LELEC2870 Machine Learning Project:
Predicting a film’s gross revenue

Tuesday 8th November, 2022

Introduction

Machine learning methods can be used to solve many practical problems in a wide range of applications such as weather forecast, customer clustering, medical diagnostics, spam blocking, financial time series prediction or signal de-noising, . . . In this project, you will apply machine learning to predict a movie’s revenue in the USA. To this end we wrote a scraper that collected data from the International Movie Database (IMDb) as well as the BoxOfficeMojo websites.

Data

You’ll find the data on a shared drive (due to Moodle size limits ...) in 3 csv files called X1.csv, Y1.csv and X2.csv. The former are the labeled dataset and the latter will be used to make your prediction. Paste these files in your working directory. The data can be loaded in your workspace by running the following commands:

```
import pandas as pd

# Use pandas to load into a DataFrame
#     Y1.csv doesn’t have a header so
#     add one when loading the file
X1 = pd.read_csv(”X1.csv”)
Y1 = pd.read_csv(”Y1.csv”, header=None, names=[’revenue’])
```

The features’ titles are self explanatory but are detailed below in Table 1 for consistency’s sake. The embeddings are generated from pretrained neural networks. The vectors are large in size and part of your work will be to handle these large vectors. This exercise will be something that you’ll very likely encounter is you ever have to process a large amount of natural images or text in your professional life.

Feature	Description
title	the (often) english title of the movie
ratings	the given rating on IMDb
n_votes	the #votes that are averaged for the rating
is_adult	a 1 signifies a movie destined for a mature audience
production_year	the year the movie was produced
runtime	how long the film lasts for
genres	a list of maximum 3 (coma-separated) genres that fit the movie best
release_year	the year the movie was released
studio	the movie-studio that produced the movie
img_url	the url to the poster image
img_embeddings	a vector of size 2048 obtained by passing the poster through a ResNet-50 [2] pretrained on the ImageNet dataset.
description	the text description of the movie
text_embeddings	a vector of size 768 obtained by passing the description to a BERT language model [1] pretrained on translation tasks
revenue	the amount in dollars (already adjusted for inflation) the film made in the USA

Table 1: All the features present in the dataset

Instructions

The project is realized by groups of two or alone. It is composed of different aspects as specified below.

Data Engineering

You will need to re-encode some categorical and integer variables differently (as discussed above). Some features may be useless or redundant, you'll need to evaluate this as well.

Model

You will build regression models that predict the number of shares of an online article. You can use any of the methods seen during the lectures. We expect you to, at least, implement linear regression *as a baseline*, KNN[1], an MLP and one other non-linear method (you can chose one outside those seen during the lectures).
Feature selection and model selection need also be part of your work. Once again you're allowed to use any tools available (e.g. statistical tests seen during other courses). Pay attention to the fact that the model selection can require a lot of computation time. You are advised to explore the metaparameters' space according to the time available.

[1] K-Nearest Neighbour : Regression model with metaparameter K that predicts the output of a sample as the mean of output of the K nearest neighbours in the features space.

Prediction

Once your model is properly selected and validated, you are asked to produce predictions Y2 on the data X2 for which we have kept secret the corresponding targets. This prediction vector should be uploaded on Moodle in a csv file named ”Y2.csv” that contains **one line per prediction and no header**, no quotation marks around your numbers either. Check that your format is correct by opening it with a text editor and compare it to ”Y1.csv”. At the *tail end* of this file you will add an additional number which is the *estimated performance of your model on the unseen data.* Thus **your file should have 1721 lines with pure floating point numbers!** No more no less!
The prediction criterion on which you will be evaluated will be the **RMSE** of your model on the test data.

Report

You will produce a report documenting your technical choices and experimental results. **We do not need a course on the methods you use. We are more interested in what you did and why. Be concise and go straight to the point!** Try to illustrate your results with graphics (with *legends* and *labeled axes*) and comment them. Be critical about what you observe and try to give a possible justification of the obtained results. Summarize your results and observations in a conclusion. A strict maximum of 7 pages (fontsize 11 or larger) will be observed. Appendices might be included in the digital version that you'll submit on Moodle, but shouldn't be resorted to unless something really interesting was found. Remember to make your main observations very clear! All your figures and computation need to be reproducible by us running your implementation code on the provided data.

Bonus question

If you feel bored, you can also play with how the embeddings were generated. A notebook is provided at the following link. It contains the code we used to generate the embeddings present in the dataset. Are you able to improve the text embeddings ? What type of information is present in those embeddings?

Programming languages

The programming language you will use is Python. You can use any toolbox/library available online. In particular, we strongly recommend using the `scikit-learn` library as it provides many useful implementations of standard machine learning approaches. For the MLP we recommend you use the one integrated into sklearn or `pyTorch`. `skorch` is a python package that links `pyTorch` with the `sklearn` syntax, some of you might find it handy, but don't hesitate to come to us with questions if you encounter bugs with these packages.

Agenda

- As soon as possible: Register your group (maximum two people) on the course website.
- Tuesday 8/11 at 10h45: Q/A session #1
- Thuesday 06/12 at 10h45: Q/A session #2
- Friday 23/12 at 23h55: **final deadline** where you submit your work as 3 separate files
  - Your report (.pdf)
  - A csv file called ”Y2.csv” **no header line: one line per prediction values**.
  - A *compressed* folder containing all scripts you wrote for the project. The code should be commented well enough and installation instructions about non-standard packages you used should be provided.

  This submission is mandatory ! You have ample time to submit something functional, so *no* extensions will be given.

Evaluation Criteria

The project will not only be evaluated by your compliance to the instructions (detailed above) and deadlines, but also by the quality of your report. You should present your approach and discuss it to show you have understood what you've done and your results.

Do not forget to think about data engineering (feature selection, how you treated a film's genre, ...) and to present your model selection and evaluation method. Concerning your results, you are expected to compare the four (or more) selected models and to discuss your models' behaviours. Why do you think one performed better than another in general or in particular cases (think about sensitivity to outliers for instance) ?

The files joined to your report (notebooks, utils.py) should be runnable and contain at least what you discussed in your report. There is no size limit, but these files should be structured, commented, and clear enough so that information can be easily found without deciphering everything ! If you used any packages that were not used during the TP sessions, or a different version of those, don't forget to mention it in the beginning of your file !

Please note that the performance of your models is not the most critical part in the evaluation: every step of your machine learning pipeline will be thoroughly evaluated:
Both your report and your points on the project question at the exam, will encompass 10 of the 20 points of your final mark (8 and 2 points on 20, respectively).

Component	**Evaluation Criteria**
Data Engineering	how did you encode non-numerical data?
	how did you approach the large embedding-vectors?
	how were mising datapoints handled?
Features Selection	which features were kept and why?
	do you vary the selected features based on the models correctly?
Model Selection	are each model's hyperparameters tuned?
	did you compare the different models on a level playing field?
	is your analysis of the differences between your models relevant?
	do you clearly discuss the strong/weak points of your model?

Table 2: Main points of evaluation

Tips

Here is a (non-exhaustive) list of tips and tricks for the project.

Before any analysis:

- Visualizing the data is always useful and should always be your first step
- Evaluate your model correctly
- Normalize or Standardize your data if necessary
- Some outliers might be excluded from the learning set (if you decide to remove some observations, explain why you removed them)
- How will you encode the different categories of genres?
- The target goes into high values, is there maybe a transformation you could apply to narrow the scope of the values?

You can discuss the project with other students, in fact, it is a great idea! You could compare your results to those obtained by other groups, but remember that it is not allowed to copy verbatim what others did ...

We will be happy to answer your questions during the Q/A sessions or on appointment. Have fun !

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.