Московский авиационный институт (национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу «Искусственый интеллект»

Студент: А. Е. Максимов
Преподаватель: А. С. Халид
Группа: М8О-306Б

Дата:
Оценка:
Подпись:

Москва, 2020

Лабораторная работа №7

Задача:

Необходимо сформировать два набора данных для приложений машинного обучения. Первый датасет должен представлять из себя табличный набор данных для задачи классификации. Второй датасет должен быть отличен от первого, и может представлять из себя набор изображений, корпус документов, другой табличный датасет или датасет из соревнования Kaggle, предназначенный для решения интересующей вас задачи машинного обучения. Необходимо провести анализ обоих наборов данных, поставить решаемую вами задачу, определить признаки необходимые для решения задачи, в случае необходимости заняться генерацией новых признаков, устранением проблем в данных, визуализировать распределение и зависимость целевого признака от выбранных признаков. В отчете описать все проблемы, с которыми вы столкнулись и выбранные подходы к их решению.

1 Описание решения

Для выполнения задачи были использованы датасеты:

1. Задача классификкации: https://www.kaggle.com/deepu1109/star-dataset
2. Задача регрессии: https://www.kaggle.com/lsind18/weight-vs-age-of-chicks-on-different-diets

Первый датасет ставит задачу многоклассовой классификации. Этот датасет был очень удобным, отсутствие пробелов в данных было очень удробным, разве что пришлось немного подумать над многоклассовостью подборки и о реализации всех необходимых преобразований. В конце обработки датасет был немного преобразован для того, чтобы содержать только самые хорошие параметры, где "хорошесть"определялась по корреляции с классом звезды.

Второй датасет был результатом долгих поисков, одним из больших его преимуществ оказалось полное отсутствие строковых и логических полей. Это позволило избавиться от большей части препроцессинга. Над вторым датасетом я решил провести линейную регрессию для предсказания веса курицы по данным. После обнаружения этого набора данных не составило труда написать все необходимые преобразования в Юпитер ноутбуке, но поиск данных оказался самым сложным моментом в подготовке второй задачи.

2 Выводы

Лабораторная работа позволила погрузиться в море данных для машинного обучения. Я с интересом рассматривал различные датасеты, но поиск отличных сетов для моих задач оказался сложнее, чем я ожидал.