Assignment 3

Introduction

In this assignment you will be using the Movie dataset provided and the machine learning algorithm you have learned in this course in order to find out, knowing only things you could know before a film was released , what the rating and revenue of the film would be. the rational here is that your client is a movie theater that would like to decide for how long should they reserve the movie theater to show a movie when it is released.

Datasets

In this assignment you will be given two datasets training.csv (https://github.com/mysilver/COMP9321-Data-Services/raw/master/20t1/assign3/training.csv) and validation.csv (https://github.com/mysilver/COMP9321-Data-Services/raw/master/20t1/assign3/validation.csv) .
(DATASETS HAVE BEEN UPDATED: Movies without known budget are removed ; more data for validation)

You can use the **training** dataset (but not validation) for training machine learning models, and you can use validation dataset to evaluate your solutions and avoid over-fitting.

Please Note:

- This assignment specification is deliberately left open to encourage students to submit innovative solutions.
- You can only use Scikit-learn to train your machine learning algorithm
- Your model will be evaluated against a third dataset (available for tutors, but not for students)
- You must submit your code and a report
- The due date is 20/04/2020 18:00 ~~(late penalty with 25% per day)~~ **UPDATE:** We will waive the late penalty but any submission submitted after the 24/04/2020 18:00 **WILL NOT** be marked.

Part-I: Regression (10 Marks)

In the first part of the assignment, you are asked to predict the "revenue" of movies based on the the information in the provided dataset. More specifically, you need to predict the revenue of a movie based on a subset (or all) of the following attributes (**make sure you DO NOT use ***rating***):

cast,crew,budget,genres,homepage,keywords,original_language,original_title,overview,production_companies,production_countries,release_date,runtime,spoken_languages,status,tagline

---

Part-II: Classification (10 Marks)

Using the same datasets, you must predict the rating of a movie based on a subset (or all) of the following attributes (**make sure you DO NOT use ***revenue***):

cast,crew,budget,genres,homepage,keywords,original_language,original_title,overview,production_companies, production_countries,release_date,runtime,spoken_languages,status,tagline

---

Submission

You must submit two files:

- A python script z{id}.py
- A report named z{id}.pdf

Python Script and Expected Output files

You code must be executed in CSE machines using the following command with three arguments:

```
$ python3 z{id}.py path1 path2
```

- **path1** : indicates the path for the dataset which should be used for training the model (e.g., ~/training.csv)
- **path2** : indicates the path for the dataset which should be used for reporting the performance of the trained model (e.g., ~/validation.csv); we may use different datasets for evaluation

For example, the following command will train your models for the first part of the assignment and use the validation dataset to report the performance:

```
$ python3 YOUR_ZID.py training.csv validation.csv
```

Your program should create 4 files on the same directory as the script:

- z{id}.PART1.summary.csv
- z{id}.PART1.output.csv
- z{id}.PART2.summary.csv
- z{id}.PART2.output.csv

For the the first part of the assignment:

" z{id}.PART1.summary.csv " contains the evaluation metrics (MSR,correlation) for the model trained for the first part of the assignment. Use the given validation dataset to compute the metrics. The file should be formatted exactly as follow:

```
zid,MSR,correlation
YOUR_ZID,6.13,0.73
```

- **MSR** : the mean_squared_error in the regression problem
- **correlation** : The **Pearson correlation coefficient** in the regression problem

" z{id}.PART1.output.csv " stores the predicted revenues for all of the movies in the evaluation dataset (not training dataset) , and the file should be formatted exactly as follow:

```
movie_id,predicted_revenue
1,7655555
2,75875765
...
```

For the the second part of the assignment:

" z{id}.PART2.summary.csv " contains the evaluation metrics (average_precision, average_recall,accuracy - the unweighted mean) for the model trained for the second part of the assignment. Use the given validation dataset to compute the metrics. The file should be formatted exactly as follow:

```
zid,average_precision,average_recall,accuracy
YOUR_ZID,6.11,0.71,0.89
```

- **average_precision** : the average precision for all classes in the classification problem
- **average_recall** : the average recall for all classes in the classification problem

" z{id}.PART2.output.csv " stores the predicted ratings for all of the movies in the evaluation dataset (not training dataset) and it should be formatted exactly as follow:

```
movie_id,predicted_rating
1,1
2,4
...
```

Marking Criteria

For **EACH** of the parts, you will be marked based on:

- **(3 marks)** You code must run and perform the designated tasks on CSE machines without problems and create the expected files.
- **(3 marks)** How well your model (trained on the training dataset) perform in the test dataset
- **(2 marks)** You must correctly calculate the evaluation metrics (e.g., average_precision - 2 decimal places) in the output files (e.g., z{id}.PART2.summary.csv)
- **(2 marks)** One page report containing:
    - Performance of your model on the validation dataset and how you evaluated the performance and improved it (e.g., relying on feature selection, switching from one machine leanring model to a more suitable one,...etc.)
    - Problems you have faced in predicting (e.g., JOSN formated columns, keywords, missing data) and how you tried to solve the problem.
- The minimum coefficient value in the regression model is 0.3 in the test dataset (not validation). As listed above, you will be marked on different aspects (e..g, report); and your submission will be compared to the rest of students to adjust marks and be fair to all. Do you best in improving your models and make sure you do not overfit because you will be marked based on a third dataset, called "test dataset". In the classification problem, your accuracy should be more than a baseline. The baseline model labels all movies with the most frequent class (e.g., assuming all movie rates are 3).
- You will be penalized if your models take more than 3 minutes to train and generate outputs
- Your assignment will not be marked (zero mark) if any of the following occur:
    - If it generates hard-coded predictions