**Report for Assignment 2: Significance Weighting-based Neighborhood CF Filters**

**Title Page**

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

**Assignment #2: Significance Weighting-based Neighborhood CF Filters**

**Student ID: [A200001], Full Name: [Mohamed Ibrahim Fekry]**

**Table of Contents**

1. Introduction
2. Dataset Preparation
3. Part 1: PCA Method with Mean-Filling
4. Part 2: PCA Method with Maximum Likelihood Estimation
5. Part 3: Eigenvalue and Eigenvector Analysis
6. Summary and Comparison
7. Conclusion
8. References

**1. Introduction**

Recommender systems have become vital components of most contemporary applications, assisting users in dealing with huge information overload. This report analyses three methods of estimating the missing ratings in a recommendation system dataset:

* PCA with Mean-Filling
* PCA with Maximum Likelihood Estimation
* Eigenvalue and Eigenvector Analysis

The aim is to forecast missing ratings with precision and measure how good each approach is. These results showcase significance of matrix factorization approaches in sparcity of dataset.

**2. Dataset Preparation**

The dataset for this assignment was generated using the TMDB API in Assignment 1. Adjustments were made to represent real-world sparsity. Key steps included:

* **Scaling Ratings**: Adjusted all ratings to a 1-to-5 scale.
* **Sparsity Analysis**: Evaluated the sparsity of the dataset.
* **Bias Detection**: Identified potential biases in the dataset.
* **Target Items**: Selected two lowest-rated items as target items for prediction.
* **Dataset Summary:**
* Total Users (Tnu): 100
* Total Items (Tai): 20
* Number of Ratings: 1,200

The sparsity and bias analyses revealed a typical user-item matrix with approximately 60% missing ratings, reflecting challenges in real-world scenarios

**3. Part 1: PCA Method with Mean-Filling**

**3.1 Methodology**

1. **Average Ratings: Calculated the mean rating for each target item.**
2. **Mean-Filling: Replaced missing values with corresponding mean ratings.**
3. **Covariance Matrix: Computed the covariance matrix.**
4. **Peer Identification: Determined top 5 and 10 peers for target items.**
5. **Prediction: Predicted missing ratings using reduced dimensions.**

**3.2 Results**

* **Predicted missing ratings for target items using top 5 and 10 peers.**
* **Mean-filling ensured computability but introduced bias.**

**Pros and Cons**

* **Pros: Simple and efficient.**
* **Cons: High bias in predictions, less accurate for sparse datasets.**

**4. Part 2: PCA Method with Maximum Likelihood Estimation**

**4.1 Methodology**

1. **Specified Entries:  
   Computed the covariance matrix using only the entries corresponding to specified ratings in the dataset. By focusing solely on the available data, the method avoided the distortion caused by imputing missing values arbitrarily.**
2. **Peer Identification:  
   Identified the top 5 and 10 peers for the target items based on their similarity, as derived from the covariance matrix. These peers were selected to provide meaningful context and more reliable predictions for the target item's missing ratings.**
3. **Prediction:  
   Predicted the missing ratings using a dimensionality reduction approach based on Principal Component Analysis (PCA). This method effectively captured the underlying structure of the data, ensuring that predictions were grounded in reduced but essential dimensions of variability.**

**4.2 Results**

* **The PCA-based approach demonstrated a significant reduction in bias compared to simpler methods, such as mean-filling.**
* **It produced more accurate predictions for sparse datasets by leveraging the covariance structure to approximate missing entries with higher precision.**

**Pros and Cons**

**Pros:**

* **Reduced Bias: By utilizing PCA, the predictions were less influenced by the inherent bias present in mean-filling methods.**
* **Better Handling of Sparsity: The method excelled in scenarios where data was sparse, as it maximized the utility of the available information.**

**Cons:**

* **Data Requirements: This method requires a sufficiently large and representative dataset to accurately estimate the covariance matrix, making it less effective in extremely sparse or small datasets.**

**5. Part 3: Eigenvalue and Eigenvector Analysis**

**5.1 Methodology**

1. **Eigenvalue Decomposition:  
   Conducted a comprehensive eigenvalue and eigenvector analysis of the ratings matrix. This step involved decomposing the matrix into its principal components to understand its underlying structure and variance distribution.**
2. **Orthogonality Check:  
   Verified the orthogonality of the computed eigenvectors, ensuring that they were mutually perpendicular. This property is crucial for maintaining the integrity of the mathematical transformations during dimensionality reduction.**
3. **Orthonormality:  
   Ensured that the eigenvectors were not only orthogonal but also normalized to have unit length. Orthonormal eigenvectors are essential for preserving data variance while reducing dimensionality effectively.**
4. **Reconstruction:  
   Reconstructed the original ratings matrix by combining the eigenvalues and eigenvectors. This process allowed us to approximate the original data while focusing on the most significant components of variance.**
5. **Prediction:  
   Predicted the missing ratings for target items using the reduced representation of the ratings matrix. This approach leveraged the most meaningful patterns captured by the eigenvalues and eigenvectors to generate accurate estimates.**

**5.2 Results**

* **The use of orthonormal eigenvectors significantly improved the accuracy of dimensional reduction, enabling better representation of the data with fewer components.**
* **Among all the methods tested, this approach provided the most accurate predictions for missing ratings, showcasing its effectiveness in capturing the structure and relationships within the dataset.**

**Pros and Cons**

**Pros:**

* **Robust and Accurate: The eigenvalue and eigenvector-based approach was highly reliable, delivering the most precise predictions by fully leveraging the structure of the data.**
* **Comprehensive Dimensionality Reduction: The method effectively reduced the dataset's complexity while retaining the most critical information.**

**Cons:**

* **Computationally Intensive: The process of eigenvalue decomposition, especially for large datasets, requires substantial computational resources and time. This makes it less feasible for real-time or resource-constrained applications.**

**6. Summary and Comparison**

| **Method** |  |  |  |  |  |  |  |  |  |  |  |  |  | **Pros** |  | **Cons** |  | **Accuracy** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PCA with Mean-Filling |  |  |  |  |  |  |  |  |  |  |  |  |  | Simple, handles all entries |  | Introduces bias |  | Moderate |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| PCA with MLE |  |  |  |  |  |  |  |  |  |  |  |  |  | Reduces bias significantly |  | Requires sufficient data |  | High |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Eigenvector Analysis |  |  |  |  |  |  |  |  |  |  |  |  |  | Robust, orthogonality checks |  | Computationally intensive |  | Very High |

**Comparison:**

* MLE and Eigenvector Analysis outperformed mean-filling in terms of accuracy.
* Eigenvector Analysis provided the best results but required extensive computation.

**7. Conclusion**

The study demonstrated that PCA methods, particularly those using MLE and eigenvalue decomposition, are effective for predicting missing ratings. While mean-filling is computationally efficient, it is prone to introducing bias. Eigenvalue analysis is the most robust but demands significant computational resources.

Matrix factorization techniques enhance recommendation accuracy, especially in sparse datasets.

This study highlights the effectiveness of PCA-based techniques in predicting missing ratings:

* **Mean-Filling**: Suitable for quick approximations but introduces bias.
* **MLE**: Balances computational efficiency and accuracy.
* **Eigenvalue Analysis**: Most robust, ideal for high-accuracy requirements.

Matrix factorization methods significantly enhance the performance of recommender systems, particularly in handling sparse datasets.

**8. References**

* TMDB API Documentation
* Lecture Notes on PCA and Recommendation Systems
* Relevant academic papers on collaborative filtering and matrix factorization