Вспоминаем KNN

Кафедра Математических методов прогнозирования, факультет Вычислительной математики и кибернетики МГУ имени М. В. Ломоносова

Каратыщев Дмитрий

26 сентября 2024 г.

Содержание

1. Ключевая идея
2. Модель
3. Гиперпараметры
4. Преимущества
5. Недостатки
6. Евклидова метрика - вычисление
7. MNIST classification
8. Особенности кросс-валидации для KNN
9. Особенности аугментации

Ключевая идея

Пусть имеется выборка объектов
$\mathbb{X} = \{(x_1, y_1), \ldots, (x_n, y_n)\},\ x_i \in X,\ y_i \in Y\ \forall i \in \overline{1, n}$

Методом **k-го ближайшего соседа** будем называть такой алгоритм метрической классификации (регрессии), в котором предсказания целевой переменной на каждом объекте определяются через его k соседей.

Предполагаем, что выполнена **гипотеза компактности** - объекты одного класса близки, а объекты разных классов далеки друг от друга

Модель

Для произвольного $x \in \mathbb{X}$ упорядочим объекты выборки по возрастанию метрики:

$$\rho(x, x_1) \leq \rho(x, x_2) \leq \ldots \leq \rho(x, x_n)$$

Модель для классификации

$$a(x; \mathbb{X}) = \underset{y \in Y}{argmax} \sum_{i=1}^{n} I[y_i = y] w(i, x)$$

Модели для регрессии

1. $a_h(x; \mathbb{X}) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{\rho(x, x_i)}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\rho(x, x_i)}{h}\right)}$ - формула Надарая-Ватсона
2. $a(x; \mathbb{X}) = \frac{1}{k} \sum_{i=1}^{k} y_i$
3. $a(x; \mathbb{X}) = median\{y_1, \ldots, y_k\}$

Гиперпараметры

Основные гиперпараметры модели

- Число ближайших соседей k
- Метрика ρ (Минковского, Махаланобиса, Евклида)
- Вес i-го ближайшего соседа w_i ($\frac{1}{\varepsilon+\rho(x,x_i)}$ или $\alpha^i,\ \alpha \in (0,1)$)
- Алгоритм поиска ближайшего соседа (*kd-tree, ball-tree, brute*)

Преимущества

Перечислим плюсы KNN

- Легко реализуем на практике
- Хорошо интерпретируемая модель
- Небольшое количество гиперпараметров
- *Lazy learning*. На обучении достаточно сохранить матрицу Грама, все вычисления происходят на инференсе
- Непараметрический подход

Недостатки

Теперь перейдём к недостаткам

- Медленно работает на большом количестве данных
- Большая чувствительность к выбору k
- *Проклятие размерности* - необходимо хранить много данных, и метрика становится неинформативной при большой размерности признакового пространства

Евклидова метрика - вычисление

Вычислим $\rho(x, z) = ||x - z||_2^2$ для каждой пары объектов из обучающей выборки $\mathbb{X} \in \mathbb{R}^{n \times d}$ и тестовой $\mathbb{Z} \in \mathbb{R}^{m \times d}$

Для этого распишем норму через скалярное произведение:

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

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

Видно, что подсчёт попарных расстояний можно сделать в самом начале и сохранить матрицу Грама, что значительно ускорит алгоритм. Кроме того, используя быстрое матричное умножение, сложность вычислений можно упростить ещё сильнее.

MNIST classification

Рассмотрим задачу классификации рукописных цифр на датасете **MNIST**

Датасет содержит $70k$ полутоновых изображений рукописных цифр от 0 до 9. Каждое изображение имеет размер 28×28. На обучающую выборку выделяется $60k$ объектов, а на тестовую - $10k$

Так как каждое изображение - это одна из цифр от 0 до 9, то перед нами задача классификации на 10 классов.

Особенности кросс-валидации для KNN

Реализацию кросс-валидации придётся реализовывать самостоятельно.

Преимуществом *KNN* для *CV* является возможность быстро перебирать различные значения k.

Поэтому для набора значений $k_1 < k_2 < \ldots < k_s$ нам необходимо посчитать матрицу расстояний лишь один раз.

Особенности аугментации

При реализации аугментации можно воспользоваться одной из следующих стратегий:

- **Аугментация обучающей выборки**. Проводим аугментацию только обучающих данных, обучаем модель на расширенной выборке и оцениваем качество на тесте.
- **Агументация тестовой выборки**. Обучаем модель на изначальных обучающих данных, аугментируем тестовую выборку, получаем предсказания для оригинальной тестовой выборки и аугментированной и предсказываем итоговую метку через обычное голосование.


Я вам
ЗАПРЕЩАЮ
выбирать k = 1 в KNN