Введение в KNN

Денисов Егор

Кафедра математический методов прогнозирования
ВМК МГУ

Содержание

Введение

Пусть имеются:

- $(x_1, y_1), \cdots (x_N, y_N)$ - обучающая выборка
- $(x_{N+1}, y_{N+1}), \cdots, (x_{N+M}, y_{N+M})$ - тестовая выборка
- $x_i \in \mathbb{R}^d$

Определение

KNN (k-Nearest Neighbors) — семейство метрических алгоритмов классификации и регрессии, предсказание которых на каждом объекте определяется через K его ближайших соседей.

Предполагаем, что выполнена *гипотеза компактности* - то есть, похожим объектам соответствуют похожие значения целевой переменной.

Предсказание модели

Пусть $\mathbb{X} = (x_i, y_i)_{i=1}^N$ - обучающая выборка, $z \in \mathcal{X}$ - новый объект. Упорядочим $x_i \in \mathbb{X}$ по их близости к z:

$$\rho(z, x_1) \leq \cdots \leq \rho(z, x_N)$$

Классификация

$y(z) = argmax_{y \in \mathbb{Y}} \sum_{i=1}^{k} w_i \cdot [y_i = y]$

Регрессия

- $y(z) = \frac{1}{k} \sum_{i=1}^{k} y_i$ - среднее
- $y(z) = median(y_1, \cdots, y_k)$ - медиана
- $y(z) = \frac{\sum_{i=1}^{k} w_i \cdot y_i}{\sum_{i=1}^{k} w_i}$ - взвешенный алгоритм

Гиперпараметры модели

- Количество ближайших соседей k
- Функция расстояния $\rho(x, y)$:
  1. L2: $\rho(x, y) = ||x - y||_2^2 = \sum_{i=1}^{D} (x_i - y_i)^2$
  2. L1: $\rho(x, y) = ||x - y||_1 = \sum_{i=1}^{D} |x_i - y_i|$
  3. Косинусное расстояние: $\rho(x, y) = 1 - \frac{\langle x, y \rangle}{||x|| \cdot ||y||}$
  4. Расстояние Чебышёва: $\rho(x, y) = ||x - y||_\infty = \max_i |x_i - y_i|$
  5. Расстояние Махаланобиса: $\rho(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$
- Веса w_i:
  1. $w_i = 1$
  2. $w_i = \frac{1}{\rho(z, x_i) + \epsilon}, \epsilon > 0$
  3. $w_i = \alpha^i, \alpha \in (0, 1)$
- Стратегия поиска соседей (*kd-tree, brute, ball-tree*)

Особенности

- Непараметрический подход
- Лёгкая реализация
- Интерпретируемость
- Небольшое число гиперпараметров
- lazy learning (при обучении происходит только сохранение выборки)
- Неэффективен по памяти
- Проклятие размерности
- Чувствителен к масштабу данных, к выбросам

Подсчёт евклидовой метрики

Пусть $X \in \mathbb{R}^{N \times D}$ - обучающая выборка, $Z \in \mathbb{R}^{M \times D}$ - тестовая выборка.
Вычисление $\rho(x, z) = ||x - z||_2^2$ для всех пар (x, z) требует $3NMD$ операций.

Раскроем норму:

$$||x - z||_2^2 = \langle x - z, x - z \rangle = \langle x, x \rangle + \langle z, z \rangle - 2\langle x, z \rangle$$

$$||x - z||_2^2 = ||x||_2^2 + ||z||_2^2 - 2\langle x, z \rangle$$

Теперь вычисления стоят $2(N + M)D + 2NMD$ операций.

Датасет MNIST

Рассмотрим задачу классификации на датасете MNIST. Датасет содержит 70 тыс. изображений размера 28×28. Обучающая выборка состоит из 60 тыс. объектов, тестовая - из 10 тыс.

Каждое изображение представляет собой рукописную цифру от 0 до 9. Таким образом, перед нами задача классификации с 10 классами.

Кросс-валидация

Кросс-валидацию необходимо будет реализовать самостоятельно.

Важно - для перебора различных $k_1 < k_2 < \cdots < k_n$ можно считать матрицу расстояний лишь один раз, что значительно ускорит работу алгоритма.

Аугментации

Вы должны будете применить две стратегии аугментации:

1. Аугментация обучающей выборки
   - Аугментируем только тренировочную выборку
   - Обучаем модель на исходных и аугментированных данных
   - Проверяем качество на тестовой выборке
2. Аугментация тестовой выборки
   - Обучаем модель на исходной тренировочной выборке
   - Аугментируем тестовую выборку
   - Получаем предсказания для исходных и полученных путем аугментаций объектов из теста
   - Проводим голосование и получаем итоговое предсказание для объекта