ЛАБОРАТОРНАЯ РАБОТА №2

Классификация табличных данных на основе нейросетевых моделей

Дан многомерный табличный размеченный набор данных. Необходимо выполнить классификационный анализ данных по указанному целевому признаку на основе полносвязной нейросетевой модели и нейросетевой модели, указанной в варианте, в соответствии со следующей последовательностью этапов.

1. Загрузить необходимые пакеты и библиотеки.
2. Загрузить данные из указанного источника.
3. Выполнить разведочный анализ данных в соответствии с этапами, описанными в файле *Этапы проекта машинного обучения в примерах.pdf:*
    a. Ознакомление с данными с помощью методов описательной статистики;
    b. Выполнить визуализацию данных одномерную для понимания распределения данных и многомерную для выяснения зависимостей между признаками;
    c. При необходимости выполнить очистку данных одним из методов.
    d. Проанализировать корреляционную зависимость между признаками;
    e. Поэкспериментировать с комбинациями атрибутов. При необходимости добавить новые атрибуты в набор данных.
    f. Выполнить отбор существенных признаков. Сформировать набор данных из существенных признаков.
    g. При необходимости преобразовать текстовые или категориальные признаки одним из методов.
    h. Выполнить преобразование данных для обоих наборов (исходного и сформированного) одним из методов по варианту.
4. Анализ выполняется для исходного набора данных, преобразованного исходного набора данных, построенного набора данных и преобразованного построенного набора данных. Во всех наборах данных выделить обучающую, проверочную (валидационную) и тестовую выборки данных.
5. Сравнить качество полносвязной нейросетевой классификационной модели и классификационной нейросетевой модели, указанной в варианте, на обучающей и валидационной выборках для всех наборов данных, включая их преобразованные варианты. Для оценки качества моделей использовать метрики: accuracy, balanced_accuracy (в случае несбалансированности классов – существенное различие численности экземпляров данных в классах), F_1 метрики (как по всей выборке, так и отдельно по классам).
6. Для лучшей модели на лучшем наборе данных оценить качество на тестовом наборе.
7. Для лучшей классификационной модели на лучшем наборе данных выполнить Grid поиск лучших гиперпараметров классификационной нейросетевой модели на обучающей и валидационной выборках. Определить значения лучших гиперпараметров.
8. Определить показатели качества полученной в результате Grid поиска классификационной нейросетевой модели на тестовом наборе. Сравнить показатели качества лучшей модели на лучшем наборе данных до поиска гиперпараметров и после поиска гиперпараметров.

9. Сделать выводы по проведенному анализу.

Варианты

1. Набор данных по раку молочной железы содержит характеристики, вычисленные на основе изображений биопсии рака молочной железы, с целью предсказать, является ли опухоль доброкачественной или злокачественной. Он включает 569 экземпляров с 30 характеристиками, такими как радиус, текстура, периметр и площадь ядер. Построить классификационную модель для целевого признака «target» - признак доброкачественной или злокачественной опухоли у пациента.
    a. Пункт 5 – одномерная сверточная сеть
    b. Пункт 3.h – Min-max масштабирование
2. Данные – результаты химического анализа вин, выращенных и произведенных в одном и том же регионе Италии тремя разными производителями. Для разных компонентов, обнаруженных в трех типах вина, проведено тринадцать различных измерений. Построить классификационную модель для целевого признака «target» - признак производства вина одним из производителей.
    a. Пункт 5 – простая рекуррентная сеть
    b. Пункт 3.h – Стандартизация
3. В наборе данных по сердечным заболеваниям содержатся различные атрибуты пациента, позволяющие определить наличие или отсутствие сердечных заболеваний. Он включает такие характеристики, как возраст, пол, тип боли в груди, кровяное давление в состоянии покоя и уровень холестерина, всего 303 экземпляра в наборе данных. Построить классификационную модель для целевого признака «target» - признак наличия или отсутствия сердечных заболеваний у пациента.
    a. Пункт 5 – одномерная сверточная сеть
    b. Пункт 3.h – Стандартизация
4. Построить классификационную модель для целевого признака «Cover_Type», определяющего тип лесного покрова (преобладающий вид древесного покрова) на основе картографических переменных. Фактический тип лесного покрова для заданной ячейки размером 30 x 30 метров был определен на основе данных Системы информации о ресурсах региона 2 Лесной службы. Затем независимые переменные были получены из данных, полученных от Геологической службы. Данные представлены в необработанном виде (не масштабированы) и содержат двоичные столбцы данных для качественных независимых переменных, таких как дикие зоны и тип почвы.
    a. Пункт 5 – LSTM рекуррентная сеть
    b. Пункт 3.h – Нормализация
5. Дан набор данных, созданный в высшем учебном заведении (полученный из нескольких разрозненных баз данных), связанный со студентами, обучающимися по разным степеням бакалавриата, таким как агрономия, дизайн, образование, сестринское дело, журналистика, менеджмент, социальная служба и технологии. Набор данных включает информацию, известную на момент зачисления студентов (академический путь, демографические и социально-экономические факторы), а также успеваемость студентов в конце первого и второго семестров. Построить классификационную

модель для целевого признака «Target» - признак, предсказывающий отсев студентов и академическую успеваемость.
    a. Пункт 5 – GRU рекуррентная сеть
    b. Пункт 3.h – Min-max масштабирование
6. В исследовании использовались семь различных типов сухих бобов с учетом таких характеристик, как форма, вид, тип и структура в зависимости от рыночной ситуации. Была разработана система компьютерного зрения для различения семи различных зарегистрированных сортов сухих бобов со схожими характеристиками с целью получения единообразной классификации семян. Для модели классификации были сделаны снимки 13 611 зерен 7 различных зарегистрированных сухих бобов с помощью камеры высокого разрешения. Изображения бобов, полученные с помощью системы компьютерного зрения, были подвергнуты этапам сегментации и извлечения признаков, и в общей сложности из зерен было получено 16 признаков: 12 измерений и 4 формы.
    a. Пункт 5 – двунаправленная LSTM рекуррентная сеть
    b. Пункт 3.h – Стандартизация
7. Дан набор данных включает данные для оценки уровней ожирения у людей из Мексики, Перу и Колумбии на основе их привычек питания и физического состояния. Данные содержат 17 атрибутов и 2111 записей, записи помечены классовой переменной NObeyesdad (Уровень ожирения), что позволяет классифицировать данные, используя значения Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. Построить классификационную модель для целевого признака «NObeyesdad» от остальных входных признаков, определяющего физическое состояние человека.
    a. Пункт 5 – двунаправленная GRU рекуррентная сеть
    b. Пункт 3.h – Min-max масштабирование
8. Дан набор данных признаков, характеризующих почтовые сообщения. Необходимо построить классификационную модель для целевого признака «Class» от остальных входных признаков, определяющего присутствие или отсутствие спама в почтовых сообщениях: 1 – спам, 0 – не спам.
    a. Пункт 5 – одномерная сверточная сеть
    b. Пункт 3.h – Стандартизация
9. Дан набор данных, которые были извлечены из изображений подлинных и поддельных банкнотоподобных образцов. Для оцифровки использовалась промышленная камера, обычно используемая для проверки отпечатков. Конечные изображения имеют размер 400x400 пикселей. Для извлечения объектов из изображений использовались инструменты вейвлет-преобразования. Необходимо построить классификационную модель для целевого признака «Class» от остальных входных признаков, определяющего подлинность банкнотоподобных образцов.
    a. Пункт 5 – простая рекуррентная сеть
    b. Пункт 3.h – Нормализация
10. Дан набор данных, который содержит разнородную информацию о статьях, опубликованных Mashable (www.mashable.com) за два года. Цель состоит в том, чтобы определить популярность опубликованных статей на основе числа репостов в социальных сетях. Построить классификационную модель для

целевого признака «Class» от остальных входных признаков, определяющего популярность (1) или непопулярность (0) опубликованных статей.
    a. Пункт 5 – LSTM рекуррентная сеть
    b. Пункт 3.h – Min-max масштабирование
11. Дан набор данных, который состоит из ряда биомедицинских измерений голоса 31 человека, из них 23 с болезнью Паркинсона (БП). Каждый столбец в таблице соответствует определенному показателю голоса, и каждая строка соответствует одной из 195 записей голоса этих людей (признак "name"). Основная цель сбора данных - отличить здоровых людей от людей с БП в соответствии с признаком "status", для которого установлено значение 0 для "здоров" и 1 для БП. Построить классификационную модель для целевого признака «status» от остальных входных признаков, определяющего состояние здоровья.
    a. Пункт 5 – GRU рекуррентная сеть
    b. Пункт 3.h – Стандартизация
12. Дан набор данных, который состоит векторов характеристик различных музыкальных произведений, подразделяющиеся на четыре разных класса музыкальных эмоций: счастливые, грустные, злые и расслабляющие. Построить классификационную модель для целевого признака «Class» от остальных входных признаков, определяющего класс музыкальной эмоции музыкального произведения.
    a. Пункт 5 – двунаправленная LSTM рекуррентная сеть
    b. Пункт 3.h – Нормализация
13. Дан набор данных для оценки точного количества людей в помещении с использованием нескольких неинтрузивных датчиков окружающей среды, таких как датчики температуры, освещенности, звука, CO2 и PIR. Экспериментальный испытательный стенд для оценки занятости был развернут в комнате размером 6 м x 4,6 м. Вектора набора подразделяются на 4 класса в соответствии со степенью занятости комнаты. Построить классификационную модель для целевого признака «Room_Occupancy_Count» от остальных входных признаков, определяющего класс – степенью занятости комнаты.
    a. Пункт 5 – двунаправленная GRU рекуррентная сеть
    b. Пункт 3.h – Нормализация
14. Дан набор данных, собранных в 2022 году в Университете короля Сауда в Эр-Рияде для распознавания действий человека с использованием датчиков IMU мобильных телефонов (акселерометра и гироскопа). Эти действия классифицируются как остановка или ходьба. Построить классификационную модель для целевого признака «Activity» от остальных входных признаков, определяющего класс – остановку или движение владельца телефона.
    a. Пункт 5 – одномерная сверточная сеть
    b. Пункт 3.h – Стандартизация