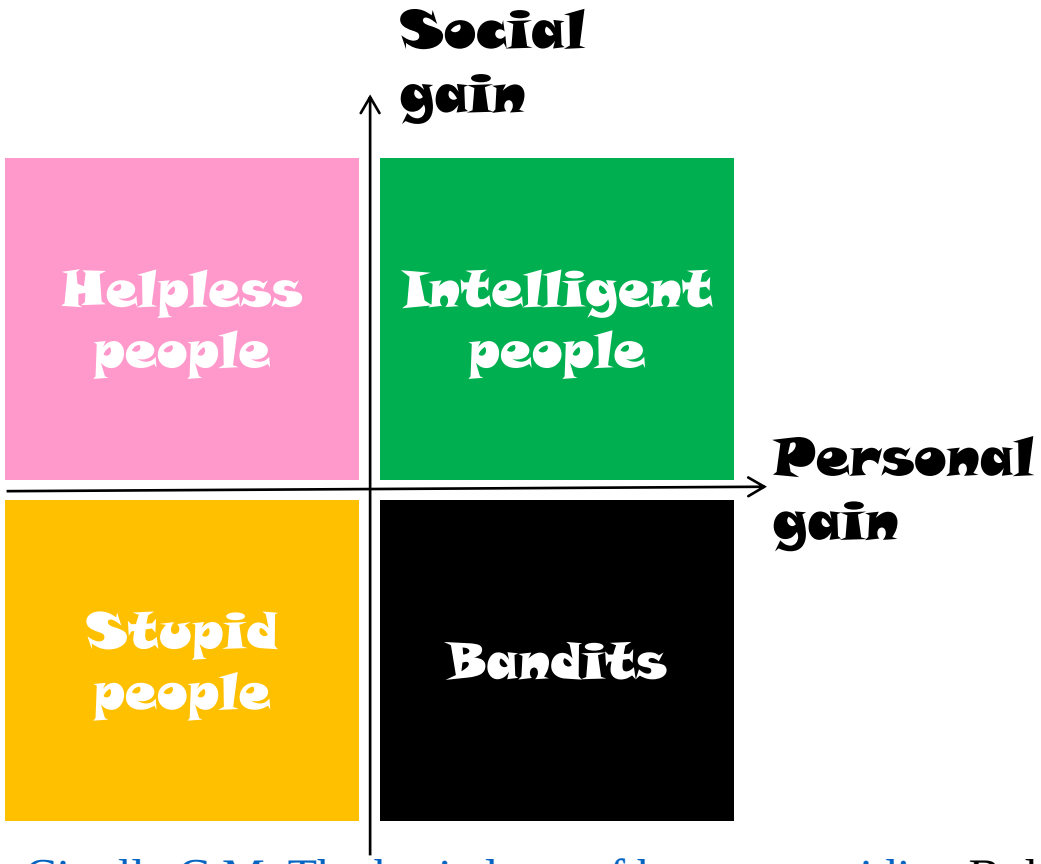
Задача классификации данных



Классификация – нить Ариадны в лабиринте природы.

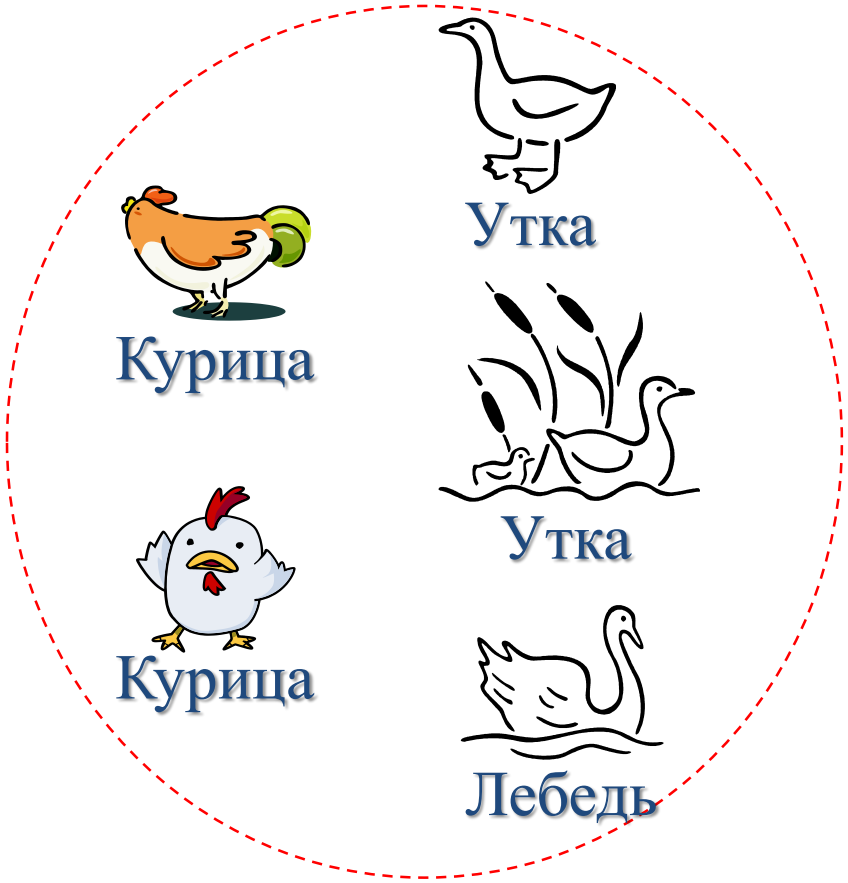
Жорж Санд

Cipolla C.M. The basic laws of human stupidity. Bologna: il Muli, 2011 (текст)

Содержание

- Основные понятия
- Деревья решений
- Байесовская классификация
- **Классификация по ближайшим соседям**
- Оценка качества классификации
- Ансамблевая классификация

Классификация по *k* ближайшим соседям (*k* Nearest Neighbors)



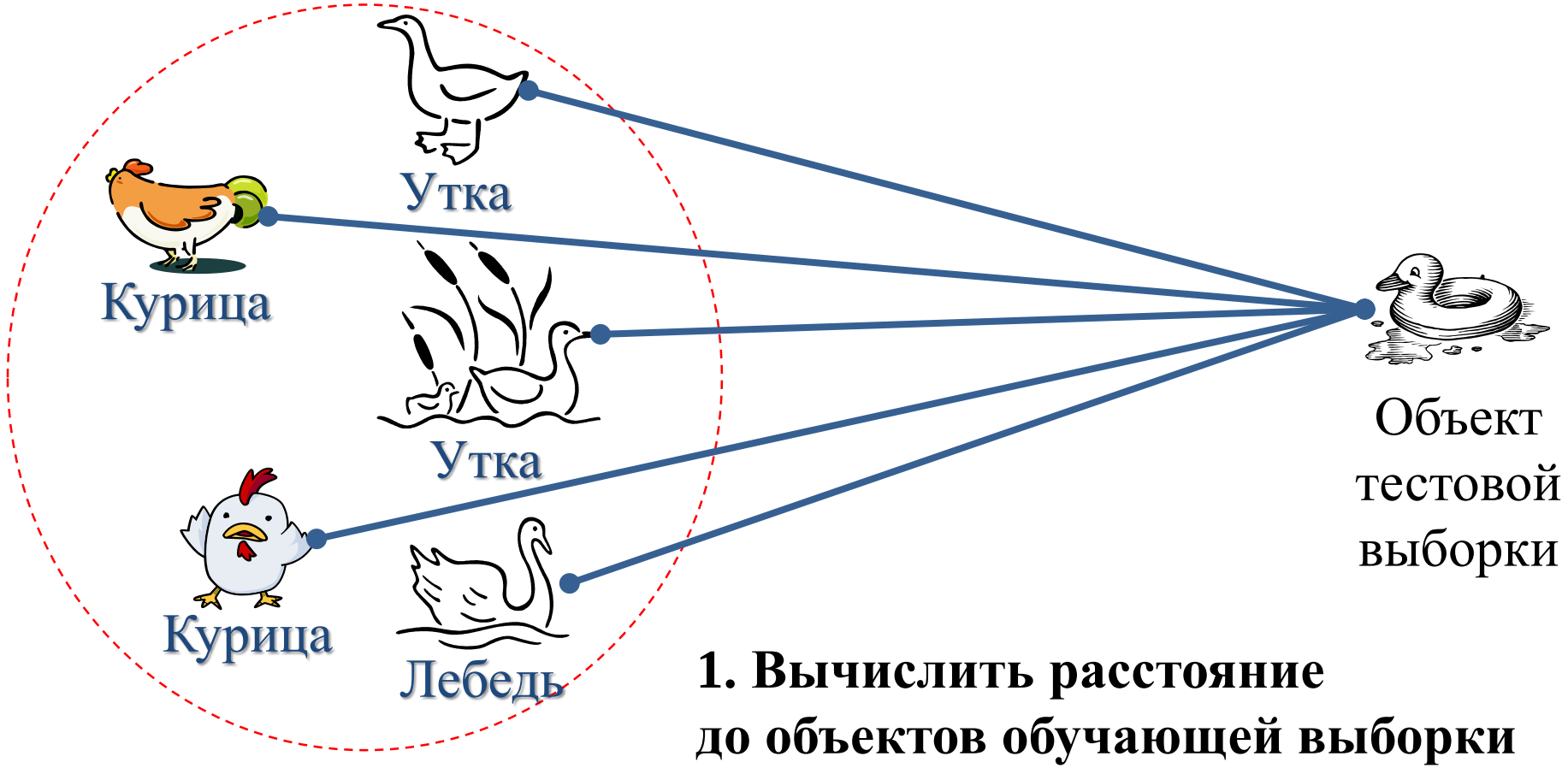
Обучающая выборка

If it looks like a duck, swims like a duck and quacks like a duck, then it probably is a duck.

Объект тестовой выборки

Число ближайших соседей
$\boldsymbol{k}$ ($\boldsymbol{k} \geq 1$)

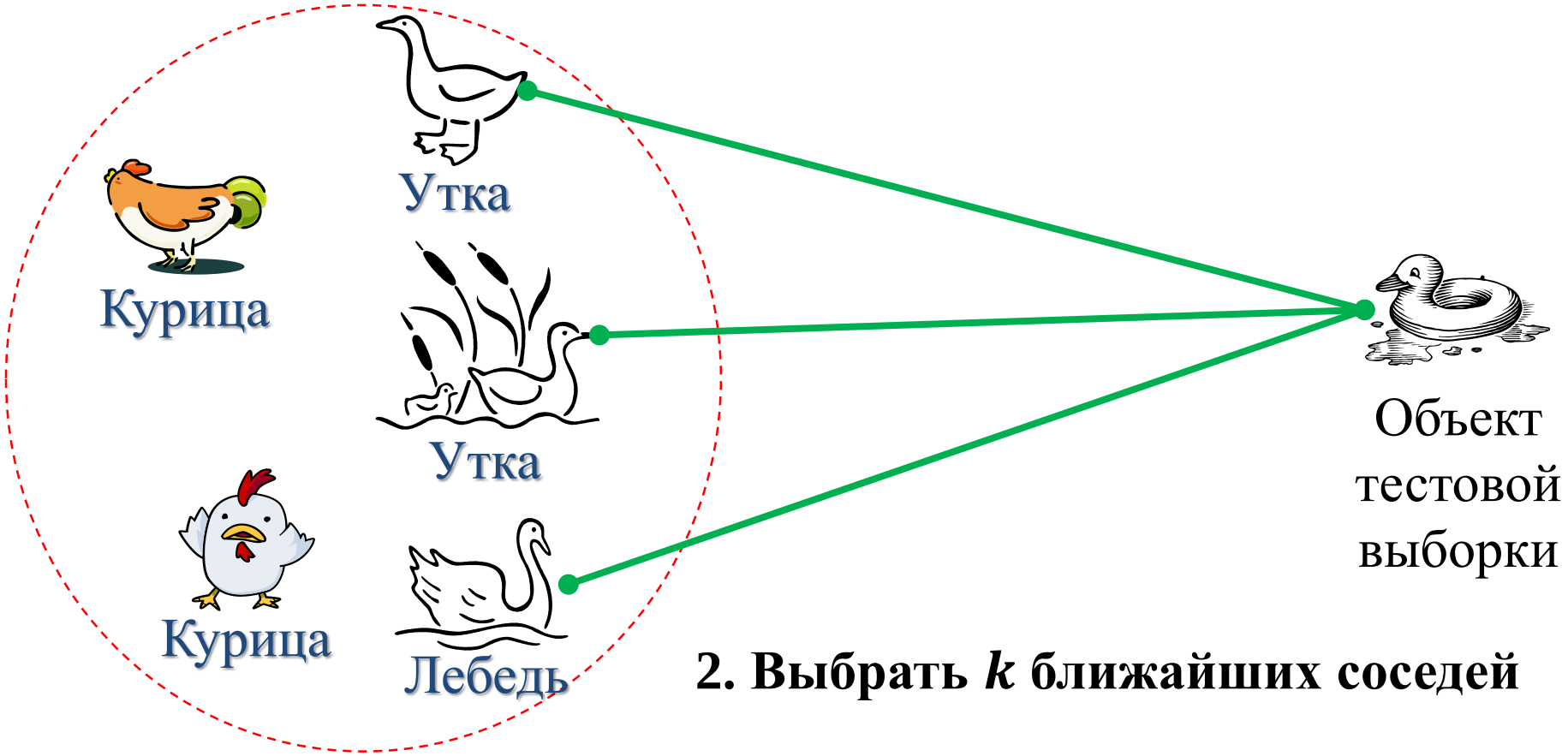
Классификация по *k* ближайшим соседям (*k* Nearest Neighbors)



Обучающая выборка

1. Вычислить расстояние до объектов обучающей выборки

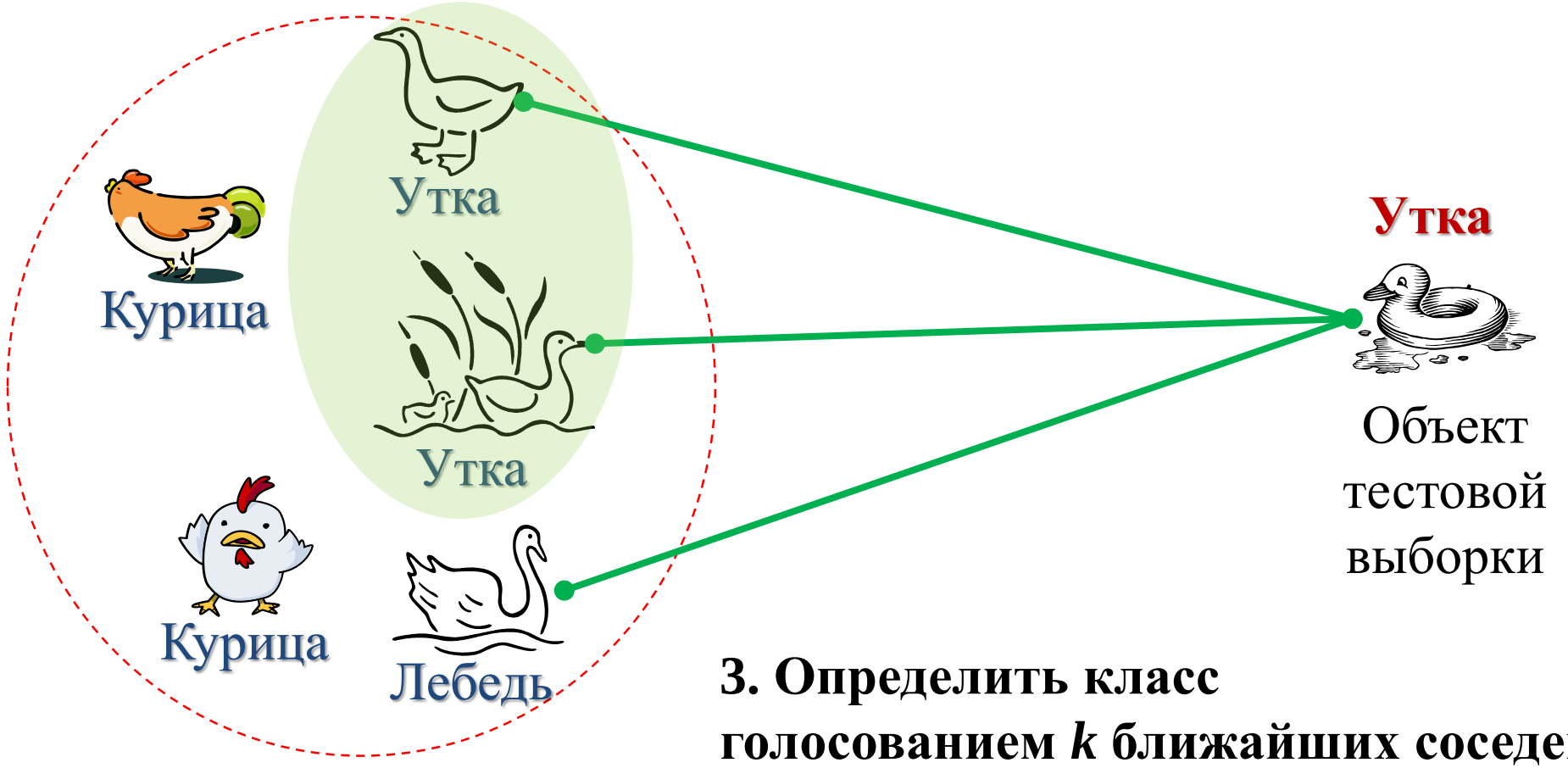
Классификация по *k* ближайшим соседям (*k* Nearest Neighbors)



2. Выбрать *k* ближайших соседей

Обучающая выборка

Классификация по *k* ближайшим соседям (*k* Nearest Neighbors)



Обучающая выборка

3. Определить класс голосованием *k* ближайших соседей

Параметры классификации по k ближайшим соседям

- Функция расстояния (мера схожести): $d: X \times X \to \mathbb{R}$
  1. Аксиома тождества: $d(x, x) = 0$
  2. Аксиома положительности: $d(x, y) \geq 0$
  3. Аксиома симметрии: $d(x, y) = d(y, x)$
  4. Аксиома треугольника: $d(x, z) \leq d(x, y) + d(y, z)$
- Количество соседей $\boldsymbol{k}$
- Способ голосования

Выбор (релевантной) функции расстояния

- Евклидово расстояние
  $$d(x,y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$
- Расстояние городских кварталов (Манхэттенское расстояние)
  $$d(x,y) = |x_1 - y_1| + \cdots + |x_n - y_n|$$
- Расстояние Чебышёва
  $$d(x,y) = \max\{|x_1 - y_1|, \ldots, |x_n - y_n|\}$$
- Метрика французского метро
  $$d(x,y) = \begin{cases} \|x\| + \|y\|, x \neq \lambda y \\ \|x - y\|, x = \lambda y \end{cases}$$

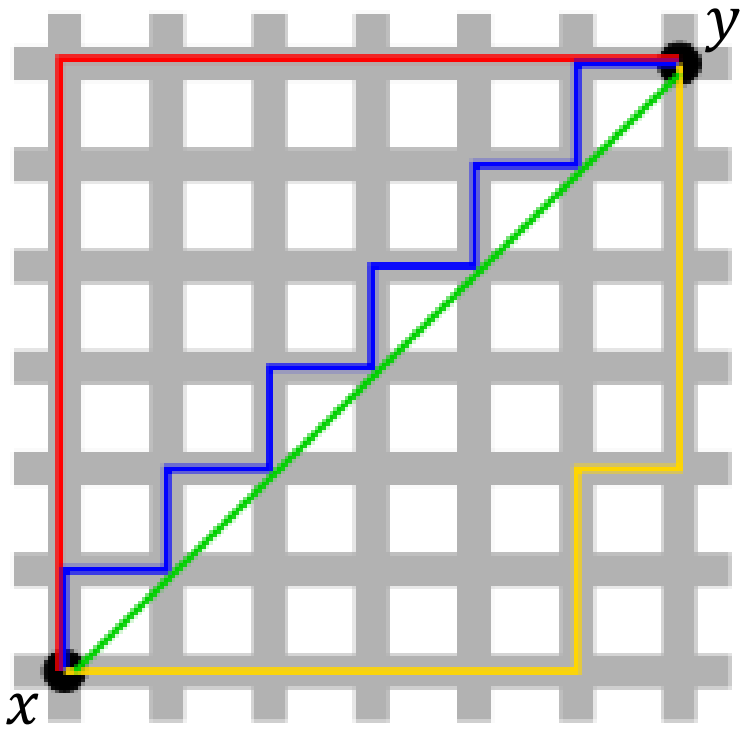
Если пункты отправления и назначения и начало координат лежат на одной прямой, то «едем» по ней, иначе «едем» через начало координат (кратчайшим путем)

- Метрика британской железной дороги
  $$d(x,y) = \begin{cases} \|x\| + \|y\|, x \neq y \\ 0, x = y \end{cases}$$

Если пункты отправления и назначения не совпадают, «едем» через начало координат

Пример: Евклидово *vs.* манхэттенское расстояние

- *Euclidean:* $d(x,y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$
- *Manhattan:* $d(x,y) = |x_1 - y_1| + \cdots + |x_n - y_n|$



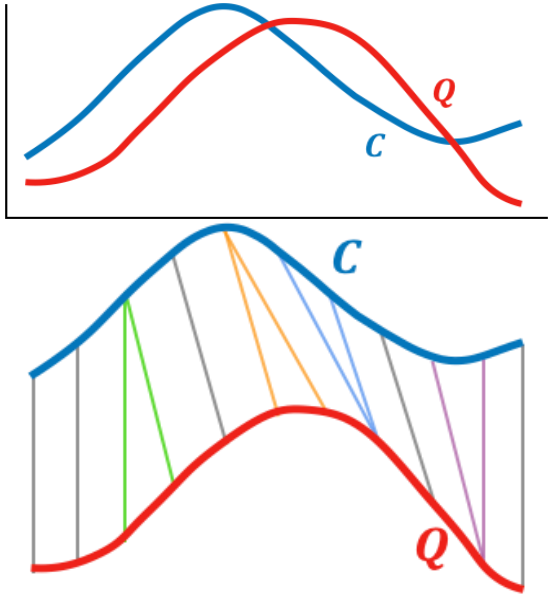
- **Euclidean**$(x,y) = 6\sqrt{2}$
- **Manhattan**$(x,y) =$ **Manhattan**(x,y)
  $=$ **Manhattan**$(x,y) = 12$

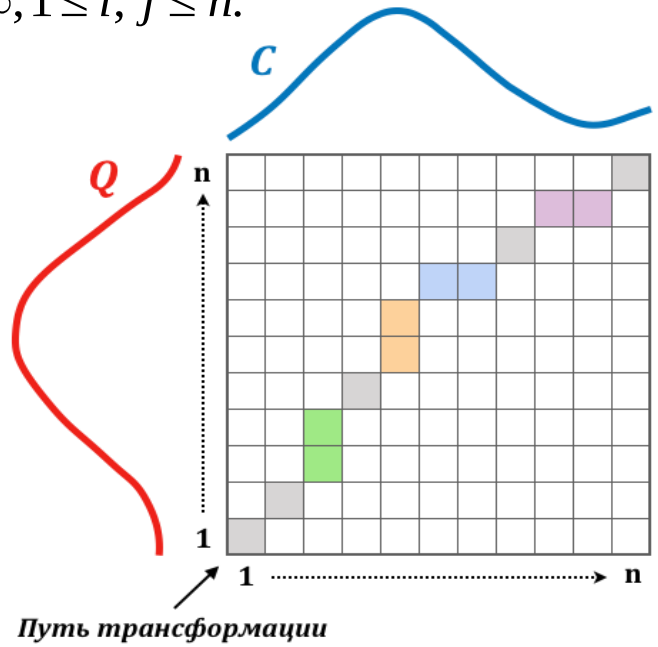
Мера DTW (Dynamic Time Warping)

$$DTW(Q,C) = d(n,n),$$

$$d(i,j) = (q_i - c_j)^2 + \min\begin{cases} d(i-1,j) \\ d(i,j-1) \\ d(i-1,j-1), \end{cases}$$

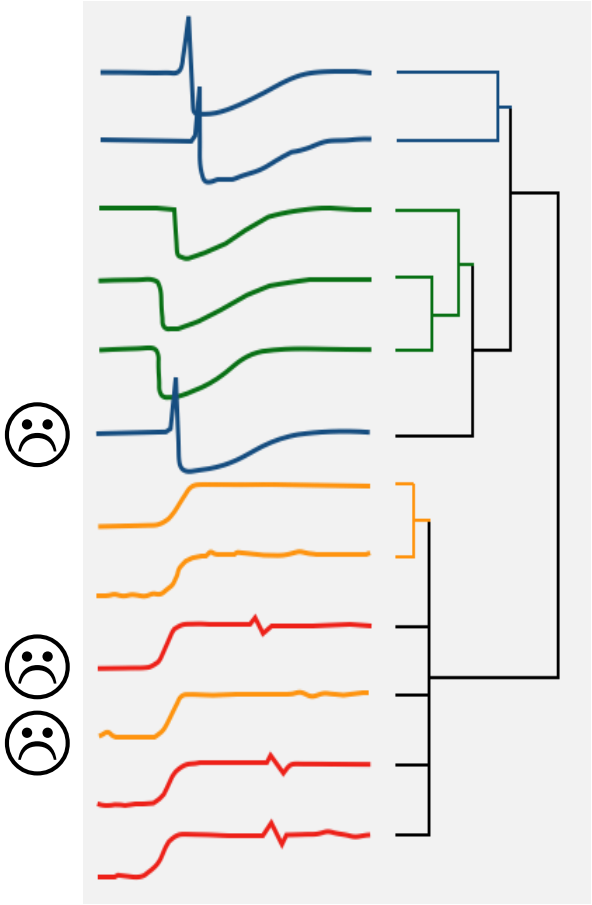
$$d(0,0) = 0;\ d(i,0) = d(0,j) = \infty;\ 1 \le i, j \le n.$$



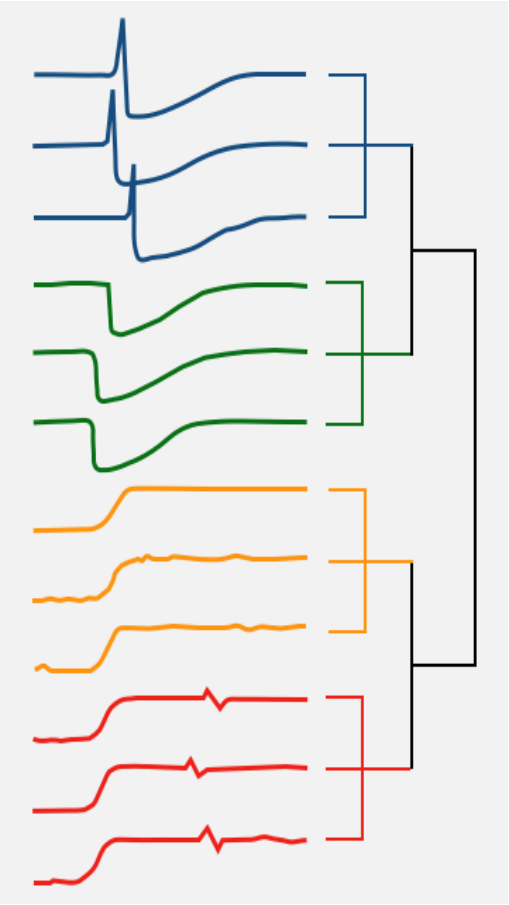


ED *vs.* DTW

Метрика Евклида

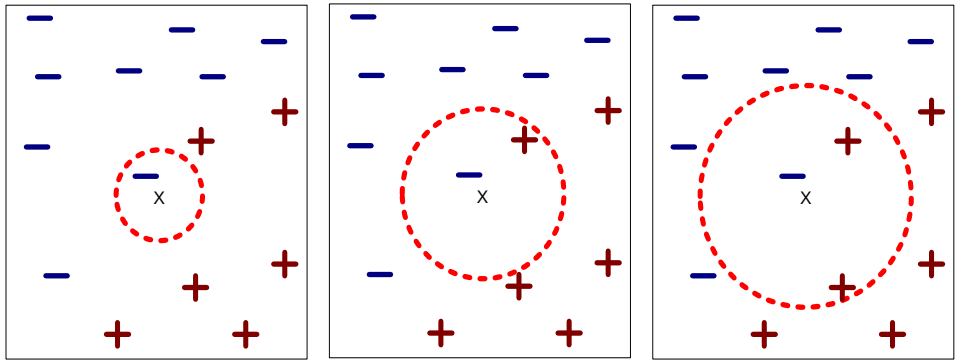


Мера DTW



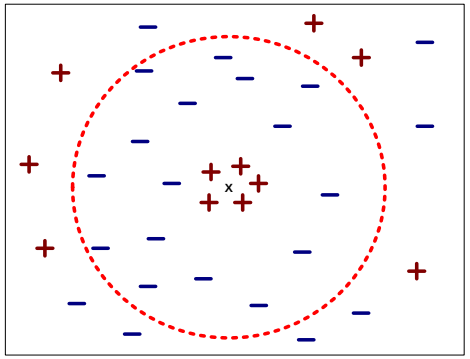
Выбор параметра *k*

- Если *k* очень мало, чувствительность к шумам



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

- Если *k* слишком большое, среди соседей могут быть представители других классов



Выбор способа голосования

- Мажоритарное голосование
- Взвешенное голосование
  - $w = \frac{1}{dist^2(neighbor, test_tuple)}$

Доминантные атрибуты

- Имеют значения, существенно большие, чем у остальных атрибутов. Например:
  Рост: 1.6..2.1, Вес: 45..100,
  Доход: 20000..100000
- $\mathrm{ED}\big((1.7, 70, 30000), (1.9, 100, 35000)\big) \approx$
  $\mathrm{ED}\big((1.7, 70, 30000), (1.7, 90, 25000)\big) \approx$
  $\mathrm{ED}\big((1.7, 70, 30000), (1.8, 80, 25000)\big)$

Нормализация данных

$$v = (v_1, \dots, v_n) \rightarrow v' = (v'_1, \dots, v'_n)$$

- *Минимаксная нормализация*

$$v'_i = \frac{(v_i - \min v_i) \cdot (\max v'_i - \min v'_i)}{\max v_i - \min v_i} + \min v'_i$$

- *Z-нормализация*

$$v'_i = \frac{v_i - \bar{v}}{\sigma}, \quad \bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i, \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v})^2}$$

- *Десятичное масштабирование*

$$v'_i = \frac{v_i}{10^{deg}}, \; deg \text{ – мин целое, дающее } \max_i \frac{|v_i|}{10^{deg}} < 1$$

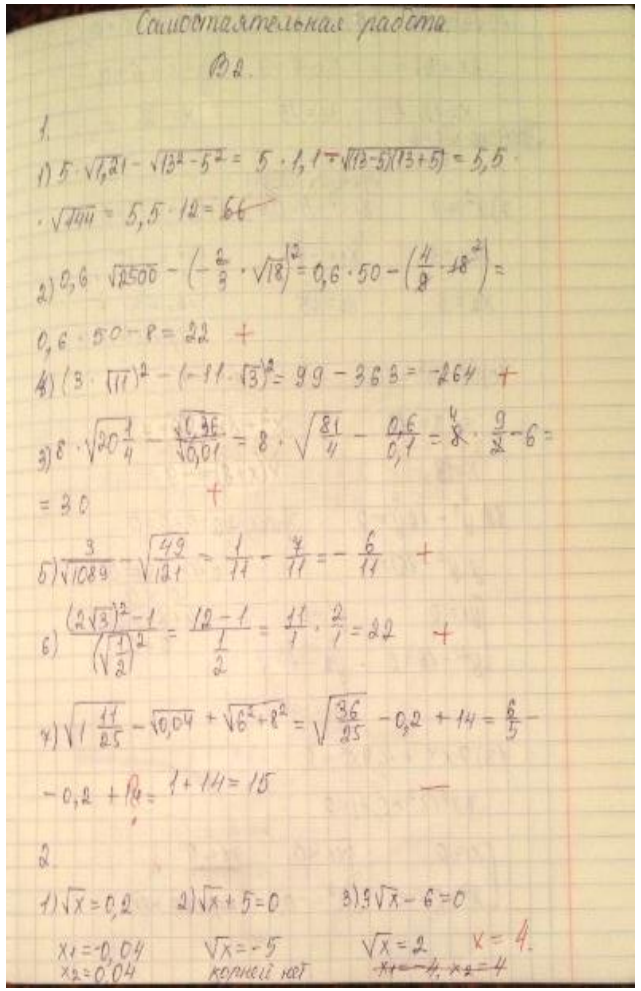
	MinMax	Z-norm	Dec. scale
v	v'	v'	v'
20	0.0000	−1.5753	0.2000
24	0.0667	−1.3653	0.2400
28	0.1333	−1.1552	0.2800
32	0.2000	−0.9452	0.3200
36	0.2667	−0.7351	0.3600
40	0.3333	−0.5251	0.4000
44	0.4000	−0.3151	0.4400
48	0.4667	−0.1050	0.4800
52	0.5333	0.1050	0.5200
56	0.6000	0.3151	0.5600
60	0.6667	0.5251	0.6000
64	0.7333	0.7351	0.6400
68	0.8000	0.9452	0.6800
72	0.8667	1.1552	0.7200
76	0.9333	1.3653	0.7600
80	1.0000	1.5753	0.8000

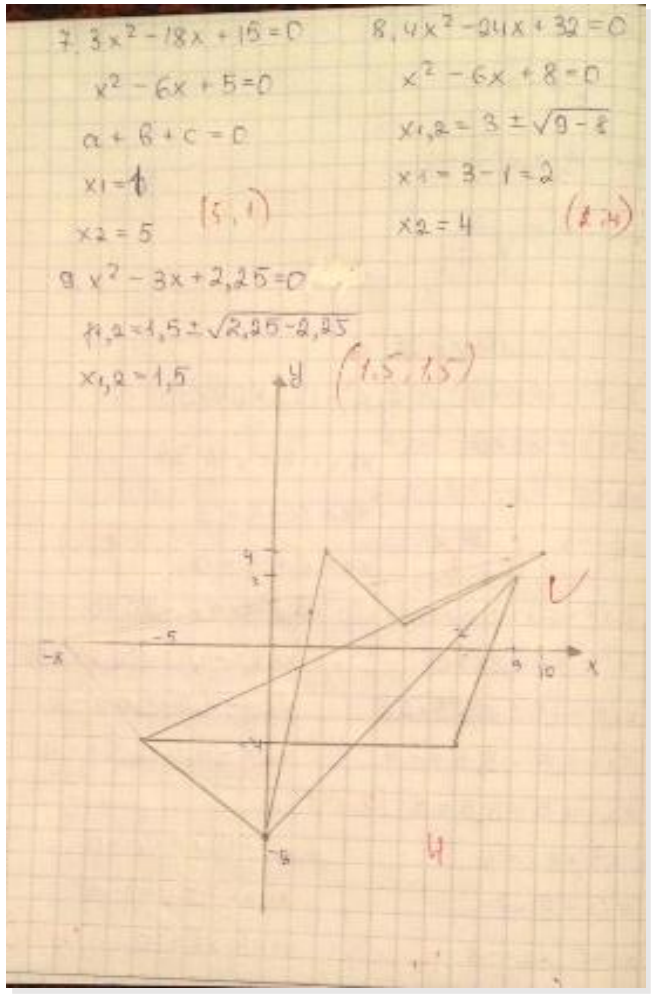
Малозначимые (избыточные) атрибуты

- Не входят в дерево решений, построенное для случайной выборки исходных данных
- Для сокращения времени вычислений малозначимые атрибуты можно отбрасывать

Избыточные атрибуты

- Учитель
  + и –
  ~~Исправления~~
  ?
  ✔
  5+, 5, 5-, 4+, …
- Ученик
  Исправления
  ~~Замазывания~~





Избыточные атрибуты

№	Ученик		Проверка учителя				Выполнение работы		ОЦЕНКА
	Пол	Класс	% правильно выполненных заданий	К-во ✓	К-во ?	К-во испр-й	К-во испр-й	К-во замазок	
1	Ж	8м	67	0	0	1	3	3	4-
2	Ж	8м	60	0	2	0	5	9	3
3	Ж	8м	100	0	0	2	0	2	5-
4	Ж	8м	100	1	0	2	3	1	4
5	Ж	8м	100	1	0	2	0	5	5
6	М	8л	80	0	1	1	0	2	4
7	Ж	8м	50	1	0	6	0	1	4-
8	Ж	8л	20	0	0	2	2	1	3-
9	М	8м	40	1	2	0	3	4	2
10	М	8м	65	0	3	4	4	8	4-
…	…	…	…	…	…	…	…	…	…
70	М	8л	15	0	0	0	0	11	2

Избыточные атрибуты

№	Ученик		…	ОЦЕНКА
	Пол	Класс		
1	Ж	8м		4-
2	Ж	8м		3
3	Ж	8м		5-
4	Ж	8м		4
5	Ж	8м		5
6	М	8л		4
7	Ж	8м		4-
8	Ж	8л		3-
9	М	8м		2
10	М	8м		4-
…	…	…		…
70	М	8л		2

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2011. 740 p. ISBN: 978-0123814791
  - 9.5 Lazy Learners (or Learning from Your Neighbors), pp. 422-425
- Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. 1st Edition. Pearson, 2014. 732 p. ISBN: 978-1-292-02615-2
  - 5.3 Bayesian Classifiers, pp. 227-240