Jooble

Тестовое задание (Data Scientist)

Описание задачи

Jooble - сервис для поиска работы. Для решения многих задач для нас важно решить проблему репрезентации вакансии в виде набора разнотипных (числовых, строковых и т.д) факторов(признаков).

В данном тестовом задании Вам требуется реализовать пакет по предобработке признаков вакансий.

Входные данные

https://drive.google.com/drive/folders/1lCVt_UtxwfLyu16rgmQm-ej68XztFEtN

В директории python-dev-test/data находится два файла: test.tsv, train.tsv одного формата.Каждая строка файла является характеристикой для одной вакансии.

В файле две колонки:

- id_job - (integer) идентификатор вакансии
- features - (string) конкатенация из признаков вакансии определенного типа (Объединенные через символ “,”).
    - Первый элемент списка - код набора признаков. (В данном примере “2”)
    - Остальные элементы - числовые характеристики данного типа. В рамках этой задачи это 256 integer чисел (итого каждый признак можно пронумеровать и название колонок может иметь вид : “feature_2_{i}”, где i - индекс элемента в массиве)(*).

Требования к выходным данным

В качестве тестирования к модулю должен быть прикреплен скрипт, который генерирует файл test_proc.tsv который содержит следующий набор колонок (признаков) для каждой вакансии из test.tsv:

1. id_job - (integer) идентификатор вакансии (размерность : 1);
2. feature_2_stand_{i} - (double) результат стандартизации (z-score нормализация) входного признака feature_2_{i} (См. Входные данные(*)) (размерность : 256);
    a. Ссылка на определение стандартизации :
        i. https://ru.wikipedia.org/wiki/Z-%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B0
        ii. https://en.wikipedia.org/wiki/Feature_scaling
    b. Для проведения этой операции требуется оценить две статистики для каждой колонки feature_2_{i} на данных из файла train.tsv:
        i. mean(feature_2_{i}) - среднее значение по всем вакансиям для признака feature_2_{i};
        ii. std(feature_2_{i}) - среднеквадратическое отклонение по всем вакансиям для признака feature_2_{i};
3. max_feature_2_index - (integer) индекс i максимального значения признака feature_2_{i} для данной вакансии (размерность : 1);
4. max_feature_2_abs_mean_diff - (double) - абсолютное отклонение признака с индексом max_feature_2_index от его среднего значения mean(feature_2_{max_feature_2_index}) (размерность : 1);

Требования к реализации

1. Реализацию требуется произвести на Python 3.7
2. При проектировании модуля требуется учитывать следующие факторы:
    a. Размеры данных файла train.tsv, test.tsv может достигать нескольких десятков миллионов строк
    b. При использовании модуля в будущем планируется добавление факторов новых типов (в нашем тесте только признаки типа “2”).
    c. При использовании модуля в будущем планируется добавление новых способов нормализации признаков (в нашем тесте только z-score нормализация).
3. Пакет требуется загрузить на github репозиторий.