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
    - 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

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 offence may include negative marks, automatic failure of the course and possibly other academic discipline. Assignment submissions will be checked using plagirisim derection 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 that is **also your duty to protect your code artifacts** . if you are using any 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 on-line 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 these. 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 months ago (Friday 03 April 2020, 12:06:06 PM), last modified 11 months ago (Tuesday 21 April 2020, 07:15:02 AM).

Comments

(/COMP9321/20T1/forums/search?forum_choice=resource/44201)
(/COMP9321/20T1/forums/resource/44201)

Add a comment

Ruoran Chen (/users/z5194816) 10 months ago (Mon May 11 2020 20:19:24 GMT+1000 (Australian Eastern Standard Time))

Hi, how could we get the feedback of this assignment.

Reply

Meishi Chen (/users/z5269035) 10 months ago (Sat May 02 2020 12:31:46 GMT+1000 (Australian Eastern Standard Time))

Hi, do we still able to have assignment3 mark out before the final?

Reply

Saksham Yadav (/users/z5164624) 11 months ago (Fri Apr 24 2020 17:21:26 GMT+1000 (Australian Eastern Standard Time))

Are average recall and average accuracy not the same by definition? I'm getting the exact same value for both, and it makes sense too.

Reply

Armaan Ravi Eshwar (/users/z5180316) 11 months ago (Fri Apr 24 2020 14:46:11 GMT+1000 (Australian Eastern Standard Time))

Hi just wanted to clarify something regarding feature engineering. Are we allowed to create columns based on revenue (such as avg revenue per director) but only based of data from the training data? You mentioned that you cannot use the revenue column in part 1, but is that referring to both datasets or just the validation set.

Reply

Mohammadali Yaghoubzadehfard (/users/z5138589) 11 months ago (Fri Apr 24 2020 14:50:47 GMT+1000 (Australian Eastern Standard Time))

You can only use the specified columns in the spec for Training your models

Reply

Armaan Ravi Eshwar (/users/z5180316) 11 months ago (Fri Apr 24 2020 14:56:56 GMT+1000 (Australian Eastern Standard Time))

Is there a reason why we can't do this though? I understand that we cant do this for the validation set for obvious reasons, but I figured this would be okay if we only referenced the training data revenues. This is part of my prediction process. Will I be penalized for doing this? Again, I would like to emphasize that I am not touching the revenue data from the test/validation dataset, only the training data.

Reply

Mohammadali Yaghoubzadehfard (/users/z5138589) 11 months ago (Fri Apr 24 2020 15:14:26 GMT+1000 (Australian Eastern Standard Time))

Since it is forbideen in the spec, if you do so your solution will not be acceptable and not be marked.

Reply

Eu Shaun Lim (/users/z5156345) 11 months ago (Fri Apr 24 2020 15:30:08 GMT+1000 (Australian Eastern Standard Time))

Other than it being not allowed in the spec, is there an actual explanation for why this would not be allowed in the real world projects? Genuinely curious for my own purposes. I believe that the idea is to have a feature such as average revenue for each directors that is created from the training dataset, and then we use the information created from the training set and apply it to the validation set.

In theory, for every new data given, we get the director name, match the name to the information created from the training set and get the revenue generated for that director in the past. We then have the prior knowledge for this particular director and can predict how any new movie with that director will fare.

I don't believe this is cheating since we're using the prior knowledge of the director and not the revenue of the new unseen movie.

Reply

Mohammadali Yaghoubzadehfard (/users/z5138589) 11 months ago (Fri Apr 24 2020 15:36:18 GMT+1000 (Australian Eastern Standard Time))

What you are explaining make sense, and in practice it may improve the performance of your models.
I did not say, this is cheating, I said: this is against the spec. And to be fair, all students need to stick to the spec of the assignment.

Reply

Eu Shaun Lim (/users/z5156345) 11 months ago (Fri Apr 24 2020 15:45:12 GMT+1000 (Australian Eastern Standard Time))

Good to know, thanks!

Reply

Armaan Ravi Eshwar (/users/z5180316) 11 months ago (Fri Apr 24 2020 15:19:01 GMT+1000 (Australian Eastern Standard Time))

The spec mentions that you cannot train your model with validation/test data.
If I have not done this will I be okay?

Could you please explain to me why what I have done is wrong?

Reply

Lijun Zhong (/users/z5243425) 11 months ago (Fri Apr 24 2020 13:30:05 GMT+1000 (Australian Eastern Standard Time)), last modified 11 months ago (Fri Apr 24 2020 13:30:18 GMT+1000 (Australian Eastern Standard Time))

Hi, should we just up load 2 files (.py and .pdf) directly?

or do we have to compress them into a .tar file then upload the tar file ?

Reply

Mohammadali Yaghoubzadehfard (/users/z5138589) 11 months ago (Fri Apr 24 2020 13:34:53 GMT+1000 (Australian Eastern Standard Time))

Upload two files

Reply

Sunreet Singh (/users/z5130780) 11 months ago (Fri Apr 24 2020 12:20:21 GMT+1000 (Australian Eastern Standard Time)), last modified 11 months ago (Fri Apr 24 2020 12:22:34 GMT+1000 (Australian Eastern Standard Time))

Does output for 'rating' also need to be up to 2 decimal places? Or just the rating itself (integer)?

Reply

Mohammadali Yaghoubzadehfard (/users/z5138589) 11 months ago (Fri Apr 24 2020 12:27:04 GMT+1000 (Australian Eastern Standard Time))

No it should be the class label which is an integer

Reply

Sunreet Singh (/users/z5130780) 11 months ago (Thu Apr 23 2020 16:48:50 GMT+1000 (Australian Eastern Standard Time))

I am using LinearDiscriminantAnalysis() as Classifier. Should that be fine? Its listed as classifier in tut 8.

I am also using average = 'macro' for recall_score and precision_score. Is that the correct way to get the scores?

Reply

Mohammadali Yaghoubzadehfard (/users/z5138589) 11 months ago (Fri Apr 24 2020 15:37:09 GMT+1000 (Australian Eastern Standard Time))

You can use any algorithim

Reply

Roshni Sesharaju Katta (/users/z5262723) 11 months ago (Thu Apr 23 2020 13:19:12 GMT+1000 (Australian Eastern Standard Time))

Hi,

Is there a particular format to follow in writing the report?

Thank you!

Reply

Mohammadali Yaghoubzadehfard (/users/z5138589) 11 months ago (Fri Apr 24 2020 15:37:17 GMT+1000 (Australian Eastern Standard Time))

No, there is not

Reply

Joseph Aczel (/users/z5194935) 11 months ago (Thu Apr 23 2020 11:33:37 GMT+1000 (Australian Eastern Standard Time))

Hi,

I have preprocessed most of my data (i.e. split out the [genres, spoken_languages, original_language, production_companies] into binary columns (encoded)). I end up with a huge Dataframe at the end of this.

	'Action'	'Fantasy'	...	'English'	...	'Walt Disney'	'Warner Bros'

987	1	1	1	1	1
132	1	1	1	0	0
664	0	1	0	0	1
357	0	0	1	0	1

Using all of this data leads to a really poor correlation. Including 'budget' makes it even worse.

I currently have a Pearson Correlation of 0.25 (i.e. *pearsonr(y_valid, y_pred)*) but to even get it to 0.25 I had to overfit my model by using specific fields that increase the correlation:

```
X_train = ['Fantasy', 'Science Fiction', 'Adventure', 'Comedy', 'Drama']
```

I've been working on this for several days and still can't get above 0.25 on the validation set (let alone the test set that will be used). Am I doing something wrong???

Reply

Sadman Shafiq (/users/z5304769) 11 months ago (Thu Apr 23 2020 12:07:41 GMT+1000 (Australian Eastern Standard Time))

Hi Joseph I think you need to make sure that the columns you are using in your X_train and X_valid are the same, and in the same order, if you can ensure that they are it will give you much better results.

Reply

Joseph Aczel (/users/z5194935) 11 months ago (Thu Apr 23 2020 13:04:56 GMT+1000 (Australian Eastern Standard Time))

Hi Sadman,

Thanks for the suggestion. Unfortunately they are in the same order and I've verified that they both contain the same columns. My code does the following to make sure they are in the same order.

```
COLUMNS = ['Fantasy', 'Science Fiction', 'Adventure', 'Comedy', 'Dr
X_train = training_df[COLUMNS]
X_valid = validation_df[COLUMNS]
y_train = training_df['revenue']
y_valid = training_df['revenue']
```

If you have any other suggestions let me know. I'm pretty stumped.

Reply

Sunreet Singh (/users/z5130780) 11 months ago (Thu Apr 23 2020 16:46:45 GMT+1000 (Australian Eastern Standard Time)), last modified 11 months ago (Thu Apr 23 2020 16:47:00 GMT+1000 (Australian Eastern Standard Time))

Try to use release_date column in some way. That might help.

Reply

Sunreet Singh (/users/z5130780) 11 months ago (Thu Apr 23 2020 10:52:51 GMT+1000 (Australian Eastern Standard Time)), last modified 11 months ago (Thu Apr 23 2020 16:52:56 GMT+1000 (Australian Eastern Standard Time))

[deleted]

Reply

Mohammadali Yaghoubzadehfard (/users/z5138589) 11 months ago (Thu Apr 23 2020 11:51:45 GMT+1000 (Australian Eastern Standard Time))

yes

Reply

Load More Comments