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: И. Д. Черненко
Преподаватель: Ахмед Самир Халид
Группа: М8О-306Б-18

Дата:
Оценка:
Подпись:

Москва, 2021

Лабораторная работа №1

Задача: 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.

1 Метод решения

Быд выбран датасет из двух классов говорящих о наличии или отсутствии диабета у индийских пим в зависимости от различных медицинских параметров.

Был проведен анализ алгоритма K-ближайших соседей с весами при различных k, для двух реализаций (своей и из sklearn). Написанная реализация работает по следующему алгоритму:

- Считаем евклидово расстояние от точек из тренировочного до тестовой точки
- Сортируем список расстояний и берём k первых элементов
- Присваиваем каждой точке вес равный обратному расстоянию до k соседей
- Возвращаем ответ с наибольшей суммой весов

Для наивного Байесовского классификатора ситуация по тестирования аналогична. Написанная реализация работает по следующему алгоритму:

- Отображаем значения классов на вероятности и вычисляем Гауссовскую функцию плотности вероятности
- Для это считаем выборочного среднее и стандартное отклонение для атрибутов
- Предварительно разделяем обучающий набор по классам
- Возвращаем ответ с наибольшей вероятностью

2 Примеры работы

```
KNN:
Статистика для стандартной реализации:
0.8311688311688312 при k = 21
Вероятность успеха реализованным мной методом: 0.8311688311688312 при k = 21
BAYES:
Статистика для стандартной реализации:
0.8051948051948052
Статистика для моей реализации:
0.8051948051948052

Без стандартизации:
KNN:
```

```
Статистика для стандартной реализации:
0.8246753246753247 при k = 27
Вероятность успеха реализованным мной методом: 0.8116883116883117 при k = 27
BAYES:
Статистика для стандартной реализации:
0.7987012987012987
Статистика для моей реализации:
0.8051948051948052
```

Как можно заметить без стандартизации данных точность падает.

3 Выводы

Выполнив данную лабораторную работу я могу сделать несколько выводов:

- Данные при машинном обучении очень важны, поэтому следует избавляться от лишних или же стараться заполнять пробелы.
- Наивный Байесовский классификатор не так чувствителен к стандартизации, как KNN.
- Оптимальный параметр k значительно зависит от того были стандартизованы данные или нет и начинает расти при их разрозненности.
- На данном датасете максимальная точность была достигнута с помощью KNN.

Список литературы

[1] *mlmastery*
URL: `https://machinelearningmastery.com/` (дата обращения:10.04.2021).