



СПБГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ

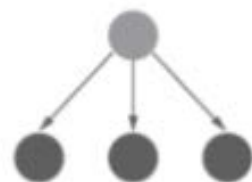
Александр Калиниченко

ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В МЕДИЦИНЕ

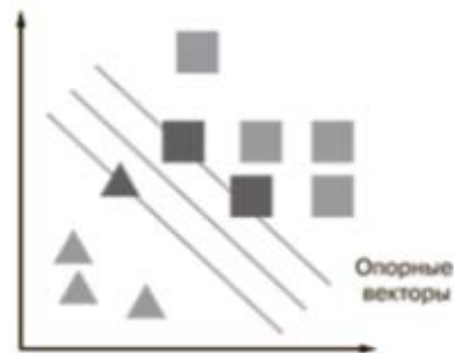
Модуль 2. Методы искусственного интеллекта

Тема 4. Обзор методов машинного обучения

ОСНОВНЫЕ АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ



Наивный
байесовский
классификатор

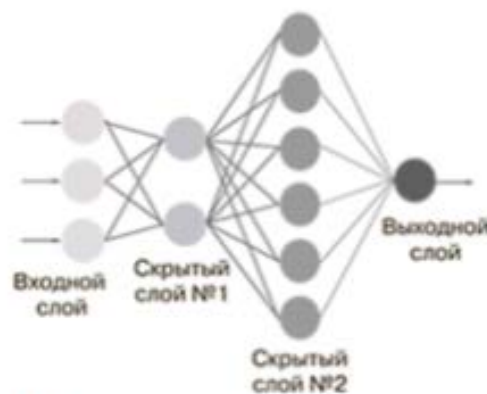


Метод опорных векторов

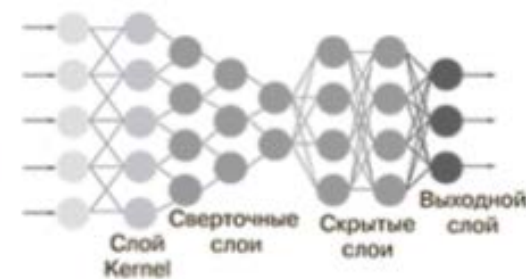


Случайные леса

Искусственные нейронные сети



Многослойный перцептрон



Сверточная нейронная сеть

АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ

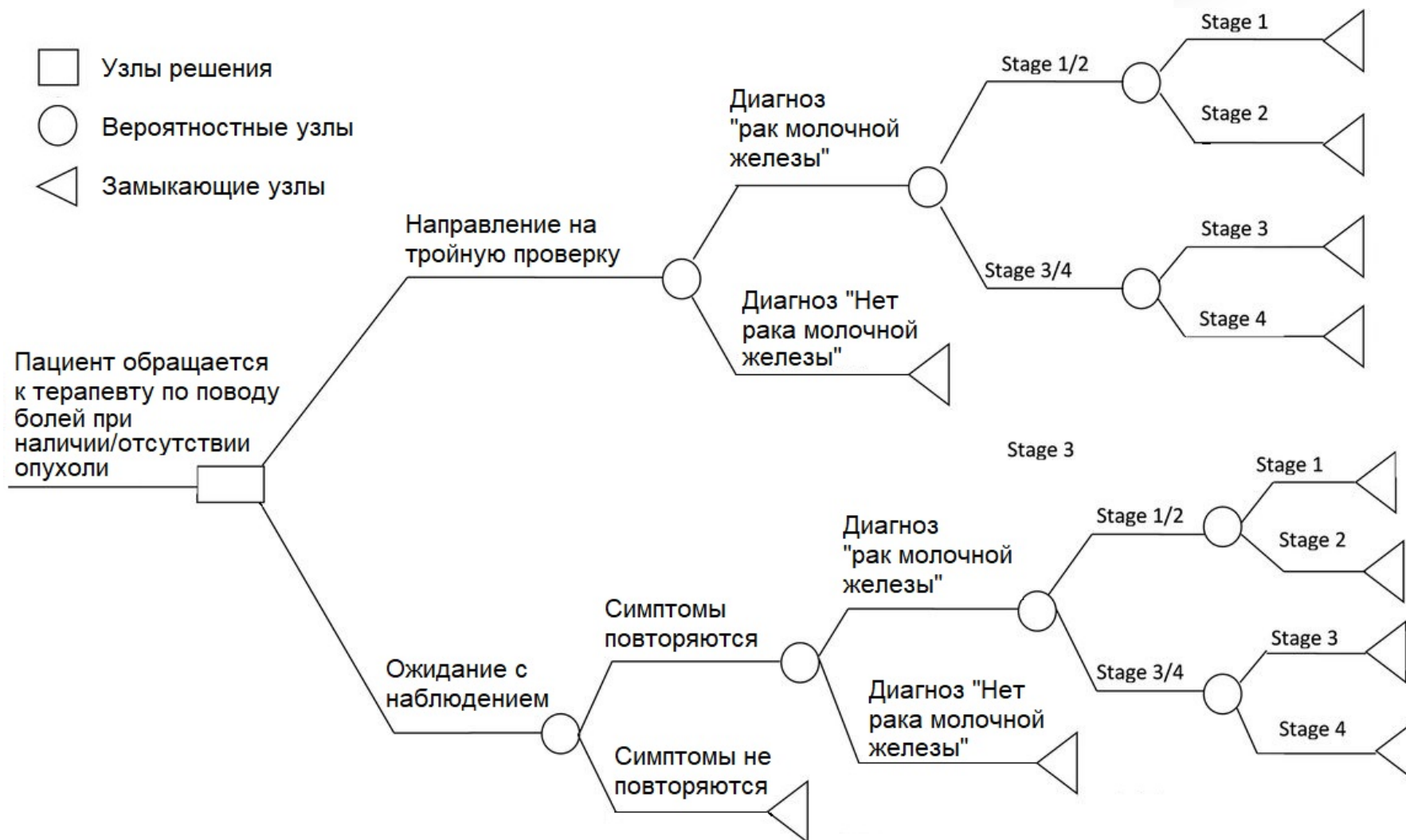
1. Дерево принятия решений

- Метод основан на использовании древовидного графа: модели принятия решений, которая учитывает их потенциальные последствия с расчётом вероятности наступления того или иного события

Дерево складывается из минимального числа вопросов, предполагающих однозначный ответ — «да» или «нет». Последовательные ответы на все эти вопросы приводят к правильному выбору

- Преимущество дерева принятия решений – оно структурирует и систематизирует проблему, а итоговое решение принимается на основе логических выводов

ПРИМЕР ДЕРЕВА ПРИНЯТИЯ РЕШЕНИЙ



НАИВНАЯ БАЙЕСОВСКАЯ КЛАССИФИКАЦИЯ

Наивные байесовские классификаторы основаны на теореме Байеса:

$$P(E | F) = \frac{P(F | E)P(E)}{P(F | E)P(E) + P(F | \neg E)P(\neg E)},$$

где F и E – независимые события, $P(E | F)$ – условная вероятность (E при условии F), а обозначение $\neg E$ – это «не E ».

Пусть E – факт наличия заболевания F – положительный результат анализа.
Тогда вероятность наличия заболевания при условии положительного результата анализа $P(E | F)$ может быть найдена по формуле Байеса.

Например: если $P(F | E) = 0.99$, а $P(E) = 0.0001$, то $P(E | F) = 0.0098$.

Только, если считать события E и F независимыми (**наивная** классификация).

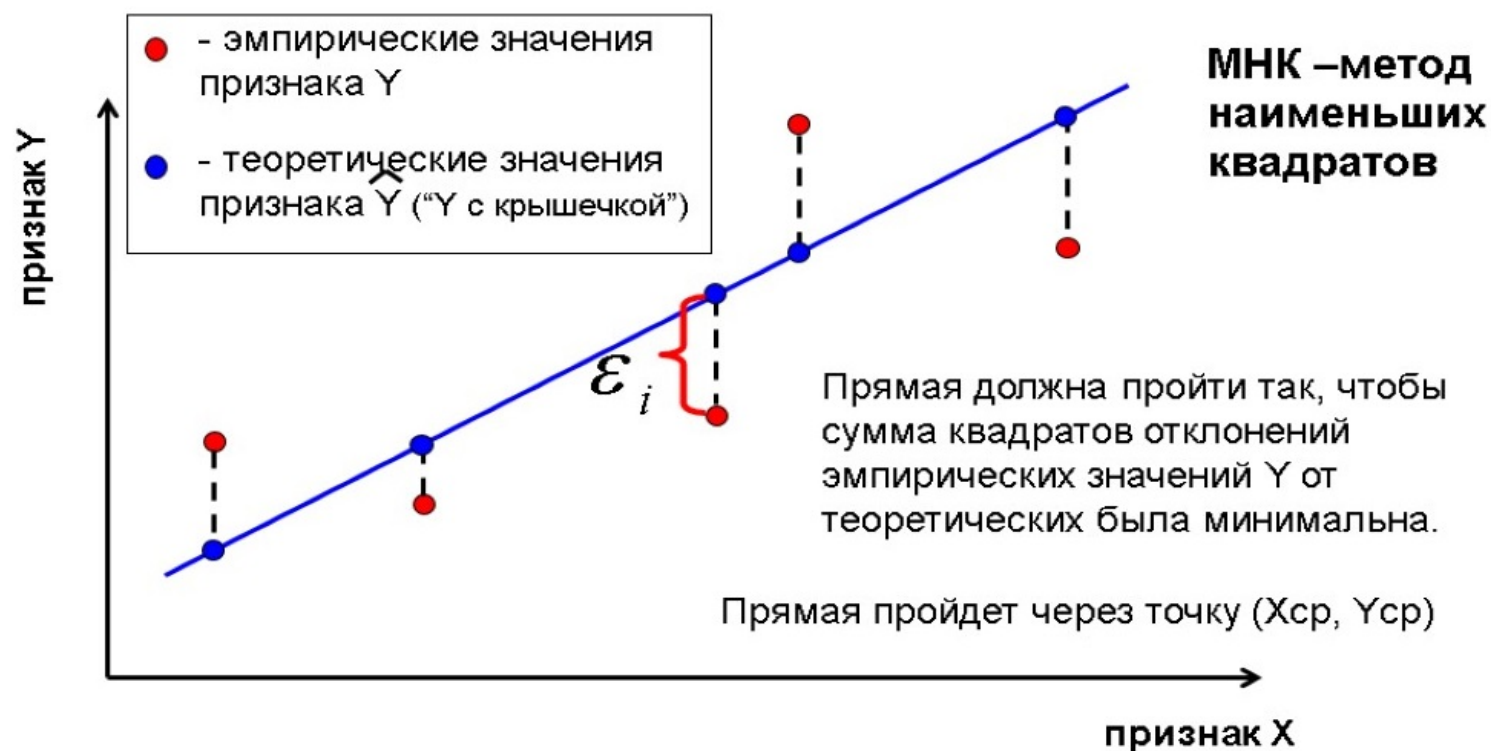
ЛИНЕЙНАЯ РЕГРЕССИЯ И МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

- С помощью линейной регрессии решают задачи по подгонке прямой, которая проходит через множество точек. Проводится прямая, измеряется расстояние от неё до каждой из точек (точки и линию соединяют вертикальными отрезками). Та кривая, в которой сумма расстояний будет наименьшей, и есть искомая.
- Линейная функция обычно используется при подборе данных для машинного обучения, а метод наименьших квадратов – в качестве метрики ошибок.

ИЛЛЮСТРАЦИЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Модель – уравнение прямой – $Y = a + b \cdot X$

Построение модели – расчет коэффициентов

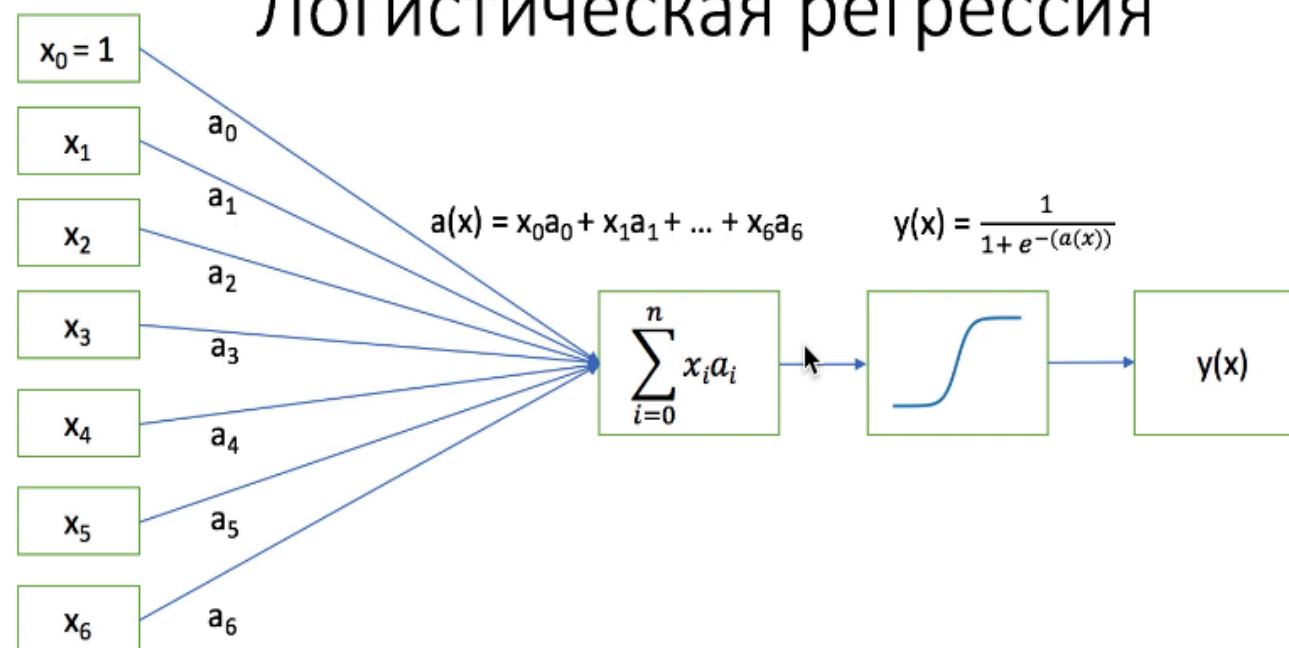


ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Это способ определения зависимости между переменными, одна из которых категориально зависима, а другие независимы. Для этого применяется логистическая функция.

Является мощным статистическим методом предсказания событий, который включает в себя одну или несколько независимых переменных.

Логистическая регрессия



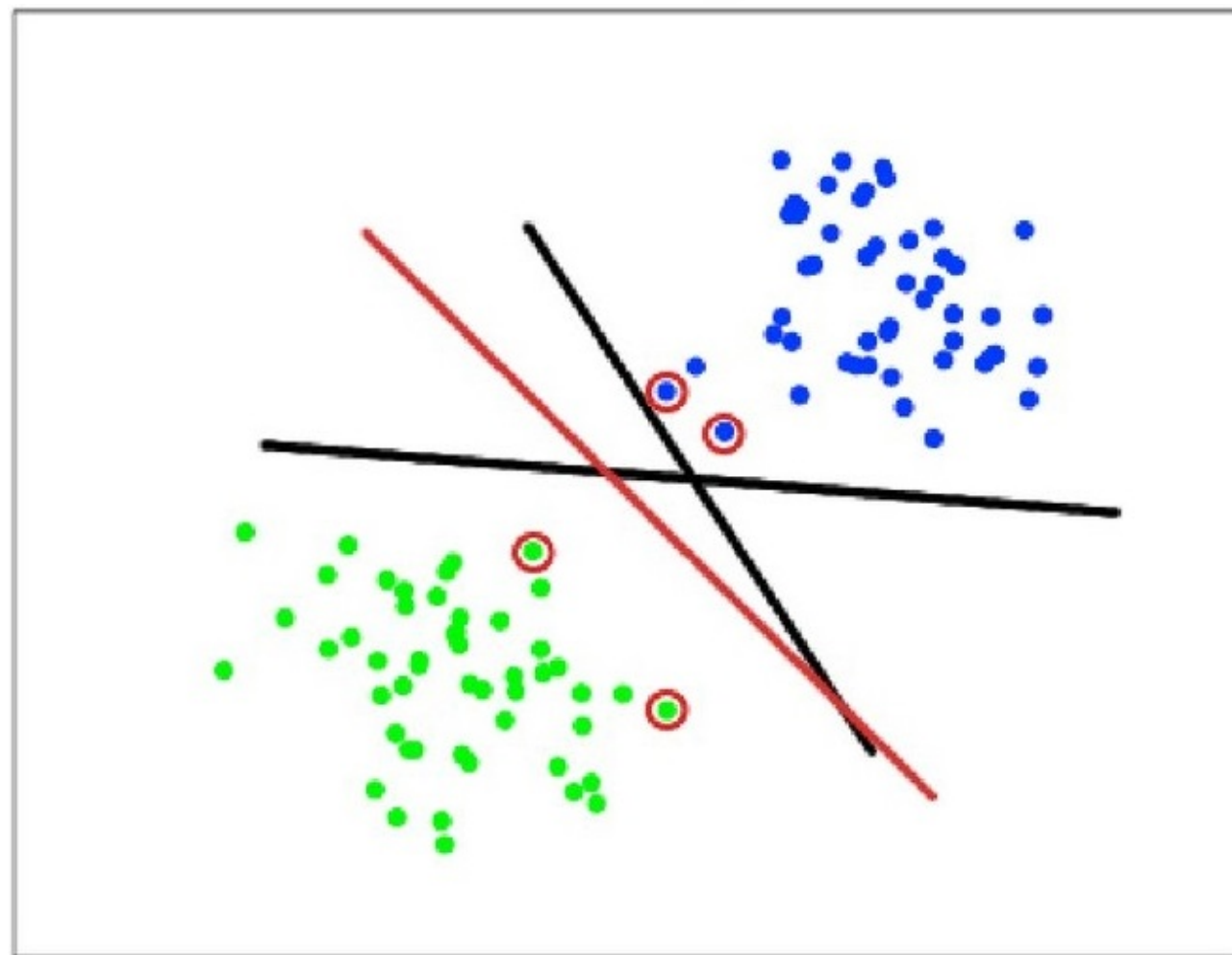
МЕТОД ОПОРНЫХ ВЕКТОРОВ

Метод опорных векторов (англ. SVM, Support Vector Machine)

Исходя из того что объект, находящийся в N -мерном пространстве, относится к одному из двух классов, строится гиперплоскость с мерностью $(N - 1)$, так чтобы все объекты оказались в одной из двух групп

Гиперплоскость генерируется таким образом, чтобы она была максимально удалена от самой близкой точки каждой группы

На практике полную линейную разделимость обычно невозможно обеспечить. Поэтому используется более «мягкий» алгоритм, допускающий некоторое количество ошибок



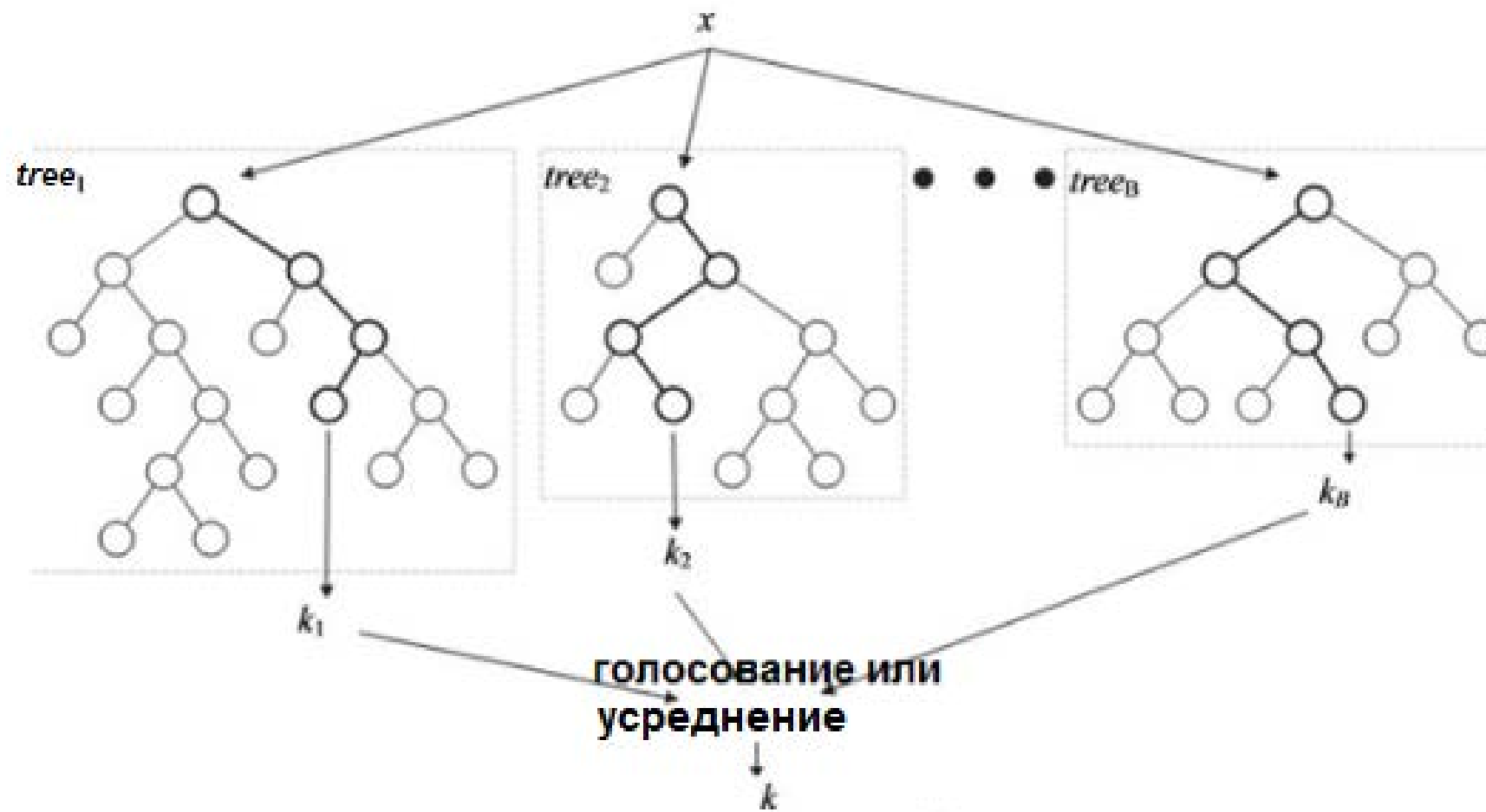
МЕТОД АНСАМБЛЕЙ

Базируется на алгоритмах машинного обучения, генерирующих множество классификаторов и разделяющих все объекты из вновь поступающих данных на основе усреднения или голосования.

Алгоритмы:

- бустинг (boosting) – преобразует слабые модели в сильные посредством формирования ансамбля классификаторов (улучшающее пересечение);
- бэггинг (bagging) – собирает усложнённые классификаторы, параллельно обучая базовые (улучшающее объединение).

ИЛЛЮСТРАЦИЯ МЕТОДА АНСАМБЛЕЙ

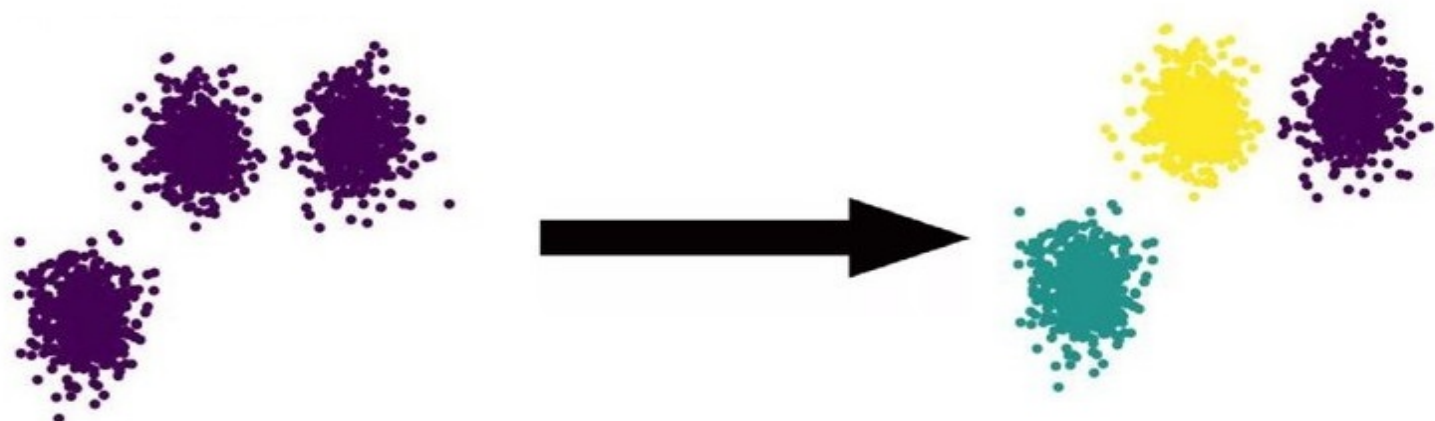


Сочетание метода ансамблей и деревьев принятия решений используется в одном из эффективных алгоритмов машинного обучения, известном как **«случайные леса»**

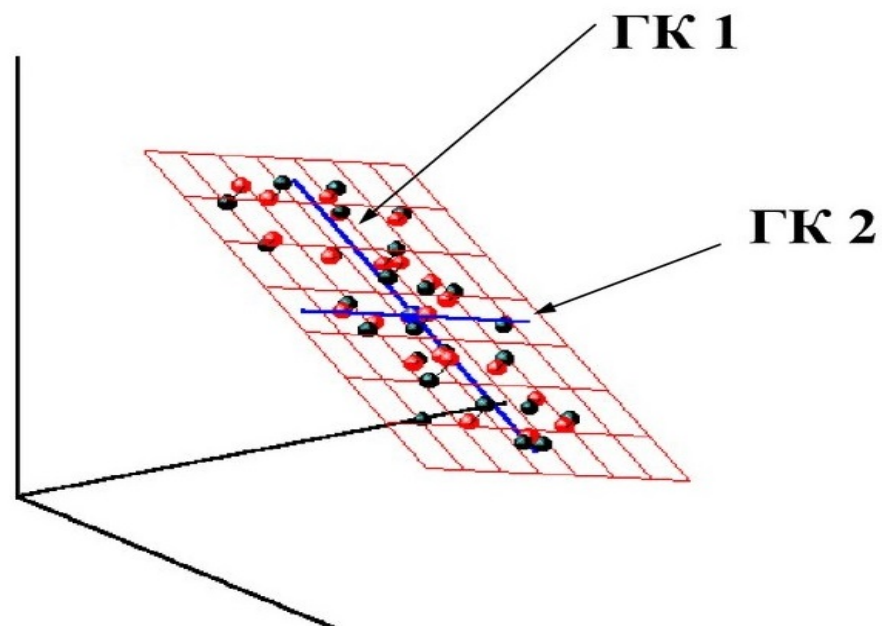
АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ

Кластеризация заключается в распределении множества объектов по категориям так, чтобы в каждой категории – кластере – оказались наиболее схожие между собой элементы.

КЛАСТЕРИЗАЦИЯ



МЕТОД ГЛАВНЫХ КОМПОНЕНТ



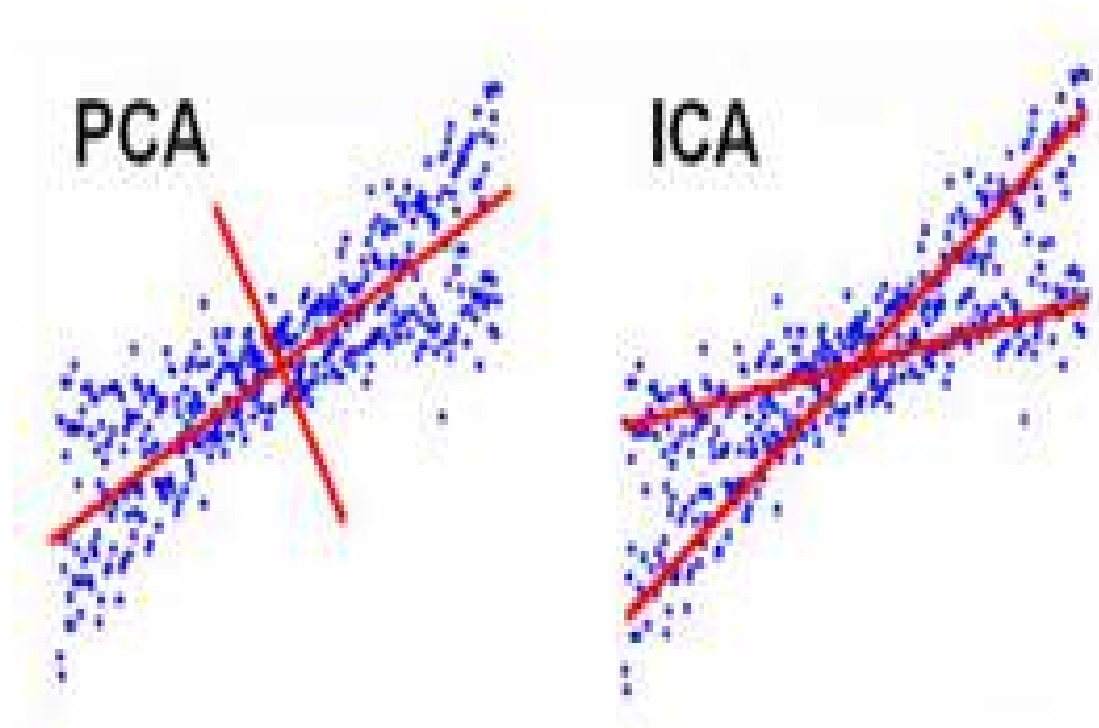
**ГК1 - направление
наибольшего разброса точек**

**ГК2 - направление
наибольшего разброса точек,
перпендикулярное ГК1**

Статистическая операция по ортогональному преобразованию, которая имеет целью перевод наблюдений за переменными, которые могут быть как-то взаимосвязаны между собой, в набор главных компонент – значений, которые линейно не коррелированы.

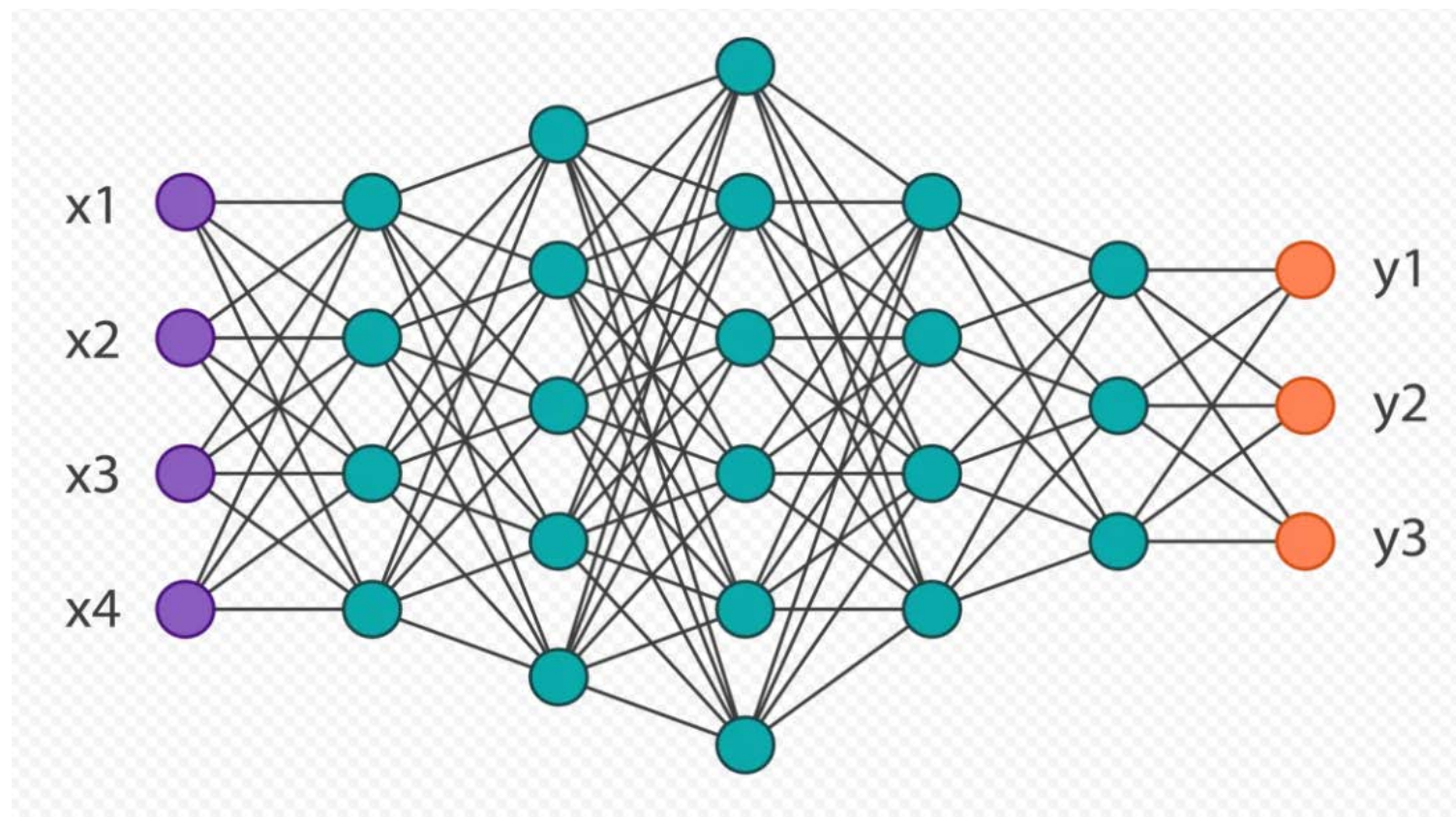
АНАЛИЗ НЕЗАВИСИМЫХ КОМПОНЕНТ

Это один из статистических методов, выявляющих скрытые факторы, оказывающие влияние на случайные величины.



ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ

Искусственная нейронная сеть (ИНС) – математическая модель, а также её программное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей — сетей нервных клеток живого организма



ПРИМЕРЫ ПРИМЕНЕНИЯ В МЕДИЦИНЕ

Пациенты в данном случае являются объектами.

Признаки

- *Бинарные*: наличие или отсутствие головной боли, кашля, сыпи и т.п.;
- *Порядковые*: оценка тяжести состояния (крайне тяжёлое, средней тяжести и др.)
- *Количественные*: объём лекарственного препарата, уровень гемоглобина в крови, показатели артериального давления и пульса, возраст, вес.

Задачи, которые могут быть решены с помощью машинного обучения:

- дифференциальная диагностика (определение вида заболевания);
- выбор наиболее оптимальной стратегии лечения;
- прогноз развития болезни, её длительности и исхода;
- оценка риска возможных осложнений;
- выявление синдромов – наборов симптомов, сопутствующих данному заболеванию.