Introduction to Artificial Intelligence

Summer 2023

Project: Movie Recommender

Version: 2
Teacher: Made Wangiyana, sandhi.wangiyana.dokt@pw.edu.pl

This project is done in pairs.

1. Project details

Develop a movie recommender system from datasets of movie statistics and description. Students may choose one or a combination of datasets from the references below. Feel free to choose any approach to develop a solution, however, it is encouraged to use more than one method for comparing performance.

The goal of this project is as follows:

- Perform exploratory data analysis (EDA) on the dataset to understand data distribution, statistical significance of each feature and observe trends. Develop hypotheses and reasoning that can help when building a model.
- Perform data preprocessing to prepare the data before feeding into the model, e.g.: feature cleaning, selection and rescale.
- Split the dataset to create separate training and test datasets.
- Train and compare regression models (may use ML libraries such as Scikit-Learn). Evaluate the model's performance metric and assess the impact of preprocessing strategy (e.g., contribution of features).

2. Datasets

1. Popular Movies of IMDB
2. IMDB 5000 Movie Dataset
3. IMDB Movies Dataset

3. References

1. Documentation and user guide from Scikit-Learn: https://scikit-learn.org/stable/user_guide.html
2. For getting started on the systematic process of Machine Learning projects, checkout Jason Brownlee's blog.

4. Project Phases and Assessment

For each phase, there will be a report to submit followed by a discussion with the instructor. You may receive up to **25 points** for this project, assessed by the following criteria. Note the expected date to complete each phase.

Phase 1: Preliminary documentation (**0-5 points**)

Assessment	Deadline	Points
Preliminary report which contains: 1. Description of the dataset based on your observation. Provide descriptive statistics, plot some samples, report interesting patterns/findings. 2. Overview of your plan to tackle this problem. This includes: a. How to split the dataset into training and validation set b. What algorithms will be used, provide a brief description. c. What are the main tools/framework/libraries used for implementation. d. Proposed evaluation methods. How to measure and compare the performance of your algorithm. e. You can take inspiration from charts in this notebook	**12 May**	2 2
Preliminary discussion 3. Discuss with the instructor regarding your proposed solution. Obtain feedbacks.	**15 May**	1

Phase 2: Midterm solution (**0-5 points**)

Assessment	Deadline	Points
Submit the following files in GitLab: 1. Python code for the current state of implementation. 2. Updated report based on your findings. Including: a. Description of your implementation b. Preliminary results from the algorithm based on proposed performance metrics. c. Experiment analysis. Based on results from your experiments so far, explain what worked and what didn't (what tweaks you made and its effect).	**31 May**	2 2
Mid discussion 3. Discuss regarding: a. Any changes from the initial plan b. Code demonstration c. Challenges or issues encountered so far	**2 June**	1

Phase 3: Final solution (**0-15 points**)

Assessment	Deadline	Points
Project submission. Submit the following files in GitLab: 1. Final Python code for implementation. 2. Final report. Please see final report guidelines below.	**16 June**	5 5

Final assessment	**19-20 June**	
3. Code demonstration:		2
a. Working training code (run for few epochs if it takes too long)		
b. Evaluation code for a trained model		
4. Analysis of the results		2
5. Conclusions: what you've learned from this project		1

5. Handing in guidelines

1. Use the GitLab platform to commit your progress and share this link instead of sending files directly (including the report). Don't forget to write clear instructions on how to use the code.
2. All communications (messages, oral discussion) will be on MS Teams.
3. Please be aware of the deadlines for each project phase.
4. Final project submission (committing the final code and report on GitLab) is on **June 16th**. Late project submissions for this final phase will result in **2 points decrease** per week overdue.

6. Final report guidelines

The following is a suggested template for the final report. However, students may add chapters and sub chapters as needed.

1. Problem definition
2. Dataset
    a. Overview (describe the dataset)
    b. Pre-processing (what is done to the data before feeding them to the model)
    c. Post-processing (mention if used, any methods applied to output prediction)
3. Technical Approach:
    a. Architecture (what kind of model is used, describe briefly)
    b. Training details (describe the hardware and software used for this project. Describe the hyperparameters used for training, e.g.: loss function, learning rate, batch size, etc.)
    c. Evaluation details (what performance metrics are used, e.g.: accuracy, precision, recall)
4. Results
5. Conclusion
6. References
    a. If you rely on algorithms obtained from works of other people, please cite the author and their work (paper, git repo, blog).
    b. Please try your best to cite quotes, facts, images used in your report that you did not create yourself. This will make your work more credible.

7. Parting wisdom

As Francois Chollet (creator of Keras) says, machine learning is an iterative process. Results might not show immediately, so be patient and test out your ideas by making minor adjustments based on your results. Software bugs are real, so check your processing pipeline if there's any unintended variable changes or incorrect use of functions. Finally, and most importantly, don't forget to **have fun while learning**.