

Введение. Задача классификации

Машинное обучение 2022

Мария Олеговна Корлякова, ктн



- Цель курса

- Знакомство с методами и алгоритмами решения интеллектуальных задач обработки информации на примере изображений и сигналов
- Пройти интервью с меткой Data Science\Machine learning

- Оценка

- Лабы (50%)
- РК (10%)
- Тесты на лекциях(20%)
- Посещение(20%)

ССЫЛКИ:

Рабочие среды:

- Python

Numpy, Pandas, SciKit-learn

Tensorflow 2.x, PyTorch

Matplotlib, Scipy, OpenCV

Ссылки:

<https://github.com/mkorlyakova/MSTU-courses>

Telegram: ???

+79109136824

Что будет :

Постановка задачи машинного обучения

Обучение с учителем. Задача классификации

Метрические алгоритмы

Считается, что компьютерная программа учится на опыте E в отношении некоторого класса задач T и метрики производительности P , если ее производительность в задачах из T , измеряемая P , улучшается с опытом E .

Т. Митчелл



Машинное обучение

Игра в шахматы

Задача Т - умение играть в шахматы

Опыт Е - обучающие игры в шахматы

Метрика Р - процент побед у соперника

Машинное обучение

Распознавание рукописных цифр


(задача MNIST)

Задача Т - распознать рукописные цифр

Опыт Е - набор данных с изображениями цифр и их реальным значением

Метрика Р - процент верно распознанных цифр

• Типы Задач

- Большие задачи
 - } Анализ качества клиентов
 - } Обработка результатов экспериментов
 - Повторяющиеся задачи
 - } Очистка сигнала от шума
 - } Поиск объектов в сигнале
 - Задачи с проблемами
 - } Пробелы
 - } Противоречия
 - } Ошибки
- 

- Плохо формализованные задачи
- Нет числовой формы
- Цель не формализована
- Нет алгоритма
- Данные **неполные, неточные, неоднозначные, противоречивые**

• Интеллект

- Intellectus – лат.
- Intelligence – англ.
- artificial intelligence – искусственный интеллект
- ИИ (AI)
- Искусственные Интеллектуальные Системы – ИИС(AIS)

- **Методы Интеллектуальной обработки информации**



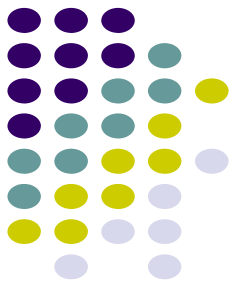
Искусственный интеллект – Artificial Intelligence – AI

The diagram consists of three nested ellipses. The outermost ellipse is blue and contains the text 'Искусственный интеллект – Artificial Intelligence – AI'. Inside it is a gray ellipse containing the text 'Machine Learning - машинное обучение'. Inside the gray ellipse is a green ellipse containing the text 'Deep Learning – Глубокое Обучение'.

Machine Learning - машинное обучение

Deep Learning – Глубокое Обучение

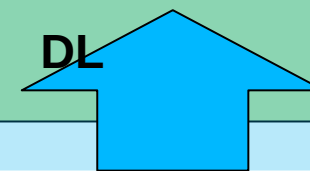
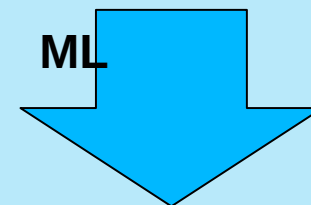
Где это работает



• Основные интеллектуальные задачи

- Представление знаний
- Решение неформализованных задач
- Создание комплексных ИИ систем
- Моделирование разума

- Интеллектуальный анализ данных
- Естественный язык и ЭВМ
- Обработка временных рядов
- Техническое зрение



- **Обучение по примерам**

- Без учителя (выделение классов)
- С учителем (отношение к классу)
- Обучение с подкреплением
- Supervised learning
- Unsupervised learning
- Reinforcement learning

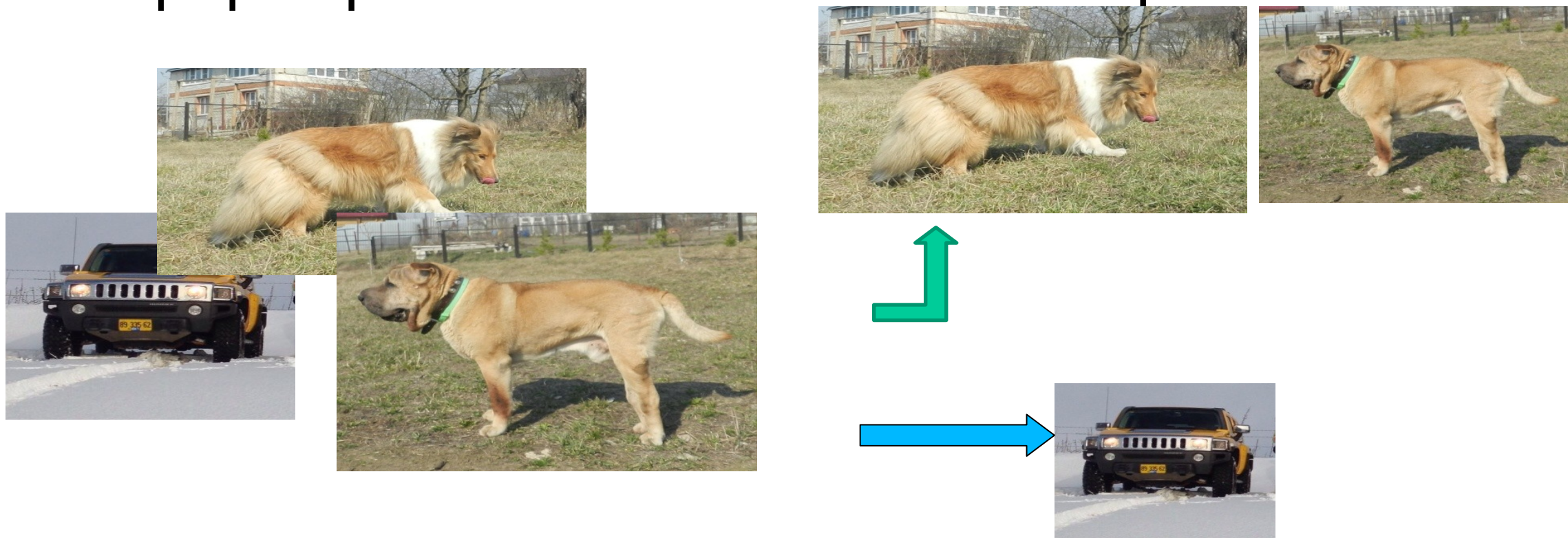
Задача **Обучение без** **учителя**

Данные вида:
“объект”

Типы задач:
Кластеризация

• Задача Кластеризации

- Необходимо определить группы, которые сформированы на основании метрики близости.



Задача **Обучение** **учителем**

Данные вида:
“объект, ответ”

Типы задач:
Классификация
Регрессия

• Задача классификации

- Разделить объекты на группы и сказать к какой из них относиться новый объект:


Dog



Car



- **Класс**

- **Классы** - это объединения объектов (явлений), отличающиеся общими свойствами, интересующими человека.
 - **Цель распознавания** – принятие решения об отнесении объекта к тому или иному классу.
- 

• Образ не объект

- Описание не полностью представляет реальный объект
- Описание зависит от задач
- Описание содержит погрешности представления
- **Machin learning**
- Любой образ представляется некоторым набором признаков
- Основное назначение описаний (образов) - это их использование в процессе установления соответствия объектов

- Описание классов по признакам
- Столы для работы

признак	Длина, м	Ширина, м	Число ящиков
Стол 1	1	0.6	3
Стол 2	1.5	0.7	5
Стол 3	3	0.7	4

- Столы для обеда

признак	Длина, м	Ширина, м	Число ящиков
Стол 1	1. 6	1.2	1
Стол 2	1.5	0.8	0
Стол 3	3	1.25	0

Типы признаков

Бинарные:

- да/нет,
- 0/1,
- черное/белое

Дискретные:

- длина, м
- вес, кг

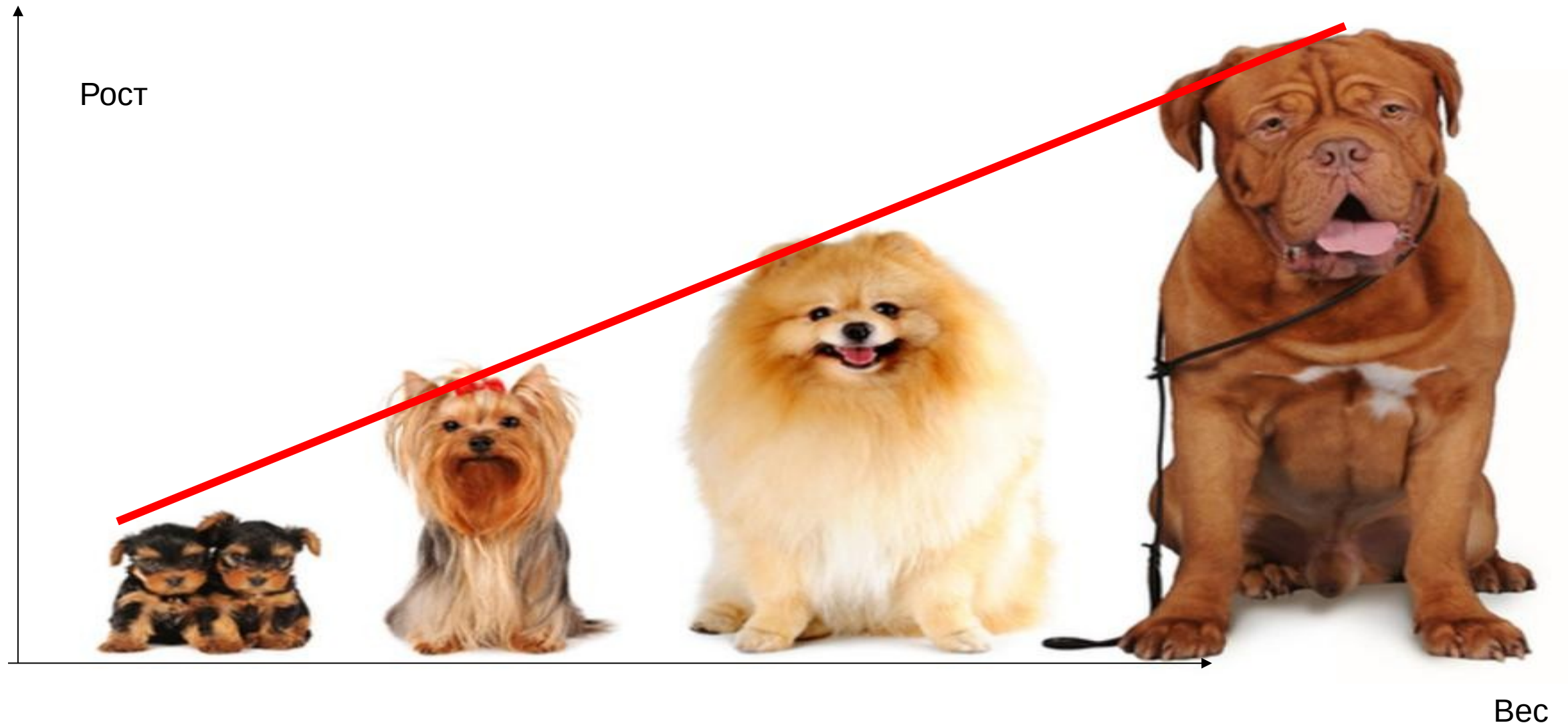
Порядковые:

- $\text{знач}_1 > \text{знач}_2$
- холодно, тепло, жарко

Категории:

- красный, зеленый, синий
- круг, квадрат, треугольник

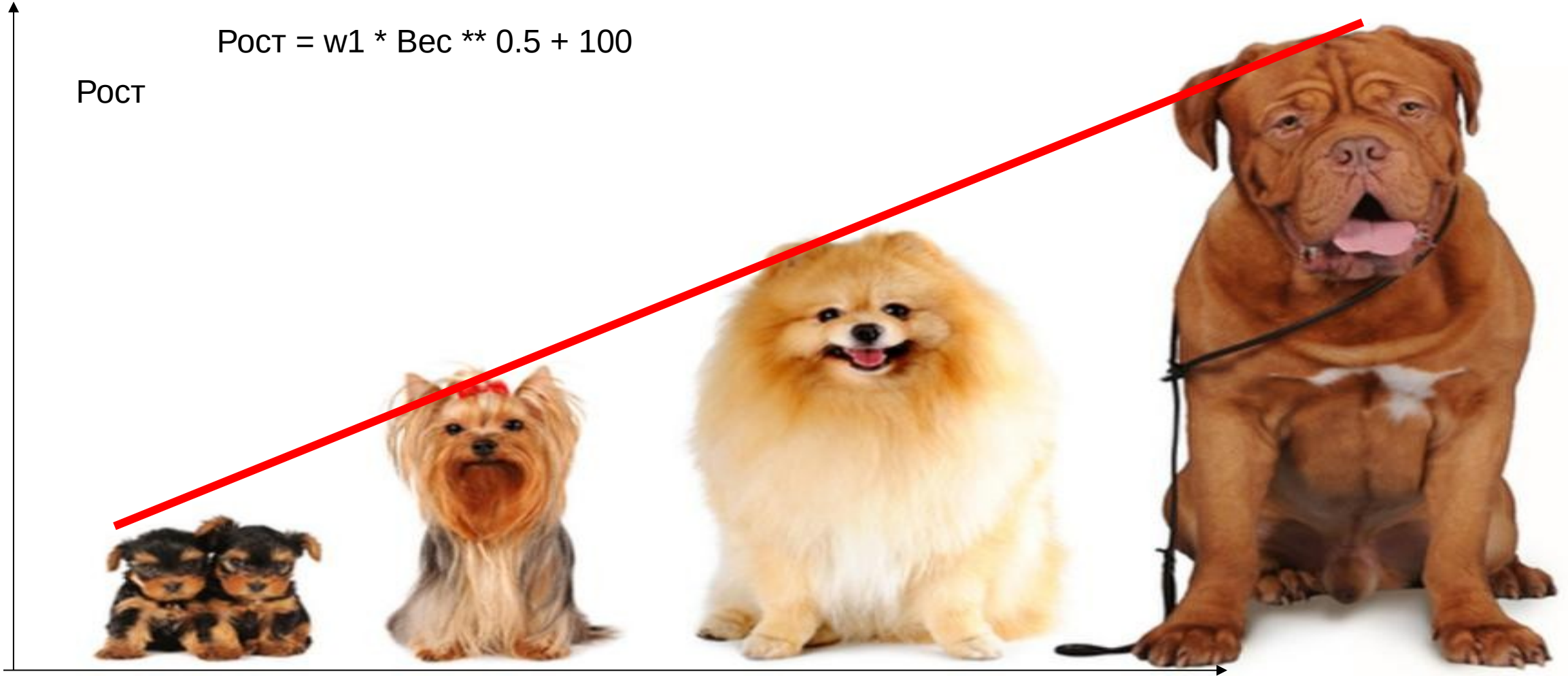
- **Задача Регрессии**



• Задача Регрессии

$$\text{Рост} = w_1 * \text{Вес}^{**} 0.5 + 100$$

Рост



Вес

- **Оценка качества Классификации**
- **Исследуем качество обучения**
- **Как?????**

- **Оценка качества Классификации**

- **Оценка по-объектная**
-
- **$Q = 1/N \sum L(a(X_i), y_i), \quad i = 1, N$**

- **Оценка качества Классификации**

$$T = \{(X_i, y_i)\}$$

- $Q = 1/N \sum L(a(X_i), y_i), \quad i = 1, N$

- **Оценка качества Классификации**

$$T = \{(X_i, y_i)\}$$

- $Q = 1/N \sum L(a(X_i), y_i), \quad i = 1, N$

Q — внутренний критерий
(потери - loss)

Train

Q — внешний
(потери или метрики)

Hold out

- **Оценка качества Классификации**

$T = \{(X_i, y_i)\}$

- $Q = 1/N \sum L(a(X_i), y_i), \quad i = 1, N$

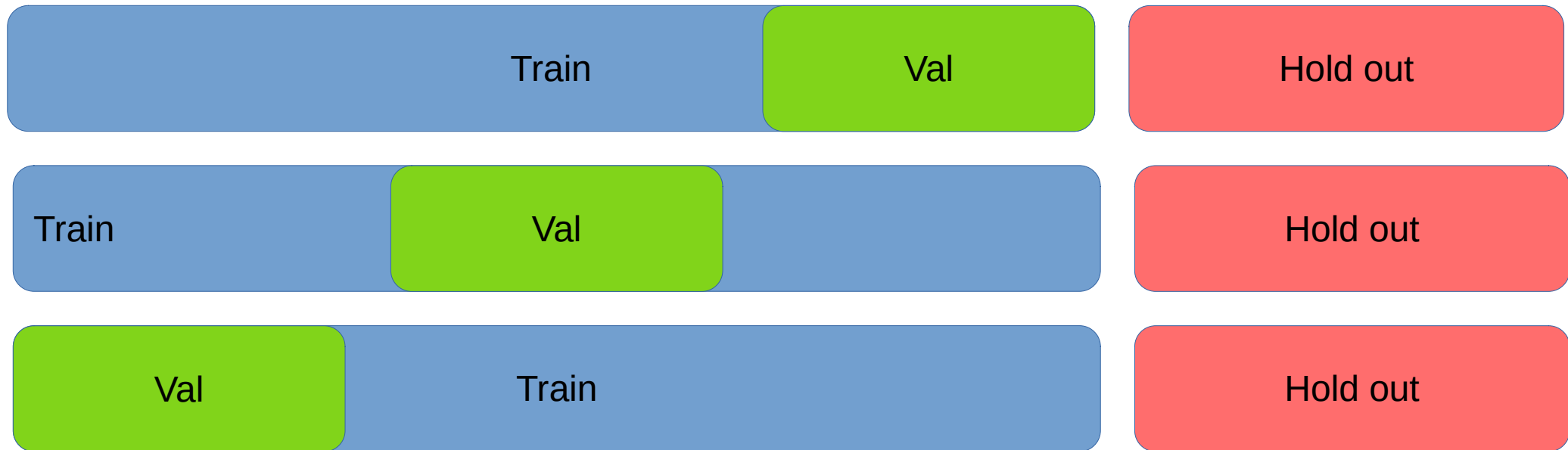
Train

Val


Hold out


- Оценка качества Классификации

Кросс-валидация



- **Оценка качества Классификации**

- **Способов введения внутренних и внешних критериев много**
 - **Можно использовать больше одного критерия**
 - **Для разных этапов можно использовать разные критерии**
 - **Нужно использовать одинаковые для всех моделей оценки на отложенной выборке**
- 

- **Оценка качества Классификации**
 - **Тренировочный, Валидационный и Отложенный наборы не пересекаются!!!!!!**
 - **Величина критерия зависит от разбиения на обучение**
- 

- **Оценка качества Классификации**

- **Тренировочный критерий монотонно падает**
- **Валидационный (тестовый) имеет минимум**
- **Недообучение - Underfitting**
- **Переобучение - Overfitting**

- Оценка качества Классификации.
 - Внешние метрики
- Исследуем качество обучения по ошибкам
- Каким?????



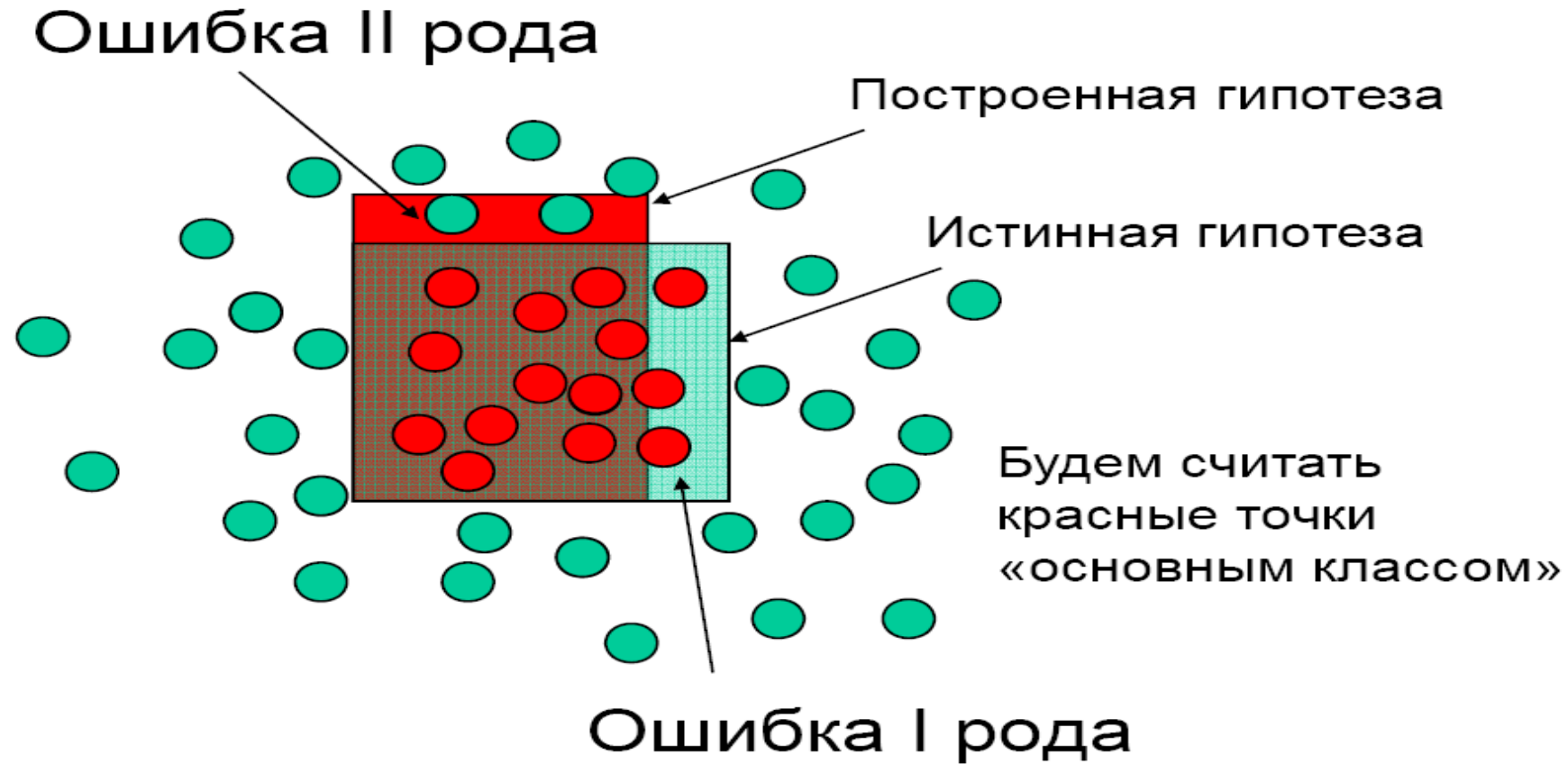
Ошибки I и II рода

Пусть, существует «основной класс» класс, при обнаружении которого, предпринимается какое-либо действие;

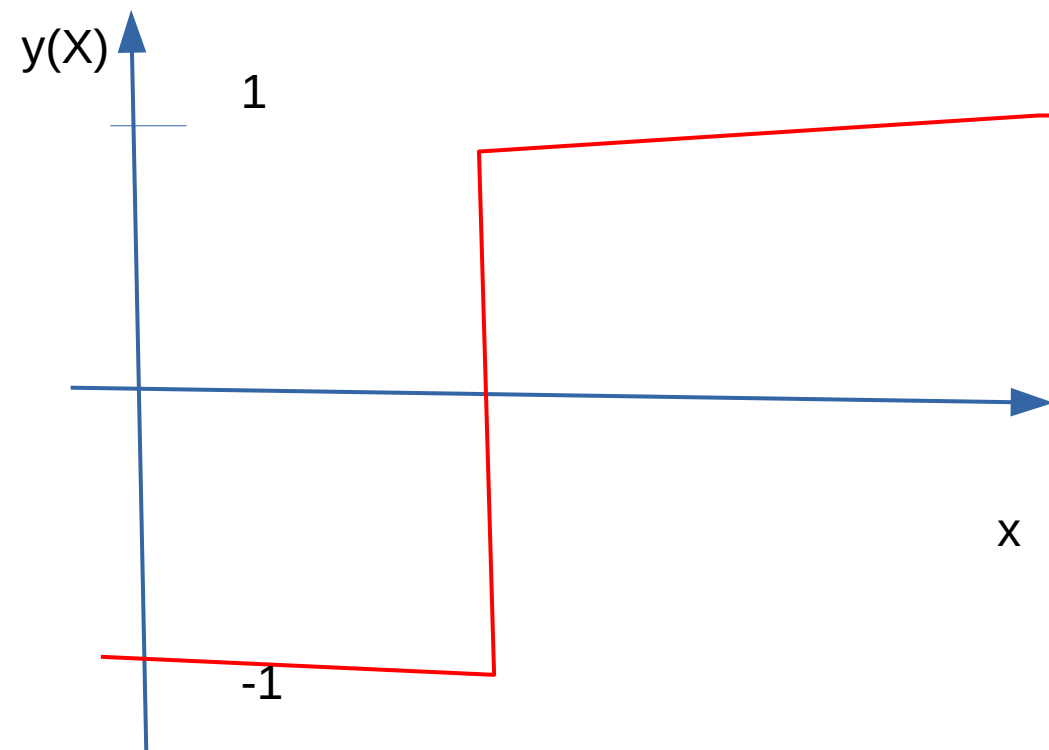
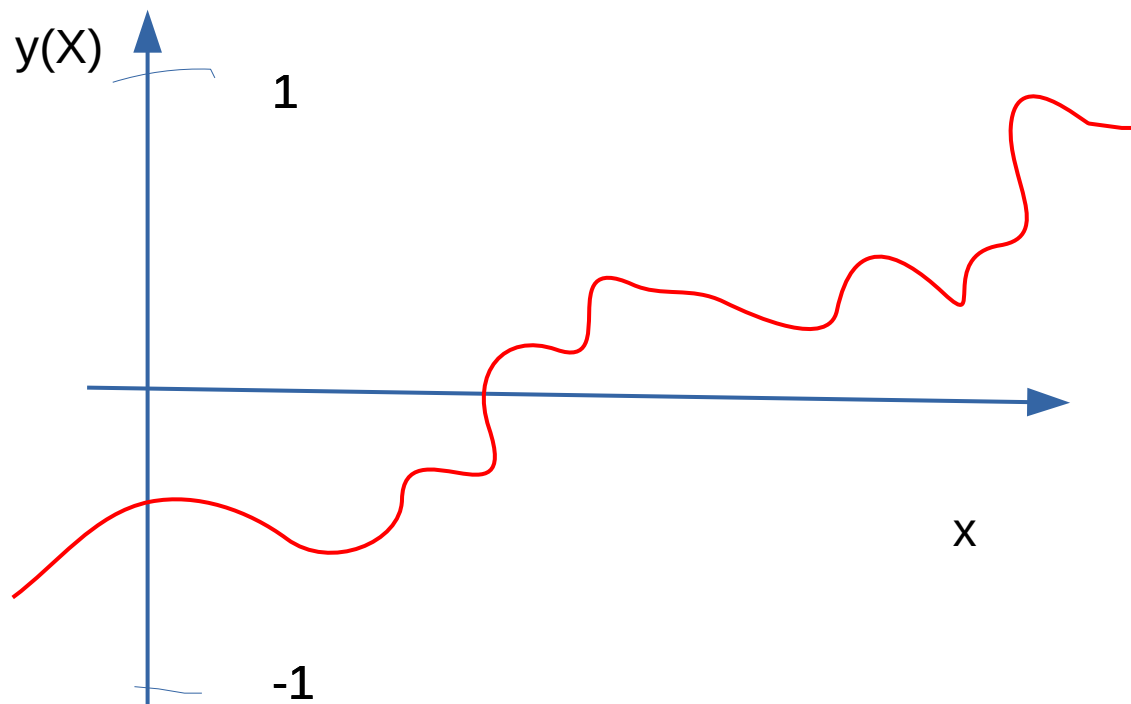
Например, «болен» и «здоров».

- Ошибка первого рода равна вероятности принять основной класс за вторичный
 - } Вероятность «промаха», когда искомый объект будет пропущен
- Ошибка второго рода равна вероятности принять вторичный класс за основной
 - } Вероятность «ложной тревоги», когда за искомый объект будет принят «фон»

• Схема ошибок I, II



Как выглядит ответ модели?



Метрики классификации

	Истина +	Истина -
Предсказано +	True positive	False positive
Предсказано -	False negative	True negative

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

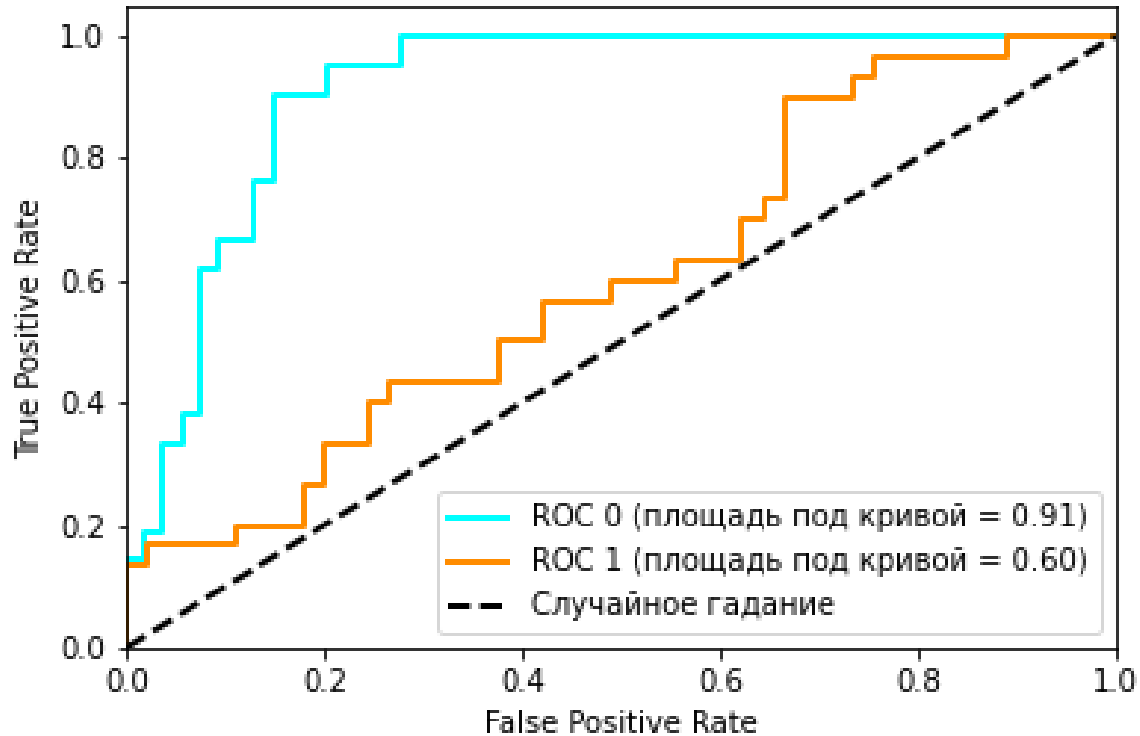
$$Precision = Sensitivity = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

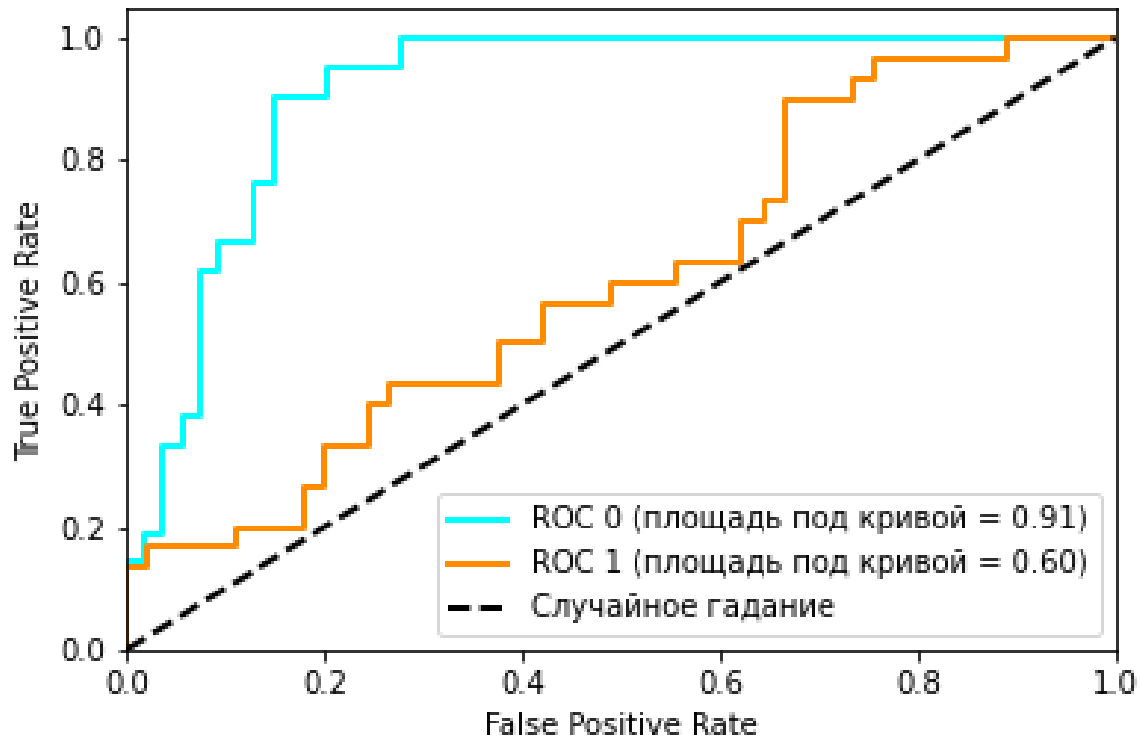
$$Specificity = \frac{TN}{TN + FP}$$

Метрики классификации



- $FPR = FP / (FP + TN)$
- $TPR = TP / (TP + FN)$

Метрики классификации



ROC кривая - зависимость верно классифицируемых объектов положительного класса (Sensitivity) от ложноположительно классифицируемых объектов негативного класса (Specificity)

AUC (Area Under Curve) - площадь под кривой

ROC-AUC - площадь под ROC кривой, численная оценка ROC метрики

Метрические алгоритмы классификации

Метрические алгоритмы - алгоритмы, основанные на вычислении оценок сходства между объектами.

$$a(x) = \arg \max_y \sum [y_i = y] w(i, x)$$

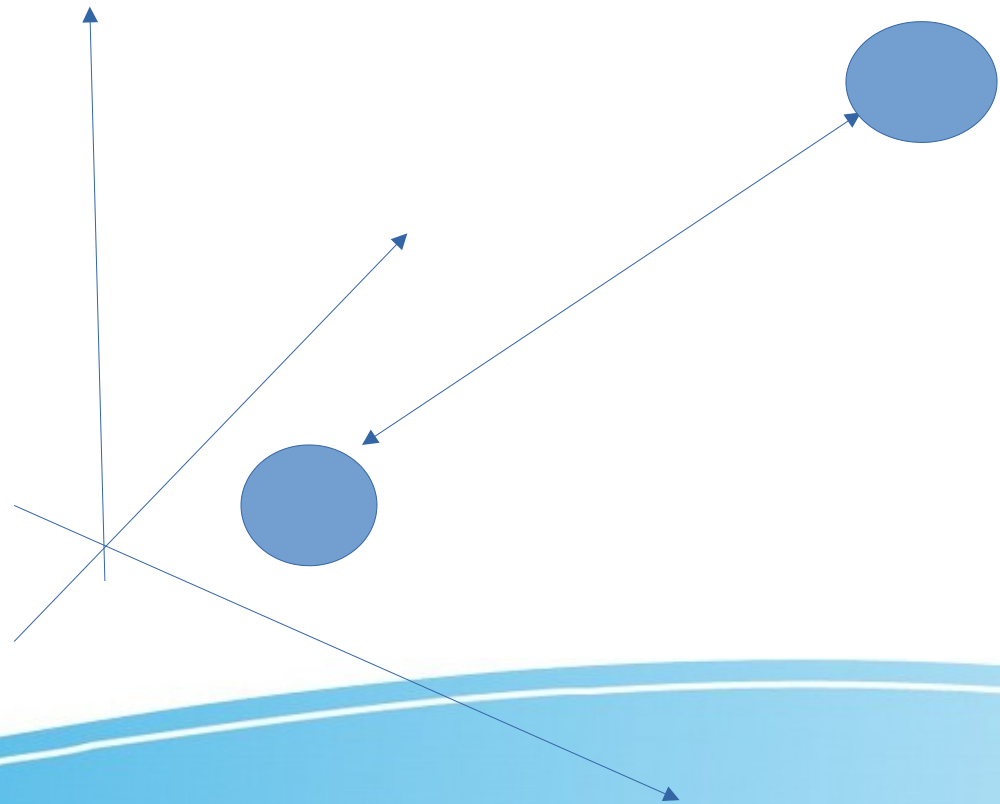
$w(i, x)$ — степень важности i -го соседа для x

$[i \leq 1]$ — ближайший сосед

$[i \leq k]$ — k ближайших соседей

Метрики, расстояния

- $\rho(x,y) \geq 0, \rho(x,y) = 0 \Leftrightarrow x=y$;
- $\rho(x,y) = \rho(y,x)$;
- $\rho(x,y) \leq \rho(x,z) + \rho(z,y)$.



- Расстояния между объектами
- Метрики : Минковский
- Меры: Хемминг
- И МНОГО ДРУГИХ МЕТОДОВ!!!!

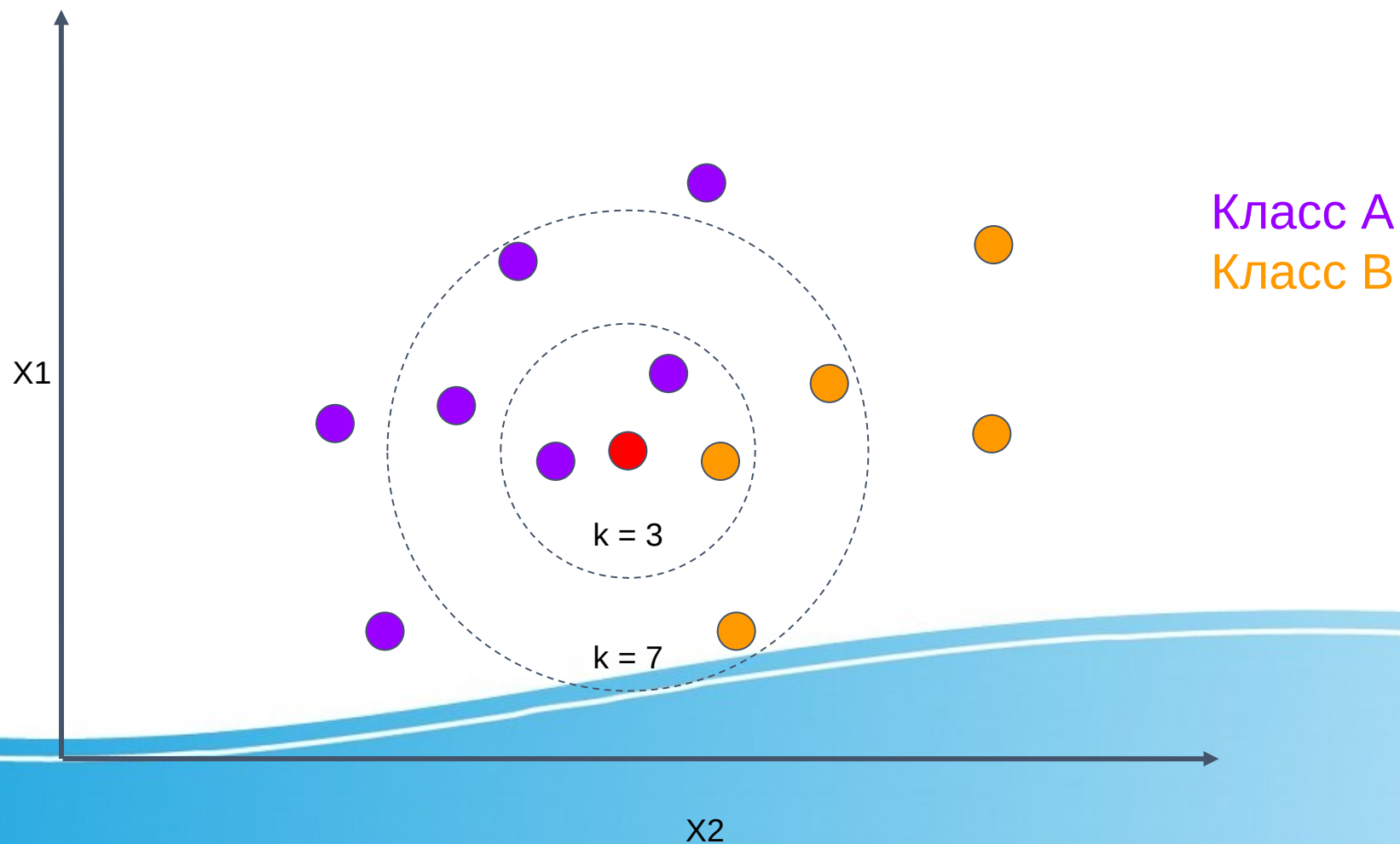
Метрические алгоритмы классификации

Метод k- ближайших соседей

Метрические алгоритмы - алгоритмы, основанные на вычислении оценок сходства между объектами.

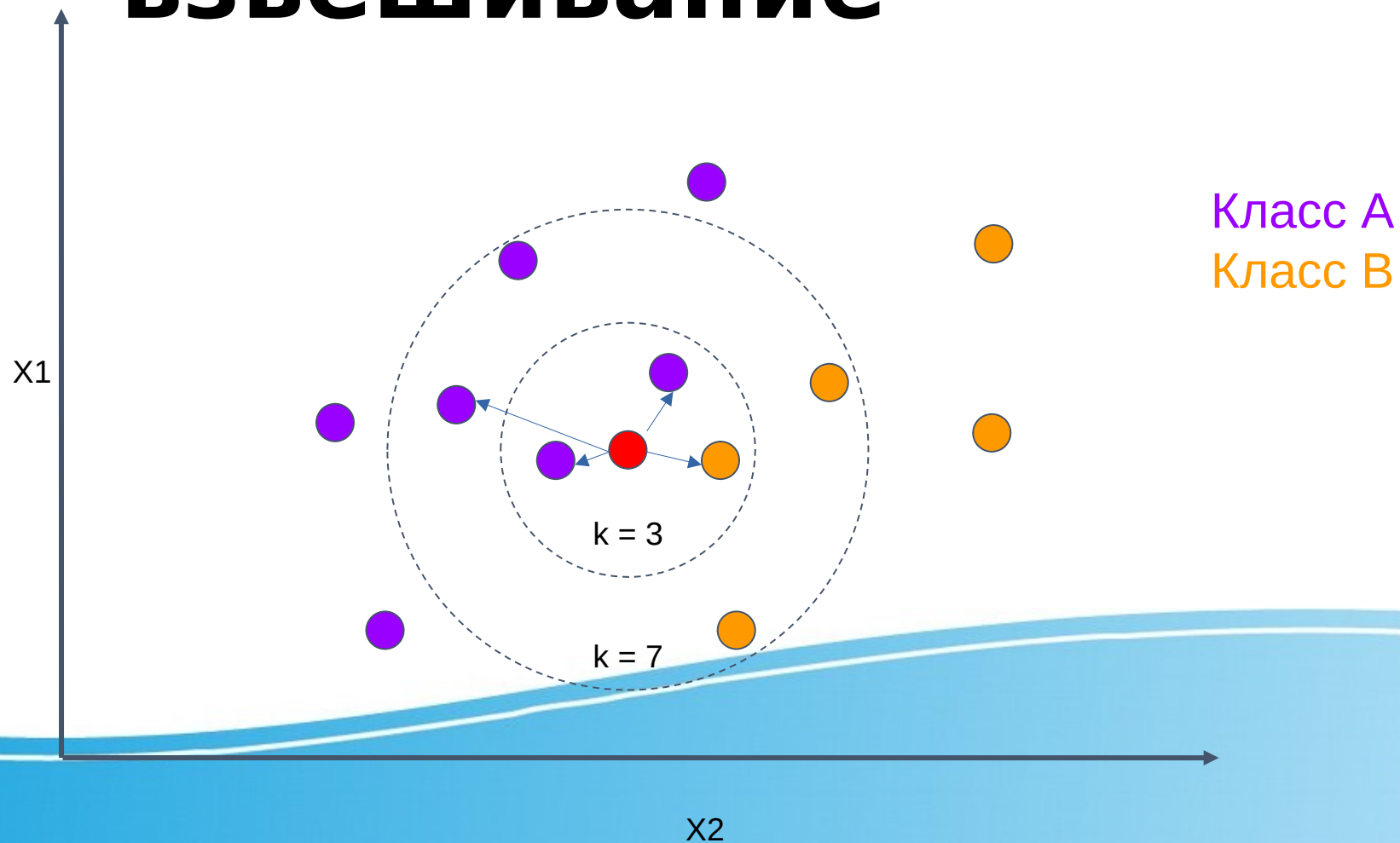
Метод k-ближайших соседей - объекту присваивается тот класс, который наиболее распространен среди его k соседей.

Метод k-ближайших соседей



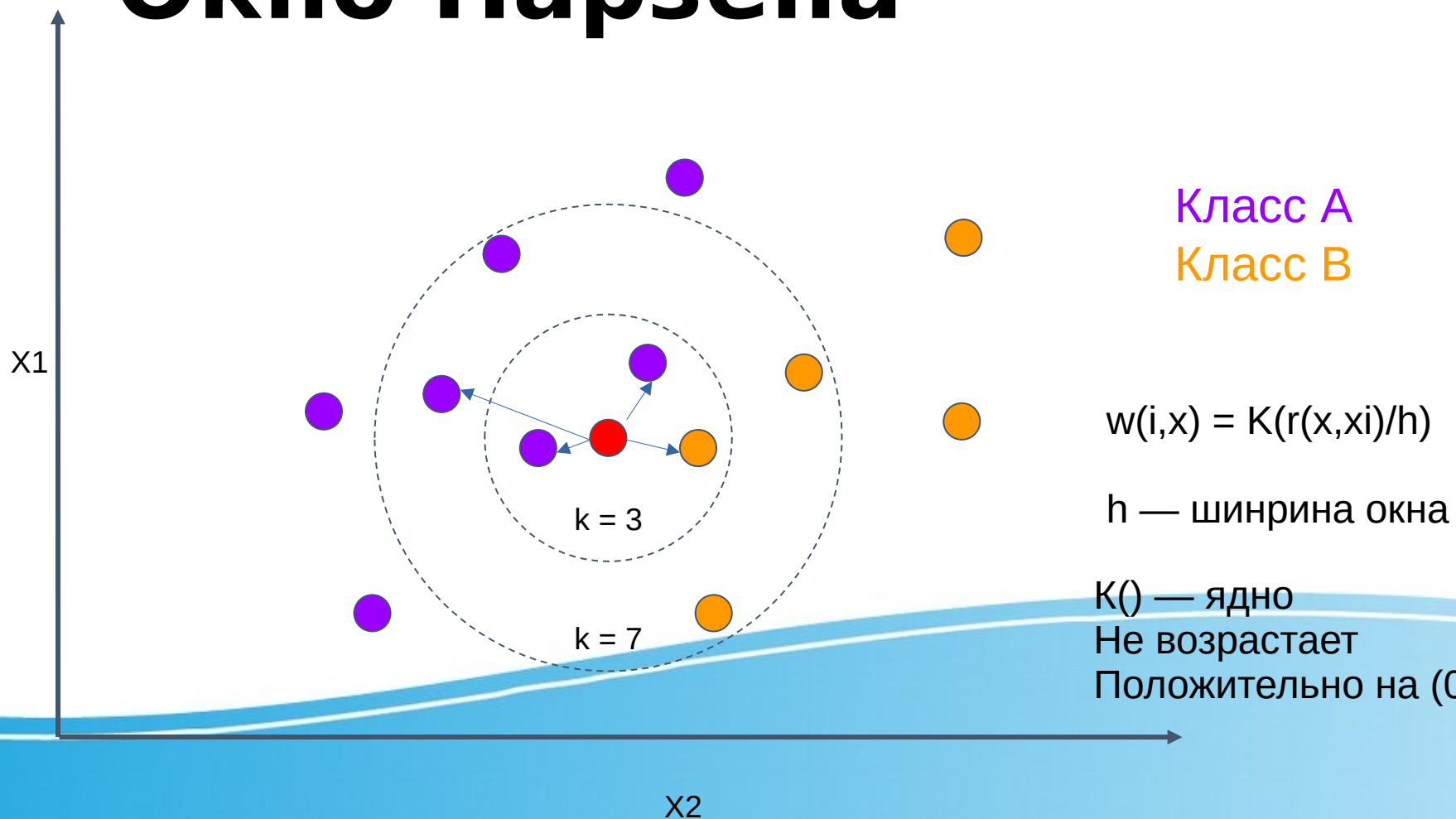
Метод k-ближайших соседей

ВЗВЕШИВАНИЕ



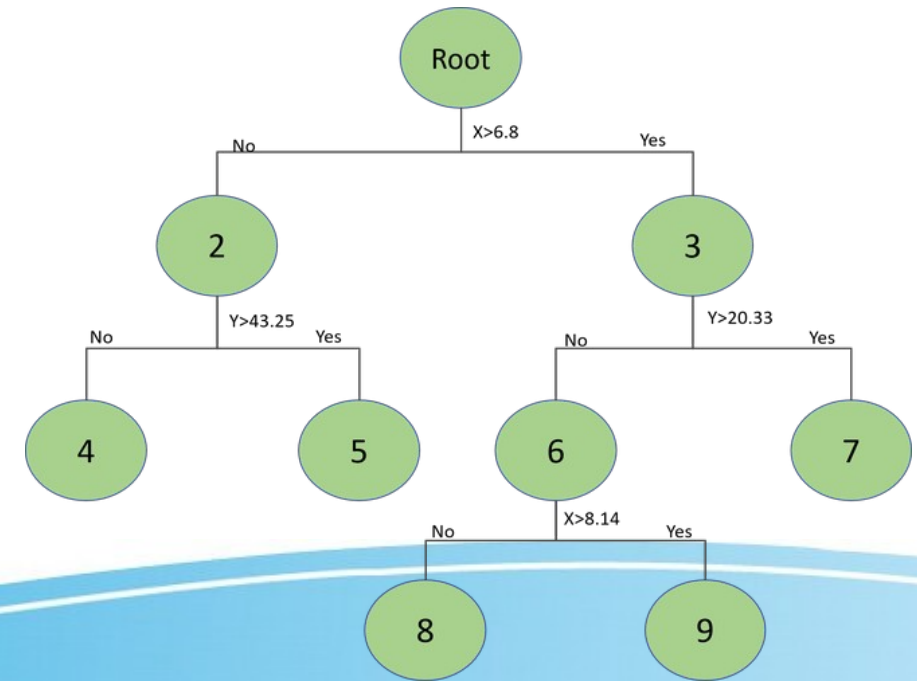
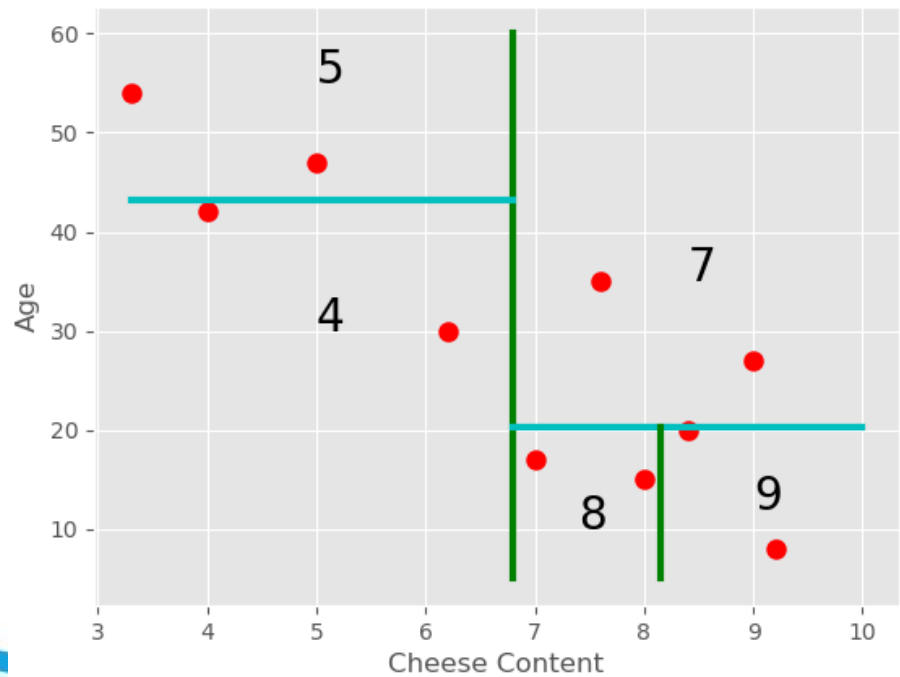
Метод k-ближайших соседей

Окно Парзена

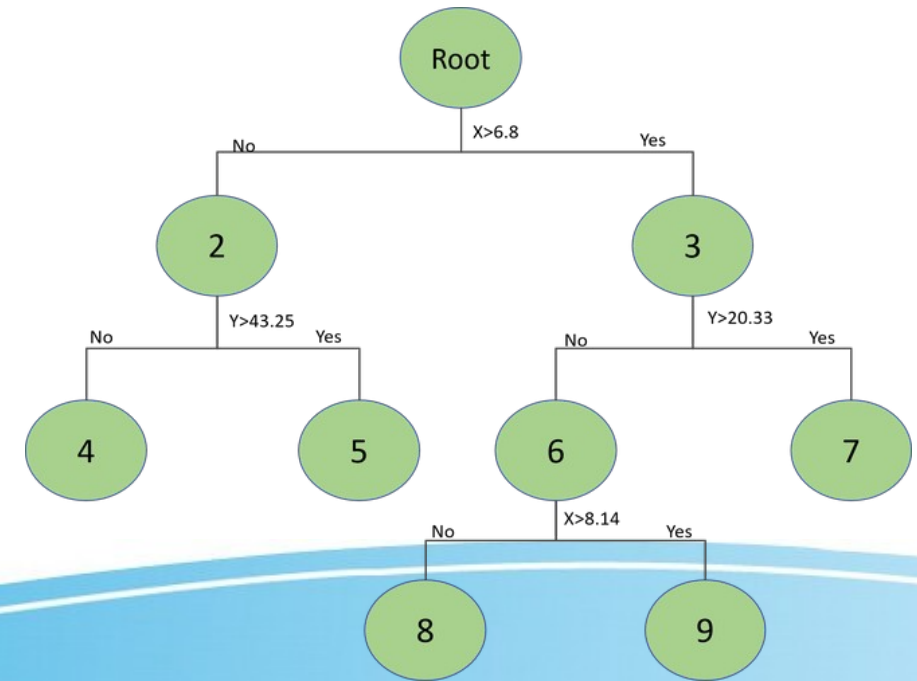
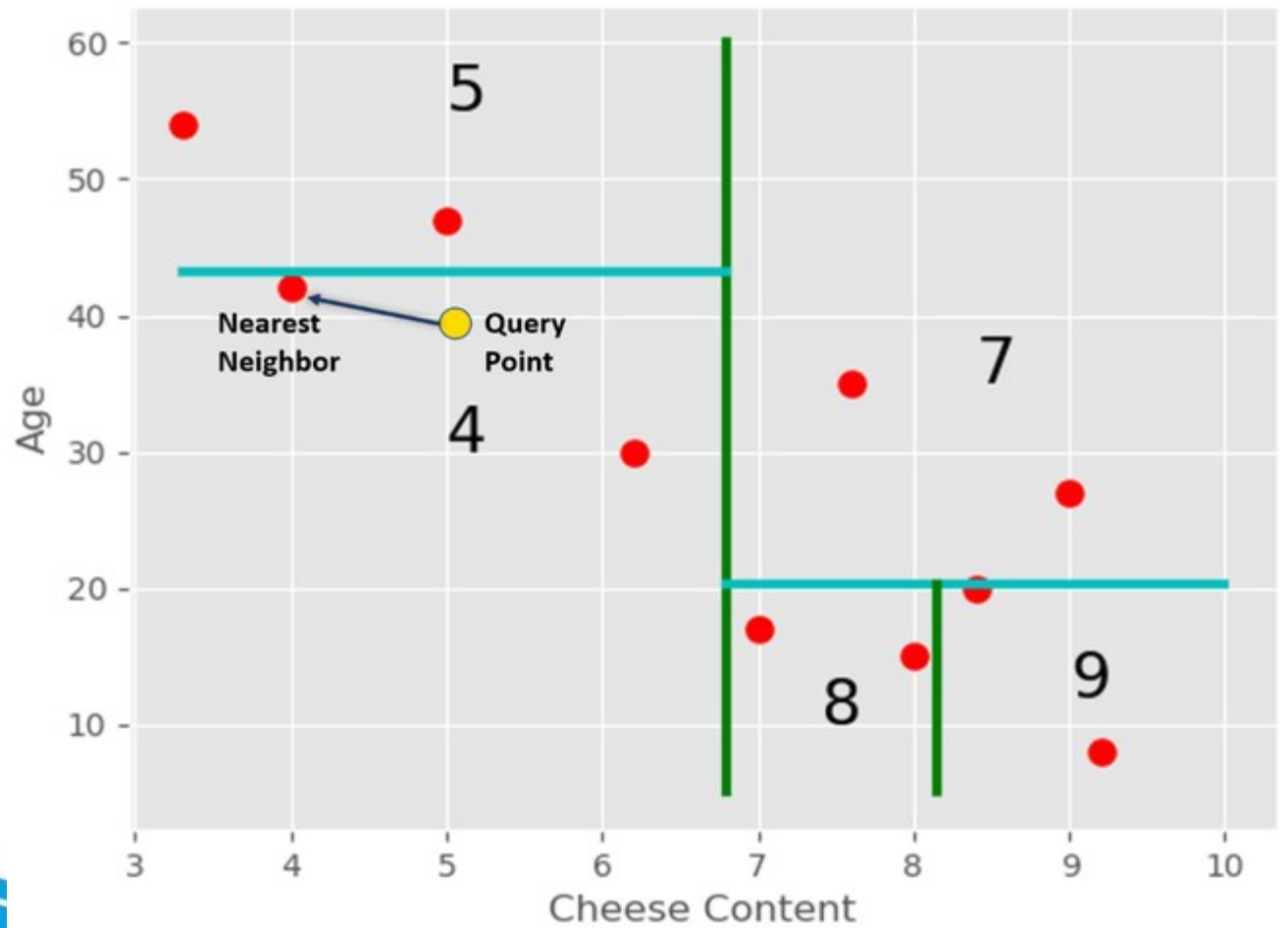


- **KdTree для поиска**

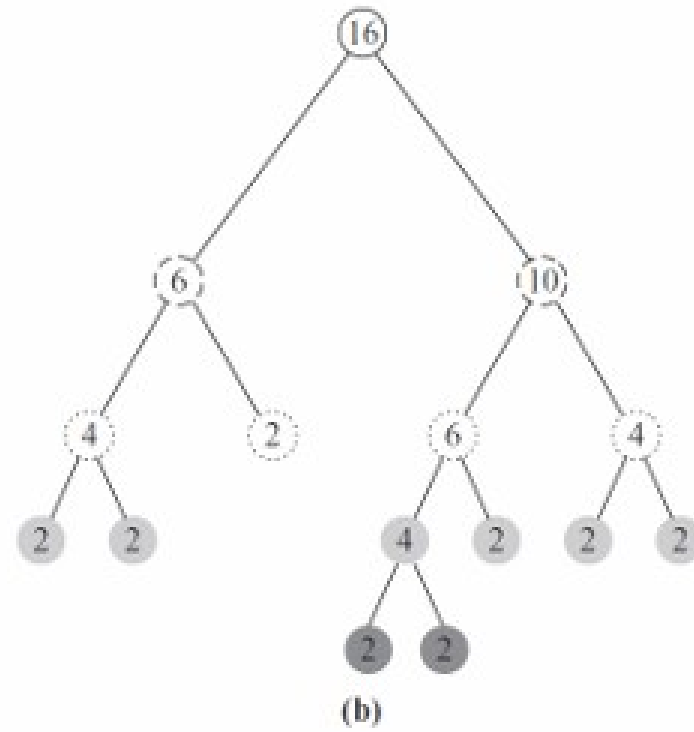
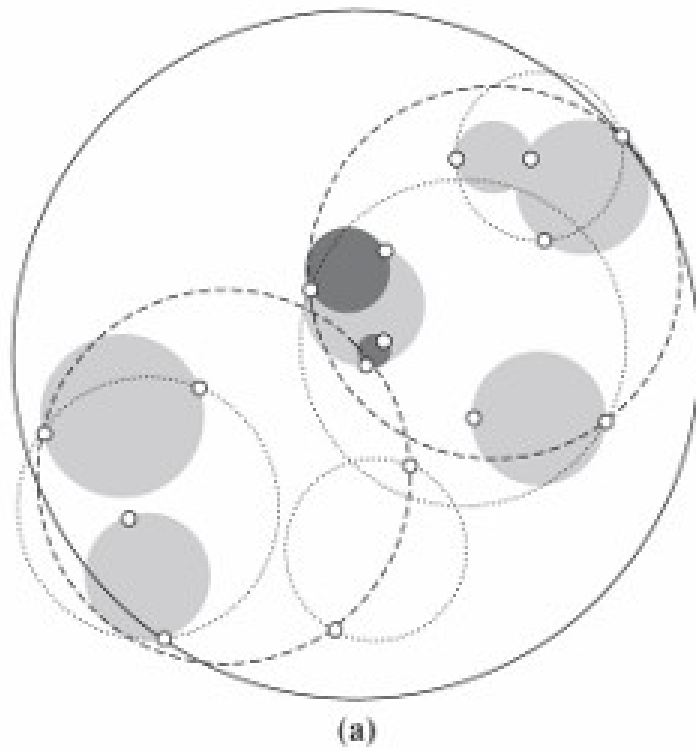
- Много данных
- Долгий поиск



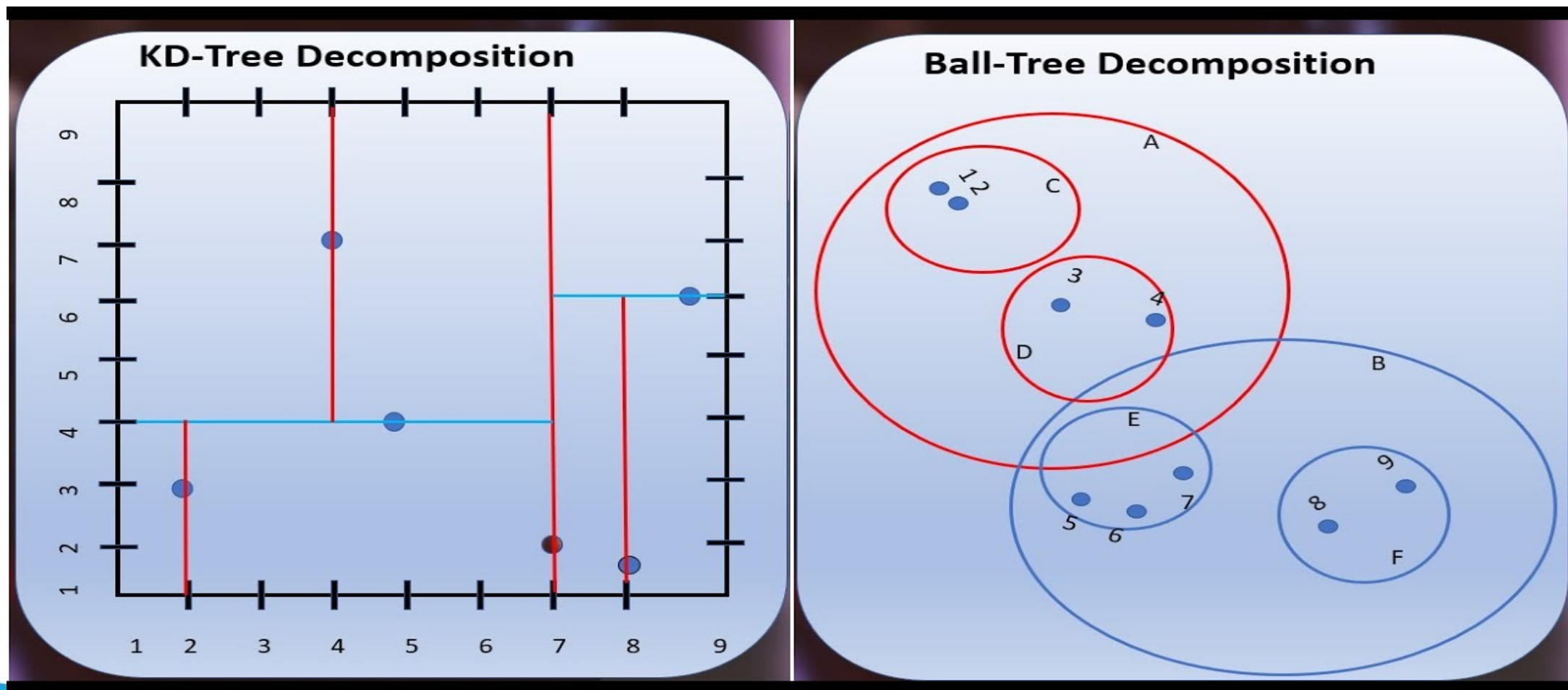
- KdTree для поиска



Дерево вложенных шаров

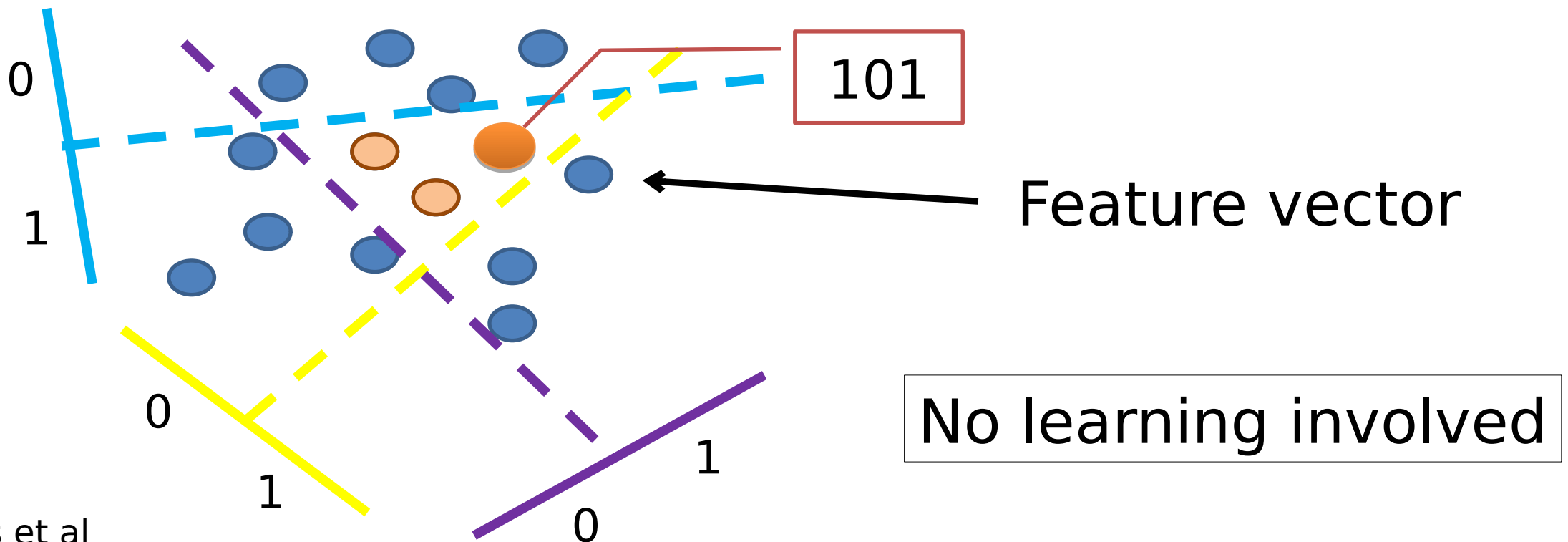


Дерево вложенных шаров



1. Locality Sensitive Hashing

- Take random projections of data $r^T x$
- Quantize each projection with few bits



• литература

- Методы современной и классической теории управления. Т5. - 2004
- Математические методы распознавания образов. Курс лекций. МГУ, ВМиК, кафедра «Математические методы прогнозирования», Местецкий Л.М., 2002–2004.
- Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин)

• ресурсы

- Wiki-портал <http://www.machinelearning.ru>
- *Воронцов К. В. Машинное обучение (курс лекций)* см. <http://www.machinelearning.ru>
- Coursera:
- <https://www.coursera.org/specializations/machine-learning-data-analysis>
- <https://www.coursera.org/browse/data-science/machine-learning>