

Лекция 6. Классификация: основные алгоритмы

Тема: *k-ближайших соседей (k-NN), Наивный Байесовский классификатор, логистическая регрессия*

1. Введение

Классификация — это один из ключевых методов интеллектуального анализа данных (Data Mining) и машинного обучения.

Она используется для автоматического отнесения объектов к заранее определённым классам на основе известных примеров.

Цель классификации — построить модель, которая, обучившись на данных с известными метками классов, сможет предсказывать метку для новых, ранее невидимых объектов.

Примеры задач классификации:

- Определение, является ли электронное письмо спамом.
 - Распознавание рукописных цифр.
 - Классификация отзывов на положительные и отрицательные.
 - Диагностика заболеваний по медицинским показателям.
-

2. Общая структура задачи классификации

Пусть у нас есть обучающая выборка:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

где

$x_i = (x_{i1}, x_{i2}, \dots, x_{im})$ — вектор признаков объекта,
 y_i — метка класса (например, 0 или 1).

Модель классификатора строится на основе обучающей выборки и затем используется для предсказания $\hat{y}$ — предполагаемого класса нового объекта.

3. Алгоритм k-ближайших соседей (k-Nearest Neighbors, k-NN)

3.1. Идея метода

Метод k -ближайших соседей — один из самых простых и интуитивно понятных алгоритмов классификации.

Он основан на предположении, что **похожие объекты имеют одинаковые метки классов**.

Для классификации нового объекта алгоритм:

1. Находит **k ближайших соседей** в обучающей выборке.
 2. Смотрит, к какому классу принадлежат эти соседи.
 3. Присваивает новому объекту **класс большинства** среди соседей.
-

3.2. Метрика расстояния

Для определения близости объектов выбирается **метрика расстояния**, чаще всего:

- **Евклидово расстояние:**

$$d(x, x') = \sqrt{\sum_{i=1}^m (x_i - x'_i)^2}$$

- **Манхэттенское расстояние:**

$$d(x, x') = \sum_{i=1}^m |x_i - x'_i|$$

- **Косинусное расстояние** (часто используется для текстов):

$$d(x, x') = 1 - \frac{x \cdot x'}{\|x\| \|x'\|}$$

3.3. Выбор параметра k

- Малое значение k → модель чувствительна к шуму.
- Большое значение k → сглаживает границы между классами, но может терять точность.

Обычно k подбирается экспериментально с помощью перекрёстной проверки (cross-validation).

3.4. Достоинства и недостатки

Преимущества:

- Простота реализации.
- Отсутствие необходимости обучения — модель «запоминает» данные.
- Гибкость при выборе метрики.

Недостатки:

- Большие вычислительные затраты при прогнозировании (нужно искать ближайших соседей).
 - Плохая работа на данных с большим количеством признаков (проклятие размерности).
 - Зависимость от масштаба признаков (нужна нормализация).
-

4. Наивный Байесовский классификатор

4.1. Теоретическая основа

Наивный Байесовский классификатор основан на **формуле Байеса**:

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)}$$

где:

- $P(C_i|X)$ — апостериорная вероятность класса C_i при наблюдаемых данных X ,
 - $P(X|C_i)$ — вероятность появления данных при данном классе,
 - $P(C_i)$ — априорная вероятность класса,
 - $P(X)$ — вероятность данных (одинакова для всех классов, можно не учитывать при сравнении).
-

4.2. «Наивное» предположение

Модель предполагает, что **все признаки независимы между собой** при условии класса:

$$P(X|C_i) = \prod_{j=1}^m P(x_j|C_i)$$

Это упрощает расчёты, даже если в реальности признаки связаны.

4.3. Пример

Пусть нужно определить, является ли письмо спамом.

Признаки — наличие слов: «скидка», «подарок», «деньги».

Если вероятность появления этих слов выше в спаме, чем в обычных письмах, модель классифицирует письмо как спам.

4.4. Типы наивных Байесовских классификаторов

- **MultinomialNB** — для текстов и счётчиков частот слов.
 - **BernoulliNB** — для бинарных признаков (например, наличие/отсутствие слова).
 - **GaussianNB** — для числовых данных, распределённых по нормальному закону.
-

4.5. Преимущества и недостатки

Преимущества:

- Простота и высокая скорость обучения.
- Хорошо работает на больших и разреженных данных (тексты).
- Устойчивость к нерелевантным признакам.

Недостатки:

- Независимость признаков редко выполняется в реальности.
 - Вероятности могут быть неточными при малых выборках.
-

5. Логистическая регрессия

5.1. Общая идея

Логистическая регрессия — это **линейная модель классификации**, которая предсказывает вероятность принадлежности объекта к классу.

Она используется, когда выходная переменная бинарная (например, 0 — «нет», 1 — «да»).

5.2. Математическая формула

$$P(y=1|x) = \frac{1}{1 + e^{-(w_0 + w_1x_1 + w_2x_2 + \dots + w_mx_m)}} \quad P(y = 1 | x) = \frac{1}{1 + e^{-(w_0 + w_1x_1 + w_2x_2 + \dots + w_mx_m)}}$$

Здесь:

- w_i — коэффициенты модели (веса признаков),
- экспоненциальная функция e^{-z} обеспечивает ограничение вероятности между 0 и 1.

5.3. Интерпретация

Логистическая регрессия вычисляет **логарифм отношения шансов (log-odds)**:

$$\ln \frac{P}{1-P} = w_0 + w_1x_1 + w_2x_2 + \dots + w_mx_m \quad \ln \frac{P}{1-P} = w_0 + w_1x_1 + w_2x_2 + \dots + w_mx_m$$

Это позволяет интерпретировать влияние каждого признака:

если $w_i > 0$, то увеличение признака повышает вероятность класса 1;

если $w_i < 0$, то снижает.

5.4. Обучение модели

Коэффициенты w_i подбираются методом **максимального правдоподобия** —

алгоритм ищет такие параметры, при которых наблюдаемые данные наиболее вероятны.

На практике используется **градиентный спуск** — итерационный метод оптимизации функции потерь.

5.5. Преимущества и ограничения

Преимущества:

- Простота интерпретации.
- Выдаёт вероятности, а не только метки классов.
- Хорошо работает на линейно разделимых данных.

Недостатки:

- Плохо работает на нелинейных зависимостях без дополнительной трансформации признаков.
- Чувствительна к мультиколлинеарности (взаимосвязи признаков).

6. Сравнительный анализ трёх методов

Характеристика	k-NN	Наивный Байес	Логистическая регрессия
Тип модели	Непараметрическая	Вероятностная	Линейная
Обучение	Не требуется	Быстрое	Требует оптимизации
Скорость предсказания	Медленная	Быстрая	Быстрая
Устойчивость к шуму	Средняя	Средняя	Высокая
Интерпретируемость	Средняя	Средняя	Высокая
Работа с большими данными	Сложно	Хорошо	Хорошо

7. Применение классификаторов

- **k-NN** — для распознавания изображений, биометрии, рекомендательных систем.
 - **Наивный Байес** — для анализа текстов, фильтрации спама, тональности отзывов.
 - **Логистическая регрессия** — для медицинской диагностики, кредитного скоринга, анализа отклика клиентов.
-

8. Заключение

Классификация — фундаментальная задача анализа данных.

Алгоритмы k-NN, Наивный Байес и логистическая регрессия представляют собой три разных подхода:

- **k-NN** — опора на близость объектов;

- **Байес** — использование вероятностей;
- **Логистическая регрессия** — построение линейной модели вероятности.

Выбор метода зависит от структуры данных, требуемой скорости и интерпретируемости.

На практике часто применяется **ансамбль моделей**, объединяющий преимущества разных подходов.

Список литературы

1. Хастие, Т., Тибширани, Р., Фридман, Дж. *Элементы статистического обучения*. — М.: Вильямс, 2018.
2. Bishop, C. M. *Pattern Recognition and Machine Learning*. — Springer, 2016.
3. Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. — MIT Press, 2012.
4. James, G., Witten, D., Hastie, T., Tibshirani, R. *An Introduction to Statistical Learning*. — Springer, 2021.
5. Гудфеллоу, Я., Бенджио, И., Курвиль, А. *Глубокое обучение*. — М.: ДМК Пресс, 2017.