Machine Learning & Data Mining
Miniproject 3

Caltech CS/CNS/EE 155
March 10, 2017

MOVIELENSVIS

by Shari Kuroyama and Carly Robison

---

1 Data Formatting

We made three data structures to hold the movie name and genre data.

- movie_names: movie ID to movie name for plotting.
- movie_genres: movie ID to the list of genres that it is categorized as.
- genres: genre name to the set of movie IDs that fit in that genre.

These worked well for our purposes.

2 Basic Visualizations

Packages Used

To plot the histograms, we used `matplotlib`'s function `pyplot.hist`. This function has options for side-by-side data as well as stacked.

Methods

For categories 1 and 4, we counted the total number of 1s, 2s, etc., and plotted their frequencies.

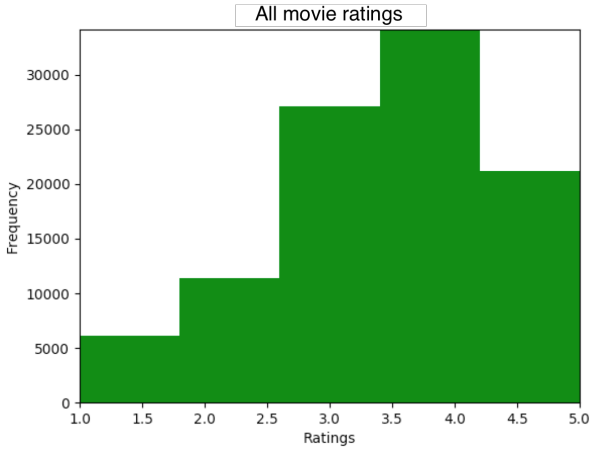
For categories 2 and 3, we calculated the top ten movies and plotted the number of 1s, 2s, etc. that each movie got side by side. For category 3, a naïve try will pick up movies with only one rating of 5. To ensure that the movies we found were actually good movies, we required our choices to have more than 20 ratings.

We decided not to normalize the frequency data, since this way the movies which had more ratings are clearly distinguishable from those with smaller numbers. This is especially relevant for category 3, since it shows how many more ratings *Star Wars* got than, for example, *Wallace & Gromit*.

Observations

1. Most of the ratings in the dataset were 3s and 4s, with very few 1 ratings.

Figure 1: All ratings in the MovieLens Dataset.

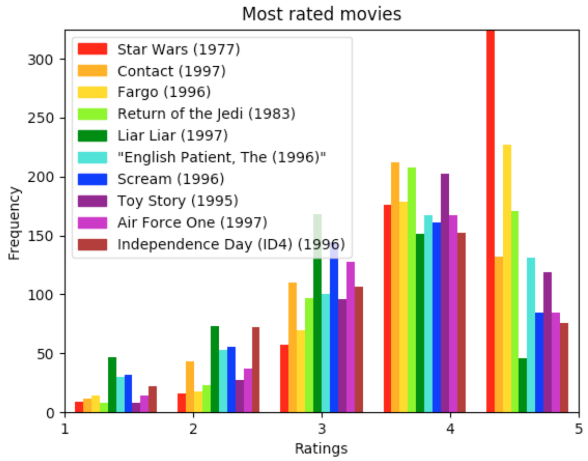


MOVIELENSVIS
by Shari Kuroyama and Carly Robison

---

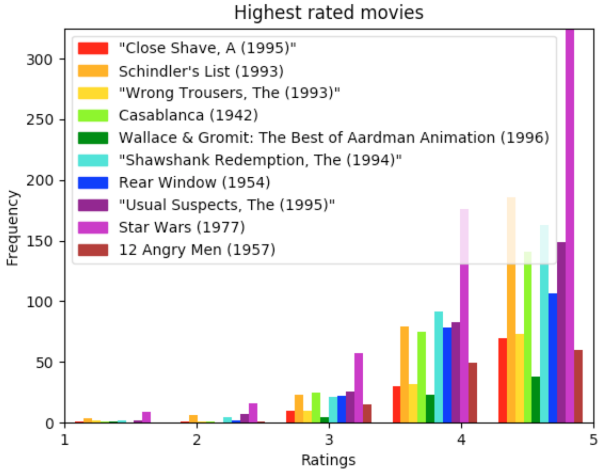
2. Most of these movies were not only rated a lot, but rated fairly highly as well. A notable exception is *Liar Liar*, which has very few 5 ratings (mostly 4s and 3s). This makes sense, because the movies which are watched the most are likely to have been promoted by people who watched them and liked them.

Figure 2: All ratings of the ten most popular movies.



3. As noted above, we restricted the movies we chose to have more than 20 ratings in order to be considered one of the "best". We agree with most of the outputs; in particular, the fact that three of these movies are from the Wallace & Gromit series was amusing.

Figure 3: All ratings of the ten best movies.



MOVIELENSVIS
by Shari Kuroyama and Carly Robison

---

4. All ratings of movies from three genres of your choice. We chose Comedy, Romance, and Western. We looked at the same genres in the Matrix Factorization analysis.

   How do the ratings of the three genres you chose compare to one another? TODO

Figure 4: All ratings of Comedy movies.

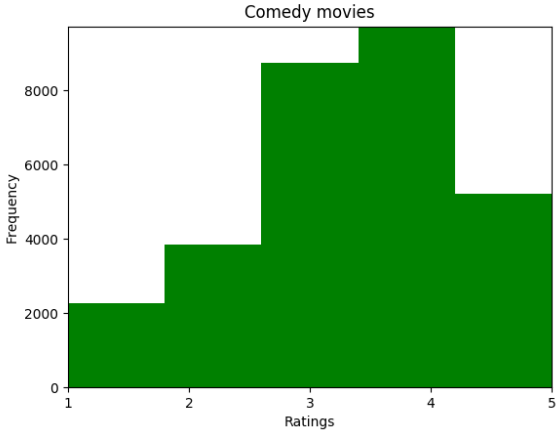


Figure 5: All ratings of Romance movies.

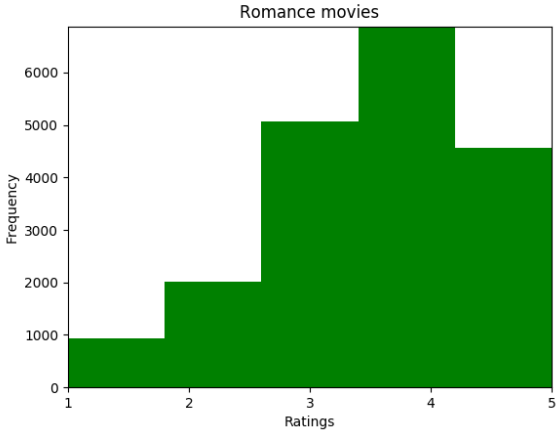
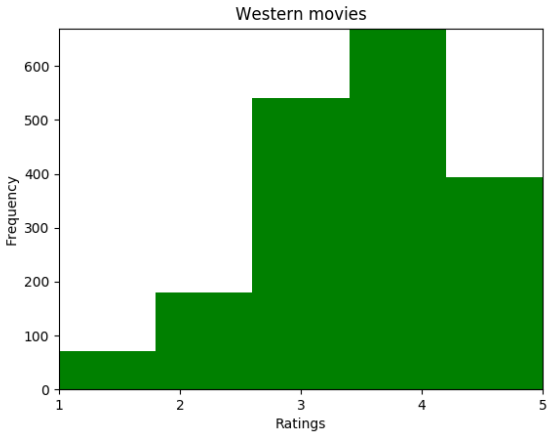


Figure 6: All ratings of Western movies.



3 Matrix Factorization Visualization

Packages Used

TODO

- We used the code from the solutions to Homework 6 for factoring a matrix.

Algorithm

TODO

Observations

Your report should contain a section dedicated to matrix factorization visualizations. What, in general, did you observe? Did the results match what you would expect to see? How does the visualization of the most popular movies compare to that of the best movies? How do the visualizations of the three genres you chose compare to one another? TODO

4 Conclusion

Division of work

Shari did basic visualization and made the pretty rainbow histograms!

Carly worked on matrix factorization and plotting the projections in 2D.

Discoveries

What are your conclusions/observations about the models you used and the visualizations generated? TODO

MOVIELENSVIS

by Shari Kuroyama and Carly Robison

Challenges

TODO

Shari forgot what a histogram was and had trouble getting the function to work, because she was trying to provide the wrong kinds of data. But we figured it out, and got very pretty rainbow graphs in the end.

Concluding Remarks

TODO

Our code is available at this github repo.