

Метод K-Nearest Neighbors

Подгруппа 2

Выполнили студенты гр. 5030102/10401
Кузнецова Юлия и Табакова Виктория

Содержание

1. Введение

2. Метод knn

3. Применение

4. Пример использования

5. Преимущества и
недостатки метода

6. Заключение

Введение

K-ближайших соседей (K-Nearest Neighbors или просто KNN) — алгоритм классификации и регрессии, основанный на гипотезе компактности, которая предполагает, что расположенные близко друг к другу объекты в пространстве признаков имеют схожие значения целевой переменной или принадлежат к одному классу.

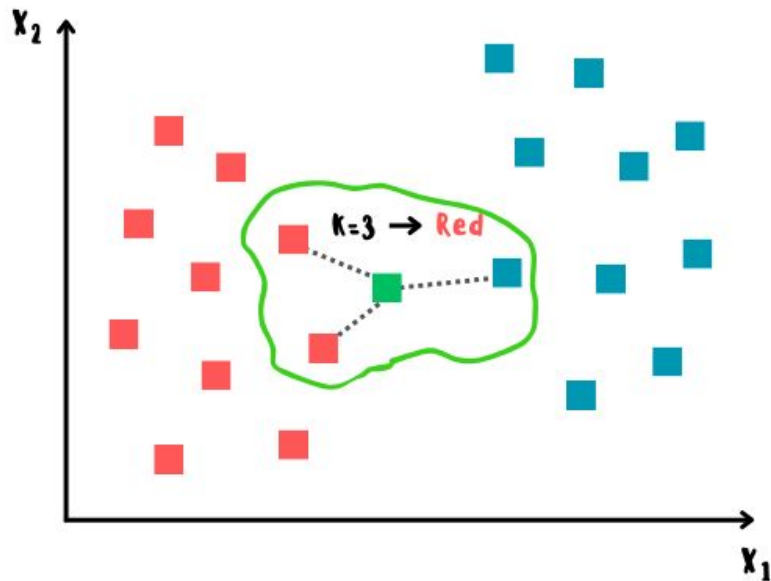
Процесс классификации начинается с появления новых немаркированных данных. Алгоритм вычисляет расстояния между размеченными объектами и неизвестными данными, чтобы определить их класс. Поскольку расстояния между объектами могут меняться для каждого нового набора данных, обучение алгоритма запоминанию этих расстояний нецелесообразно, так как число возможных позиций бесконечно, и хранить их все невозможно.

Для оценки точности модели используют метрику доли правильных ответов — отношение числа верных предсказаний к общему числу предсказаний. Значения этой метрики варьируются от 0 (модель бесполезна) до 1 (модель абсолютно точна).

Метод knn

Алгоритм:

1. Сначала вычисляется расстояние между тестовым и всеми обучающими образцами;
2. Далее из них выбирается k -ближайших образцов (соседей), где число k задается заранее;
3. Итоговым прогнозом среди выбранных k -ближайших образцов будет мода в случае классификации и среднее арифметическое в случае регрессии;
4. Предыдущие шаги повторяются для всех тестовых образцов.



Метод kNN

Параметры

- **К**

Определяет количество ближайших соседей, которые будут использоваться для классификации. Маленькое значение К может сделать модель чувствительной к шуму, а большое значение К может сгладить границы между классами.

- **Метрика**

С её помощью вычисляются расстояния между новым объектом и всеми объектами в обучающем наборе данных. Наиболее часто используемая метрика расстояния - евклидово расстояние.

Применение

Классификация:


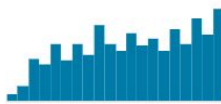
- **Распознавание изображений:**
Классификация изображений по категориям (например, собаки, кошки, птицы).
- **Анализ текста:** Классификация текстов по темам (например, новости, спорт, технологии).
- **Медицинская диагностика:**
Классификация пациентов по группам риска (например, диабет, рак).

Создание рекомендаций:

- **Рекомендательные системы:**
Предложение товаров или услуг, которые могут заинтересовать пользователя.
- **Системы поиска:** Выдача наиболее релевантных результатов поиска.
- **Прогнозирование поведения пользователей:** Предсказание будущих действий пользователей, например, покупки товара.

Пример использования

В качестве примера, мы подберем максимально похожую музыку. Для этого воспользуемся датасетом, который предоставляет список песен с 1950 по 2019 год, описывающих музыкальные метаданные как грусть, жанр, танцевальность, громкость, акустика и т. д.

#	artist_name	track_name	# release_date
 0 82.5k	5426 unique values	23689 unique values	 1950 2019
0	mukesh	mohabbat bhi jhoothi	1950
4	frankie laine	i believe	1950
6	johnnie ray	cry	1950

Преимущества и недостатки

Преимущества:

- Простота в реализации и интерпретации;
- Применяется во многих задачах, особенно в рекомендательных системах;
- Высокая точность прогнозов при правильном подборе k и метрики расстояния

Недостатки:

- Большое потребление памяти и низкая скорость работы из-за хранения и вычисления расстояний между всеми обучающими и тестовыми образцами
- Чувствительность к выбросам и шуму, а также к несбалансированным классам в данных

Заключение

Метод k-ближайших соседей (kNN) предлагает простой и интуитивно понятный подход к классификации и рекомендации. Его основное преимущество - легкость реализации. kNN особенно полезен для задач с нелинейными зависимостями и неопределенным распределением данных.

Однако, kNN также имеет свои недостатки. Чувствительность к размерности и затраты на вычисление могут сделать его неэффективным для больших наборов данных с высокой размерностью. Несмотря на ограничения, kNN остается популярным и ценным алгоритмом. Он может быть успешно использован в различных задачах, особенно когда данные имеют сложную структуру или доступны ограниченные ресурсы для обучения.