Predicting movie ratings and viewer's sentiment analysis based on Movielens dataset

(By Manisha Chandran, Shwetha Panampilly, Tanmay Sawaji, Vishwas Desai - Group 10)

Problem Statement

Our project uses the movielens dataset. This dataset contains key attributes like movie titles, genres, ratings, IMDb ID. Our primary goal is to predict ratings of a movie based on the above attributes and additional attributes will be retrieved by making use of the IMDb API. The secondary goal is to binarize the reviews attribute by applying sentiment analysis which is trained on a different dataset containing reviews and sentiment. Through this project we intend to predict the performance of the movie and the viewer's sentiment.

Tasks Completed

1. Retrieval of attributes that were not available in the original dataset.
2. Cleaning performed on the newly retrieved data.
3. Finding correlation coefficients for each attribute against the target (ratings).
4. Maintaining a separate review dataset for sentiment analysis.
5. Written code for bag-of-words Naive Bayes classifier for sentiment analysis.

Future Challenges:

- The correlation found between budget and ratings was very low, but we decided to keep these attributes and train different models and compare their performance.
- The list of cast for each movie contains a lot of actors. We need to decide whether we include all actors or just top 'k' actors in the cast. We plan on training different models for both cases and compare their performance.
- We need to decide whether to keep the attribute 'positive review percentage' as it will be calculated by our Naive Bayes classifier.
- Determining which activation functions to use for each layer. We also need to determine how many layers we need in the ANN and how many neurons should be present in each layer.

Tasks performed by each member:

Tanmay Sawaji:

- Wrote code for fetching director, actors, budget, box office and reviews from the IMDB API for each movie in the Movielens dataset and stored them against the imdb_id.
- Wrote Bag-of-Words Naive Bayes Classifier for sentiment analysis. (testing and implementation is still remaining).

Manisha Chandran:

- Downloaded the dataset from Movielens that contains 3 files and IMDb dataset.
- Merged them into a single csv file.
- Did initial data preprocessing such as dropping unwanted attributes like IMDb ID, TMDb ID, timestamp.

Vishwas Desai:

- Did further data pre processing part on the merged dataset
- Changed ratings.csv (from Movielens) to avgRatings which contained an average rating for each movie in the dataset. Earlier there were multiple ratings for each movie which would have been difficult to use ir for further processing
- Found the correlation between attributes like box office, budget with our target variable(ratings)

Shwetha Panampilly:

- Wrote code to return the leading cast from the entire cast given.
- Wrote an API call to get the leading cast from the entire cast of a given movie.
- Created a csv file to store the above attribute.

Answering comments to Milestone 1:

1. **I'm not sure how easy it will be to predict movie ratings from the mentioned attributes. Are there other attributes that haven't been listed?**
   Ans. Yes, there are attributes other than the ones listed in the movielens dataset. The movielens dataset contains the following attributes that are useful to us 'movie_id', 'genre', 'ratings' and 'imdb_id'. Using the 'imdb_id' attribute in the movielens dataset, we are calling the IMDB API to fetch attributes like 'budget', 'box office earnings', 'director', 'cast' and 'reviews'. Using sentiment analysis on the reviews, we will form a new attribute 'positive sentiment percentage'.

2. **Also, what is the size of the movielens dataset?**
   Ans. The movielens dataset that we are using has 9742 unique movies.

3. **Will there be enough data to train your learning algorithm?**
   Ans. We believe that there will be sufficient data to train our learning algorithm.

4. **The second goal is also reasonable. How, will you binarize the reviews?**
   Ans. We are using a different dataset called 'IMDB 50k reviews'. This dataset contains around 50,000 reviews with their sentiments (positive or negative). We will use this dataset to train our sentiment analysis bag-of-words Naive Bayes classifier and then classify (or binarize) the review attribute for our main dataset. The final attribute in the dataset will be the percentage of positive reviews in that specific movie based on our classification.

5. **Also, do you have data that already has the "performance" for each movie?**
   Ans: Wrt scope of the project the "performance" of a movie is the rating we intend to predict and yes, we have enough data to predict ratings.

6. **You have four people in the team and how will the work be divided amongst the four of you?**
   Ans. Right now, we are forming smaller goals and dividing the work among ourselves by assigning each goal to each member of the team. We have mentioned challenges that we might face in the future in the section above. Solving these problems might be considered as smaller goals and the analysis of these problems will also be divided among the teammates along with progression on the modelling of the ANN to be used.