Movie Preferences Markov Model

Ishaan Sathaye

December 10, 2020

1 Real-World Phenomenon

Recommendations and preferences in technology services are very important to companies and people, as it can help companies build profiles and help people find their favorites. Many services like Netflix, Youtube, and Amazon Prime Video depend on their recommendation systems to engage their customers. Personalizing a user's preferences on movies can help movie theaters and advertisements to be more focused. Using genres and movie ratings in a recommendation chain program can provide users with similar movies, based on what they have watched or liked. Movies that users tend to enjoy, usually have a genre in common and they all come with a rating. Using a combination of these aspects would allow a program to recommend users to watch certain movies over others.

2 Derivation and Assumptions

2.1 Derivation

The Markov model was derived by finding characteristics of movies that can be used to recommend the user more movies.

Genres(19): Adventure, Animation, Children, Comedy, Fantasy, Romance, Drama, Action, Crime, Thriller, Horror, Mystery, SciFi, IMAX, Documentary, War, Musical, Western, and Film-Noir
Ratings: On a scale from 1.0 to 5.0

The initial state matrix would entail the movie name and the rating of the user. This data would be organized with the name following the rating in the program. This matrix below could be a sample of what the user has watched or liked.

$$S_0 = \begin{bmatrix} ToyStory : 3.5 \\ Jumanji : 2.0 \\ Akira : 4.5 \\ Rampage : 3.1 \\ AvergersInfinityWar : 4.2 \end{bmatrix}$$

The program would then search the dataset for these movies and find all their genres. The transition matrix would include the quantitative data values consisting of a weighted average between genres and ratings for all the movies in the dataset. This data will be organized in a tabular format, and then converted to a list to find the preferences. Each position shows the probability of watching one movie over another, as shown in the matrix below. The matrix is calculated and continued for all movies in the dataset.

$$T = \begin{bmatrix} 1 & 0.1 & 0 & 0 & ... \\ 0 & 0.1 & 0 & 0.1 & ... \\ 0 & 0.5 & 0.3 & 0 & ... \\ 0 & 0.3 & 0.7 & 0 & ... \\ ... & ... & ... & ... & ... \end{bmatrix}$$

2.2 Assumptions

Assumptions need to be made in order to simplify the created recommendation system. Outside recommendations will not be considered, therefore the program will only take in the current user's preferences. This makes the system less complex and easier to track the sole user's preferences. Another assumption would be that there will be only two characteristics that the preferences will be based. The movie dataset will only include ratings and genres. Also, the assumption will be made that well-known and popular movies will be considered as a recommendation, so this may exclude foreign movies.

3 Data and Parameters

Data of movies, ratings, and genres will be taken from GroupLens (link) and IBM. Using `!wget -O`, the program will download and unzip the dataset from an IBM API that contains the GroupLens movie data. All types of data will be first put into a CSV file, then into a data frame, and preprocessed.

GroupLens Link: http://grouplens.org/datasets/movielens/

The movie dataset will be then put into a data frame and preprocessed to include the movieID, title, and the genres only. The table below shows the first five rows of the move and genre data frame:

-	movieID	title	genres
0	1	Toy Story (1995)	Adventure,Animation,Children,Comedy,Fantasy
1	2	Jumanji (1995)	Adventure,Children,Fantasy
2	3	Grumpier Old Men (1995)	Comedy—Romance
3	4	Waiting to Exhale (1995)	Comedy—Drama—Romance
4	5	Father of the Bride Part II (1995)	Comedy
...	...	...	...

The same thing will apply to the ratings dataset, which will be first be in a CSV file then put into a data frame to be preprocessed. The table below shows the first five rows the ratings dataset that are organized by the movieID, which correspond to the movie and genre data frame:

-	userID	movieID	rating
0	1	169	2.5
1	1	2471 (1995)	3.0
2	1	48516	5.0
3	2	2571	3.5
4	2	109487	4.0
...	...	...	...

Parameters for the initial state matrix will include the user's watched or liked movies that will be inputted. The Pandas data frame of the input will look simple with only the title of the movies that the user likes and the ratings of those movies:

-	title	rating
0	Breakfast Club, The	5.0
1	Toy Story	3.5
2	Jumanji	2.0
3	Pulp Fiction	5.0
4	Akira	4.5

4 Benchmarking

5 Results

6 Discussion

7 Matlab Code