Movie Recommendation System with Python and Machine Learning

Maoze Wang mwang373@wisc.edu
Shixuan Song ssong85@wisc.edu
Xuchen Xue xxue8@wisc.edu
Tianchang Li tli289@wisc.edu

1. Introduction

The goal of this project is to create a film recommendation system based on the rating data. During the last few decades, the rise of Internet and web services such as Amazon, Youtube, Netflix, and many other companies had led development in the algorithm of recommendation system. The recommendation system not only helps companies and merchants gain a significant amount of benefits but also make people's lives easier and more efficient than before. Generally speaking, the recommendation system is an algorithm strived to suggest relevant products, such as films, goods, or services, to every customer (Rocca, 2019).[2] The recommendation system is crucial in some industries and can help companies receive a huge income. For instance, Netflix held a challenge, which to find a recommendation system from the public that performs better than its system. The prize for the winner is 1 million dollars.[1].

The online film market is large. Many large companies including Hulu, Amazon Prime Video, and Netflix are using a recommendation system to suggest movies to the customers. The video streaming industry is growing and is expected to grow at approximately USD 82 Billion by 2023, at 17% of the compound annual growth rate (CAGR) between 2017 and 2023 (Vikash,2019). [4] Therefore, it will be meaningful to produce a recommendation system for movies by using the techniques of machine learning.

2. Motivation

Recommendation algorithm matches the demand and the supply in a market which are the two key sides in commerce. As the needs grow and products diversify, it is hard for either side to navigate their target groups. Using merely the past record, an accurate recommendation system could eliminate the information barrier and deliver the most needed products/customers to smooth the commercial chain and reduce the waste of extra resource for navigation.

Movie market is a great application area of such recommendation system, given the considerable number of movies coming up each year and the huge amount of audience. In fact, a lot of online streaming websites already have their own recommendation algorithm. Despite the accuracy of their algorithms, most of them leverage the personal information and watching history of individual users. However, in the future, privacy will be more and more concerned and such personal information may not be as easy to access. Or for a third party who does not have such access at all, a recommendation algorithm with only the features of movies themselves would be necessary. This is what this project is trying to achieve, recommending movies to people only based on the features of movies.

From a more personal angle, our group members all want to gain practical experience of building a recommendation algorithm as it is so popular these days. We hope to under the required machine learning approaches for this application better by working through this project together. This is another motivation.

3. Evaluation

For our item-based recommendation system, we would like to estimate how a specific user would rate a movie by looking at the other movies with similar user rating patterns. In the Figure.1 below, we can see that The Hobbit, The Godfather and Jurassic Park have been rated similarly. To predict how User A would think about The Hobbit, we can refer to the ratings of the later two movies and weighted scores (Saluja, 2018). [3]

Rather than using numeric values for movie ratings, the ratings in our dataset will be binarized into three levels: bad, moderate and good. Then, we would predict the rating level by taking a majority vote of the k most similar rated movies.

To evaluate the model, we are going to compare the difference between the actual ratings in our dataset and the predicted rating levels. After that, we would be able to evaluate our model(s) by calculating the probability of correct predictions.

4. Resources

In this project, we are using The Movie Database (TMDB) 5000 Movie Dataset posted on Kaggle and MovieLens Datasets compiled by Harper and Konstan. We will

[1] https://www.kaggle.com/laowingkin/netflix-movie-recommendation

Ratings Matrix	The Godfather	Rocky	The Hobbit	Fight Club	Jurassic Park
⇨ User A	5	3	?	1	4
User B	4	3	2	?	?
⇨ User C	5	3	5	1	4
User D	1	1	1	4	1
⇨ User E	4	2	5	1	5

Figure 1. Example illustrating how to get the result of recommended movies and their ratings.[3]

primarily work on MovieLens datasets and might need to combine some movie features from TMDB dataset.

MovieLens Datasets is separated into 4 csv files. Ratings.csv contains 100,000 ratings on 9,000 movies by about 600 users, which is one row per user per movie. TMDB movie dataset contains two comma-separated values (csv) files.

Tmdb_5000_movies.csv is one row per movie for 4803 distinct movies with 20 features including movie budget, genre, keywords, original language, popularity, production companies, production countries, release date, run time, average rating and counts of total ratings. Tmdb_5000_movies.csv file is also one row per movie with a list of cast and crew members that might support our further analysis.

We will be using Python with scikit-learn library to conduct our matrix factorization. If necessary, we might need to use R Studio for data cleaning and join our datasets together.

Link to dataset:

TMDB Dataset `https://www.kaggle.com/tmdb/tmdb-movie-metadata/download`

MovieLens Dataset `https://grouplens.org/datasets/movielens/latest/` [1]

Computer hardware:

Python 3.7 Jupyter Notebook

Computational tools:

Regression, Supervised Machine Learning (TBD)

5. Contributions

The overall works of the project, including project paper writings, code files writings, editing powerpoint and presentation, will be assigned evenly to each member in our group.

All members of the group will put effort and participate during each step in our project compiling.To be more specific, the coding part will be assigned equally to each member according to personal interest. Besides, Maoze Wang and Xuchen Xue will write the introduction and background part in the writing paper. Shixuan Song will write the data manipulation part. Tianchang Li will handle the conclusion and discussion parts. All works are divided temporarily and it will be re-assigned properly if there are any changes.

References

[1] F. M. Harper and J. A. Konstan. The movielens datasets: History and context., 2015.

[2] B. Rocca. Introduction to recommendation system,toward data science. 2019.

[3] C. Saluja. Why recommendation systems? 2018.

[4] Vikash. Netflix inc, competitive position and analysis, medium. 2019.