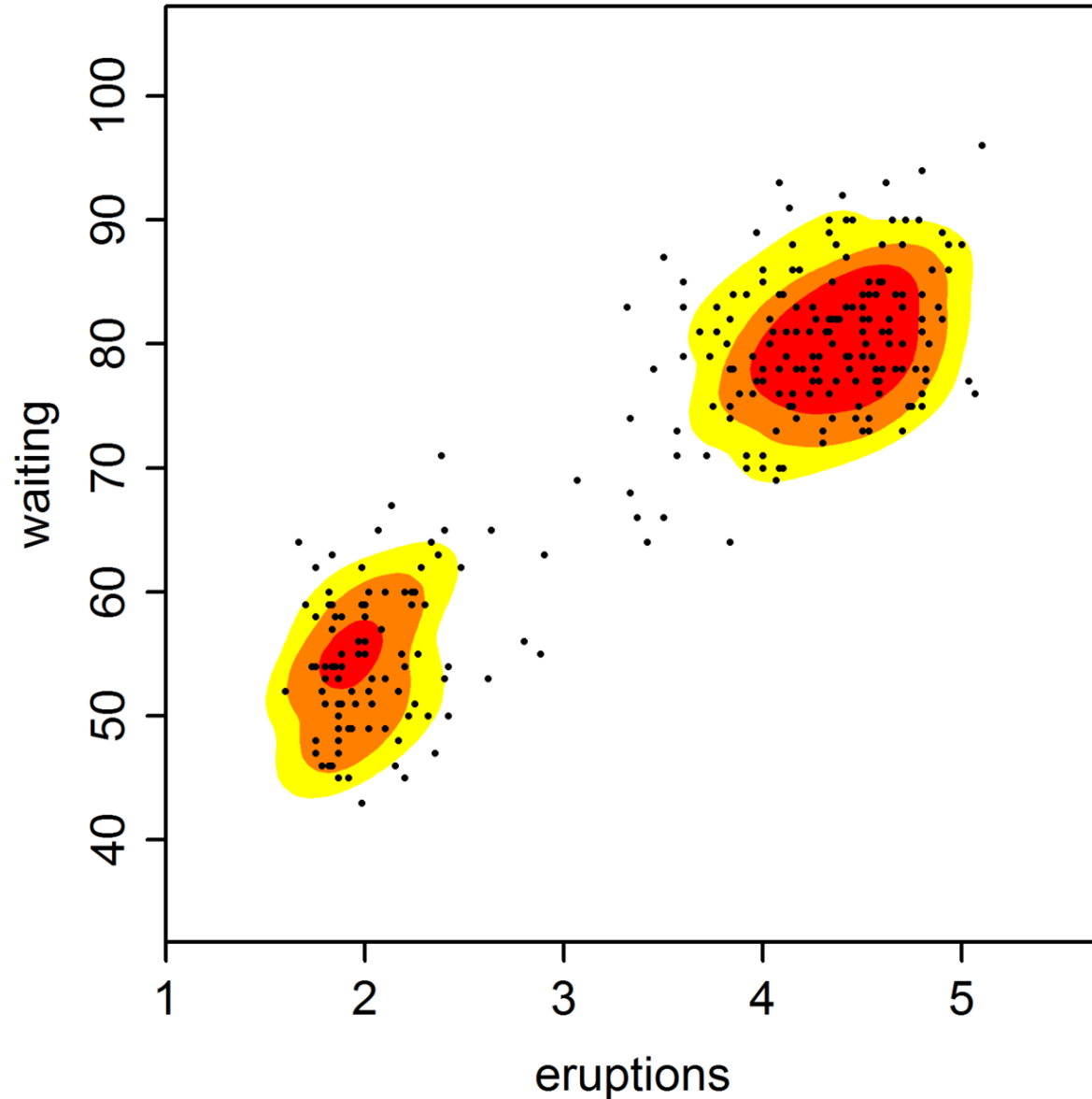
- задачи, в которых оценить близость легче, чем ввести признаки
- небольшие выборки
- fall-back алгоритм

Ограничения:

- ленивое обучение
- выбор расстояния
- чувствителен к выбросам в обучении

Вероятностный подход

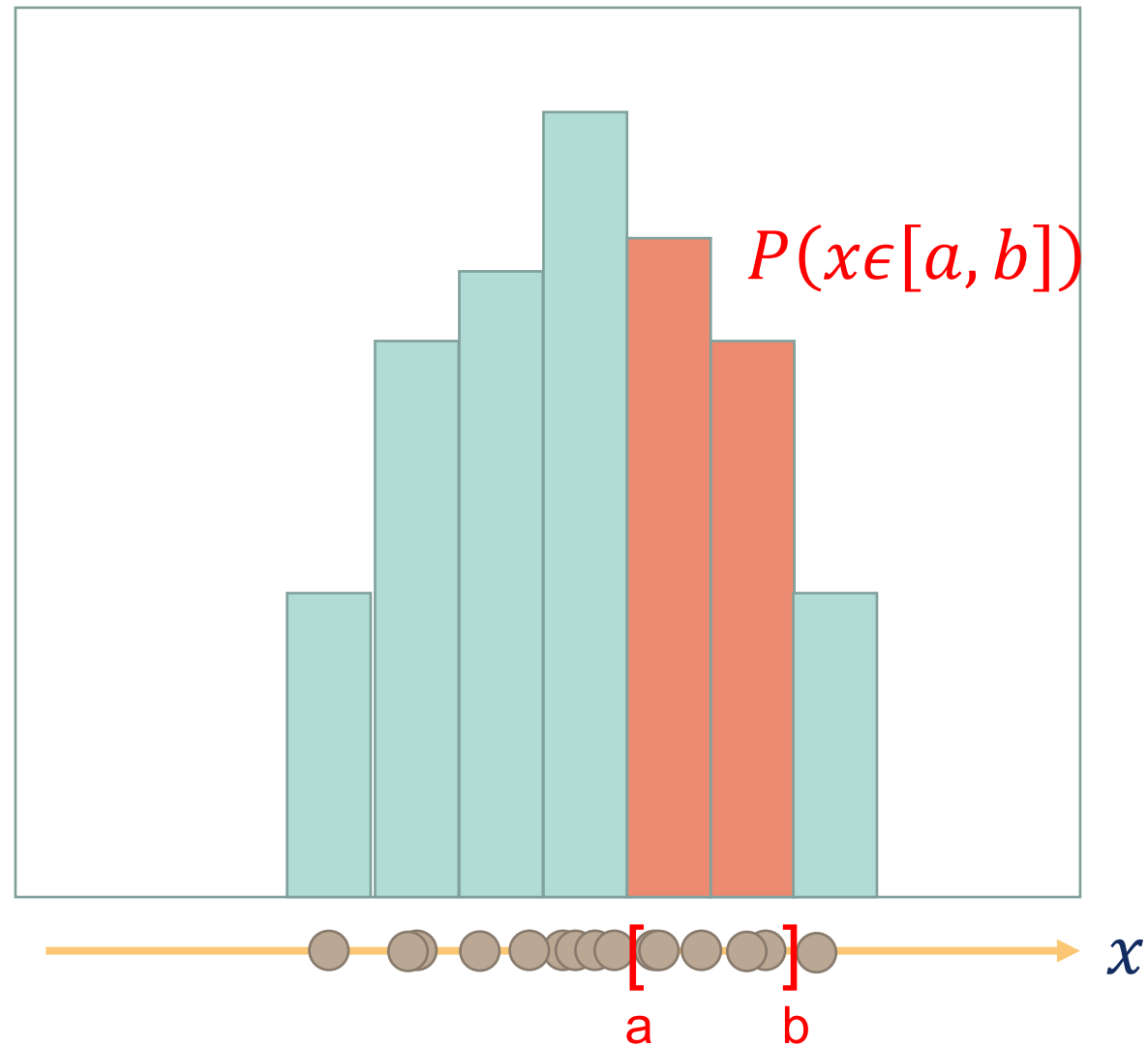
Бинарная классификация



Вероятностный
подход

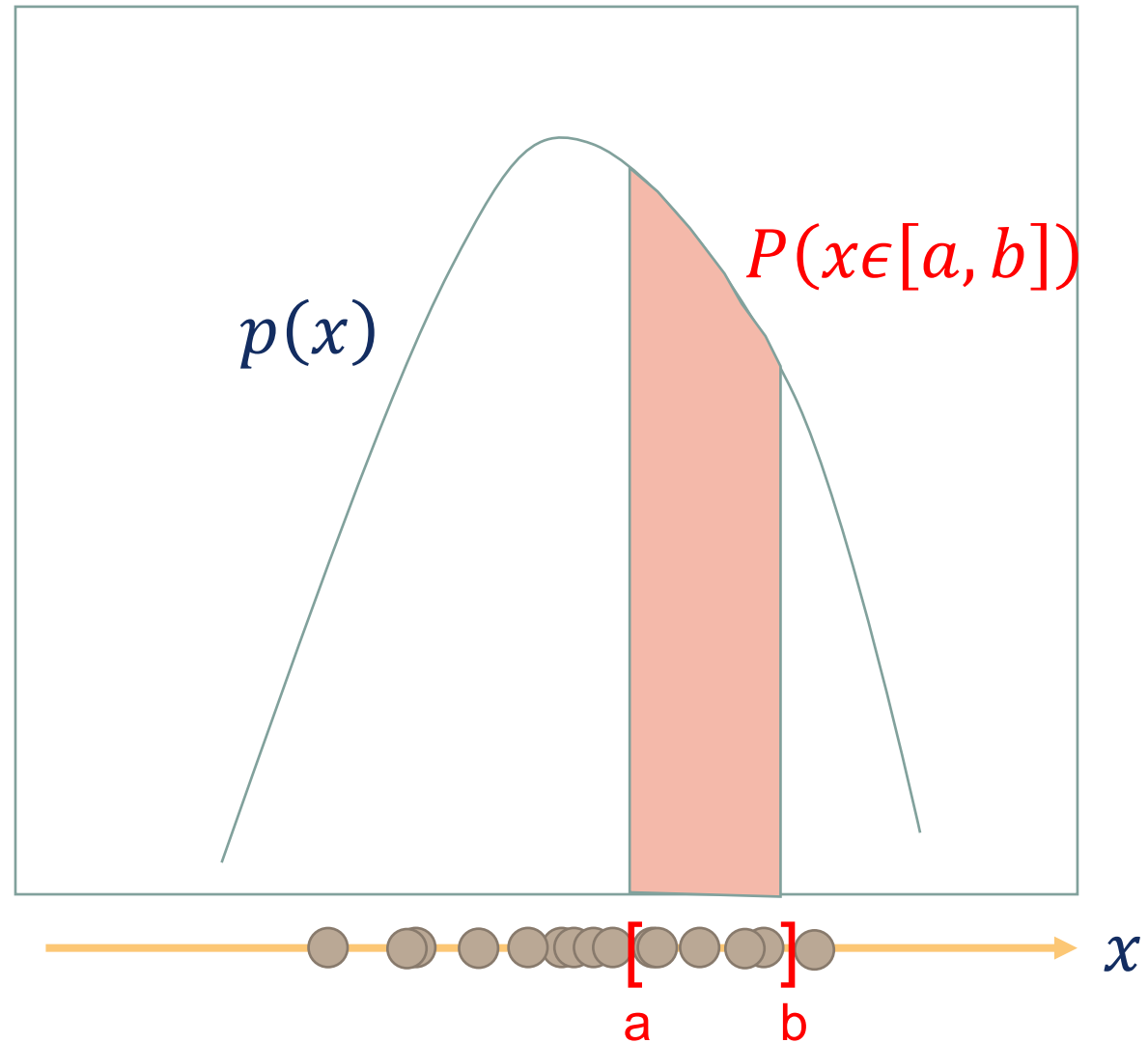
Одномерный случай

Вероятностный
подход



Плотность распределения

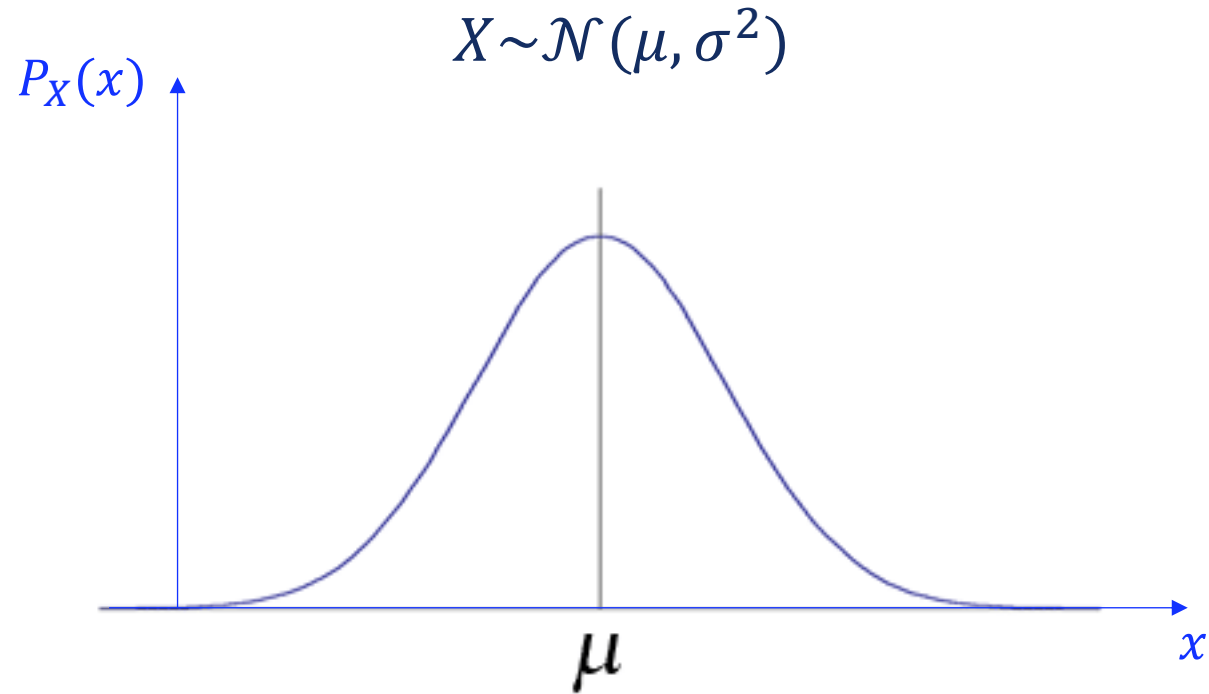
Вероятностный
подход



Оценка плотности

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

Нормальное распределение



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

Вероятностный
подход

Нормальное распределение

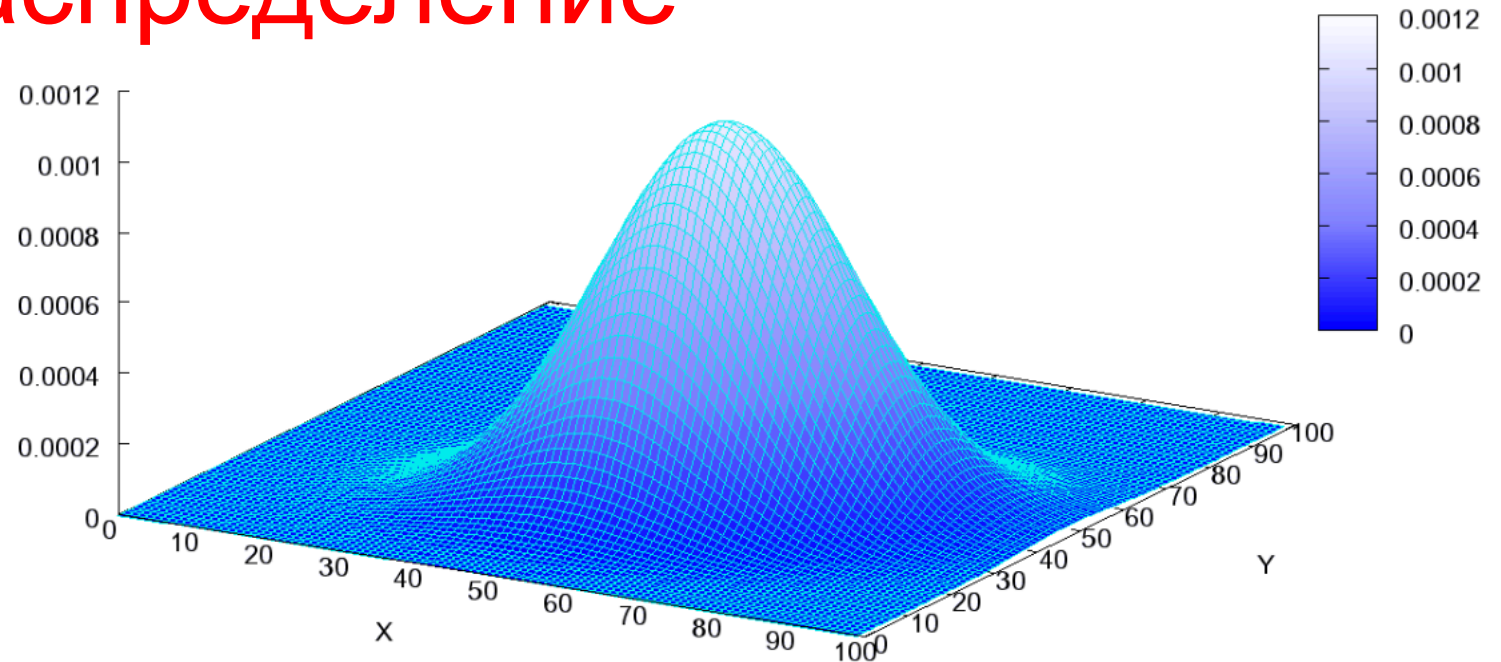
$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

другой вариант оценки для σ^2 :

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

Вероятностный
подход

Многомерное нормальное распределение



Вероятностный
подход

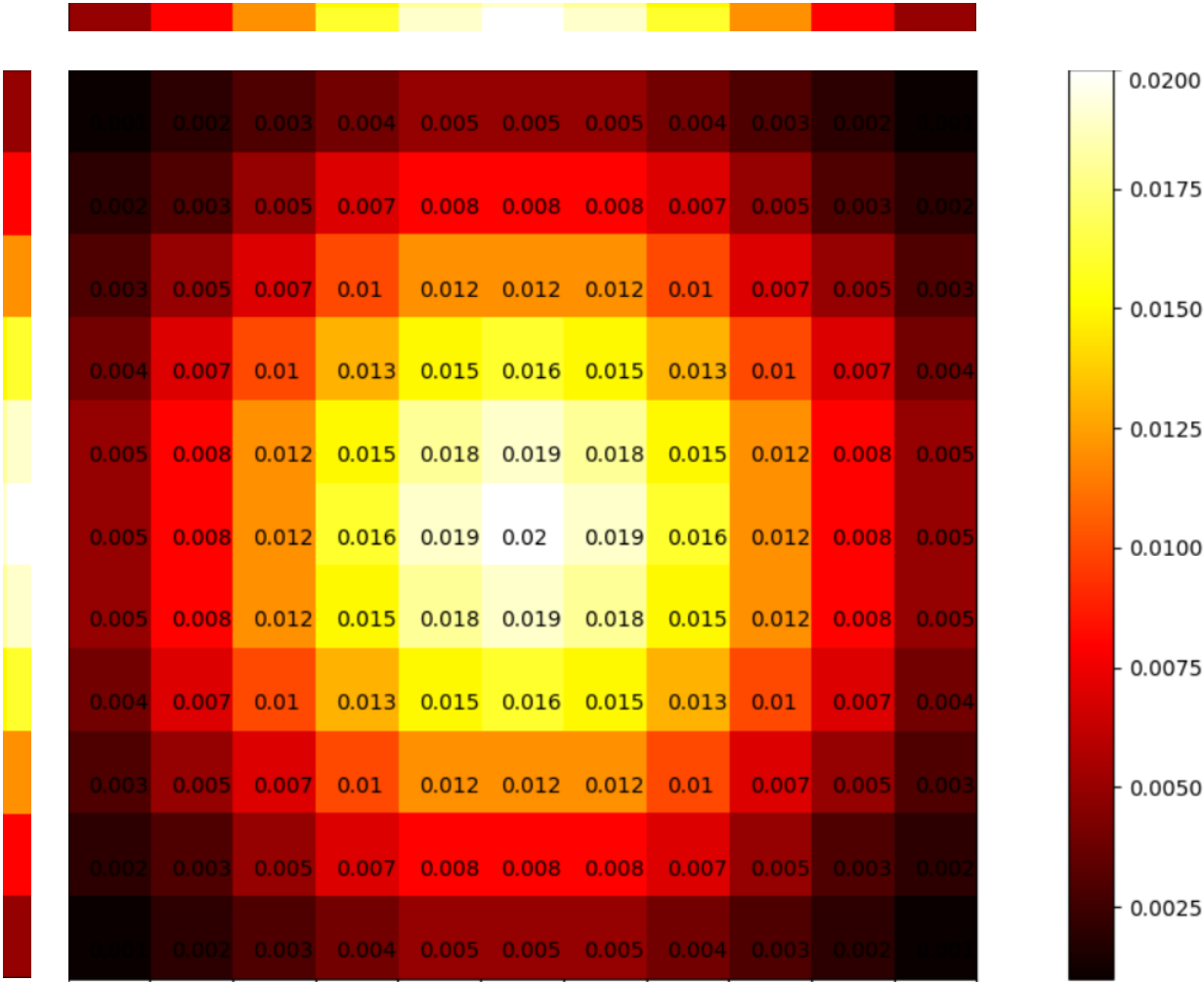
$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Очень много параметров: вектор средних μ и матрица ковариаций Σ

Вероятностный
подход

Наивное предположение

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



Наивное предположение

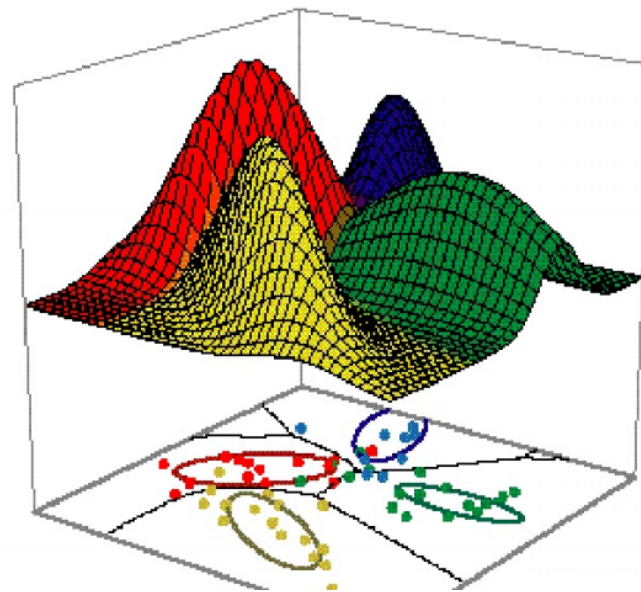
Если признаки $x^{(1)}, \dots, x^{(d)}$ распределены независимо:

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

- в общем случае это не так
- если признаки не независимы, это можно предположить и всё равно воспользоваться этим свойством
- отсюда название **наивный байесовский классификатор**

Наивный Байесовский классификатор

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



Вероятностный
подход

Вероятностный подход

Наивный Байесовский классификатор

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

Вероятностный подход

Наивный Байесовский классификатор

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

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

Наивный Байесовский классификатор

$p(x, y) = P(y)p(x|y) = P(x)p(y/x)$ – формула Байеса

$P(y/x) = P(y)p(x/y)$ – так как $p(x) = 1$, мы наблюдаем x

$p(x|y) = p(x^{(1)}|y) \dots p(x^{(d)}|y)$ - наивное предположение

Обучение модели:

1. Оцениваем **для каждого класса y** каждую из одномерных плотностей $p(x^{(k)}|y)$ по выборке
2. Оцениваем **для каждого класса y** его априорную вероятность $P(y)$
3. Классифицируя объект x выбираем класс с максимальной $P(y/x)$

Применение модели:

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

Вероятностный
подход

Наивный Байесовский классификатор

Особенности:

- требуется восстановление плотности
- предположение о независимости признаков работает не всегда (но иногда работает!)

Машинное обучение: простые алгоритмы обучения с учителем

Спасибо!
Эмили Драль