Collaborative Filtering - Group 2C

Animesh Danayak, Dave Arno, Dhwani Mehta, Edwin Liu, Letian (William) Ma and Max Gordon

BAX-401-002 Information, Insight, and Impact

TABLE OF CONTENTS

Executive Summary

We have been provided with user ratings of 45 classmates and 139 user ratings from DBMI for 50 movies. The objective of this assignment is to use different filters to arrive at a model that predicts user ratings for our team mates as well as for three new users (Amy, Shachi and Camille).

We used a variety of filters to test predictions for our team members. These included User based methods where users' preferences are compared and Item based methods where the movies are compared with each other. We also used different forms of normalization to decrease biases. These included Mean Centered and Z-score normalization. We chose cosine similarity as the measure to compare user and item vectors as it gave us the most accurate predictions. Below are the filters with the best prediction for each problem.

- *Predicting User Scores for certain movies: We used user-user mean centered*
- *Predicting User Scores for movies with few ratings: user-user mean centered*
- *Predicting movie ratings for new users: item-item mean centered*
- *Predicting movie ratings for users with few data points: user-user mean centered*

Apart from movies, user recommendation systems can be applied to other industries which require selection of a product or individual basis their past preferences. For eg: Finding the ideal life partner basis the inputs provided on a dating website, identifying promotional products to display to a customer basis past purchases, music recommendations based on similar singers or genres, etc.

Problem statement and ratings

1. A recommendation system encompasses algorithms to suggest similar products to users. The similarities can be calculated by using many different methods, such as Cosine, Euclidian, Manhattan distances, or correlation between users. We applied cosine similarity to predict movie ratings and compared and contrast the ratings using original data, mean-centered, and z-scored.

We have been given a list of movie ratings from 45 students in our MSBA 2021 class along with movie ratings from 139 DMBI users on 50 movies. The ratings are ordinal and range from 1 to 5, with 1 being the lowest and 5 being the highest rating. Three types of collaborative filtering - User-User cosine similarity, User-User Mean Centered, and User-User Z-Scored - have been applied to build a movie recommendation system to predict our group members' ratings for three movies. Root Mean Squared Error (RMSE) is then used to evaluate and compare the performances of the three methods against predicted ratings and observed ratings.

Based on the RMSE, a measure of the error rates, User-User Mean Centered filtering yielded the lowest error rates for four out of six of our group members and thus the best model for predicting the ratings of the three movies. The prediction results for Life of Pi, Toy Story 3, and The Imitation Game for our group members are shown in the chart below.

User-User Mean Centered												
	Edwin Actual	Edwin Predict	Animesh Actual	Animesh Predict	Dhwani Actual	Dhwani Predict	Letian Actual	Letian Predict	Dave Actual	Dave Predict	Max Actual	Max Predict
Life of Pi	2	2.66	4	4.53	4	3.16	2	2.38	3	3.48		3.16
Toy Story 3	3	2.34	4	4.13	4	2.79		1.84	5	3.22	5	2.96
The Imitation Game	2	2.43	4	4.29	4	2.91	3	2.19		3.20	3	3.06

	Actual Ratings		Predicted Ratings	
Movie	Harshest Reviewr	Kindest Reviewer	Harshest Reviewer	Kindest Reviewer
Life of Pi	Edwin/Letian	Animesh/Dhwani	Letian	Animesh
Toy Story 3	Edwin	Dave/Max	Letian	Animesh
The Imitation Game	Edwin	Animesh/Dhwani	Letian	Animesh

Edwin seems to be the harshest critic in actual ratings while Letian is predicted to be the harshest critic in predicted ratings. Animesh is predicted to be the kindest reviewer, but we have mixed results in actual ratings.

2. We first investigated the three films to be predicted. We found that out of 180 users, Son of Saul only had two ratings, while both A Serious Man and Winter's Bone only had one rating each. A single user, User.12, was responsible for three of these ratings. We decided on user-based collaborative filtering using cosine similarity. User-based filtering lets us tailor predictions to each individual user. We chose to standardize the ratings using both mean-centered and z-score. A large reason for this was because User.12 rated movies at a mean rating of 1.5, which caused an abnormal bias in the data. By normalizing the data we could remove this bias. Below are our best predictions for all three movies per person. As you can see, the ratings cluster around each team member's average rating. This is due to the lack of ratings for each movie as well as the two users who did rate the movies having very little variance within their overall ratings.

	Arno	Danayak	Gordon	Liu	Ma	Mehta
Winter's Bone	3.305667	4.266267	4.269898	2.998131	3.198939	3.622715
A Serious Man	3.308555	4.266296	4.270642	2.996057	3.197761	3.620399
Son of Saul	3.305398	4.266263	4.269592	2.998036	3.198830	3.622447

3. The recommendation system suffers from Cold Start, a problem faced by companies with new users or items, where there isn't enough historical data to make accurate recommendations. Therefore we used Item-item mean score to predict the ratings for the new users.

	Avatar	The.Wolf.of.Wall.Street	Inception
Shachi.Govil	3.993865	4.071942	4.486486
Amy.Russell	3.993865	4.071942	4.486486
Camille.Mack	3.993865	4.071942	4.486486

4. We first combined the 45 students in our MSBA 2021 class along with movie ratings from 139 DMBI users. With the additional rating information, we decided on user-based collaborative filtering using cosine similarity. We have explored user-user mean-centered collaborative filtering without normalization, which returned similar average ratings for three movies across the users. However, it can be observed that average ratings vary wildly across users. Thus, we decided on using user-user mean-centered collaborative filtering with normalization to predict the movie ratings for Camille, Shachi and Amy with the additional rating information provided.

	Avatar	Inception	The Wolf of Wall Street
Shachi Govil	4.17	4.19	4.41
Amy Russell	3.05	3.10	3.14
Camille Mack	2.74	2.67	2.75

With the additional information from the new customers, we are able to find out the user preferences, thus we obtained significantly different results for the predicted ratings of the three movies compared to part 3.

Managerial recommendations

5. E-Commerce firms are able to take advantage of the “Long Tail” effect which allows them to provide a wide array of products at minimal extra costs. Physical retailers on the other hand, are unable to provide products that are not profitable enough to secure limited shelf space. Collaborative Filtering allows online retailers to advertise specific items within their large catalog to specific consumers. As shown above, we were able to reliably predict how users would rate movies. This allows the firm to target consumers with products they have a higher probability of enjoying and purchasing. Firm’s can easily expand this model outside of movies across music, books and video suggestions, product recommendations based on past searches and purchases. Additionally, the model can also be employed as a way of strengthening a cross-selling strategy, as customers can be recommended to other products they did not have intention to buy otherwise. The system could also be customized used to customize partner search for dating, job requirements and even MSBA students.

Conclusion

All companies, traditional and high-tech firms, want to have personalized products for their customers. Recommendation systems help companies understand their customers and provide the most relevant products that satisfy their needs. However, recommendation systems do have limitations and have to be remedied with domain knowledge and more data-informed decisions.

References

R. Vidiyala. (2020) *How to Build a Movie Recommendation System*. Retrieved from:

How to Build a Movie Recommendation System

H. Gaspar. (2015) *The Cold Start Problem for Recommender Systems*. Retrieved from:

https://yuspify.com/blog/cold-start-problem-recommender-systems/

S. Prabhakaran. (n.d.) *Cosine Similarity – Understanding the math and how it works (with python codes)*. Retrieved from:

https://www.machinelearningplus.com/nlp/cosine-similarity/

Leskovev, J., Rajaraman, A., & Ullman, J. (2014). *Mining of Massive Datasets.* Cambridge University Press.

Appendix

Formulas Used:

Cosine similarity formula:

$$Cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \, \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \sqrt{\sum_1^n b_i^2}}$$

where, $\vec{a} \cdot \vec{b} = \sum_1^n a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ is the dot product of the two vectors.

User-User Collaborative Filter:

$$P_{u,i} = \frac{\sum_v (r_{v,i} * s_{u,v})}{\sum_v s_{u,v}}$$

Root Mean Squared Error Formula:

Root Mean Squared Error (RMSE)

RMSE is the square root of the average of squared errors and is given by the below formula.

$$RMSE = \sqrt{\frac{(r - r\hat{})^2}{N}}$$

Where:
r is the actual rating,
r^ is the predicted ratings and
N is the total number of predictions

RMSE measures the squared loss and can be used to evaluate a recommender system's performance. Lower values mean lower error rates and thus better performance.

1.1

RMSE						
Method	Edwin	Animesh	Dhwani	Letian	Dave	Max
Orignal Data User-User	1.240872572	2.674001843	1.806215167	0.677942718	1.948570413	2.564129485
User-User Mean Centered	1.215346215	0.979810887	1.414548433	0.814192678	1.286188475	1.994134639
User-User Zscored	1.427093089	0.987992666	1.435939131	1.61159941	1.303916047	1.928139723

User-User Mean Centered performed the best among the three methods.

4.1

	Avatar	Inception	The Wolf of Wall Street
Shachi Govil	3.45	3.49	3.06
Amy Russell	3.46	3.52	3.03
Camille Mack	3.46	3.50	3.03

User-User without normalization

4.2

	ShachiGovil <dbl>	AmyRussell <dbl>	CamilleMack <dbl>
Inception	4.199836	3.051592	2.675243
Avatar	4.173337	3.005182	2.748138
The Wolf of Wall Street	4.372169	3.129403	2.720995

3 rows

Z-Score Standardized User-User Cosine Similarity