

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: Э. Д. Орозбакиев
Преподаватель: Ахмед Самир Халид
Группа: М8О-406Б
Дата:
Оценка:
Подпись:

Москва, 2021

Лабораторная работа №1

Задача: Найти себе набор данных и проанализировать его. Выявить проблемы набора данных и устранить их. Визуализировать зависимости, показать распределение некоторых признаков. Реализовать алгоритмы К ближайших соседей с использованием весов и наивный Баесовский классификатор и сравнить с реализацией библиотеки sklearn.

1 Метод решения

Для анализа и классификации я взял классический датасет с информацией о крушении Титаника. Классификация здесь бинарная: нужно предсказать, выжил ли человек в катастрофе.

Первой задачей было убрать пропуски в данных. Я посчитал соотношение пропусков к количеству наблюдений в процентах и построил тепловую карту для пропусков с помощью seaborn. Далее путём анализа таблиц я устранил пропуски в тренировочных и тестовых данных. Далее я преобразовал данные для применения к ним моделей: убрал несущественные для предсказания признаки и объединил два признака в один для уменьшения размерности.

Далее я реализовал алгоритм К ближайших соседей с использованием весов. Краткое описание алгоритма:

1. Для данной точки считаем евклидовы расстояния от неё для каждой точки из тренировочного набора данных и добавляем их в список.
2. Сортируем этот список и берём К первых элементов.
3. Присваиваем каждой такой точке вес, равный обратному расстоянию до рассматриваемой точки, делённому на сумму таких обратных расстояний для всех К соседей.
4. Для каждого класса находим сумму весов точек, чей класс равен данному.
5. Возвращаем класс с наибольшей суммой весов.

Вторым алгоритмом был наивный байесовский классификатор:

1. Считаем выборочные средние и дисперсии для каждого признака в зависимости от класса.
2. Для каждого класса находим оценку вероятности того, что случайное наблюдение принадлежит данному классу путём деления количества наблюдений с данным классом на общее число наблюдений.
3. Находим условную вероятность признаком при условии данного класса с помощью плотности вероятности нормального распределения.
4. Предсказываем класс, для которого вероятность по формуле Байеса наибольшая.

2 Результаты

Accuracy for weighted KNN with $K = 7$: 0.8729016786570744

Best accuracy for sklearn's KNN: 0.6714728297362111

Best accuracy for sklearn's KNN with normalized data: 0.9760191846522782

Accuracy for custom Naive Bayes: 0.7961630695443646

Accuracy for sklearn Naive Bayes: 0.7991021324354658

3 Выводы

Мы видим, что нормализация данных в таблице по столбцам значительно увеличивает точность предсказаний. Мои модели не сильно отстают в точности предсказаний от моделей sklearn, но работают медленнее.