Data Analytics 1

Assignment 3

Association Rule Mining

Release : 30th September 2023
Deadline : 13th October 2023 (11:55 pm)

The objective of the assignment is to Build an association rule-based movie recommender system.

Provided is a dataset comprising 100,836 ratings and 3,683 tag applications across 9,742 movies. This dataset captures user ratings, with a 5-star rating system, from MovieLens, a movie recommendation service. For additional information, please consult the README.

All the ratings are stored in the 'ratings.csv' file. Each line in this file, following the header row, corresponds to a single user's rating of a movie and includes the following details: userId, movie, rating, and timestamp.

Movie-related data can be found in the 'movies.csv' file. Each line in this file, following the header row, represents information about a specific movie, and it includes the following information: movieId, title, and genre

Data preprocessing :

- Form the transactional data set, which consists of entries of the form <user id, {movies rated above 2}>.
- Consider only those users who have rated more than 10 movies
- Divide the data set into 80% training set and 20% test set. Remove 20% of movies watched from each user and create a test set using the removed movies

Association rule mining:

- From the training set, extract the set of all association rules of form X→Y, where X contains a single movie and Y contains the set of movies from the training set by employing the apriori or FPgrowth approach and set some minsup and minconf (eg : 50 and 0.1 respectively)
- Recommendation: Generate two sets of lists. The initial list includes the top 100 association rules, arranged in order of their support. The second list comprises the top 100 rules, prioritizing them according to confidence. Identify the rules that appear in both lists, and then arrange these shared rules based on their confidence score.
- For each user in the test set, select association rules of the form X→Y, where X is the movie in the training set. Compute the average precision and average recall by varying the number of rules from 1 to 10 and plot the graphs. For example, consider rule X→Y, where X is the movie from the training set. The set of movies in set Y is recommended movies. In this manner, if we consider N rules, combining the movies on the right side of each N rule constitutes the set of recommendations, say R. The intersection of R with the test set is called the hit set. The ratio of the hit set and test set is equal to recall. The ratio of the hit set and recommendation set is equal to the precision.
- Take a sample example of users and their movie ratings from the test set and Display precision and recall graphs.
- Include the plots in Report (md or pdf) along with your justification of selection of your algorithm and also briefly explain how you built the entire recommendation system in the report .

Submission format:

- Mention any inference from plots or results in notebook Markdown itself
- Submit a zip folder named <assignment3_teamId> containing the following files(you can include any other implementation files as well):
    - <TeamId>_report (.pdf or .md)
    - <TeamId>_recommender(.py or etc)
    - <TeamId>_top100RulesByConf (.txt)
    - <TeamId>_top100RulesBySup (.txt)