Netflix Prize Dataset Recommender System

Aaron Gauthier, Pedro Uria Rodriguez, Zach Buckley

The George Washington University
Data Science Program

DATS6103: Intro to Data Mining

Submitted
22 April 2019

Introduction

Netflix Prize Competition

- Open Competition for team or individual participants
- $1 million grand prize
- > 10% Improvement over 2006 Netflix Cinematch algorithm
  - Based on Root Mean Squared Error (RMSE)
- Competition started October 2, 2006
- Grand Prize awarded September 29, 2009
  - Team “Belkor’s Pragmatic Chaos” with 10.6% better RMSE

EDA & Preprocessing

- Original Data Dimensionality
  - 100,480,507 Ratings
  - 480,189 Users
  - 17,770 Movies
- Sample
  - 510,852 Ratings
  - 1,934 Users
  - 11,866 Movies

movie_id	user_id	rating	date
1	1488844	3	2005-09-06

Modeling each user

- For each user, found a dataset of similar users B such as d(A, B) <= threshold

$$d(A,B) = \frac{1}{n \cdot 5^2} \sum_i (r_{A_i} - r_{B_i})^2$$

movie	`cluster_avg_rating`	`user_A_rating`
movie_1	Known	Maybe Known
movie_2	Known	Maybe Known
...	....	...

- Run linear regression and other regression models.
- Tried different distance thresholds. Smaller threshold $\rightarrow$ Better accuracy but less movies with predictions and smaller training & testing set.
- Also, small threshold $\rightarrow$ some users with no other similar users $\rightarrow$ cannot predict nor recommend any movie.

Modeling each user

- We found threshold = 0.02 to be a good middle ground. Mean R^2 = 0.65, and enough data to train and evaluate for most users (73 movies on average).
- This threshold means a 0.07 unit distance on average. For example, if A rated a movie with a 5, B will be similar to A if he rated the same movie with at least a 4.3.
- We got perfect R^2 for some users and even negative for others.
- This model could be used for those users for which it performs well.
- It can also be used to identify abnormal users (negative or very small R^2), and take them out when training other more sophisticated models.
- The threshold could also be selected automatically for each user.
- The dataset could be augmented with IMDb features, which could drastically improve the results.

Item-Item Collaborative Filtering

- Method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).
- Assumption of the collaborative filtering approach is such that if person A has the same opinion of person B about something, then A is more likely to have B’s opinion on a different issue than that of a randomly chosen person.
- Considered Pearson R, Kendall-Tau Rank, and Spearman Ranking Correlation.
- Pearson R and Spearman Ranking were selected for analysis.

Analysis of Recommendations

- “Miss Congeniality” results included “Another Stakeout” for both correlations
- “The Godfather” results included “Latter Days” for both correlation
- Comparison between Collaborative Filtering (correlation) and Linear Regression:
  - No overlap between Linear Regression and Collaborative Filtering for “Flubber” and “Ice Age”
  - Overlap between Pearson R and Spearman Ranking Correlation
- Different approaches definitely yield different results
- Netflix “perturbed” the dataset
  - Statistical modification
  - Protection of users identity

Conclusion

- Learned about Recommender Systems
- Recommender Systems are very complex!
- Experienced “The Curse of Dimensionality”!
- Improvements for the Future:
    - Utilize IMDB, MovieLens, Gross Revenue, and other metadata
    - Parallel computing, GPUs, and Neural Networks
    - Experimentation with Self Organizing Maps or Self Organizing Feature Maps (SOM/SOFM) & Sparse Matrices

References

https://www.kaggle.com/netflix-inc/netflix-prize-data

https://www.netflixprize.com/faq.html

https://en.wikipedia.org/wiki/Netflix_Prize

https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

https://github.com/amir-jafari/Data-Mining

Stanford Lecture Series 41-45: Overview on Recommender Systems:

- https://www.youtube.com/watch?v=6BTLobS7AU8
- https://www.youtube.com/watch?v=2uxXPzm-7FY
- https://www.youtube.com/watch?v=1JRrCEgiyHM
- https://www.youtube.com/watch?v=h9gpufJFF-0
- https://www.youtube.com/watch?v=VZKMyTaLI00