Методические указания к практическим работам по машинному обучению

Ксемидов Б.С.

27 ноября 2019 г.

Оглавление

1. Введение

Данные методические указания предназначены для усвоения теоретического материала по машинному обучению на практике. Здесь рассматриваются базовые методы, необходимые для решения задач машинного обучения. Предполагается, что в результате ознакомления с содержимым данной методической разработки читатель научится анализировать поставленную задачу, выбирать наилучший метод решения для неё и реализовывать соответствующую модель машинного обучения.

2. Практическое работа №1

2.1 Задание

Реализовать алгоритм k ближайших соседей, подобрав оптимальный параметр k, использовав для этого кросс-валидацию leave-one-out.

2.2 Теоретический материал

2.2.1 Алгоритм ближайшего соседа

Алгоритм ближайшего соседа - метрический способ классификации объектов, являющийся методом "ленивого обучения" (lazy learning), так как для его использования необходимо постоянно хранить всю выборку данных в памяти для последующей классификации новых объектов.

Пусть задана некоторая выборка X^l, где l - размер выборки.

1. Для нового классифицируемого объекта считается его расстояние до всех уже известных объектов из выборки X^l, используя выбранную метрику (см. 2.2.5)
2. Объекты упорядочиваются по возрастанию расстояния
3. Выбирается первый объект
4. Классифицируемому объекту присваивается тот же класс

2.2.2 Алгоритм k ближайших соседей

Алгоритм k ближайших соседей является модификацией предыдущего алгоритма. Идея модификации проста - вместо использования для классификации одного самого близкого соседа берутся первые k соседей, так вероятность ошибки классификации может быть снижена.

Сам алгоритм описывается следующим образом:

1. Для нового классифицируемого объекта считается его расстояние до всех уже известных объектов из выборки X^l, используя выбранную метрику (см. 2.2.5)
2. Объекты упорядочиваются по возрастанию расстояния
3. Выбираются первые k объектов

4. Среди данных k объектов выбирается тот класс, который встречается чаще

2.2.3 Методы оценки модели машинного обучения

Существует два способа оценки качества классификации модели машинного обучения:

- На отложенных данных
- Кросс-валидация

Алгоритм оценки модели на отложенных данных:

1. Исходная выборка (X^l) делится на две части - обучающую и контрольную
2. Производится обучение модели машинного обучения с помощью обучающей части исходной выборки
3. Производится оценка полученной модели с помощью контрольной части выборки (например, в виде количества/доли правильных ответов)

Алгоритм оценки модели с помощью кросс-валидации leave-one-out:

1. Выборка делится на заданное количество частей
2. Каждая часть по очереди участвует в оценке качества модели машинного обучения (см. рисунок 2.1)
3. Для общей оценки можно взять среднее арифметическое от долей правильных ответов за каждую часть

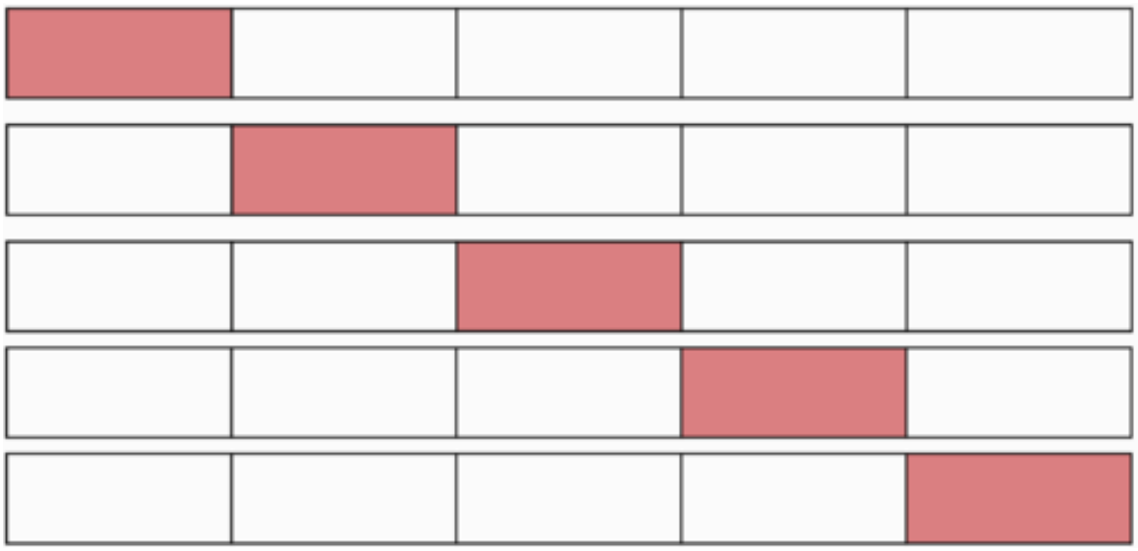

Рис. 2.1: Кросс-валидация

2.2.4 Алгоритм подбора оптимального параметра k

1. Задаются различные возможные значения k (например, 1, 2, 4, 8, 16, 32)
2. Для всех k
   (a) Производится кросс-валидация (см. алгоритм в 2.2.3). Выборка разбивается на заданное количество частей.
   (b) Объекты каждой из частей (контрольной выборки) классифицируются с помощью других частей (обучающей выборки) и для каждого случая считается доля правильных ответов.
   (c) Считается среднее арифметическое всех долей(AV_k)
3. Среди AV_k для каждого k ищется максимальное (это и будет оптимальным значением)

2.2.5 Метрики

- $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ - евклидово расстояние
- $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$ - расстояние городских кварталов
- $d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$ - расстояние Минковского (p выбирается самостоятельно, $p > 0$)

3. Критерии оценки

Работа считается выполненной, если частота правильных ответов модели машинного обучения больше вероятности случайного угадывания. В частности, для набора данных ирисов Фишера необходимо получить не менее 90% правильных ответов.

Рекомендуемая литература

[1] Курс лекций К.В. Воронцова. — URL: https://bit.ly/1bCmE3Z (дата обращения: 26.11.2019). - Текст: электронный.