ЛАБОРАТОРНАЯ РАБОТА №1

Разведочный и регрессионный анализ данных

на основе нейросетевых моделей

Дан многомерный размеченный набор данных. Необходимо выполнить регрессионный анализ данных 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 регрессионной модели и регрессионной нейросетевой модели, указанной в варианте, на обучающей и валидационной выборках для всех наборов данных, включая их преобразованные варианты. Для оценки качества моделей использовать метрики: корень из среднеквадратичной ошибки, коэффициент детерминации R^2.
6. Для лучшей модели на лучшем наборе данных оценить качество на тестовом наборе.
7. Для лучшей модели на лучшем наборе данных выполнить Grid поиск лучших гиперпараметров регрессионной нейросетевой модели на обучающей и валидационной выборках. Определить значения лучших гиперпараметров.
8. Определить показатели качества полученной в результате Grid поиска регрессионной 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 электронной коммерции содержит данные о клиентах, которые покупают одежду в Интернете. Магазин предлагает консультации по стилю и одежде в магазине. Покупатели приходят в магазин, проводят сеансы/встречи с личным стилистом, затем они могут пойти домой и заказать через мобильное приложение или веб-сайт ту одежду, которую они хотят. Построить регрессионную модель для целевого признака «Yearly Amount Spent» (годовая сумма расходов покупателя) от остальных входных признаков.
    a. Пункт 5 – одномерная сверточная сеть
    b. Пункт 3.h – Min-max масштабирование
2. Набор данных цен на недвижимость. Построить регрессионную модель для целевого признака «Y house price of unit area» (цена объекта недвижимости) от остальных входных признаков.
    a. Пункт 5 – простая рекуррентная сеть
    b. Пункт 3.h – Стандартизация
3. Набор данных прибыли стартапов в зависимости от трех типов расходов. Построить регрессионную модель для целевого признака «Profit» (прибыль стартапа) от остальных входных признаков.
    a. Пункт 5 – LSTM рекуррентная сеть
    b. Пункт 3.h – Нормализация
4. Для набора данных потенциальных покупателей разработать модель для прогнозирования общей суммы, которую клиенты готовы заплатить за новый автомобиль. Построить регрессионную модель для целевого признака «Car Purchase Amount» (сумма покупки автомобиля) от остальных входных признаков.
    a. Пункт 5 – GRU рекуррентная сеть
    b. Пункт 3.h – Min-max масштабирование
5. Набор данных телемониторинга болезни Паркинсона. Построить регрессионную модель для целевого признака «total_UPDRS» от остальных входных признаков. Признак «motor_UPDRS» исключить из набора.
    a. Пункт 5 – двунаправленная LSTM рекуррентная сеть
    b. Пункт 3.h – Стандартизация
6. Набор данных телемониторинга болезни Паркинсона. Построить регрессионную модель для целевого признака «motor_UPDRS» от остальных входных признаков. Признак «total_UPDRS» исключить из набора.
    a. Пункт 5 – двунаправленная GRU рекуррентная сеть
    b. Пункт 3.h – Min-max масштабирование
7. Набор данных прочности бетона на сжатие. Прочность бетона на сжатие является сильно нелинейной функцией возраста и ингредиентов: цемент, доменный шлак, летучую золу, воду, суперпластификатор, крупный заполнитель и мелкий заполнитель. Построить регрессионную модель для целевого признака «concrete_compressive_strength» (компрессионная прочность бетона) от остальных входных признаков.
    a. Пункт 5 – одномерная сверточная сеть
    b. Пункт 3.h – Стандартизация

8. Набор данных схемы пирамиды – определение прибыли или убытка. Схемы пирамид, запущенные в разных странах, часто соблазняют простых людей делать деньги в краткосрочной перспективе. Построить регрессионную модель прогностической оценки схемы пирамиды для целевого признака «profit» (выгода от схемы) от остальных входных признаков.
    a. Пункт 5 – простая рекуррентная сеть
    b. Пункт 3.h – Нормализация
9. Набор обезличенных данных. Построить регрессионную модель прогностической оценки схемы пирамиды для целевого признака «EQV1» от остальных входных признаков.
    a. Пункт 5 – LSTM рекуррентная сеть
    b. Пункт 3.h – Min-max масштабирование
10. Набор данных расхода топлива в городском цикле автомобилями в милях на галлон. Построить регрессионную модель прогностической оценки целевого признака mpg расхода топлива от трех многозначных дискретных и пяти непрерывных входных атрибутов (в файле датасета отсутствуют заголовки столбцов, использовать соответствие по нумерации из описания признаков ниже):
    1) миль на галлон: числовой
    2) цилиндры: категориальный дискретный
    3) пробег: числовой
    4) мощность: числовой
    5) вес: числовой
    6) ускорение: числовой
    7) модельный год: категориальный дискретный
    8) происхождение: категориальный дискретный
    9) название автомобиля: строка (уникальная для каждого экземпляра)
    a. Пункт 5 – GRU рекуррентная сеть
    b. Пункт 3.h – Стандартизация
11. Набор данных для прогноза площади лесных пожаров на основе метеорологических и других признаков аргументов. Построить регрессионную модель прогностической оценки целевого признака area (целое число, в гектарах) площади лесных пожаров:
    1) X целочисленная пространственная координата по оси X на карте
    2) Y целочисленная пространственная координата оси Y на карте
    3) Месяц категориальный месяц года от «января» до «декабря»
    4) День категориальный день недели: от «понедельника» до «воскресенья»
    5) FFMC непрерывный индекс из системы FWI
    6) DMC целочисленный индекс из системы FWI
    7) DC непрерывный индекс из системы FWI
    8) ISI непрерывный индекс из системы FWI
    9) temp непрерывный показатель температуры в градусах Цельсия
    10) RH целочисленный показатель относительной влажности в %
    11) wind непрерывный показатель скорости ветра в км/ч
    12) rain непрерывный показатель уровня осадков в мм/м2
    13) area целевой показатель площади лесных пожаров в гектарах
    a. Пункт 5 – двунаправленная LSTM рекуррентная сеть
    b. Пункт 3.h – Min-max масштабирование

12. Набор данных различных форм зданий, различающихся по площади остекления, распределению площади остекления, ориентации и другим параметрам. в зависимости от параметров здания других признаков аргументов. Построить регрессионную модель оценки целевого признака требований к тепловой нагрузке зданий:
    1) X1 вещественный, относительная компактность
    2) X2 вещественный, площадь поверхности
    3) X3 вещественный, площадь сплошной стены
    4) X4 вещественный, зона крыши
    5) X5 вещественный, общая высота
    6) X6 целочисленный, ориентация
    7) X7 вещественный, площадь сплошного остекления
    8) X8 целочисленный, распределение площади остекления
    9) Y1 целевой признак, вещественный, отопительная нагрузка
        a. Пункт 5 – двунаправленная GRU рекуррентная сеть
        b. Пункт 3.h – Нормализация
13. Набор данных различных форм зданий, различающихся по площади остекления, распределению площади остекления, ориентации и другим параметрам. в зависимости от параметров здания других признаков аргументов. Построить регрессионную модель оценки целевого признака требований к охлаждающей нагрузке зданий:
    1) X1 вещественный, относительная компактность
    2) X2 вещественный, площадь поверхности
    3) X3 вещественный, площадь сплошной стены
    4) X4 вещественный, зона крыши
    5) X5 вещественный, общая высота
    6) X6 целочисленный, ориентация
    7) X7 вещественный, площадь сплошного остекления
    8) X8 целочисленный, распределение площади остекления
    9) Y2 целевой признак, вещественный, нагрузка непрерывного охлаждения
        a. Пункт 5 – одномерная сверточная сеть
        b. Пункт 3.h – Стандартизация
14. Набор данных энергопотребления распределительной сети города. Построить регрессионную модель оценки целевого признака уровня энергопотребления распределительной сети города:
    1) DateTime дата, время, периодичность 10 мин.
    2) Temperature вещественный, температура
    3) Humidity вещественный, влажность
    4) Wind Speed вещественный, скорость ветра
    5) General diffuse flows вещественный, общие диффузные потоки
    6) Diffuse flows вещественный, диффузные потоки X7 вещественный, площадь сплошного остекления
    7) Power consumption (Zone1) целевой признак, вещественный, уровень энергопотребления
        a. Пункт 5 – GRU рекуррентная сеть
        b. Пункт 3.h – Min-max масштабирование