Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу Искусственный интеллект (Машинное обучение)

Студент: Д. О. Петрухин
Преподаватель: Ахмед Самир Халид
Группа: М8О-301Б
Дата:
Оценка:
Подпись:

Москва, 2021

Общая постановка задачи

Найти себе набор данных (датасет) для следующей лабораторной работы и проанализировать его. Выявить проблемы набора данных, устранить их. Визуализировать зависимости, показать распределения некоторых признаков. Реализовать алгоритмы K ближайших соседей с использованием весов и Наивный Байесовский классификатор и сравнить с реализацией библиотеки sklearn.

Используемый датасет

В качестве набора данных был выбран датасет, содержащий информацию физико-химических тестов о белом вине.

Описание признаков:

- Входные данные
    1. fixed acidity (фиксированная кислотность) - вещественный признак.
    2. volatile acidity (летучая кислотность) - вещественный признак.
    3. citric acid (лимонная кислота) - вещественный признак.
    4. residual sugar (остаточный сахар) - вещественный признак.
    5. chlorides (хлориды) - вещественный признак.
    6. free sulfur dioxide (свободный диоксид серы) - вещественный признак.
    7. total sulfur dioxide (общий диоксид серы) - вещественный признак.
    8. density (плотность) - вещественный признак.
    9. pH (водородный показатель) - вещественный признак.
    10. sulphates (сульфаты) - вещественный признак.
    11. alcohol (алкоголь) - вещественный признак.
- Выходные данные
    1. quality (качество) - от 0 до 10.

Анализ датасета

В ходе анализа датасета были сделаны следующие выводы:

- Датасет не содержит пропусков.
- Данные датасетов нор.
- Имеется сильная корреляции пары признаков.

Для визуализации распределений признаков использовался метод hist() библиотеки pandas. Этот метод вызывает метод matplotlib.pyplot.hist() для каждого признака датафрейма.

Работа с данными

Основная задача предсказания: предсказать качество вина, исходя из его физико-химических свойств.

В ходе работы с данными я разделил выходные данные на два класса:

- качество вина более 5
- качество вина не более 5.

Также отнормировал входные признаки для корректной работы некоторых моделей классификации. От одного коррелируемого признака избавился (density)

Расчет оценки качества классификации

Качество определяется как доля правильных ответов, то есть алгоритм соотнес объект к истинному классу, к общему числу объектов.

KNN

Алгоритм:

1. Загрузить данные.
2. Для каждого примера данных рассчитать расстояние между примером запроса и текущим примером данных. А затем это расстояние в упорядоченную коллекцию.
3. Отсортировать коллекцию по расстоянию в порядке возрастания.
4. Выбрать первые K элементов коллекции.
5. Получить лейблы K элементов коллекци.
6. Вернуть наиболее встречающийся лейбл.

Результаты работы:

```
Расчет точности реализации KNN на обучающей выборке:  85.77 %
Расчет точности реализации KNN на тестовой выборке:  74.11999999999999 %
Результат sklearn реализации KNN на обучающей выборке: 85.77%
Результат sklearn реализации KNN на тестовой выборке:  74.12%
```

Наивный байесовский классификатор

Алгоритм:

1. Вычисляем вероятности $P(C_i)$ каждого класса.
2. Вычисляем условные вероятности $P(F_i|C_i)$ для каждого признака. В моей реализации, поскольку признаки являются непрерывными величинами, используется нормальное распределение.
3. По следующей формуле вычисляем класс, к которому относится объект:

$$classify(f_1, f_2, ..., f_n) = arg\max_c[ln(P(C = c)) + \sum_{i=1}^{n} ln(P(F_i = f_i|C = c))]$$

Результаты работы:

```
Результат собственной реализации наивного байесовского классификатора на обучающей выбор
Результат собственной реализации наивного байесовского классификатора на тестовой выборк

Результат sklearn реализации классификатора на обучающей выборке: 71.85%
Результат sklearn реализации классификатора на тестовой выборке: 71.09%
```

Выводы

В ходе данной лабораторной работы я проанализировал 1 датасет и реализовал алгоритмы KNN и наивный байесовский классификатор.