

**Московский авиационный институт
(национальный исследовательский университет)**

**Факультет информационных технологий и прикладной
математики**

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу Машинное обучение

Студент: С. Я. Симонов
Преподаватель: Ахмед Самир Халид
Группа: М8о-406Б-18
Дата:
Оценка:
Подпись:

Москва, 2021

1. Условие 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 К ближайших соседа с использованием весов и Наивный Байесовский классификатор и сравнить с реализацией библиотеки `sklearn`.

2. Метод решения

Для анализа и классификации, я взял датасет с съедобными и несъедобными грибами. Классификация здесь соответствующая.

Сперва я проверил дата сет на пропуски, их не оказалось. Затем удалил столбец содержащий в себе одинаковую информацию для всех строк. Далее я преобразовал количественные признаки из буквенной формы в численную. И наконец преобразовал строковые столбы в фиксированные столбцы индикаторов.

Далее я реализовал KNN с использованием весов. Расстояние определялось функцией обратного квадрата расстояния. Для каждого класса находим сумму весов точек, чей класс равен данному. Возвращаем класс.

Второй алгоритм – это наивный Байесовский классификатор. Сначала считаются выброшенные средние и дисперсии для каждого признака в зависимости от класса, для каждого класса находим оценку вероятности того, что наблюдение принадлежит данному классу. Находим условную вероятность с помощью плотности вероятности данного распределения. Предсказываем класс, для которого вероятность по формуле Байеса наибольшая.

3. Результаты

accuracy KNN: 1.0
accuracy for sklearn's KNN: 1.0
accuracy Naive Bayes: 1.0
accuracy for sklearn's Naive Bayes: 0.9101538461538462

4. Вывод

Как видно из результатов, самописный KNN, показал такой же результат, как и из sklearn. Скорее всего, такие конечные результаты были достигнуты по следующим двум причинам:

- 1) в рамках поставленной задачи достаточно весомый датасет
- 2) большое число признаков, которые представляют четкое разбиение на два класса

Наивный Байесовский классификатор показал результат лучше, чем из sklearn.