

# Машинное обучение Лекция 1. Простые методы

# Планы по курсу

#### Ближайшие темы

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

#### Ближайшие задания

- 1. Введение: теория + контест
- 2. Линейные модели: теория + реализовать алгоритм
- 3. Ансамбли: теория + контест

#### Оценка:

- 1. Домашние задания (допуск к зачету, просрочка сдачи == доп. задание)
- 2. Баллы за семинары (если есть задание на семинар) и «листки» (+балл на зачете)
- 3. Устный дифф. зачет (теорминимум + вопросы по билету)

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

	Средний заказ	Жалоб за неделю	Дней не активен		Лет с нами	Ушел в отток?
$x_1$	<b>x</b> 11	$x_{12}$	$x_{13}$		0.1	Да
$x_2$	600 p	1	8	•••	1.5	Нет
$x_3$	200 p	0	6	•••	0.2	Да

• • •

	Средний заказ	Жалоб за неделю	Дней не активен	•••	Лет с нами	Ушел в отток?
$\mathcal{C}_1$	$x_{11}$	$x_{12}$	$x_{13}$	• • •	0.1	Да
£2	$x_{21}$	$x_{22}$	8		1.5	Нет
£3	200 p	0	6		0.2	Да

	Средний заказ	Жалоб за неделю	Дней не активен		Лет с нами	Ушел в отток?
$x_1$	$x_{11}$	$x_{12}$	$x_{13}$		0.1	$y_1$
$x_2$	$\infty_{21}$	$x_{22}$	8	•••	1.5	$y_2$
$x_3$	200 p	0	6	•••	0.2	<i>y</i> <sub>3</sub>

6

# Терминология

Выборка:  $\{x_1, ..., x_n\}$ 

Объекты выборки:

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

 $x_{i1}, ..., x_{id}$  - признаки і-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:

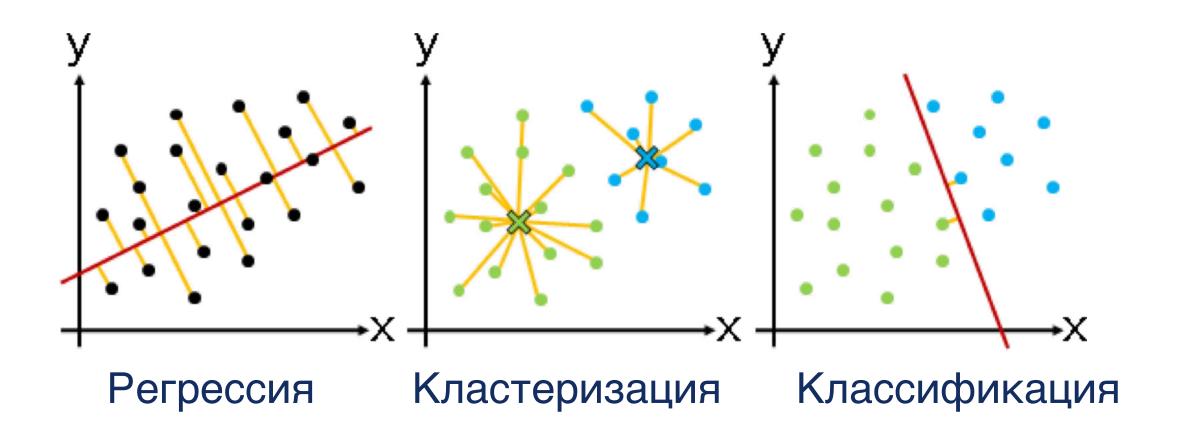
$$(X^{\ell}, Y^{\ell}) = (\{x_1, \dots, x_{\ell}\}, \{y_1, \dots, y_{\ell}\})$$

#### Тестовая выборка:

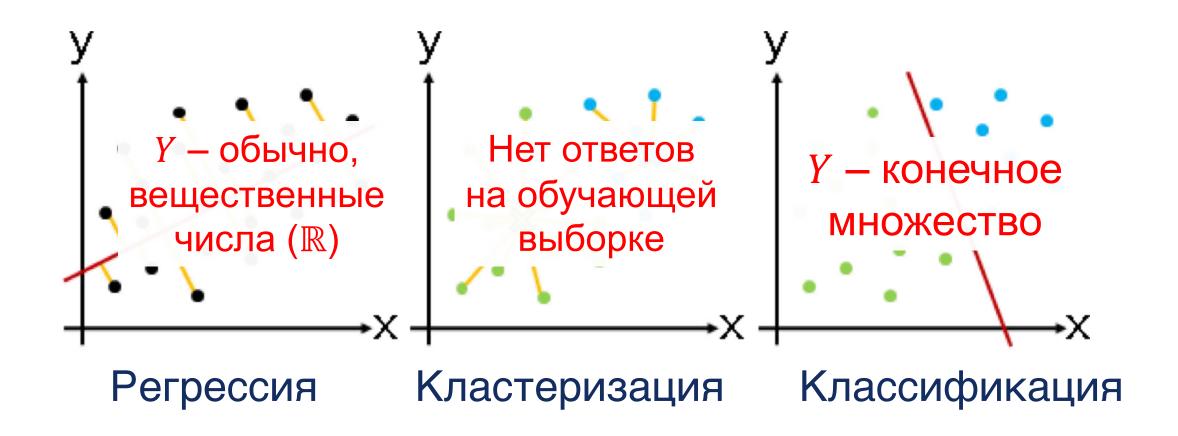
$$(X^t, Y^t) = (\{x_{\ell+1}, \dots, x_{\ell+t}\}, \{y_{\ell+1}, \dots, y_{\ell+t}\})$$

На обучающей выборке настраиваем алгоритм На тестовой – оцениваем качество прогнозирования

# Стандартные задачи



# Стандартные задачи



1. Порог по одному признаку

2. От пней к деревьям

3. Сложные границы и соседи

4. Плотность и наивный байес

## План лекции

# 1. Порог по одному признаку



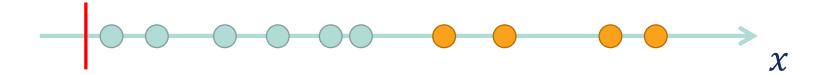
#### Выборка

Как подобрать порог по признаку в задаче бинарной классификации?



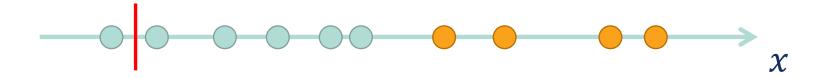
#### Выборка

Как подобрать порог по признаку в задаче бинарной классификации?



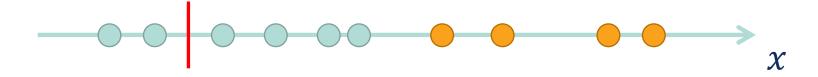
#### Выборка

Как подобрать порог по признаку в задаче бинарной классификации?



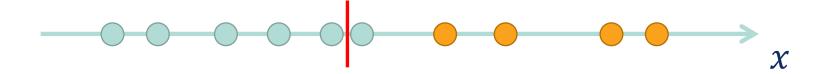
#### Выборка

Как подобрать порог по признаку в задаче бинарной классификации?



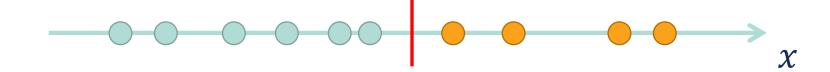
#### Выборка

Как подобрать порог по признаку в задаче бинарной классификации?



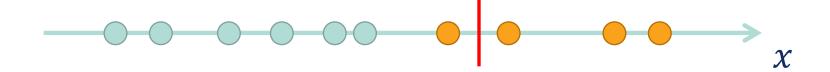
#### Выборка

Как подобрать порог по признаку в задаче бинарной классификации?



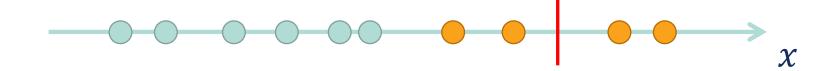
#### Выборка

Как подобрать порог по признаку в задаче бинарной классификации?



#### Выборка

Как подобрать порог по признаку в задаче бинарной классификации?



#### Выборка

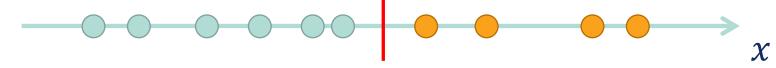
Как подобрать порог по признаку в задаче бинарной классификации?



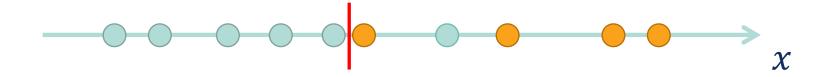
#### Выборка

Как подобрать порог по признаку в задаче бинарной классификации?

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

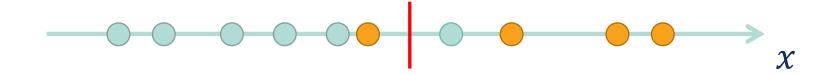


Часто есть несколько неплохих порогов:



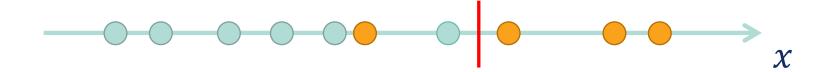
## Проблема выбора

Часто есть несколько неплохих порогов:



#### Проблема выбора

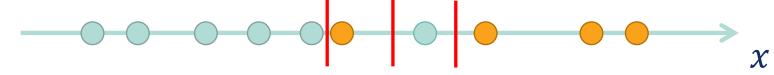
Часто есть несколько вариантов деления:



### Проблема выбора

### Усложнение модели

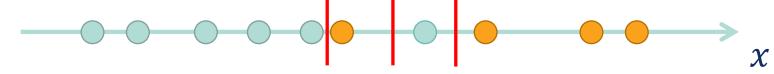
Если потребовать от модели максимальной точности



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

## Усложнение модели

Если потребовать от модели максимальной точности

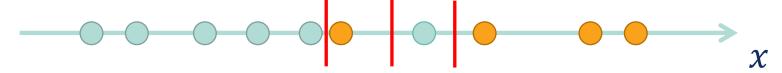


и не ограничивать количество порогов, можно было бы разделить выборку идеально

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

## Усложнение модели

Если потребовать от модели максимальной точности



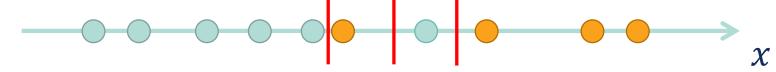
и не ограничивать количество порогов, можно было бы разделить выборку идеально

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



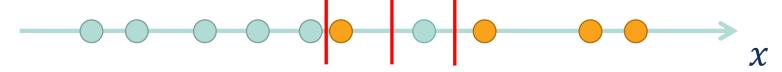
и не ограничивать количество порогов, можно было бы разделить выборку идеально

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



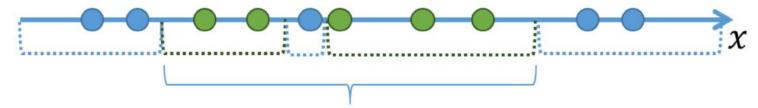
# Случай множества порогов

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:



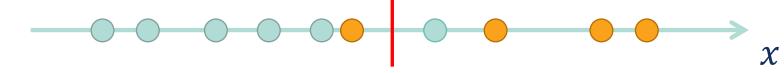
Возможно какие-то интервалы лучше объединить

#### Разделение неразделимо й выборки

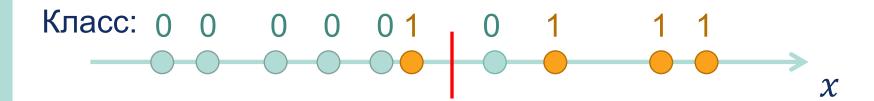
Предположим, выборка не разделима идеально:



Но нужно по-прежнему адекватно ее разделить:

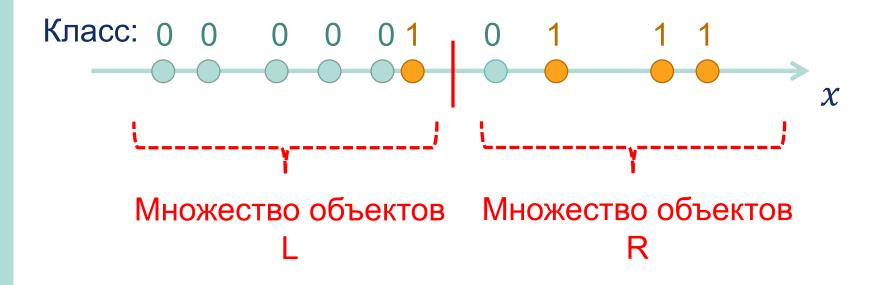


Как это записать?

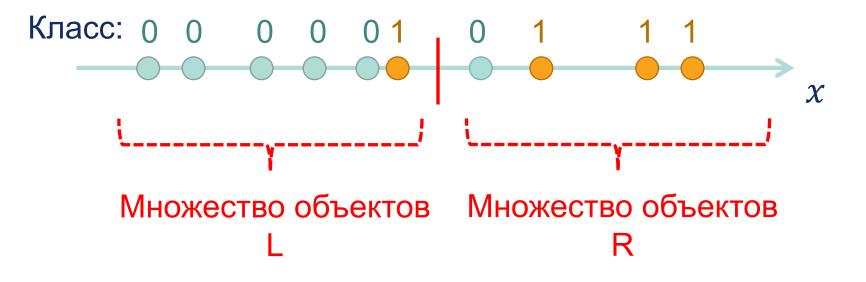


# Задача оптимизации

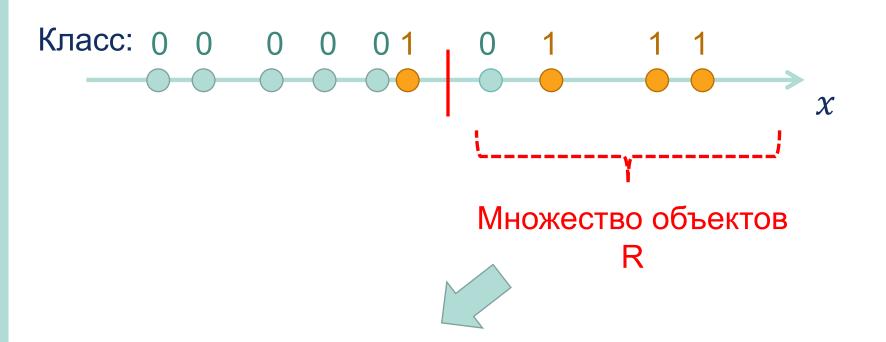
# Задача оптимизации



# Задача оптимизации

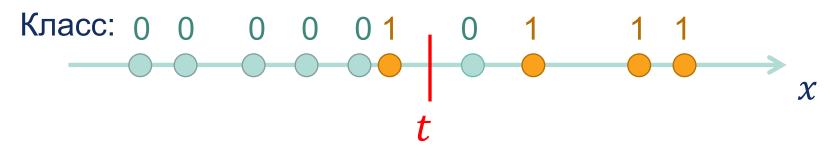


Чтобы разделить классы хорошо – нужно, чтобы и в L и в R преобладал только один класс



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  $p_1=\frac{3}{4}$ 

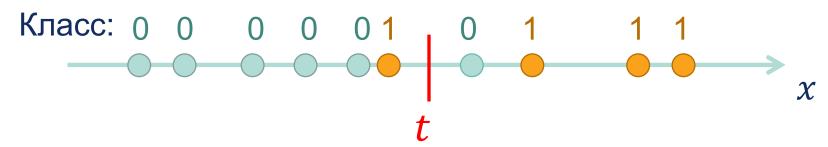
Как записать, что один из классов преобладает?



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

Например, так:

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



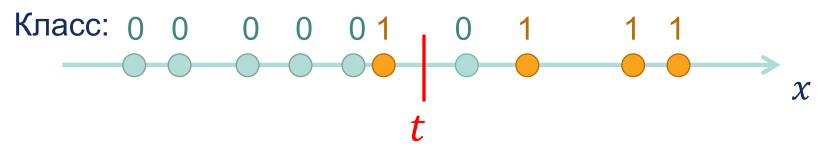
Как записать, что один из классов должен преобладать в R?

Например, так:

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

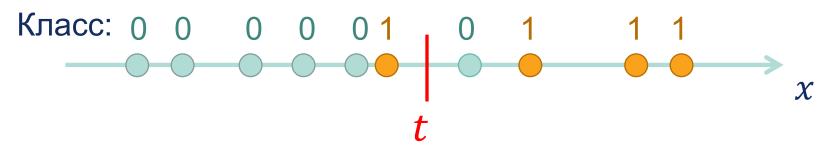
Или так:

$$1 - p_{max} \rightarrow \min_{t}$$



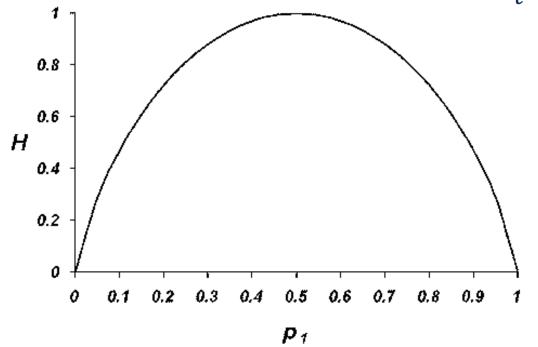
Другой вариант:

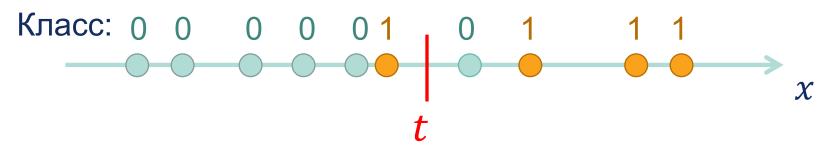
$$H(R) = -p_0 \ln p_0 - p_1 \ln p_1 \to \min_t$$



Другой вариант:

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



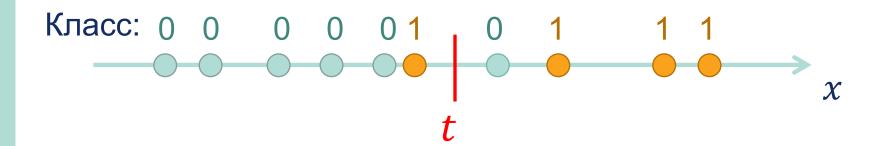


Другой вариант:

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

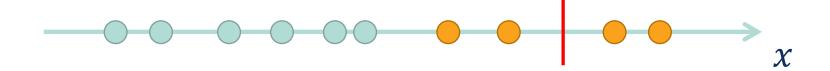
вопрос: какое основание у
логарифма в энтропии на
этом графике?

Р1



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

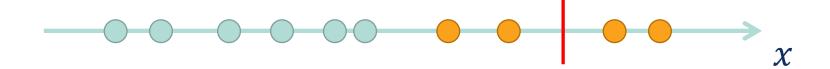
Но если смотреть только на R, можем нечаянно разделить выборку так:



#### **Уточнение**

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

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

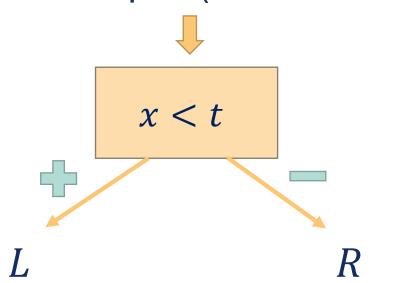


#### **Уточнение**

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

Значит надо учитывать обе части: R и L

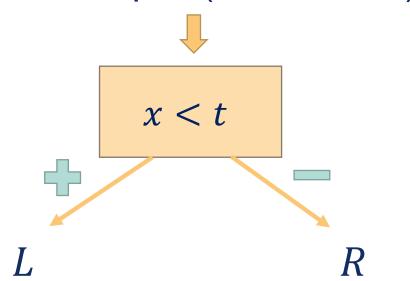
Выбор разбиения Вся выборка (п объектов)



$$G(t) = H(L) + H(R) \rightarrow \min_{t}$$

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

Выбор разбиения Вся выборка (п объектов)

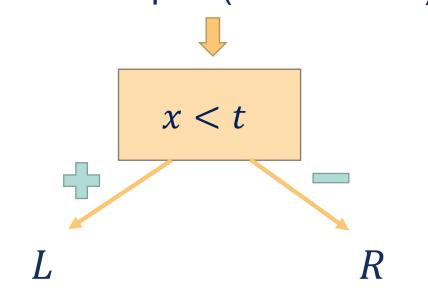


$$G(t) = H(L) + H(R) \rightarrow \min_{t}$$

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

#### Выбор разбиения

Вся выборка (п объектов)



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

H(R) — мера «неоднородности» множества R

Критерии построения разбиений

# Критерии построения разбиений

H(R) — мера «неоднородности» множества R

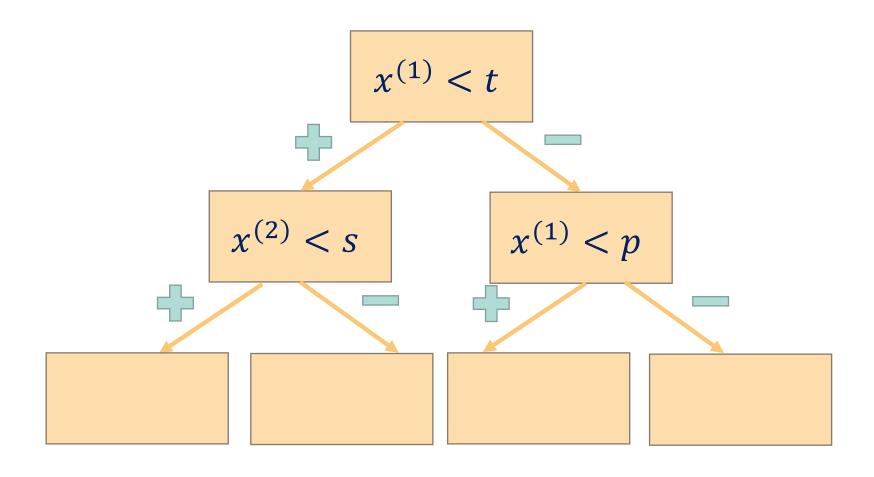
Варианты этой функции:

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

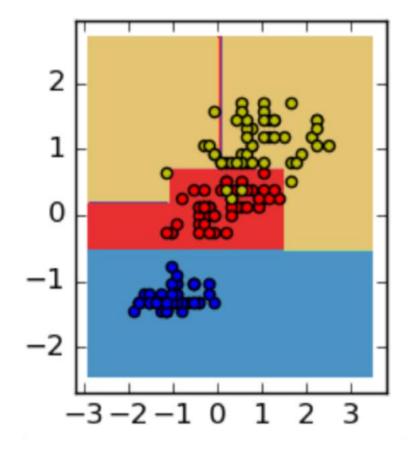
Если признаков и порогов больше



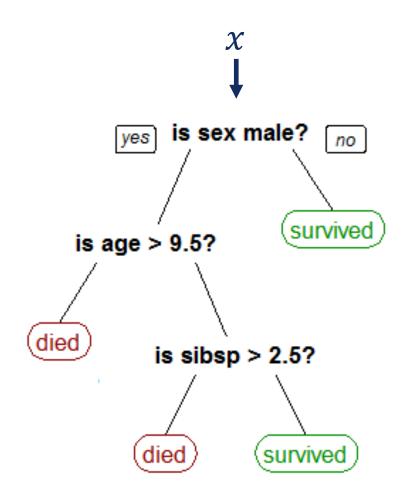
### 2. От пней к деревьям

#### Пример границ для 3 классов при 2 признаках:

### **Границы** классов



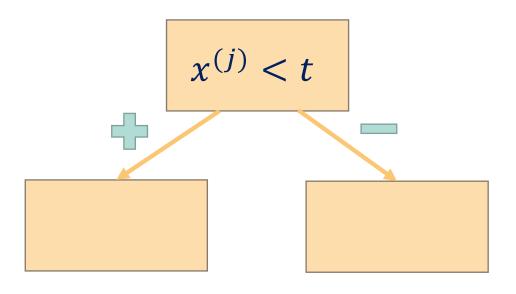
#### Решающее дерево



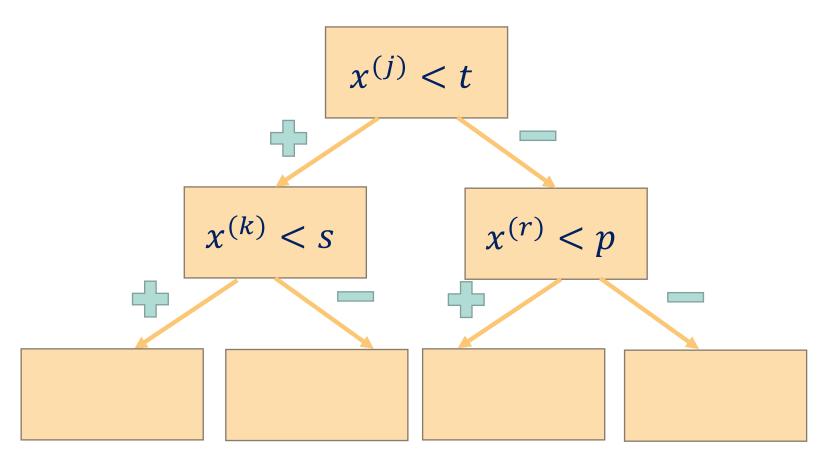
 $x^{(j)} < t$ 

#### Рекурсивное построение

#### Рекурсивное построение

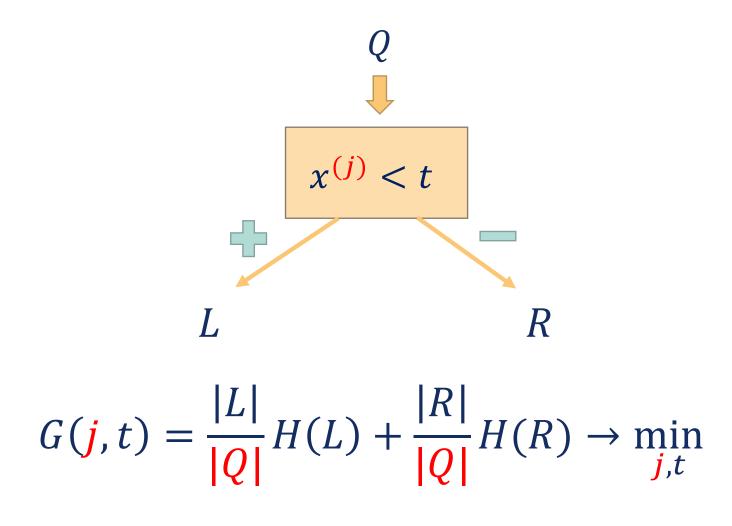


#### Рекурсивное построение



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

Выбор разбиения



# Критерии построения разбиений

H(R) — мера «неоднородности» множества R

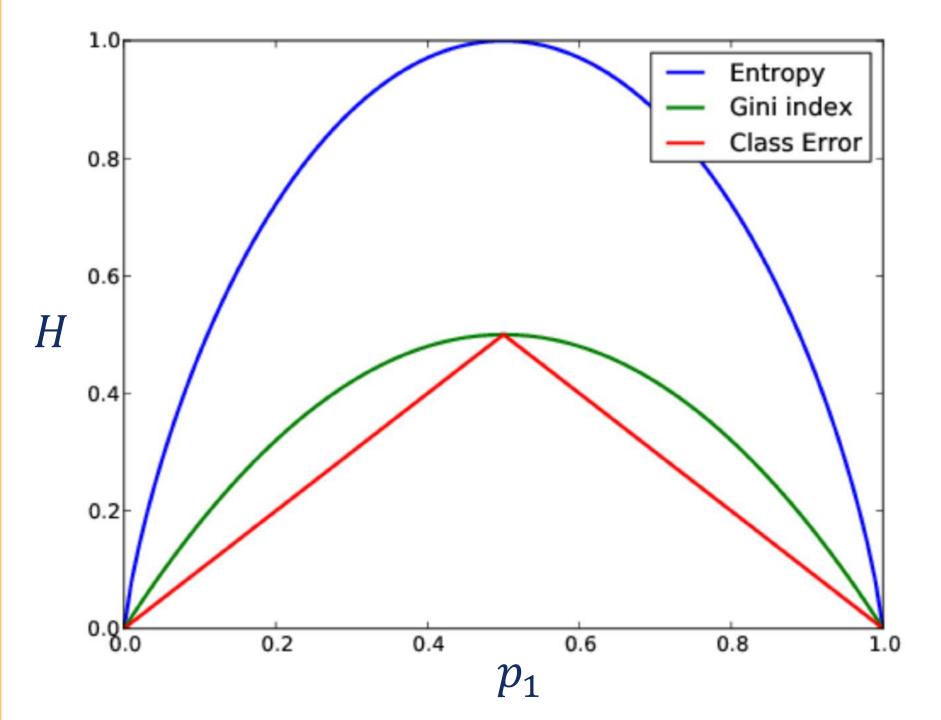
Варианты этой функции:

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

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

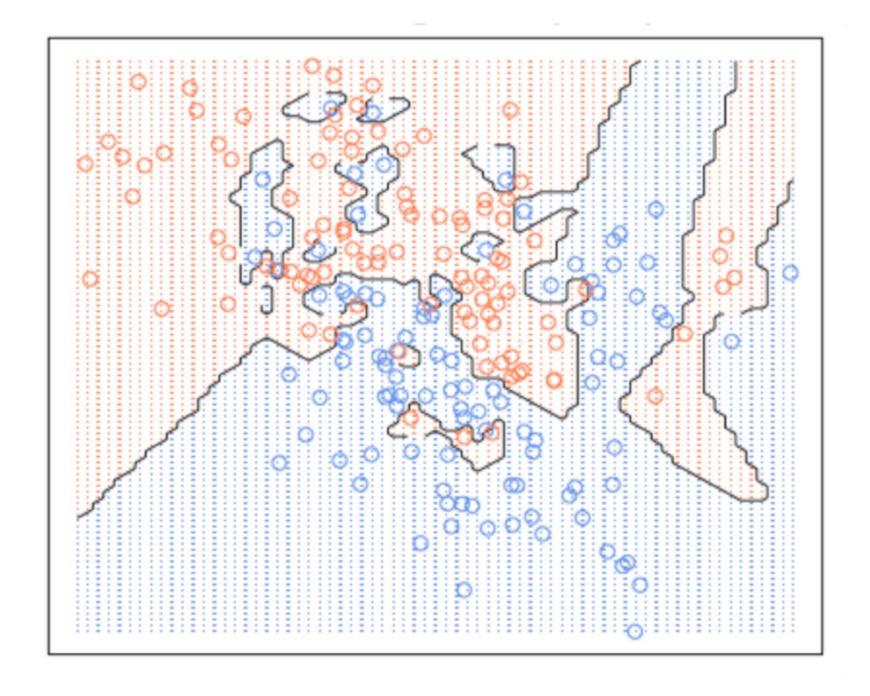
3) Gini criteria:  $H(R) = 1 - p_0^2 - p_1^2 = 2p_0p_1$ 

Критерии построения разбиений

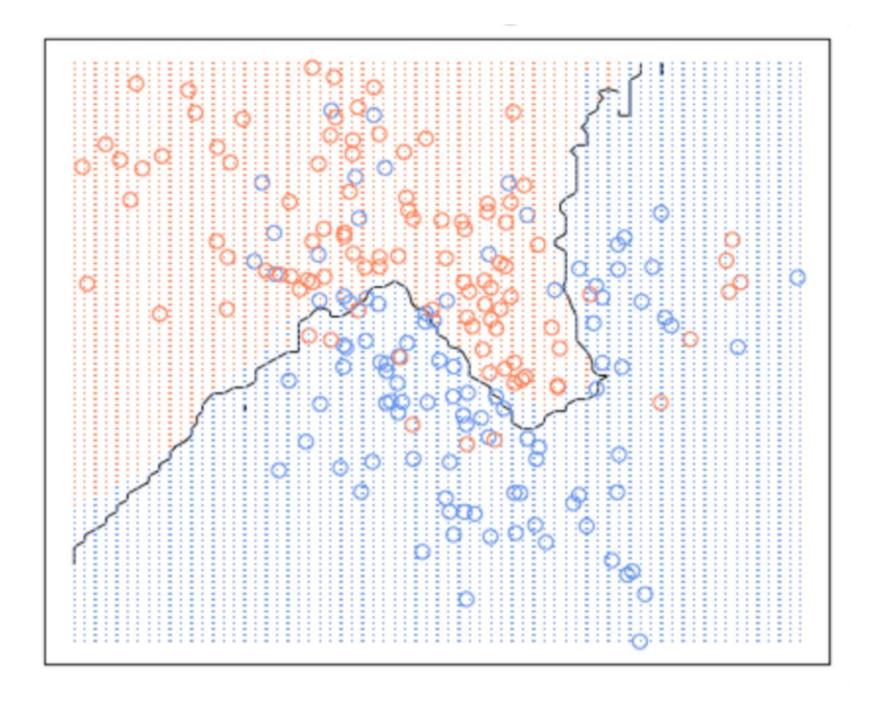


## 3. Сложные границы и соседи

## Сложные границы



## Сложные границы

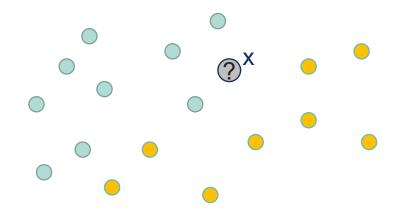


## Сложные границы



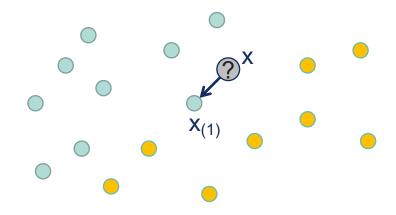
#### Пример классификации:

Метод ближайшего соседа



#### Пример классификации:

Метод ближайшего соседа (1NN)



#### 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:  $d(x_1, x_2) = d(x_2, x_1) = ||x_2 - x_1||$ 

## Что такое расстояние

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

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

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):

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

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

## Примеры норм

#### Примеры норм:

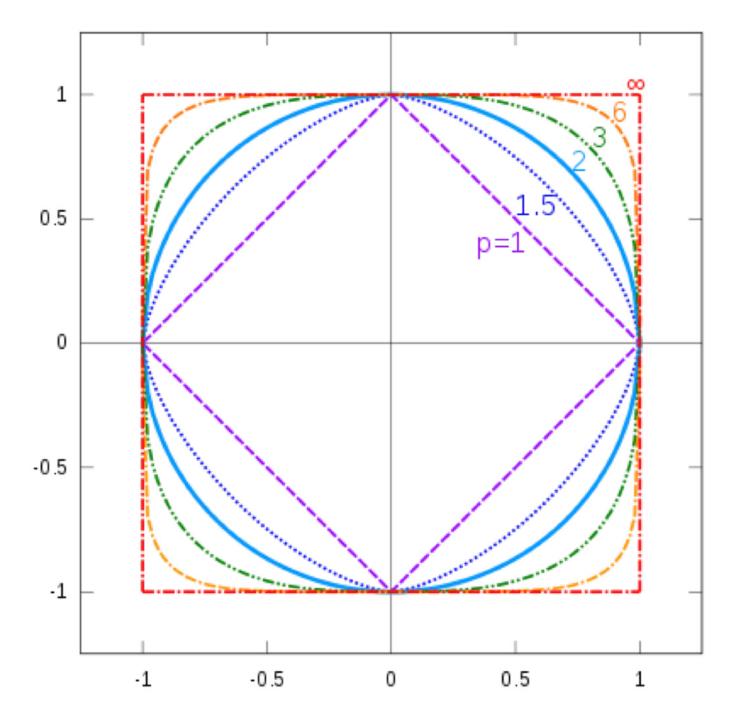
$$||x||_{\ell_2} = \sqrt{(x^{(1)})^2 + \dots + (x^{(d)})^2}$$

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

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

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

#### Примеры норм



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

#### Функция 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

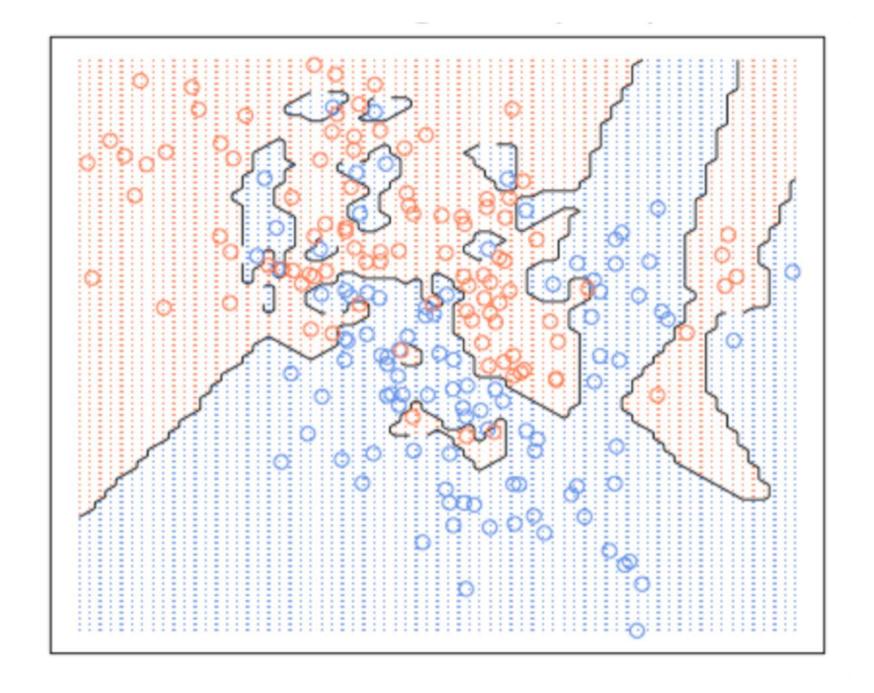
Но можно ввести расстояние каким-то своим особым способом или вообще ввести не расстояние, а функцию близости

Пример: Косинусная мера близости (cosine similarity)

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

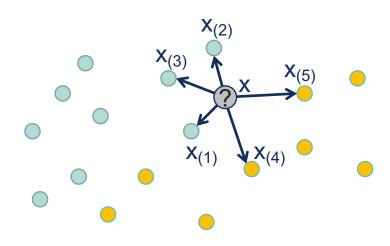
$$x_1 = (x_1^{(1)}, \dots, x_1^{(d)})$$
  $x_2 = (x_2^{(1)}, \dots, x_2^{(d)})$  
$$\alpha = sim(x_1, x_2) = cos \alpha$$

#### Границы классов в 1NN



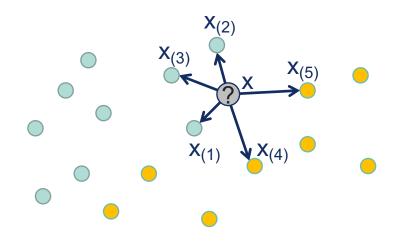
Пример классификации (k = 5):

Метод k ближайших соседей (kNN)



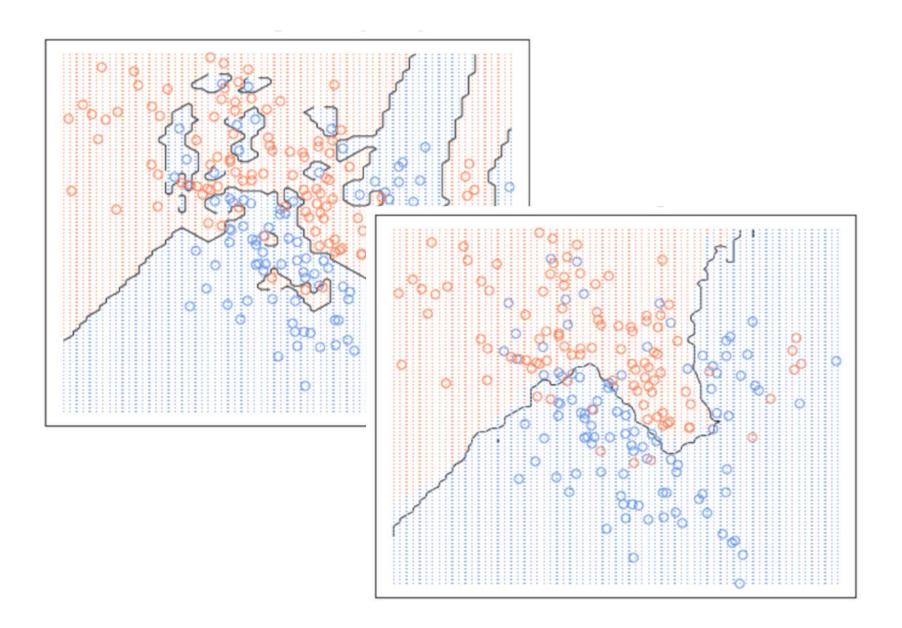
## Метод k ближайших соседей (kNN)

Пример классификации (k = 5):



Выбираем класс, который преобладает

## Сглаживание границ



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

Вопрос про настройку параметров

## 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

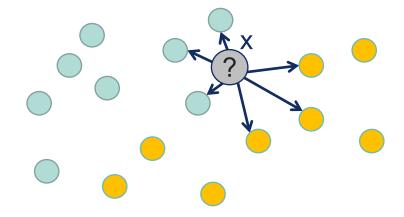
Какое количество соседей оптимально брать с точки зрения **качества работы на обучающей выборке?** 

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

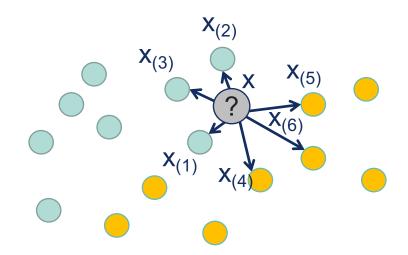
Пример классификации (k = 6):

#### kNN с весами



Пример классификации (k = 6):

#### kNN с весами



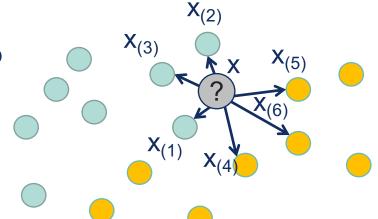
#### kNN с весами

#### Пример классификации (k = 6):

Веса можно определить как функцию от соседа или его номера:

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

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



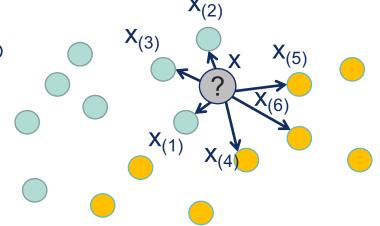
#### kNN с весами

#### Пример классификации (k = 6):

Веса можно определить как функцию от соседа или его номера:

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

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



$$Z_{\bullet} = \frac{W(X_{(1)}) + W(X_{(2)}) + W(X_{(3)})}{W(X_{(1)}) + W(X_{(2)}) + W(X_{(3)}) + W(X_{(4)}) + W(X_{(5)}) + W(X_{(6)})}$$

$$Z_{\bullet} = \frac{W(X_{(4)}) + W(X_{(5)}) + W(X_{(6)})}{W(X_{(1)}) + W(X_{(2)}) + W(X_{(3)}) + W(X_{(4)}) + W(X_{(5)}) + W(X_{(6)})}$$

#### kNN с весами

#### Пример классификации (k = 6):

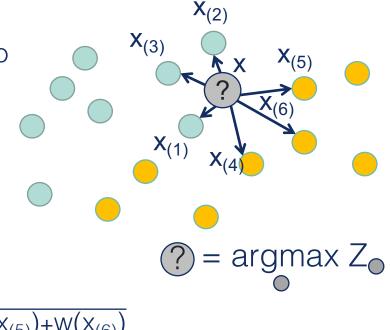
Веса можно определить как функцию от соседа или его номера:

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

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

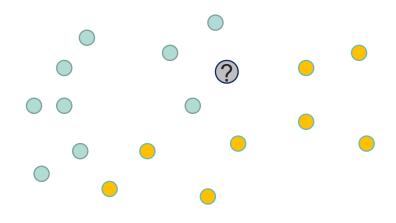
$$Z_{\odot} = \frac{W(X_{(1)}) + W(X_{(2)}) + W(X_{(3)})}{W(X_{(1)}) + W(X_{(2)}) + W(X_{(3)}) + W(X_{(4)}) + W(X_{(5)}) + W(X_{(6)})}$$

$$Z_{\bullet} = \frac{W(X_{(4)}) + W(X_{(5)}) + W(X_{(6)})}{W(X_{(1)}) + W(X_{(2)}) + W(X_{(3)}) + W(X_{(4)}) + W(X_{(5)}) + W(X_{(6)})}$$



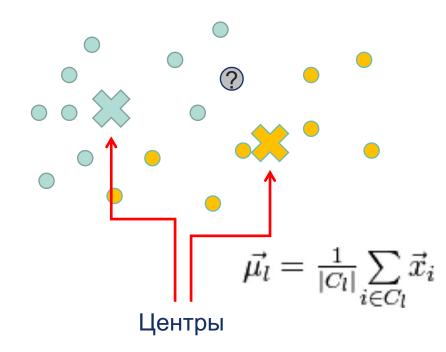
## Центроидный классификатор

Другой похожий алгоритм



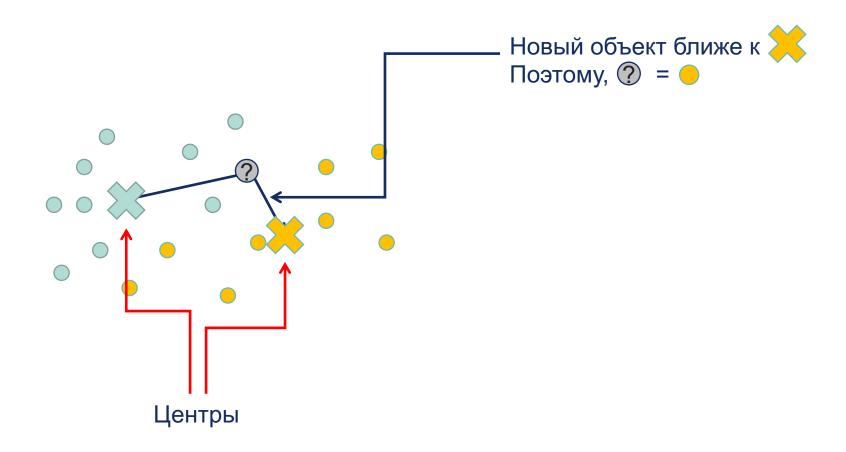
## Центроидный классификатор

Другой похожий алгоритм



## Центроидный классификатор

Другой похожий алгоритм



## 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

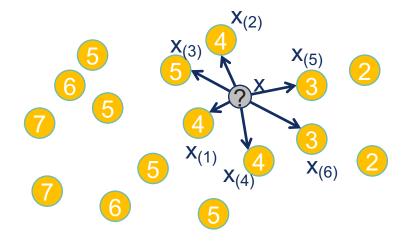
Можно брать расстояния до самих объектов класса, можно до центров класса.

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

## Все это применимо и в регрессии

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

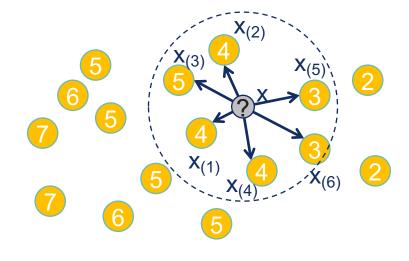
### Обобщение для регрессии



#### 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:

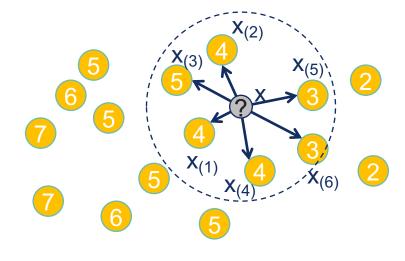
$$w(x_{(i)}) = w(i)$$

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

## 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:

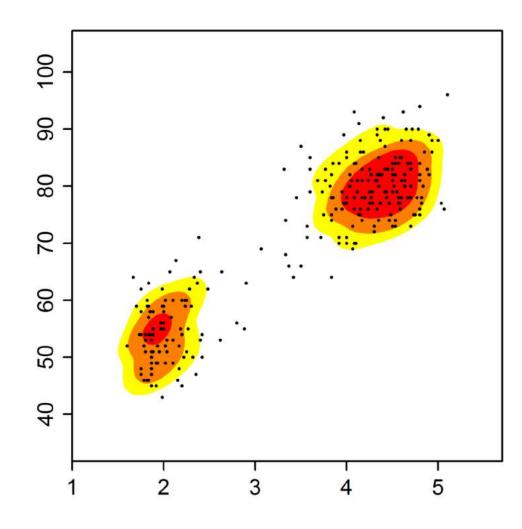
$$w(x_{(i)}) = w(i)$$

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

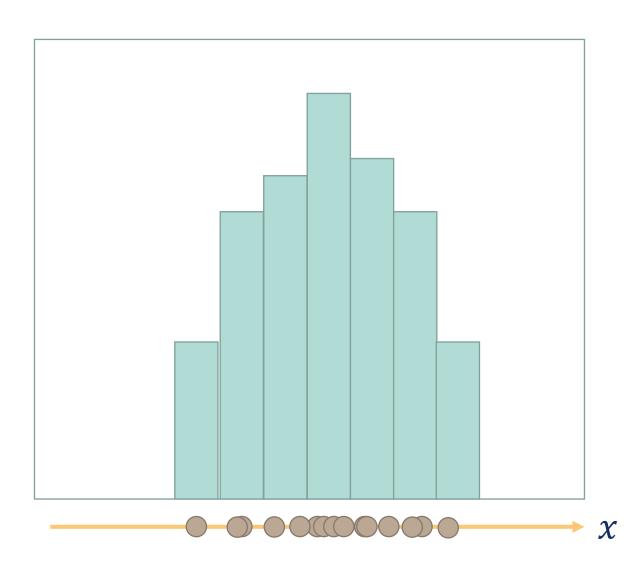
$$= \frac{4 \cdot w(x_{(1)}) + 4 \cdot w(x_{(2)}) + 5 \cdot w(x_{(3)}) + 4 \cdot w(x_{(4)}) + 3 \cdot w(x_{(5)}) + 3 \cdot w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

# 4. Плотность и наивный байес

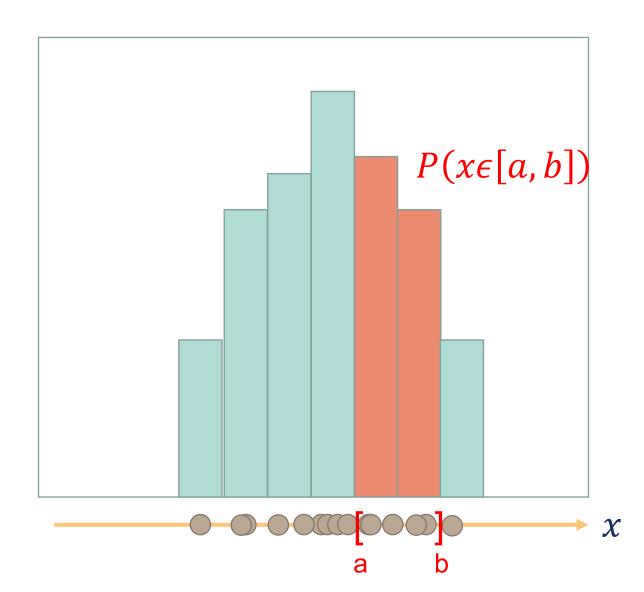
## Пример для двух признаков



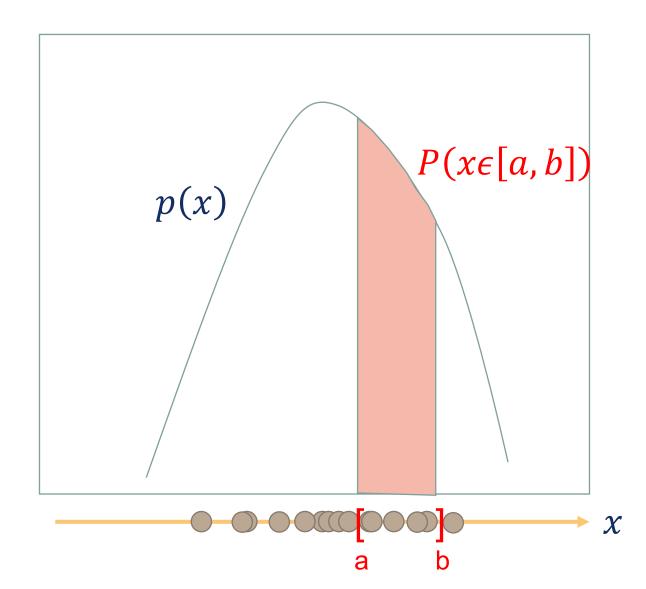
## Одномерный случай



## Одномерный случай



## Плотность распределения

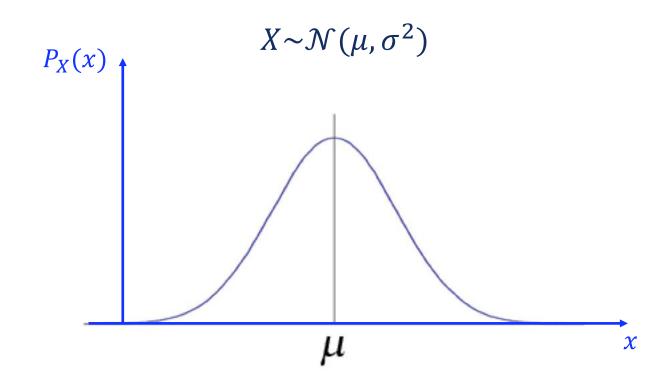


#### Подходы к оценке плотности

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

#### Пример оценки параметров

#### Нормальное распределение



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

#### Пример оценки параметров

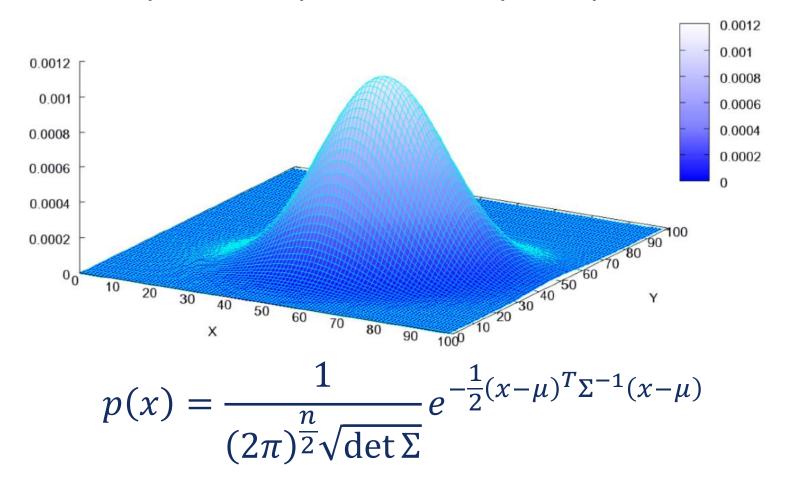
## Нормальное распределение

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

другой вариант оценки для 
$$\sigma^2$$
: 
$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

#### Многомерный пример

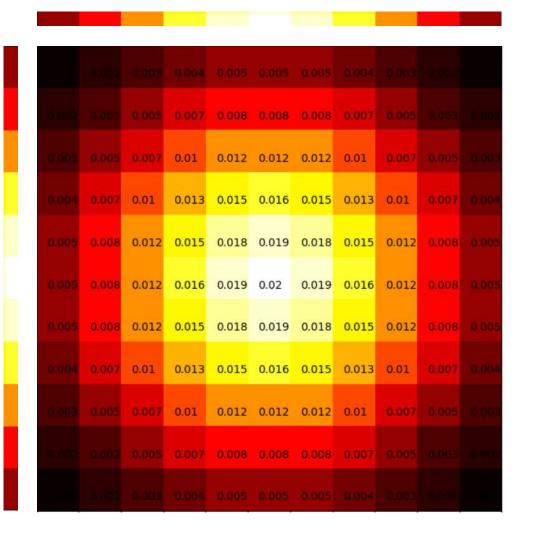
#### Многомерное нормальное распределение

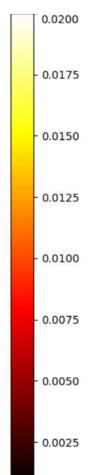


Очень много параметров: вектор средних  $\mu$  и матрица ковариаций  $\Sigma$ 

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

Произведение одномерных плотностей





## На самом деле так можно не всегда

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

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

#### Наивная гипотеза

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?

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

#### 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?

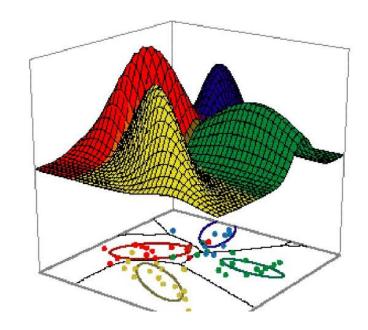
OTBET: 
$$\frac{1}{2} \cdot \frac{7}{10}$$

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

#### 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:

Классификация



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

#### Наивный байесовский классификатор

$$p(x|y)$$
 - Плотность класса у   
Считаем, что  $p(x|y) = p\big(x^{(1)}|y\big) ... p\big(x^{(d)}|y\big)$ 

#### 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:

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

1. Порог по одному признаку

2. От пней к деревьям

3. Сложные границы и соседи

4. Плотность и наивный байес

#### План лекции

### Домашнее задание 1

#### Теория

Теоретическое задание из 5 обязательных задач и одной задачи со звездочкой Оформлять в LaTeX или Word, присылать PDFкой

#### Контест

Сделайте как минимум одну осмысленную посылку в контест Titanic, применив Python библиотеку sklearn для построения модели

Дедлайн: 25 сентября

#### Работа «на опережение»

- 1. Прорешайте «листок» по линейной регрессии
- 2. Сравните в Titanic разные модели и выясните, какой алгоритм работает лучше всего

Решения теории (вкл. листок) + отчет текстом в свободной форме про решение контеста присылайте на <u>viktor.kantor@phystech.edu</u> **одним письмом** с заголовком [MIPT ML 23] HW1 Ваша Фамилия

# Машинное обучение

Лекция 1



https://t.ly/YeiWy - группа для обсуждений

https://t.ly/11AZK - канал для анонсов

Приложение Минимизация риска и байесовская классификация



## Байесовский классификатор

По известному вектору признаков х алгоритм относит объект к классу а(х) по правилу:

$$a(x) = argmax_y P(y|x)$$

## Байесовский классификатор

$$a(x) = argmax_y P(y|x)$$

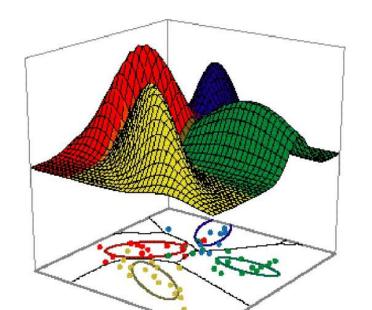
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$a(x) = argmax_y P(x|y) P(y)$$

## Байесовский классификатор

$$a(x) = argmax_y P(x|y) P(y)$$

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



# Зачем нам понадобилась теорема Байеса

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

# 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:

$$a(x) = argmax_y P(x|y) P(y)$$

### 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 х довольно трудно

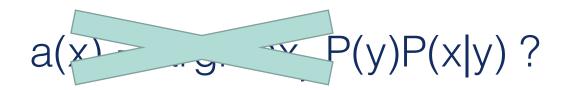
## Байесовская регрессия

 $a(x) = argmax_y P(y)P(x|y) ?$ 

х – признаковое описание

у – прогнозируемая величина

### Байесовская регрессия



х – признаковое описание

у – прогнозируемая величина

Вряд ли получится восстановить P(x|y)

### Байесовская регрессия

$$a(x) = P(y)P(x|y)$$
?

х – признаковое описание

у – прогнозируемая величина

Вряд ли получится восстановить P(x|y)

 $a(x) = argmax_y P(y|x)$  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:

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

• Сумма модулей отклонений:

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

### Более общий подход

- Для объекта х мы делаем прогноз а(х)
- Правильный ответ на этом объекте у
- 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 х ответ, который минимизирует ожидаемую ошибку:

$$a(x) = argmin_s R(s, x)$$

# Оптимальный байесовский классификатор

Для классификации:

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

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

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

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

# Оптимальный байесовский регрессор

Для регрессии:

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

$$a(x) = arg \min_{S} R(s, x) = arg \min_{S} \int_{y \in Y} L(y, s) p(y|x) dy$$

### Функционал среднего риска

- Можно рассмотреть  $R(a) = \mathbb{E}_x R(a(x), x)$  (по всем х 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:

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

### Функционал среднего риска

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

• Можно оценить R(a) снизу:

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

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

# Оптимальный байесовский классификатор

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

Если 
$$L(s, y) = [y != s]$$
:

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

Если L(s, y) = [y != s]:

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

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

Если L(s, y) = [y != s]:

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

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

$$P(s|x) \to \max_{s}$$

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

$$\int_{Y} (t - y)^2 p(y|x) dy \to \min_{t}$$

$$\int_{Y} (t - y)^{2} p(y|x) dy \to \min_{t}$$

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

$$\int_{Y} (t - y)^{2} p(y|x) dy \to \min_{t}$$

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

$$= 2 \left( \int_{Y} t p(y|x) dy - \int_{Y} y p(y|x) dy \right) = 0$$

$$\int_{Y} (t - y)^{2} p(y|x) dy \to \min_{t}$$

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

$$= 2 \left( \int_{Y} t p(y|x) dy - \int_{Y} y p(y|x) dy \right) = 0$$

$$a(x) = t = \int_{Y} y p(y|x) dy = \mathsf{E}(y|x)$$

#### Абсолютное отклонение

$$\int\limits_{Y} |t-y|p(y|x)dy \to \min\limits_{t}$$

$$\int_{Y} |t - y| p(y|x) dy \to \min_{t}$$

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

$$\begin{split} \int\limits_{Y} |t-y| p(y|x) dy &\to \min_{t} \\ \frac{\partial}{\partial t} \int\limits_{Y} |t-y| p(y|x) dy &= \frac{\partial}{\partial t} \int\limits_{Y\backslash\{t\}} |t-y| p(y|x) dy = \\ &= \int\limits_{Y\backslash\{t\}} \mathrm{sign}(t-y) p(y|x) dy = \int\limits_{\{t>y\}} p(y|x) dy - \int\limits_{\{t< y\}} p(y|x) dy = \end{split}$$

$$\begin{split} \int\limits_{Y} |t-y| p(y|x) dy &\to \min_{t} \\ \frac{\partial}{\partial t} \int\limits_{Y} |t-y| p(y|x) dy &= \frac{\partial}{\partial t} \int\limits_{Y \backslash \{t\}} |t-y| p(y|x) dy = \\ &= \int\limits_{Y \backslash \{t\}} \mathrm{sign}(t-y) p(y|x) dy = \int\limits_{\{t>y\}} p(y|x) dy - \int\limits_{\{t< y\}} p(y|x) dy = \\ &= \mathsf{P}(\{t>y\}|x) - \mathsf{P}(\{t< y\}|x) = 0. \end{split}$$

$$\begin{split} \int\limits_{Y} |a(x)-y| p(y|x) dy \\ \frac{\partial}{\partial t} \int\limits_{Y} |t-y| p(y|x) dy &= \frac{\partial}{\partial t} \int\limits_{Y\backslash \{t\}} |t-y| p(y|x) dy = \\ &= \int\limits_{Y\backslash \{t\}} \operatorname{sign}(t-y) p(y|x) dy = \int\limits_{\{t>y\}} p(y|x) dy - \int\limits_{\{t< y\}} p(y|x) dy = \\ &= \mathsf{P}(\{t>y\}|x) - \mathsf{P}(\{t< y\}|x) = 0. \end{split}$$

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

$$\begin{split} \int\limits_{Y}|a(x)-y|p(y|x)dy \\ \frac{\partial}{\partial t}\int\limits_{Y}|t-y|p(y|x)dy &= \frac{\partial}{\partial t}\int\limits_{Y\backslash\{t\}}|t-y|p(y|x)dy = \\ &= \int\limits_{Y\backslash\{t\}}\mathrm{sign}(t-y)p(y|x)dy = \int\limits_{\{t>y\}}p(y|x)dy - \int\limits_{\{t< y\}}p(y|x)dy = \\ &= \mathsf{P}(\{t>y\}|x) - \mathsf{P}(\{t< y\}|x) = 0. \end{split}$$

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

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

$$Y = \{0; 1\} \qquad L(y, a(x)) = -y \ln a(x) - (1 - y) \ln(1 - a(x))$$
$$\sum_{y \in Y} (-y \ln t - (1 - y) \ln(1 - t)) P(y|x) \to \min_{t}$$

$$Y = \{0; 1\} \qquad L(y, a(x)) = -y \ln a(x) - (1 - y) \ln(1 - a(x))$$
$$\sum_{y \in Y} (-y \ln t - (1 - y) \ln(1 - t)) P(y|x) \to \min_{t}$$
$$P(1|x) = p$$

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

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

$$P(1|x) = p \qquad -(1 - p) \ln(1 - t) - p \ln t \to \min_{t}$$

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

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

$$-(1 - p) \ln(1 - t) - p \ln t \to \min_{t}$$

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

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

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

$$P(1|x) = p \qquad -(1 - p) \ln(1 - t) - p \ln t \to \min_{t}$$

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

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

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

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

$$P(1|x) = p \qquad -(1 - p) \ln(1 - t) - p \ln t \to \min_{t}$$

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

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

#### 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

## Машинное обучение

Лекция 1



https://t.ly/YeiWy - группа для обсуждений

https://t.ly/11AZK - канал для анонсов