Team Proposal: Deep Blue Deep Learning

Team members: Haoyu Zhang; Dingyuan Liu; Yunzhou Liu; Chuanrui Liu; Yiyao Qu; Yumeng Zhang

Date:9/11/2020

Dataset Description

a. **Main Dataset: Movie Recommendation System & Dataset**

- This data set consists of all movies released from 1996 to 2016. It has four separate files, including link, moives.csv, ratings.csv, and tags.csv. UserID, moiveID, ratings, and genres are features that could potentially be used to predict the likelihood of certain users to like certain movies.
- Link to the dataset: https://www.kaggle.com/bandikarthik/movie-recommendation-system

b. **Backup Dataset: Amazon Product Review**

- The Amazon Product Review dataset contains various products' and users' ratings on different products. This dataset contains only one file with four columns, and our current idea for this dataset would be to transpose the dataset by UserID use FNN & PNN to build a product recommendation system.
- Link to the dataset: https://www.kaggle.com/saurav9786/amazon-product-reviews

Motivation and Goals

Motivated by the various recommendation systems in current apps, such as Amazon, YouTube, and Spotify, we have decided to focus on building a recommendation system for a user-based online service. To accomplish this goal, we analyze users' feedback on different objects to build a well-defined recommendation system and improve their overall experience. Specifically, we try to develop a list of potential items that a user might like based on their previous ratings.

Techniques

We have searched for some commonly used recommendation systems and decided that both classification model-based(supervised) and clustering (unsupervised learning) could work.

In the classification model, we would use the users' ratings and features of the movies to decide whether our users will be interested in the film or not. Currently, we are considering applying some classification algorithms like decision trees. The basic idea is to build a model to predict each user's likelihood to choose any given movie and then recommend the movie that produces the highest likelihood. We are also considering applying K-nearest-neighbor (KNN) on the dataset for prediction. In theory, it will sift out new data points according to the k number of the closest data points we preselected. As in our case, we could take each user's rating for a particular movie as our preselected data points. And then we select k users who lie close to those points (who have similar tastes as out preselected users.) The distance between users will be calculated by movie features, such as movie genres, when users' ratings are given. Additionally, we could try to build a Neural Network to achieve this goal. In particular, we sequence the user's previous history and calculate "which movie will likely be followed by a particular movie." For instance, if we found a trend that many users tend to watch Titanic after Romeo and Juliet, we would recommend Titanic after they watch Romeo and Juliet.

Unsupervised learning would be our alternative approach, and we tentatively assume to use the Mixture of Gaussian model to achieve this goal. With this method, we hope to identify clusters of users who have similar interests in movies based on features such as ratings, genres, and movie tags. For a given new user, we will first group him or her into some individual cluster. Within the individual cluster, we could offer recommendations by recognizing the movies in the movie lists of the user's neighbors but have not been seen by the user himself. Furthermore, as there may be more than one movie in the neighbors' lists that the user has not watched, in this case, we may utilize the average ratings of the candidate movies as the standard to sort and recommend the film with the highest ratings.