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 K 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 грабами. Классификация здесь соответсвующая.

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. Растояние определялось функцией обратного квадрата растояния. Для каждого класса находим сумму весов точек, чей класс равен данному. Ввозвращаем класс.

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

3. Результаты

accuracy KNN: 0.9046153846153846
accuracy for sklearn's KNN: 0.9046153846153846
accuracy Naive Bayes: 0.8935384615384615
accuracy for sklearn's Naive Bayes: 0.8935384615384615

4. Вывод

Как видно из результатов, оба самописных алгоритма показали одинаковые результаты с модулем sklearn.

Из существенных различий выделяется только время исполнения кода.

Немало важно упомянуть о значимости обработки данных. В ходе выполнения лабораторной работы возникли определенные сложности при отчистке данных, по данной причине пришлось искать несуществующую ошибку.