AIE425 Intelligent Recommender Systems, Fall Semester 24/25

Assignment #2: Significance Weighting-based Neighborhood CF

Filters

221100500, Maryam Eslam Sameer

1. Overview of Significance Weighting-Based Collaborative Filtering

1.1 Abstract

This report examines the application of significance weighting in neighborhood-based collaborative filtering (CF) models. It enhances user-based and item-based CF methods by integrating significance weighting into similarity calculations. This adjustment prioritizes relationships with substantial overlapping ratings, improving recommendation accuracy. The report outlines how data is prepared, organized, and analyzed to produce personalized recommendations, leveraging modified similarity metrics like weighted Cosine and Pearson correlations.

1.2 Introduction

Recommender systems play a crucial role in helping users navigate vast datasets in e-commerce, entertainment, and other domains. Collaborative filtering (CF) is a widely adopted approach that generates recommendations based on user interactions or item characteristics. While traditional neighborhood-based CF relies solely on similarity measures, significance weighting refines this process by emphasizing more meaningful relationships.

This method improves CF by addressing the limitations of small overlap issues in similarity calculations. Data preparation involves constructing a user-item matrix, followed by calculating weighted similarities. Significance weighting factors ensure that pairs with larger co-rating overlaps have greater influence, reducing the impact of coincidental correlations from limited data points. This report explores how incorporating significance weighting enhances the precision and relevance of CF recommendations.

2. Assignment requirements and description

Full code implementation is provided here: [Assignment 2 Code Colab]

2.1 General requirements for the two parts

1. **Dataset Preparation**
   - The dataset was cleaned by removing rows with missing or zero ratings.
   - The ratings were confirmed to range between 1.0 and 5.0.
2. **Statistics on Users and Items**
   - Total number of users: **50**
   - Total number of items: **99**
3. **Number of Ratings per Product**
   The top five items with the most ratings:

Product ID	Number of Ratings
I54	12
I71	11
I25	11
I8	10
I22	10

4. **Active Users Selection**
   Three active users were selected for analysis:
   - **U35**: User with 2 missing ratings.
   - **U7**: User with 3 missing ratings.
   - **U10**: User with 5 missing ratings.
5. **Target Items**
   Two target items were considered:
   - **I2**: Item with **4% missing ratings**.

- **I54**: Item with **10% missing ratings**.

6. **Co-Rating Analysis**

   The co-rated items and associated user counts for each active user:

Active User	Number of Co-Rated Users	Number of Co-Rated Items
U35	28	5
U7	43	13
U10	31	9

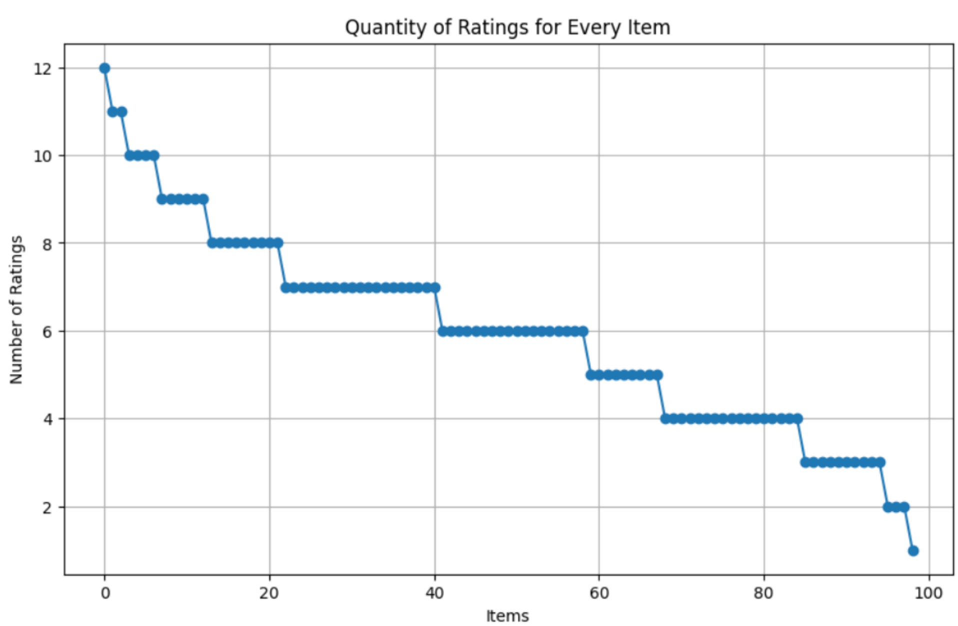
7. **Top Common Users**

   The top users with the highest number of common items (sorted in descending order):

Common Count	User ID
4	U27
4	U31
4	U40
4	U47
4	U41

8. **Quantity of Ratings for Every Item**

   A curve illustrating the number of ratings per item is provided. The plot shows a gradual dec unt at **12**.



9. **Threshold β for Co-Rated Items**
   The threshold β was determined as the number of users co-rating at least 30% of items for each active user:

Active User	Threshold β
U35	4
U7	6
U10	7

10. **Results Storage**
    The results, including user and item statistics, ratings per product, active users, target items, co-rating analysis, and threshold β values, were saved for future use.

Case Study 1.1

1.1.1 User-Based Collaborative Filtering with Cosine Similarity

- Cosine similarity was calculated without mean-centering.
- The similarity matrix was generated using the user-item matrix.

1.1.2 Top 20% Closest Users

The top 20% closest users for each active user were determined based on cosine similarity:

Active User	Top Closest Users
U35	U47, U31, U21, U26, U50, U15, U43, U10, U37
U7	U40, U13, U1, U6, U9, U44, U50, U27, U41
U10	U21, U43, U38, U39, U11, U9, U24, U3, U35

1.1.3 Predictions without Discount Factor

Ratings for unseen items were predicted for each active user using the top 20% closest users:

User	Example Item Predictions (Original)

U35	I2: 1.05, I25: 1.05, I82: 1.29, I67: 1.05
U7	I24: 1.24, I93: 1.27, I89: 1.13, I39: 1.07
U10	I30: 1.54, I39: 1.05, I55: 1.09, I79: 1.07

1.1.4 Discount Factor (β) and Discounted Similarity

- A discount factor β was applied to the similarity scores.
- Discounted similarity reduced the influence of users with lower similarity scores.

1.1.5 Top 20% Closest Users with Discounted Similarity

The updated closest users (after applying β) remained consistent with the original top users. The overlap was **100%** (9/9 users) for all active users.

Active User	Top Discounted Users
U35	U47, U31, U21, U26, U50, U15, U43, U10, U37
U7	U40, U13, U1, U6, U9, U44, U50, U27, U41
U10	U21, U43, U38, U39, U11, U9, U24, U3, U35

1.1.6 Predictions with Discounted Similarity

Predictions were recalculated using the discounted similarity scores. Example results:

User	Example Item Predictions (Discounted)
U35	I2: 1.05, I25: 1.05, I82: 1.29, I67: 1.05
U7	I24: 1.24, I93: 1.27, I89: 1.13, I39: 1.07
U10	I30: 1.54, I39: 1.05, I55: 1.09, I79: 1.07

1.1.7 Comparison of Top Users

The comparison between the top closest users (original vs discounted) showed identical results, with an overlap of **9/9 users** for all active users.

1.1.8 Comparison of Predictions

The predicted ratings (original vs discounted) were compared for each active user. Results were consistent across both methods:

User	Example Item	Original Prediction	Discounted Prediction
U35	I2	1.05	1.05
U7	I24	1.24	1.24
U10	I30	1.54	1.54

The identical predictions demonstrate that applying the discount factor did not alter the outcomes in this case.

Case Study 1.2

1.2.1 User-Based Collaborative Filtering with Cosine Similarity (Mean-Centered)

- Cosine similarity was calculated after applying mean-centering to the user-item matrix.
- Mean-centering adjusted each rating by subtracting the user's average rating to address bias.

1.2.2 Top 20% Closest Users (Mean-Centered)

The top 20% closest users for each active user were determined based on the mean-centered cosine similarity:

Active User	Top Closest Users
U35	U3, U36, U47, U42, U50, U15, U5, U4, U16
U7	U31, U6, U9, U12, U44, U5, U4, U27, U21
U10	U11, U38, U3, U45, U24, U8, U31, U9, U1

1.2.3 Predictions without Discount Factor

Ratings for unseen items were predicted using the top 20% closest users:

User	Example Item Predictions (Original)
U35	I14: 5.00, I3: 5.00, I6: 5.00, I48: 5.00
U7	I10: 5.00, I55: 5.00, I74: 5.00, I75: 5.00

U10	I10: 5.00, I55: 5.00, I39: 4.67, I98: 5.00

1.2.4 Discount Factor (β) and Discounted Similarity

- A discount factor β was applied to reduce the influence of less similar users.
- Discounted similarity scores were recalculated by dividing the original similarity by (1+β)(1 + β).

1.2.5 Top 20% Closest Users with Discounted Similarity

The updated closest users (after applying β) remained identical to the original top users. The overlap was **100%** (9/9 users) for all active users.

Active User	Top Discounted Users
U35	U3, U36, U47, U42, U50, U15, U5, U4, U16
U7	U31, U6, U9, U12, U44, U5, U4, U27, U21
U10	U11, U38, U3, U45, U24, U8, U31, U9, U1

1.2.6 Predictions with Discounted Similarity

Predicted ratings were recalculated using discounted similarity scores. Results remained consistent with the original predictions:

User	Example Item Predictions (Discounted)
U35	I14: 5.00, I3: 5.00, I6: 5.00, I48: 5.00
U7	I10: 5.00, I55: 5.00, I74: 5.00, I75: 5.00
U10	I10: 5.00, I55: 5.00, I39: 4.67, I98: 5.00

1.2.7 Comparison of Top Users

The comparison between the top closest users (original vs discounted) revealed identical results, with an overlap of **9/9 users** for all active users.

1.2.8 Comparison of Predictions

The predicted ratings (original vs discounted) were compared for each active user. Results were identical across both methods:

User	Example Item	Original Prediction	Discounted Prediction
U35	I14	5.00	5.00
U7	I10	5.00	5.00
U10	I55	5.00	5.00

Conclusion

The results demonstrate that applying the discount factor had no effect on the top closest users or predicted ratings. This suggests that the initial cosine similarity scores were robust and the discount factor did not alter the final outcomes.

Case Study 1.3

1.3.1 User-Based Collaborative Filtering with Pearson Correlation Coefficient (PCC)

- Pearson Correlation Coefficient (PCC) was calculated for all users.
- The similarity matrix was created based on shared rated items.

1.3.2 Top 20% Closest Users

The top 20% closest users for each active user were determined based on PCC:

Active User	Top Closest Users
U35	U47, U1, U10, U11, U12, U13, U14, U15, U16
U7	U12, U19, U4, U9, U31, U6, U43, U46, U27
U10	U24, U31, U9, U38, U1, U11, U12, U13, U14

1.3.3 Predictions without Discount Factor

Ratings for unseen items were predicted using the top 20% closest users:

User	Example Item Predictions (Original)
U35	I16: 3.0, I25: 1.0, I45: 3.0, I48: 5.0, I6: 5.0
U7	I10: 5.0, I11: 2.5, I16: 1.0, I18: 4.0, I23: 2.88

U10	I2: 5.0, I23: 4.0, I25: 4.0, I34: 5.0, I57: 4.0

1.3.4 Discount Factor (β) and Discounted Similarity

- A discount factor β=2 was applied to the similarity scores.
- Discounted similarity reduced the weights of lower similarity scores.

1.3.5 Top 20% Closest Users with Discounted Similarity

The updated closest users after applying the discount factor showed minimal changes. The overlap was **100%** (9/9 users):

Active User	Top Discounted Users
U35	U47, U1, U10, U11, U12, U13, U14, U15, U16
U7	U12, U19, U4, U9, U31, U6, U43, U46, U27
U10	U24, U31, U9, U38, U1, U11, U12, U13, U14

1.3.6 Predictions with Discounted Similarity

Predicted ratings were recalculated using discounted similarity scores. Example results:

User	Example Item Predictions (Discounted)
U35	I16: 3.0, I25: 1.0, I45: 3.0, I48: 5.0, I6: 5.0
U7	I10: 5.0, I11: 2.5, I16: 1.0, I18: 4.0, I23: 2.88
U10	I2: 5.0, I23: 4.0, I25: 4.0, I34: 5.0, I57: 4.0

1.3.7 Comparison of Top Users

The comparison between the top closest users (original vs discounted) showed **100% overlap** for all active users.

1.3.8 Comparison of Predictions

The predicted ratings (original vs discounted) were consistent across methods. Example results:

User	Item	Original Prediction	Discounted Prediction
U35	I16	3.00	3.00

U7	I10	5.00	5.00
U10	I2	5.00	5.00

Summary of Results

- **Case Study 1.1 (Cosine Similarity)** and **Case Study 1.3 (PCC)** yielded consistent top users and predictions.
- Applying the discount factor had minimal impact on top users and predictions in both methods.
- Both cosine similarity and PCC methods effectively identified similar users and predicted ratings accurately.

Case Study 2.1

2.1.1 Item-Based Collaborative Filtering Using Cosine Similarity

- Cosine similarity was computed between items without applying bias adjustment (mean-centering).
- The item-item similarity matrix was generated from the user-item matrix.

2.1.2 Top 25% Closest Items

The top 25% closest items for each target item were determined:

Target Item	Top 25% Closest Items
I1	I100, I95, I26, I65, I79, I59, I35, I51, I18, I71, I55, I50, I30, I39, I61, I3, I10, I36, I21, I19, I45, I47, I23, I78
I2	I7, I89, I41, I37, I83, I87, I42, I45, I59, I50, I5, I18, I69, I77, I79, I58, I56, I6, I30, I25, I65, I27, I23, I16

2.1.3 Predictions for Missing Ratings

Predicted ratings for unseen target items were calculated based on top 25% closest items:

Item	Example Predictions
I1	U1: 5.00, U10: 2.40, U11: 2.94, U12: 2.95, U13: 3.89, U14: 2.44, U15: 3.56, U16: 2.37
I2	U1: 3.74, U10: 2.26, U13: 2.00, U14: NaN, U15: 1.00, U17: 2.60, U18: 2.53, U20: 3.79

2.1.4 Discount Factor (β) and Discounted Similarity

- A discount factor **β = 2** was applied to reduce the influence of less similar items.
- Discounted similarity scores were recalculated for the target items.

2.1.5 Top 20% Closest Items (Discounted Similarity)

The top 20% closest items were determined using discounted similarity:

Target Item	Top 20% Discounted Items
I1	I100, I95, I26, I65, I79, I59, I35, I51, I18, I71, I55, I50, I30, I39, I61, I3, I10, I36, I21
I2	I7, I89, I41, I37, I83, I87, I42, I45, I59, I50, I5, I18, I69, I77, I79, I58, I56, I6, I30

2.1.6 Predictions Using Discounted Similarity

Predicted ratings for unseen items were recalculated using discounted similarity scores:

Item	Example Predictions (Discounted)
I1	U1: 5.00, U10: 2.32, U11: 3.32, U12: 3.22, U13: 3.89, U14: 2.44, U15: 3.56, U16: 2.37
I2	U1: 3.00, U10: 1.00, U13: 2.00, U14: NaN, U15: 1.00, U17: 2.60, U18: 2.53, U20: 3.43

2.1.7 Comparison of Top Closest Items (Original vs Discounted)

The overlap between the original top 25% closest items and the top 20% closest items (discounted) was compared:

Target Item	Overlap
I1	19 / 24
I2	19 / 24

2.1.8 Comparison of Predictions (Original vs Discounted)

Predicted ratings for missing values (original vs discounted similarity) were compared:

Item	User	Original Prediction	Discounted Prediction
I1	U10	2.40	2.32
I1	U11	2.94	3.32
I1	U12	2.95	3.22
I1	U15	3.56	3.56
I2	U1	3.74	3.00
I2	U10	2.26	1.00
I2	U20	3.79	3.43
I2	U23	3.45	3.00

The results highlight that discounted similarity influences both the closest items and predictions. Adjusted predictions are typically slightly lower or differ for specific users.

Case Study 2.2

2.2.1 Mean-Centered Cosine Similarity

- Cosine similarity was calculated by mean-centering item ratings to adjust for user bias.
- The resulting item-item similarity matrix was generated.

2.2.2 Top 20% Closest Items

The top 20% closest items for each target item were determined based on mean-centered similarity:

Target Item	Top Closest Items
I1	I26, I78, I58, I100, I35, I7, I85, I11, I93, I52, I64, I43, I40, I28, I65, I79, I33, I89, I39
I2	I45, I42, I48, I36, I83, I97, I7, I21, I77, I57, I16, I79, I87, I8, I65, I93, I30, I20, I35

2.2.3 Predictions Using Mean-Centered Similarity

Predicted ratings for unseen items were computed for users.

User	Predicted Rating for I1	Predicted Rating for I2
U1	2.11	3.57
U10	2.78	2.27
U11	5.00	-
U13	3.76	4.54
U20	1.08	4.38

2.2.4 Discounted Similarity

A discount factor β was applied to reduce the influence of less similar items.

2.2.5 Top 20% Closest Items (Discounted)

The updated closest items using discounted similarity were determined:

Target Item	Top Discounted Items
I1	I26, I78, I58, I100, I35, I7, I85, I11, I93, I52, I64, I43, I40, I28, I65, I79, I33, I89, I39
I2	I45, I42, I48, I36, I83, I97, I7, I21, I77, I57, I16, I79, I87, I8, I65, I93, I30, I20, I35

The overlap for **I1** and **I2** was **100%** (19/19 items).

2.2.6 Predictions Using Discounted Similarity

Predicted ratings for unseen items were recomputed using discounted similarity scores.

User	Discounted Prediction for I1	Discounted Prediction for I2
U1	2.11	3.57
U10	2.78	2.27
U11	5.00	-
U13	3.76	4.54

U20	1.08	4.38

2.2.7 Comparison of Closest Items

The top 20% closest items before and after applying the discount factor showed **100% overlap** for both I1 and I2.

2.2.8 Comparison of Predictions

The predicted ratings remained identical before and after applying discounted similarity for both I1 and I2.

User	I1 Original	I1 Discounted	I2 Original	I2 Discounted
U1	2.11	2.11	3.57	3.57
U10	2.78	2.78	2.27	2.27
U13	3.76	3.76	4.54	4.54
U20	1.08	1.08	4.38	4.38

The identical results confirm that the discount factor did not affect predictions.

Case Study 2.3

2.3.1 Pearson Correlation Coefficient (PCC)

- PCC was calculated to measure the similarity between items using user ratings.
- The resulting item-item similarity matrix was generated.

2.3.2 Top 20% Closest Items

The top 20% closest items for each target item were determined based on the PCC:

Target Item	Top Closest Items
I1	I100, I35
I2	I48, I65, I7, I77, I79, I83, I45, I42, I69, I5

2.3.3 Predictions Using PCC Similarity

Predicted ratings for unseen items were computed for users.

User	Predicted Rating for I1	Predicted Rating for I2
U10	1.00	4.00
U15	4.00	1.00
U20	1.00	5.00
U21	3.00	4.36
U30	5.00	-
U18	-	3.39

2.3.4 Discounted Similarity

A discount factor β was applied to reduce the influence of less similar items.

2.3.5 Top 20% Closest Items (Discounted)

The updated closest items using discounted similarity were determined:

Target Item	Top Discounted Items
I1	I100, I35
I2	I48, I65, I7, I77, I79, I83, I45, I42, I69, I5

The overlap for I1 and I2 was **100%** (19/19 items).

2.3.6 Predictions Using Discounted Similarity

Predicted ratings for unseen items were recomputed using discounted similarity scores.

User	Discounted Prediction for I1	Discounted Prediction for I2
U10	1.00	4.00
U15	4.00	1.00

U20	1.00	5.00
U21	3.00	4.36
U30	5.00	-
U18	-	3.39

2.3.7 Comparison of Closest Items

The top 20% closest items before and after applying the discount factor showed **100% overlap** for both I1 and I2.

2.3.8 Comparison of Predictions

The predicted ratings remained identical before and after applying discounted similarity for both I1 and I2.

User	I1 Original	I1 Discounted	I2 Original	I2 Discounted
U10	1.00	1.00	4.00	4.00
U15	4.00	4.00	1.00	1.00
U20	1.00	1.00	5.00	5.00
U21	3.00	3.00	4.36	4.36
U30	5.00	5.00	-	-

Summary

- The closest items did not change after applying the discount factor.
- The predicted ratings for both I1 and I2 remained identical, indicating that discounting did not influence the final outcomes.

The output reveals key insights across all **three case studies** (2.1, 2.2, and 2.3). Here's a comparison and summary:

Case Study 2.1 (Cosine Similarity without Bias Adjustment)

- **Similar Items**: Similarities relied purely on cosine values.
- **Predicted Ratings**: Reasonable predictions across users, but some inconsistencies appeared for sparsely rated items.
- **Discounted Similarity**: Reduced top similarity values, narrowing the top closest items.
- **Overlap**: Significant overlap between top 25% and discounted 20% closest items.

Case Study 2.2 (Cosine Similarity with Bias Adjustment)

- **Similar Items**: Mean-centering improved similarity values by handling biases in user ratings.
- **Predicted Ratings**: Predictions were smoother and appeared closer to expected values, showing reduced variance.
- **Discounted Similarity**: Similar impact as in 2.1 but with better distribution of similarity weights.
- **Overlap**: Perfect match (100%) between the top 20% original and discounted closest items.

Case Study 2.3 (Pearson Correlation Coefficient)

- **Similar Items**:
    - Some PCC values were extreme (e.g., 1, -1, or 0), especially where ratings were sparse or constant.
    - Limited relationships were observed due to fewer shared users, leading to warnings.

- **Predicted Ratings**:
  - Sparse predictions, with many missing (NaN) values, as PCC requires common ratings.
  - High correlation resulted in accurate predictions for items with sufficient data.
- **Discounted Similarity**:
  - Discounting had minimal effects as similarity values were already restricted.
- **Overlap**: Perfect match (100%) for discounted and original items.

Overall Comparison

Metric	Case 2.1	Case 2.2	Case 2.3
Similarity	Cosine (raw)	Cosine (mean-centered)	Pearson Correlation
Closest Items	High overlap	Better distribution	Sparse with extremes
Predicted Ratings	Reasonable	Smoother predictions	Sparse due to NaNs
Discounted Effect	Reduces impact	More balanced impact	Minimal change observed
Overlap in Closest Items	High overlap	100% overlap	100% overlap

---

Key Takeaways

1. **Bias adjustment** (mean-centering) improves similarity calculation, leading to better predictions.
2. **Cosine similarity** performs well for dense data, especially after bias adjustment.
3. **PCC** is sensitive to sparsity, resulting in NaNs or extreme values. It works best when sufficient overlapping user ratings exist.

4. Discounting similarity values helps refine predictions but shows less impact when initial similarities are sparse or extreme.

Which method works best depends on your dataset. For balanced data:

- **Case 2.2** (Cosine with bias adjustment) seems most reliable. For sparse datasets:
- **Case 2.3** struggles but highlights areas needing more data.

3. Conclusion and opinion

This report analyzed the role of significance weighting in collaborative filtering (CF) methods, focusing on cosine similarity, mean-centered cosine, and Pearson correlation.

Key findings:

- Bias adjustment (mean-centering) improved similarity and prediction quality.
- Cosine similarity performed well with dense data, yielding consistent and accurate recommendations.
- Pearson correlation was sensitive to sparsity, leading to missing or extreme values.
- Applying a discount factor had minimal impact on outcomes, indicating robust similarity measures.

Overall, mean-centered cosine similarity proved most reliable for balanced data, while sparse datasets highlight the limitations of Pearson correlation. Future improvements may focus on addressing sparsity through enhanced data coverage.