Movie Recommender System - Project proposal

Agnes Huang, Wendy Chiang, Wei Wang

March 2019

1 Outline

This project proposal is prepared for the course Applied Analytics: Machine Learning Pipeline(95-845 AAML). This project will focus on a supervised learning process for Movie Recommendation System. The likely outcome will be a recommender system giving recommendation based on users current movie preference, including the movie they have watched and the features of these movies.

Recommendation systems are widely applied in areas such as book, music, video recommendation. We believe our recommendation system could leverage the large existing user dataset to generate robust, accurate and reproducible work for our research purpose.

There are many existing methods building recommendation system, among which the most popular one is SVD algorithm. This method require minimal knowledge engineering efforts to produce good enough results, by analyzing similarity among users, its recommendation system is built based on user's network[1]. In our study, we will put more emphasis on user him/herself, since for some website, there are lots of user don't friend with other user but still need to get proper recommendation. In addition, we can figure out if our method has a higher accuracy for this group of user.

2 Approach

Our approach can be broken down into dataset, treatment, covariates, population, and outcome:

- The ideal dataset is composed by several tables: movies (movieId, title), genres, ratings (userId, movieId, rating, timestamp), tags (userId, movieId, tag, timestamp). In the end, we will compare and contrast the performance of different machine learning methods and see which one performs best on our test data.

[1] Maya.hristakeva. "Overview of Recommender Algorithms – Part 2." A Practical Guide to Building Recommender Systems. November 19, 2015. https://buildingrecommenders.wordpress.com/2015/11/18/overview-of-recommender-algorithms-part-2/

- The treatment includes movie type, the year that the movie is produced, movie titles, the move tags. The potential values for these features include: tag - boring, dentist, short, dull story, documentary, etc; genres - Drama, Comedy, Romance, etc.

- The covariates can be the year, age, of the audience. Usually the movie preference of one person can somewhat relate to his or her age. For example, young people might have a higher probability of accepting movie produced late recently. While for this project, our potential data does not contain personal information such as age. While we do have the year that the movie is produced.

- The target population is movie audience from 1995 - 2016.

- The outcome is top 10 movies recommended for user. Firstly our recommend system will calculate the rating of particular movie for certain user, and choose the top 10 scores movie.

We consider using measures like MSE(Mean Square Error) or RMSE(Root Mean Square Error) for this project. Since we decide to use item-based error by comparing the difference between top 10 recommending list and the actual list, we can easily get MSE or RMSE for this type of predictive analysis.

SVD is going to be our baseline model. Since, during AAML lectures we were exposed with other machine learning approaches like Latent Dirichlet Allocation (LDA) and ensembling and boosting techniques like Gradient Boosting, we would like to explore the opportunity to apply them on real-world dataset and study more about diverse methods through implementation.

In terms of the size of this analysis, we did not form a hypothesis to specify the Type I and Type II errors in our study, but luckily we found movie lens data has 20 million ratings and 465,000 tag applied to 27,000 movies, which was updated till 2016 by 138,000 users. We believe this data size should be sufficient for our analysis.

3 Remainder

The limitation of our study may be in the recommender system evaluation. The best way to evaluate the system is to run A/B testing on an actual website to see if our recommendations increase click rates, usage, etc. In our study, we do not try to hold an actual survey or to build a website to collect users' feedback. Instead, we will split the data into training/test data set to evaluate the system.

In addition, a real-world recommender system may not only consider end-user, but also consider the (movie) marketer. For example, marketer want the recommender system help them to push the product. In our study, we only consider if the recommender system could add value to end-user.

The use of this analysis is for people who want to build a recommender system from scratch. For example, people could refer to our study to see the performance of different machine learning models in building a recommender system, and see if those algorithms are suitable for their analysis. In addition, the products of the recommender system won't be limited to movies. Our analysis could help people who want to build any products of recommender systems.