

[Resources](#) / [Assignment 3](#)

Assignment 3

[Specification](#)[Make Submission](#)[Check Submission](#)[Collect Submission](#)

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 rationale here is that your client is a movie theater that would like to decide how long should they reserve the movie theater for 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>).

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 **21/04/2021 18:00**
-

Part-I: Regression (10 Marks)

In the first part of the assignment, you are asked to predict the "revenue" of movies based on 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

Your 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 first part of the assignment:

"`z{id}.PART1.summary.csv`" contains the evaluation metrics (MSR, correlation) for the model trained in 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 (a floating number between -1 and 1)

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

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

For 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 in the second part of the assignment. Use the given validation dataset to compute the metrics. The file should be formatted exactly as:

```
zid,average_precision,average_recall,accuracy
YOUR_ZID,0.69.71,0.89
```

- **average_precision** : the average precision for all classes in the classification problem (a number between 0 and 1)
- **average_recall** : the average recall for all classes in the classification problem (a number between 0 and 1)

" z{id}.PART2.output.csv " stores the predicted ratings for all of the movies in the evaluation dataset (not the 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)** Your 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) performs 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 learning model to a more suitable one,...etc.)
 - Problems you have faced in predicting (e.g., JSON formatted columns, keywords, missing data) and how you tried to solve the problems.
- 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 the students to adjust marks and be fair to all. Do your 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 output.
- Your assignment will not be marked (zero marks) if any of the following occur:
 - If it generates hard-coded predictions
 - If it also uses the second dataset (test/validation) to train the model
 - If it does not run on CSE machines with the given command (e.g., python3 zid.py training_dataset.csv test_dataset.csv)
Do NOT hard-code the dataset names

FAQ

- **Can we define our own feature set?**

Yes, you can define any features; make sure your features do not rely on the validation (or test) datasets

- **What is the difference between validation and test datasets?**

The validation dataset is provided for you to tune your models; the test dataset will not be provided to students, instead, it will be used to evaluate your model.

- **For the average precision/recall functions, should we use the unweighted ('macro') mean or the weighted mean?**

use the unweighted ('macro') mean

- **Should we calculate metrics to 1 Decimal Place?**

2 Decimal Places

- **Can we use any machine learning algorithm?**

Yes, as long as it is provided in sklearn.

- **What python modules can we use for developing our solutions?**

You can use any modules presented in the lab activities; if it is a one that not in the labs, you may get permission by asking ...

- **How should we calculate the Pearson correlation coefficient?**

It is calculated between your predictions and the real values for the validation (or test) dataset.

Plagiarism

This is an *individual assignment*. The work you submit must be your own work. Submission of work partially or completely derived from any other person or jointly written with any other person is not permitted. The penalties for such offense may include negative marks, automatic failure of the course, and possibly other academic disciplines. Assignment submissions will be checked using plagiarism detection tools for both code and the report and then the submission will be examined manually.

Do not provide or show your assignment work to any other person - apart from the teaching staff of this course. If you knowingly provide or show your assignment work to another person for any reason, and work derived from it is submitted, you may be penalized, even if the work was submitted without your knowledge or consent. Pay attention to that is **also your duty to protect your code artifacts**. If you are using an online solution to store your code artifacts (e.g., GitHub) then make sure to keep the repository private and do not share access to anyone.

Reminder: Plagiarism is defined as (<https://student.unsw.edu.au/plagiarism>) using the words or ideas of others and presenting them as your own. UNSW and CSE treat plagiarism as academic misconduct, which means that it carries penalties as severe as being excluded from further study at UNSW. There are several online sources to help you understand what plagiarism is and how it is dealt with at UNSW:



- Plagiarism and Academic Integrity (<https://student.unsw.edu.au/plagiarism>)
- UNSW Plagiarism Procedure (<https://www.gs.unsw.edu.au/policy/documents/plagiarismprocedure.pdf>)

Make sure that you read and understand this. Ignorance is not accepted as an excuse for plagiarism. In particular, you are also responsible for ensuring that your assignment files are not accessible by anyone but you by setting the correct permissions in your CSE directory and code repository, if using one (e.g., Github and similar). Note also that plagiarism includes paying or asking another person to do a piece of work for you and then submitting it as your own work.

UNSW has an ongoing commitment to fostering a culture of learning informed by academic integrity. All UNSW staff and students have a responsibility to adhere to this principle of academic integrity. Plagiarism undermines academic integrity and is not tolerated at UNSW.

Resource created 11 days ago (Sunday 28 March 2021, 07:53:59 AM), last modified about 9 hours ago (Thursday 08 April 2021, 07:37:24 AM).

Comments

 [Q \(/COMP9321/21T1/forums/search?forum_choice=resource/59350\)](/COMP9321/21T1/forums/search?forum_choice=resource/59350)  [\(/COMP9321/21T1/forums/resource/59350\)](/COMP9321/21T1/forums/resource/59350)

 Add a comment



Hanxin Wu (/users/z5196075) about 2 hours ago (Thu Apr 08 2021 14:14:02 GMT+1000 (澳大利亚东部标准时间))

There are (2 marks) in ' You must correctly calculate the evaluation metrics', so this means we can not use sklearn.metrics.mean_squared_error or others ?

Reply



Xiaoyi Meng (/users/z5202244) about 3 hours ago (Thu Apr 08 2021 13:46:51 GMT+1000 (澳大利亚东部标准时间))

Must all the data in the 4 files be consistent or similar to the assignment specification?

Reply



Danny Ly (/users/z5020225) about 4 hours ago (Thu Apr 08 2021 12:58:44 GMT+1000 (澳大利亚东部标准时间))

Hi - can we expect the final testing set to be a similar size to the validation set?

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) about 3 hours ago (Thu Apr 08 2021 13:35:15 GMT+1000 (澳大利亚东部标准时间)), last modified about 3 hours ago (Thu Apr 08 2021 13:35:24 GMT+1000 (澳大利亚东部标准时间))

it would be slightly bigger.

Reply



Hanxin Wu (/users/z5196075) about 22 hours ago (Wed Apr 07 2021 18:49:20 GMT+1000 (澳大利亚东部标准时间))

hi,

for part 1, we are asked to predict the "revenue", but why can not use **rating** ?

Thanks

Reply



May Altulyan (/users/z5131400) [about 20 hours ago \(Wed Apr 07 2021 20:31:33 GMT+1000 \(澳大利亚东部标准时间\)\)](#)

From the 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** "

Reply



Yuchen Yang (/users/z5189310) [a day ago \(Wed Apr 07 2021 10:48:52 GMT+1000 \(澳大利亚东部标准时间\)\)](#)

Hello, I would like to ask is an official version? Will there be other versions?

Reply



May Altulyan (/users/z5131400) [a day ago \(Wed Apr 07 2021 11:20:07 GMT+1000 \(澳大利亚东部标准时间\)\)](#)

sorry didnt get your question

Reply



Yukun Yin (/users/z5199930) [a day ago \(Wed Apr 07 2021 00:04:40 GMT+1000 \(澳大利亚东部标准时间\)\)](#)

Hi

seaborn is ok?

Reply



May Altulyan (/users/z5131400) [a day ago \(Wed Apr 07 2021 11:10:07 GMT+1000 \(澳大利亚东部标准时间\)\)](#)

you can use all packages that we used during the labs

Reply



Runqi Liu (/users/z5241723) [2 days ago \(Tue Apr 06 2021 20:43:28 GMT+1000 \(澳大利亚东部标准时间\)\)](#),
last modified [2 days ago \(Tue Apr 06 2021 21:31:33 GMT+1000 \(澳大利亚东部标准时间\)\)](#)

1. Could we import sys module to pass in parameter?

2. Could we import numpy?

3. In addition, in the below picture

" `z{id}.PART1.output.csv` " stores the predicted revenues for all of the movies in the evaluation dataset (not the training dataset),

the requirement asks us to output predicted revenues for all of movies in the evaluation dataset, but where could we find the evaluation dataset?

Thanks

Reply



May Altulyan (/users/z5131400) a day ago (Wed Apr 07 2021 11:12:14 GMT+1000 (澳大利亚东部标准时间)), last modified a day ago (Wed Apr 07 2021 11:25:35 GMT+1000 (澳大利亚东部标准时间))

1. you can use all the packages that we used during the labs
2. yes
3. means the validation dataset

Reply



Di Wu (/users/z5247036) 2 days ago (Tue Apr 06 2021 19:07:13 GMT+1000 (澳大利亚东部标准时间))

Can you give download files for these two csv files?

Reply



Vishal Bondwal (/users/z5278101) 2 days ago (Tue Apr 06 2021 20:20:16 GMT+1000 (澳大利亚东部标准时间))

I could download them by right-clicking and selecting 'Save Link As...'. You can also click to open them, and then right-click the page, and choose 'Save Page As...'

Reply



Daniel Fan (/users/z5114117) 2 days ago (Tue Apr 06 2021 17:01:36 GMT+1000 (澳大利亚东部标准时间))

Hi,

What is the meaning of this sentence in the Marking Criteria section: "The minimum coefficient value in the regression model is 0.3 in the test dataset (not validation)"?

- Doesn't the minimum coefficient value depend on what features we choose? E.g. if we scale a feature, won't it potentially be much higher?
- Is this a restriction imposed for our model? But how can we even determine this if we can't train on the test set?

Thanks in advance

Reply



May Altulyan (/users/z5131400) a day ago (Wed Apr 07 2021 11:19:43 GMT+1000 (澳大利亚东部标准时间))

The coefficient value will be a metric to evaluate the performance of your model!

you have to ensure that the coefficient value is less than 0.3 for both train and validation parts

because your model will be tested on the test dataset

Reply



Daniel Fan (/users/z5114117) a day ago (Wed Apr 07 2021 12:05:17 GMT+1000 (澳大利亚东部标准时间))

Why is the coefficient value used as a metric? Is there a link between this and the performance of a model?

Reply



May Altulyan (/users/z5131400) a day ago (Wed Apr 07 2021 13:01:27 GMT+1000 (澳大利亚东部标准时间))

you will learn about it in Linear Regression lecture

Reply