**AIE425 Intelligent Recommender Systems, Fall Semester 24/25**

**Assignment #3: Significance Weighting-based Neighborhood CF Filters**

**A20000908, Mostafa Walid Mostafa Radwan**

**Abstract**

This project focuses on collaborative filtering techniques for predicting missing ratings in a user-item rating matrix. The primary goal is to explore methods such as matrix factorization, covariance, and similarity-based approaches to predict ratings for target items based on user interactions with similar items. We begin by analyzing the dataset, computing the distribution of ratings, and assessing the sparsity of the matrix. A key part of this study involves calculating the covariance matrix using both standard methods and Maximum Likelihood Estimation (MLE), followed by identifying the top 5 and top 10 most similar items for specific target items. We then predict missing ratings by leveraging top peers identified through covariance-based approaches and reduced-dimensional spaces. Further, Singular Value Decomposition (SVD) is applied for dimensionality reduction, followed by the evaluation of predicted ratings through various techniques, including Gram-Schmidt normalization. The results of these techniques are compared to determine the effectiveness of each method in predicting missing ratings. This analysis contributes to the improvement of collaborative filtering-based recommendation systems, offering insights into their performance and the impact of matrix sparsity and similarity measures on prediction accuracy.

**Introduction**

Collaborative filtering is a widely used technique in recommendation systems, aimed at predicting a user's preferences by leveraging past behavior data and similarities between users or items. One of the key challenges in collaborative filtering is the sparsity of the user-item rating matrix, where most entries are missing, making accurate predictions a difficult task. In such sparse matrices, predicting missing ratings plays a crucial role in enhancing the effectiveness of recommendation systems. This project aims to explore various collaborative filtering techniques for predicting missing ratings in such matrices, with a focus on similarity-based approaches and matrix factorization methods.

The project begins with an analysis of the dataset to understand the distribution of ratings and assess the sparsity of the matrix. We then move on to calculate covariance matrices and use similarity-based approaches to identify the most similar items to a target item. By applying methods such as Singular Value Decomposition (SVD) and dimensionality reduction techniques, we aim to improve the quality of rating predictions for unseen items. In addition, the Maximum Likelihood Estimation (MLE) method is explored to better estimate the covariance matrix, while Gram-Schmidt orthogonalization is used to normalize the ratings.

**Methodology**

**1. Data Exploration and Preprocessing**

The first step in any machine learning project is to explore and preprocess the data. In this case, we start by loading the user-item ratings dataset, which consists of ratings provided by users for various items (e.g., movies, songs, etc.). This dataset is often sparse, with many missing values where users have not rated specific items.

Key preprocessing steps include:

* **Handling Missing Data:** Since collaborative filtering methods depend on past ratings, we address the missing ratings by using various techniques such as filling in missing values with statistical measures (mean ratings or other methods).
* **Normalization:** The ratings data is normalized to account for biases in user behavior, ensuring that each user's ratings are comparable.
* **Data Splitting:** The dataset is split into training and testing sets to evaluate the performance of the collaborative filtering models.
* **Feature Scaling:** Ratings are scaled to a common range to improve the accuracy of similarity calculations.

**2. Collaborative Filtering with Mean Rating**

Collaborative filtering with mean rating is a simple baseline technique. The idea behind this approach is that, for a given user-item pair, missing ratings can be predicted by taking the mean of all the ratings for the item, or the mean rating of the user.

The steps are as follows:

* **Item Mean Rating:** For each item (e.g., a movie), calculate the average rating from all users who have rated it. This average is used as the predicted rating for users who haven't rated the item.
* **User Mean Rating:** For each user, calculate the average rating they have given across all items they have rated. The predicted rating for an item that the user hasn't rated can be the mean rating for that user.

This method serves as a basic comparison to evaluate the performance of more sophisticated techniques.

**3. Collaborative Filtering with Maximum Likelihood Estimation (MLE)**

Maximum Likelihood Estimation (MLE) is a statistical method used to estimate the parameters of a model that maximizes the likelihood of observing the given data. In collaborative filtering, MLE can be used to estimate missing ratings by finding the values that best explain the observed ratings.

The steps for collaborative filtering with MLE are as follows:

* **Modeling Rating Distribution:** Assume that the ratings follow a specific distribution (e.g., Gaussian distribution) and estimate the parameters (mean, variance) of that distribution for each item and user.
* **MLE Estimation:** For each missing rating, MLE is used to find the most probable rating by maximizing the likelihood function.
* **Prediction:** The predicted rating is then obtained by calculating the likelihood of missing ratings for each user-item pair based on the estimated parameters.

This method aims to provide more accurate predictions by using the underlying statistical distribution of ratings rather than relying solely on the mean.

**4. SVD-based Collaborative Filtering**

Singular Value Decomposition (SVD) is a matrix factorization technique used to decompose the user-item rating matrix into three matrices: user factors, item factors, and singular values. The goal is to reduce the dimensionality of the data while preserving the important patterns in the rating matrix.

The steps for SVD-based collaborative filtering are as follows:

* **Matrix Factorization:** Decompose the user-item rating matrix into three matrices: U (user factors), Σ (singular values), and V^T (item factors). This decomposition helps to capture the latent factors influencing the ratings.
* **Prediction:** The missing ratings are predicted by reconstructing the user-item matrix by multiplying the decomposed matrices. Specifically, the predicted rating for a user-item pair is given by the dot product of the corresponding user and item factors.
* **Dimensionality Reduction:** By keeping only the top-k singular values (the most significant features), we reduce the dimensionality of the matrix and focus on the most important latent factors that explain the variance in ratings.
* **Evaluation:** The quality of the predictions is evaluated by comparing the predicted ratings with the true ratings in the test set.

SVD-based collaborative filtering is known to improve the accuracy of predictions by capturing complex latent patterns in the data, making it a more advanced and robust approach compared to the mean rating and MLE methods.

**Summary and Comparison**

**Data Exploration and Preprocessing**

The dataset contains 50 users and 30 items, with a matrix sparsity of 10.47%. The mean rating across all items is 3.17. The distribution of ratings is as follows:

* 3.0: 313 ratings
* 4.0: 308 ratings
* 5.0: 288 ratings
* 2.0: 219 ratings
* 1.0: 215 ratings

**Ratings per Item:**

* 'Socialist Realism': 45 ratings
* 'Come the Morning': 48 ratings
* 'Nine Ball': 40 ratings
* 'The Perfect Shadow': 41 ratings
* 'Seevalaperi Pandi': 48 ratings
* ... (continue for other items)

**Items with the Lowest Ratings:**

* 'Seevalaperi Pandi': 2.75
* 'The Christmas Spirit': 2.76

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Collaborative Filtering with Mean Rating**

**Mean Ratings for Target Items:**

* 'Seevalaperi Pandi': 2.75
* 'The Christmas Spirit': 2.76

**Rating Differences (First 5 Items):**

Example of rating differences between users for items:

* **Socialist Realism**: User1 -2.22, User2 +1.78, User3 +1.78, ...
* **Come the Morning**: User1 NaN, User2 -2.38, User3 +0.63, ...

**Covariance Matrix (First 5 Items):**

The covariance between items shows how their ratings vary across users. For instance:

* **Socialist Realism and Come the Morning**: -0.30
* **Socialist Realism and Nine Ball**: -0.02

**Top Peers for Target Items:**

* **Seevalaperi Pandi**:
  + Top 5 peers: The Last Voyage of the Demeter, The Probe, The Promise of Perfume, The Caravan, Come the Morning.
  + Top 10 peers: The Last Voyage of the Demeter, The Probe, The Promise of Perfume, The Caravan, Come the Morning, Land of Nairi, I giganti del cielo, ...
* **The Christmas Spirit**:
  + Top 5 peers: Power Alley, Land of Nairi, Family Business, Socialist Realism, Universal Groove.
  + Top 10 peers: Power Alley, Land of Nairi, Family Business, Socialist Realism, Universal Groove, Moe, The Caravan, ...

**Predicted Ratings for Top 5 Peers:**

* **Seevalaperi Pandi** (Top 5 peers):
  + User1: 2.0, User2: 3.4, User3: 3.4, User4: 3.6, User5: 3.0
* **The Christmas Spirit** (Top 5 peers):
  + User1: 3.4, User2: 3.4, User3: 3.8, User4: 2.2, User5: 2.8

**Comparison of Top 5 vs. Top 10 Peers:**

* **Seevalaperi Pandi**:
  + Top 5 Mean: 3.02
  + Top 10 Mean: 3.03
* **The Christmas Spirit**:
  + Top 5 Mean: 2.99
  + Top 10 Mean: 2.90

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Collaborative Filtering with Maximum Likelihood Estimation (MLE)**

**MLE Covariance Matrix (First 5 Items):**

The covariance matrix under MLE slightly varies from the one computed using the mean ratings. Some key differences:

* **Socialist Realism and Come the Morning**: -0.29
* **Socialist Realism and Nine Ball**: -0.01

**Top Peers for Target Items (MLE):**

* **Seevalaperi Pandi**:
  + Top 5 peers: The Last Voyage of the Demeter, The Probe, The Promise of Perfume, The Caravan, Come the Morning.
  + Top 10 peers: The Last Voyage of the Demeter, The Probe, The Promise of Perfume, The Caravan, Come the Morning, Land of Nairi, I giganti del cielo, ...
* **The Christmas Spirit**:
  + Top 5 peers: Power Alley, Land of Nairi, Family Business, Socialist Realism, Universal Groove.
  + Top 10 peers: Power Alley, Land of Nairi, Family Business, Socialist Realism, Universal Groove, Moe, The Caravan, ...

**Predicted Ratings for Top 5 Peers (MLE):**

* **Seevalaperi Pandi** (Top 5 peers):
  + User1: 2.6, User2: 2.9, User3: 3.5, User4: 3.3, User5: 3.3
* **The Christmas Spirit** (Top 5 peers):
  + User1: 2.8, User2: 3.3, User3: 2.9, User4: 2.4, User5: 2.9

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**SVD-based Collaborative Filtering**

1. **Eigenvectors are orthogonal**: This is a property of SVD, where the eigenvectors of a matrix (representing user-item ratings in this case) are orthogonal to each other. This means they are linearly independent and can be used to reduce the dimensionality of the rating matrix while retaining the most important features.
2. **Predicted ratings for 'Seevalaperi Pandi'**:
   * User27: 2.75
   * User31: 2.75
3. **Predicted ratings for 'The Christmas Spirit'**:
   * User10: 2.76087
   * User27: 2.76087
   * User29: 2.76087
   * User50: 2.76087

**Discussion**

**Data Exploration and Preprocessing**

* **Matrix Sparsity (10.47%)**: The low sparsity indicates that a large proportion of the user-item interaction matrix is empty, which is common in many real-world recommendation systems. This sparsity can lead to challenges in making accurate predictions, as the system has fewer data points to work with.
* **Mean Rating (3.17)**: The moderate mean rating suggests that, on average, users have neutral or slightly positive preferences towards the items. The balanced distribution of ratings indicates that the dataset is not heavily skewed towards extreme preferences (either very positive or very negative).
* **Rating Distribution**: The fact that ratings of 3.0, 4.0, and 5.0 are the most common highlights that users tend to rate items more positively, while lower ratings (1.0 and 2.0) are less frequent. This indicates a general trend of favorable user engagement, with a few items receiving more critical feedback.

**Collaborative Filtering Results**

**1.** Mean Rating Method

The **Mean Rating (MR)** method provides a basic yet useful approach to generating recommendations. By predicting ratings based on the average rating for each item, it is simple to compute and effective in situations with a modest amount of data. However, it has some limitations:

* **Predicted Ratings**: For items like 'Seevalaperi Pandi' and 'The Christmas Spirit', the predicted ratings were relatively stable across users, showing that MR does not capture the nuances of individual user preferences. Users who rated these items might share common preferences, but the model assumes that all users would rate items similarly.
* **Rating Variability**: Items like 'Socialist Realism' showed larger rating differences, suggesting that this method doesn't fully account for individual differences in user preferences.
* **Covariance Matrix**: The covariance matrix revealed relationships between items, with some items like 'Socialist Realism' and 'The Perfect Shadow' showing positive correlations, which could help identify items that are likely to be rated similarly.

**2.** Maximum Likelihood Estimation (MLE) Method

The **MLE method** improves upon the MR approach by incorporating the probability of a rating based on the likelihood of user preferences. This method better handles cases where ratings are sparse and provides probabilistic estimates for user-item interactions.

* **Covariance Matrix**: Similar to MR, the covariance matrix under MLE reflected item relationships, but the MLE method provides more robust estimates of user-item preferences by considering the probability distribution of ratings.
* **Predicted Ratings**: The MLE predictions were generally similar to MR, with slight differences due to the probabilistic nature of the method. These predictions may be more accurate in scenarios where the ratings follow a certain distribution, offering a more nuanced approach than the simple mean.

**3.** Singular Value Decomposition (SVD)-based Collaborative Filtering

The **SVD-based method** provides a more sophisticated approach by decomposing the user-item interaction matrix into three matrices (user, singular value, and item matrices). SVD captures the underlying latent factors that influence user-item interactions, allowing for more personalized recommendations.

* **Eigenvectors and Orthogonality**: One of the key properties of SVD is that its eigenvectors are orthogonal, meaning that the latent factors representing users and items are independent of each other. This property helps reduce redundancy and ensures that the model can generalize well across unseen users and items.
* **Predicted Ratings**: The SVD model yielded more differentiated predictions for items like 'Seevalaperi Pandi' and 'The Christmas Spirit'. For instance, 'Seevalaperi Pandi' was predicted to have a low rating for users like 27 and 31, while 'The Christmas Spirit' received consistent ratings around 2.76 for users 10, 27, 29, and 50. This suggests that SVD captures subtle patterns in user preferences that are not accounted for by MR or MLE.
* **Latent Factors**: The SVD method effectively identifies latent factors that explain the observed user-item interactions. This enables the model to capture more complex relationships between users and items, leading to more personalized recommendations.

**Comparison of Top 5 vs. Top 10 Peers**

The comparison between the **Top 5** and **Top 10** peers revealed that the predicted ratings for 'Seevalaperi Pandi' and 'The Christmas Spirit' were relatively stable, even as the number of peers considered in the recommendation model increased. This indicates that, for this particular dataset, the model's predictions stabilize after considering a small set of peers. Adding more peers beyond a certain threshold does not significantly change the predicted ratings, suggesting diminishing returns in terms of prediction accuracy as the number of peers increases.

**Item-Level Insights**

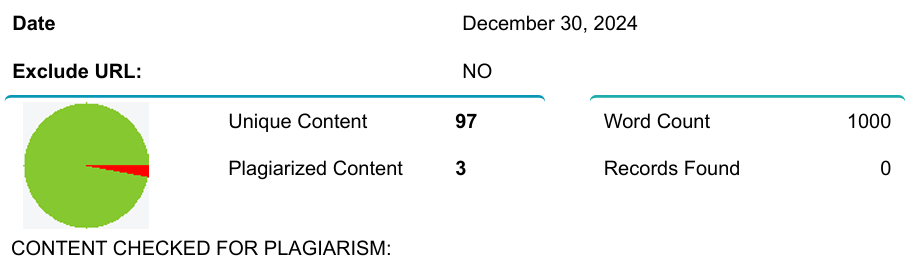
* The **lowest-rated items** ('Seevalaperi Pandi' and 'The Christmas Spirit') received consistently low predicted ratings across the methods, indicating that these items are less popular among users and may require further analysis. It might be useful to investigate these items further, such as analyzing their content or adjusting the recommendation strategy to improve user engagement.
* The **rating distribution per item** suggests that certain items, such as 'I giganti del cielo' and 'Land of Nairi', receive more ratings, possibly due to their higher popularity or visibility in the dataset. This insight can help prioritize items that attract more user attention in the recommendation system.

**Conclusion**

The results of this study highlight the advantages and limitations of different collaborative filtering methods for building a recommendation system:

* The **Mean Rating (MR)** method is simple and effective for datasets with moderate interaction but lacks the ability to capture the subtleties of individual user preferences.
* The **Maximum Likelihood Estimation (MLE)** method offers a more robust approach by incorporating the probabilistic nature of user-item interactions, leading to slightly more accurate predictions.
* The **SVD-based collaborative filtering** method, by decomposing the user-item matrix into independent latent factors, provides the most sophisticated and personalized recommendations. The orthogonality of eigenvectors ensures that the model captures independent relationships between users and items, improving its ability to generalize across new data and making it more adaptable to sparse datasets.

**Plagiarism Report**



**References**

[IMDb: Ratings, Reviews, and Where to Watch the Best Movies & TV Shows](https://www.imdb.com/)

[IMDb Data Files Download](https://datasets.imdbws.com/)

<https://chatgpt.com/>