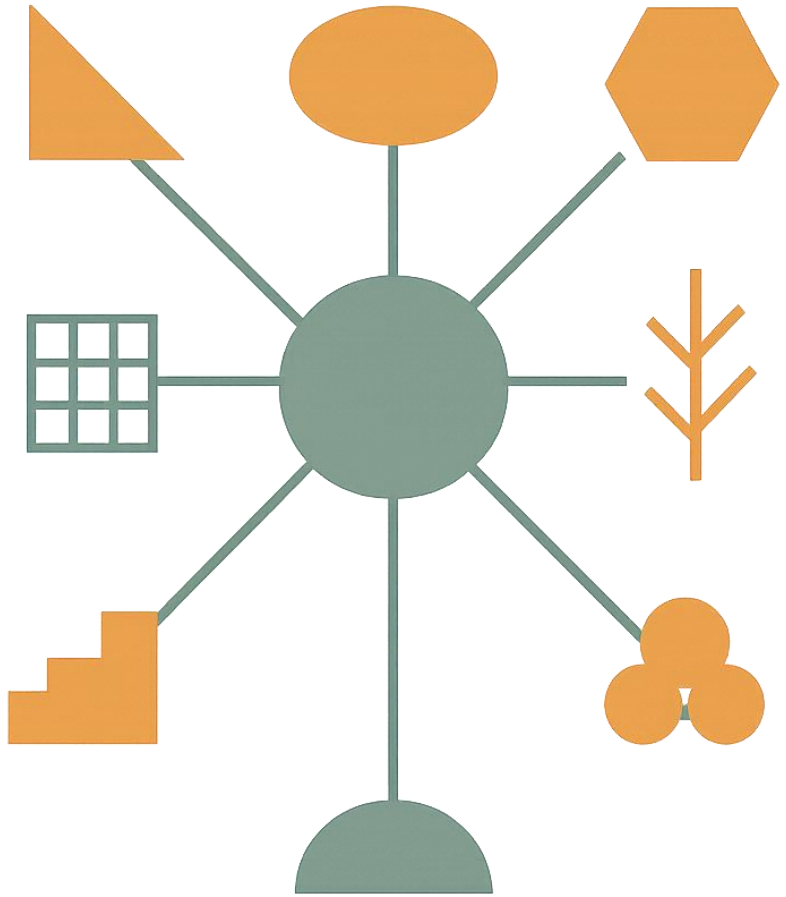
Методы классификации

Лекция 7

Методы классификации: принципы, различия и осознанный выбор

- Классификация реализуется через различные алгоритмические подходы
- Методы отличаются по предпосылкам, устойчивости и интерпретируемости
- Выбор алгоритма влияет на свойства модели и поведение при обучении
- Понимание методов — ключ к корректному применению и анализу

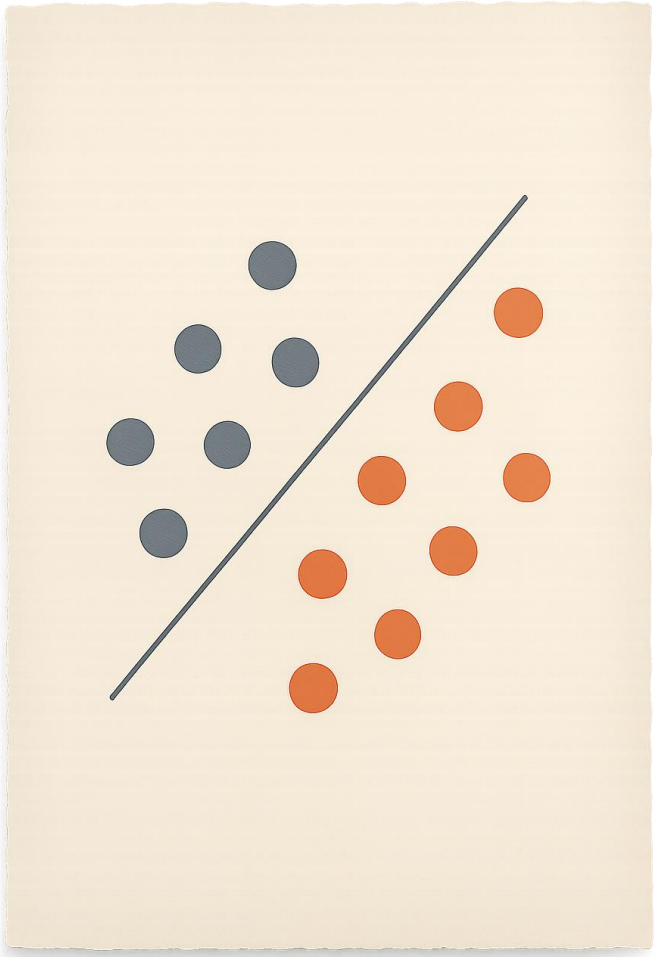


Группы методов классификации и их концептуальные различия

- Методы классификации формируются вокруг разных принципов построения решений
- Основные направления: линейные, вероятностные, на основе расстояний, деревья решений, ансамбли, нейросети
- Критерии различия: представление данных, способ обучения, интерпретируемость, устойчивость
- Классификация подходов помогает системно понимать архитектуру задач

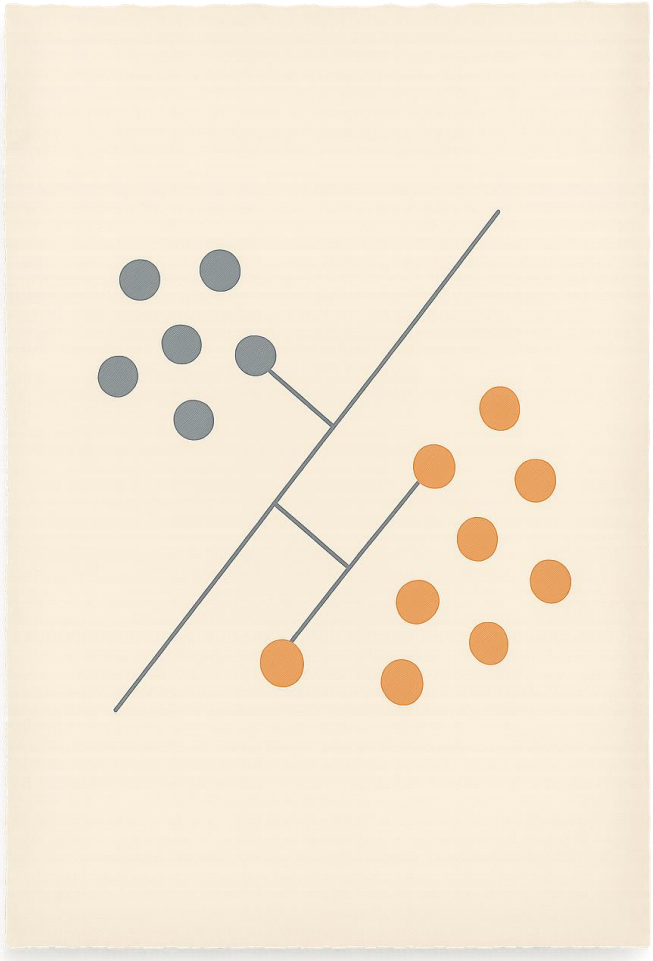
Линейные методы классификации: структура и свойства

- Линейные методы разделяют классы гиперплоскостью в пространстве признаков
- Примеры: логистическая регрессия, линейный SVM, персептрон
- Особенности: простота, интерпретируемость, чувствительность к структуре данных
- Применимы при условной линейной разделимости классов



Метод опорных векторов (SVM)

- SVM строит разделяющую гиперплоскость с максимальным отступом между классами
- Основная идея — использование опорных точек для определения границы
- Позволяет контролировать баланс между ошибками и устойчивостью
- Может быть расширен через ядровые функции для нелинейных задач



Вероятностные методы классификации: наивный байесовский классификатор

- Классификация формулируется как вычисление вероятности принадлежности объекта к классу
- Основная идея: использование априорных и условных вероятностей
- Наивное допущение — независимость признаков
- Простота, устойчивость и эффективность при ограниченных данных

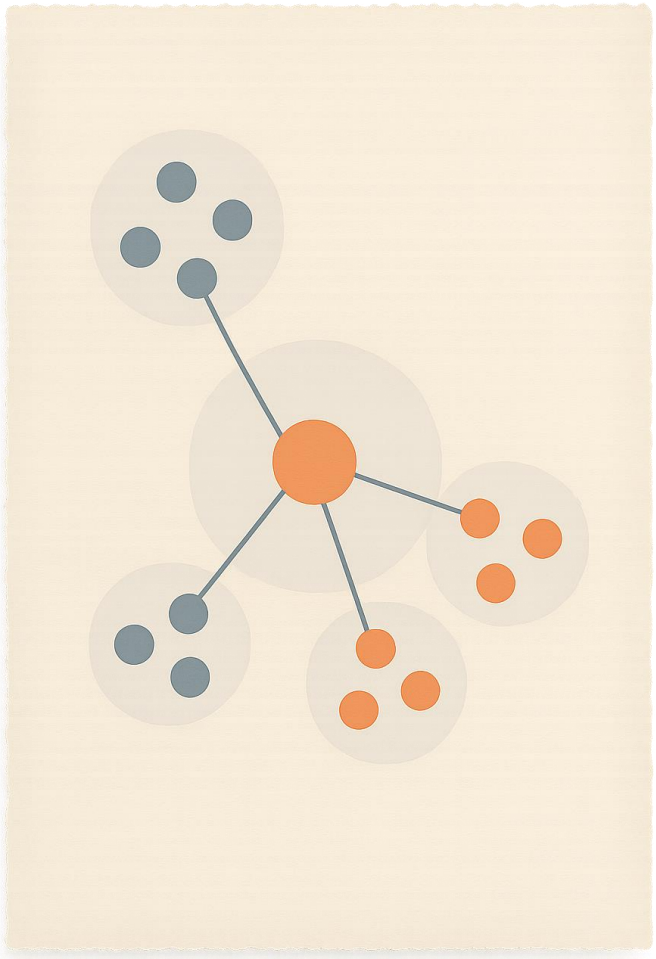
Разновидности наивных байесовских классификаторов

- Различаются по типу предположений о распределении признаков
- Гауссовский классификатор — моделирует непрерывные признаки с помощью нормального распределения (классификация объектов по числовым измерениям - рост, температура, интенсивность сигнала)
- Мультиномиальный классификатор — основан на счётных признаках и частотных распределениях событий (анализ текстов на основе частоты слов или терминов)
- Бернуллиевский классификатор — описывает бинарные признаки через распределение Бернулли (определение наличия или отсутствия слова в документе)
- Выбор модели определяется природой данных и типом наблюдаемых признаков

Байесовские сети

- Представляют зависимости между переменными в виде ориентированного ациклического графа
- Узлы соответствуют случайным величинам, рёбра — условным зависимостям
- Наивный байесовский классификатор — частный случай байесовской сети с независимыми признаками
- Используются для моделирования сложных вероятностных отношений и причинных связей

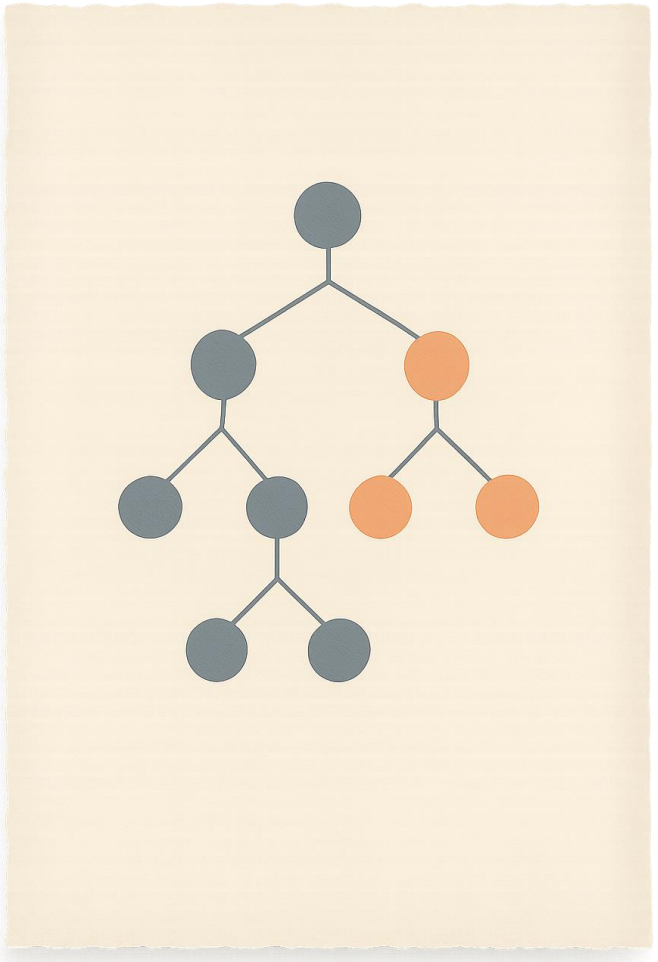
Методы на основе расстояний: k-ближайших соседей (kNN)



- Классификация основана на сходстве объектов в пространстве признаков
- Новый объект получает класс по меткам ближайших обучающих примеров
- Отсутствует явная стадия обучения, решение формируется при предсказании
- Качество зависит от метрики, параметра k и размерности признаков

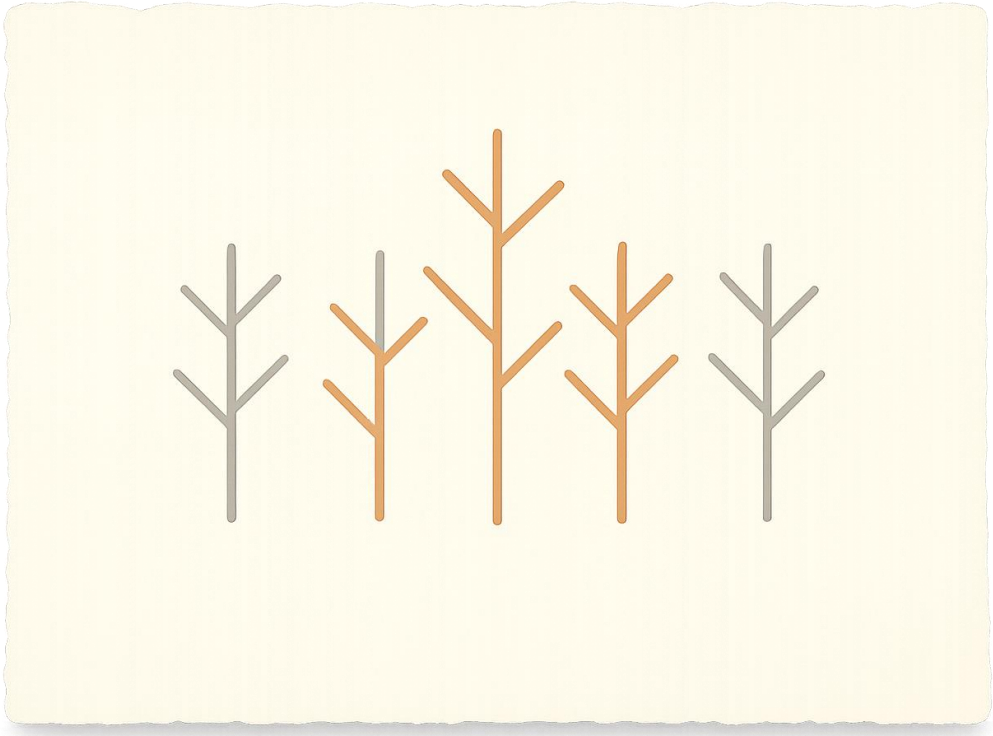
Деревья решений: иерархическая структура классификации

- Алгоритм последовательно делит данные по признакам, формируя древовидную структуру
- Каждое разветвление соответствует условию, каждая вершина — подмножество данных
- Простая интерпретация и визуальная наглядность
- Склонность к переобучению при избыточной глубине дерева



Ансамблевые методы классификации

- Основная идея: совмещение нескольких базовых моделей для улучшения обобщающей способности
- Ансамбль снижает влияние ошибок отдельных алгоритмов
- Основные подходы: бэггинг и бустинг
- Примеры: случайный лес, градиентный бустинг, AdaBoost



Нейросетевые методы классификации

- Нейросетевые классификаторы строят решение через последовательные обучаемые преобразования признаков
- Многослойные архитектуры моделируют нелинейные зависимости между входными данными и классами
- Каждый слой формирует скрытые (латентные) представления, отражающие обобщённые свойства данных
- Обладают высокой способностью к обучению сложных распределений, но требуют значительных вычислительных ресурсов и объёма данных

Сравнительный обзор методов классификации

Метод	**Преимущества**	**Ограничения**	**Типичные гиперпараметры**	**Примеры применения**
Наивный Байес	Простота, стабильность на малых и разреженных данных	Сниженная точность при коррелирующих признаках	Практически не требует настройки	Классификация текстов, фильтрация спама, анализ тональности
Байесовские сети	Интерпретируемость, моделирование причинных связей	Высокая вычислительная сложность, редкое практическое применение	Структура графа, априорные вероятности	Диагностические системы, биоинформатика, прогнозирование событий
Методы на основе расстояний (kNN)	Простая логика, адекватны при локальных структурах данных	Плохо масштабируются, чувствительны к размерности	Количество соседей (k), метрика расстояния	Рекомендательные системы, определение аномалий
Метод опорных векторов (SVM)	Хорош на умеренных данных с чёткой границей классов	Требует настройки ядра и нормализации признаков	Параметр регуляризации (C), тип и параметры ядра	Распознавание изображений, биоинформатика
Деревья решений	Интерпретируемость, быстрое обучение	Быстро переобучаются, слабые результаты на сложных данных	Глубина дерева, критерий разбиения	Анализ рисков, медицина, прогнозирование
Ансамблевые методы	Высокое качество на табличных данных, устойчивость	Потеря интерпретируемости, рост ресурсоёмкости	Количество деревьев, скорость обучения (для boosting)	Финансовый анализ, задачи прогнозирования, соревнования по ML
Нейросетевые методы	Лучшие результаты на сложных, неструктурированных данных	Требуют больших выборок и вычислительных ресурсов	Архитектура сети, скорость обучения, число эпох	Компьютерное зрение, обработка текста, генеративные модели

Тенденции развития методов классификации

- Гибридные подходы — объединение нейросетевых признаков и классических моделей (эмбеддинги текста + градиентный бустинг)
- Байесовские нейросети — учёт неопределённости и доверия к предсказаниям (медицинская классификация с вероятностной оценкой уверенности)
- Генеративные модели — синтез данных и самообучение (дополнение редких классов с помощью автоэнкодеров)
- Объяснимое и причинное ML — анализ вклада признаков и причинных связей
- Интеграция в системы принятия решений — классификация как часть цепочки интеллектуальных действий

Итоги лекции

- Классификация — основа большинства задач машинного обучения
- Существуют разные подходы: вероятностные, геометрические, алгоритмические, нейросетевые
- Каждый метод отражает своё предположение о структуре данных
- Универсального классификатора не существует
- Выбор модели определяется данными, целью и требованиями к интерпретации
- Современные тенденции: гибридные, объяснимые и вероятностные модели