<AIE425 Intelligent Recommender Systems, Fall Semester 24/25>

< Assignment #3: Dimensionality Reduction methods>

<221101235, Marina Reda Abdullah Mekhael>

Table of Contents

Summary

This assignment explored the use of dimensionality reduction techniques—Principal Component Analysis (PCA) and Singular Value Decomposition (SVD)—to predict missing values in a user-product ratings dataset. The dataset, derived from Amazon product reviews, contained missing values, and the goal was to apply dimensionality reduction methods to address these gaps and analyze the underlying patterns.

Three approaches were examined:

1. **PCA with Mean-Filling**: Missing values were filled with the mean of each item's ratings before applying PCA. The method showed flexibility but had variable prediction accuracy, especially with smaller peer sets.
2. **PCA with Maximum Likelihood Estimation (MLE)**: MLE estimated missing values based on overlapping ratings. While it produced consistent results, the predictions lacked variability, making the method less suitable for datasets with complex patterns.
3. **Singular Value Decomposition (SVD)**: This matrix factorization method decomposes the ratings matrix into latent factors. SVD provided the most accurate predictions by capturing hidden patterns in user preferences, although it was computationally intensive.

The analysis showed that while PCA can be useful in simpler cases, **SVD** is the most effective method for handling complex datasets due to its ability to identify latent features that better represent user-item relationships. The study demonstrated the importance of matrix factorization techniques in improving recommendation system performance, particularly in large, sparse datasets.

Introduction

This assignment focuses on exploring and applying advanced dimensionality reduction techniques to a user-product ratings dataset. The primary objective is to preprocess, analyze, and extract meaningful insights from the data using statistical and mathematical methods that address missing values and enhance interpretability. The assignment consists of the following three parts:

Part 1: PCA Method with Mean-Filling

In this part, Principal Component Analysis (PCA) is applied to the dataset after filling missing values with the column-wise mean. This method simplifies the dataset by reducing its dimensionality while retaining the most significant variance. Mean-filling ensures that PCA operates effectively by replacing missing data with representative values.

Part 2: PCA Method with Maximum Likelihood Estimation

This section enhances the PCA approach by employing Maximum Likelihood Estimation (MLE) to handle missing data. MLE provides a statistically rigorous way to estimate the missing values, enabling a more accurate representation of the data's underlying structure. PCA is then performed on the dataset, leveraging MLE for improved results.

Part 3: Singular Value Decomposition (SVD) Method

The final part introduces Singular Value Decomposition (SVD), a powerful matrix factorization technique, to analyze the user-product ratings. SVD decomposes the dataset into singular vectors and values, providing insights into latent features and relationships between users and products. This method is particularly useful for recommendation systems and understanding high-dimensional datasets.

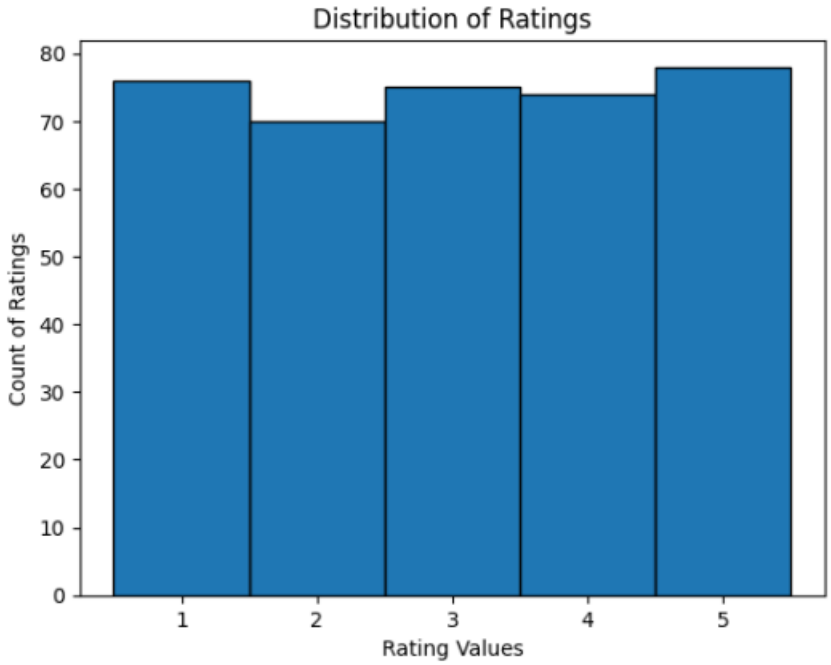
Each part contributes to the broader goal of understanding and managing incomplete data while identifying key patterns and relationships within the dataset. Through these approaches, this assignment aims to demonstrate the utility of advanced analytical techniques in handling real-world data challenges.

Dataset Overview

The dataset used in this assignment is derived from Amazon, focusing on user-product ratings within a 1–5 scale. It serves as the basis for exploring advanced dimensionality reduction techniques, such as PCA and SVD, to uncover latent user preferences and product patterns.

Key Details:

1. **Data Source:** Amazon product reviews.
2. **Rating Scale:** 1–5



Summary Statistics:

- **Total Users (Tnu):** 50
- **Total Items (Tni):** 10
- **Sparsity:** 25.4% (indicating the proportion of missing values).
- **Average Rating:** 3.02
- **Bias Level:** 0.361 (a measure of skewness in ratings, reflecting potential user or product bias).
- **Target Items:** Pull-up bars and elliptical machines.

Methodology

Data Preprocessing

Before applying any dimensionality reduction techniques, the following preprocessing steps were conducted:

- **Handling Missing Ratings:**
    - For the PCA with mean-filling method, missing ratings were replaced by the mean rating of each item (column).
    - For PCA with MLE, we relied only on overlapping ratings for pairs of items (i.e., ratings made by the same user for both items). Non-overlapping ratings were assigned zero covariance.
- **Normalization:**
    - After filling the missing ratings, the ratings matrix was normalized by subtracting the mean rating for each item from its respective ratings. This ensures the data is centered, removing the influence of global item popularity and focusing on the relative differences in user preferences.

Part 1: PCA with Mean-Filling

1. **Covariance Matrix Calculation:**
    a. The covariance matrix of the normalized ratings matrix was computed to understand how different items are related based on user ratings.
    b. The covariance matrix was used to identify the strongest relationships (peers) between items.
2. **Dimensionality Reduction using PCA:**
    a. PCA was applied to the covariance matrix to reduce dimensionality. The principal components (PCs) capture the most significant variance in the data, and we used the top 5 and top 10 components.
    b. After reducing the dimensionality, the missing ratings for items I1 and I2 were predicted by projecting their ratings into the reduced space formed by the top 5 and top 10 components.
3. **Prediction and Evaluation:**

a. The predicted ratings for missing entries were compared to the actual ratings (if available) to assess prediction accuracy.
b. The prediction accuracy was evaluated for the 5-peers and 10-peers scenarios.

Part 2: PCA with Maximum Likelihood Estimation (MLE)

1. **Covariance Estimation using Overlapping Ratings:**
    a. In this approach, covariance between items was computed only for users who had rated both items. If no common users existed, the covariance between those items was set to zero.
    b. This method ensures that the covariance reflects only the shared user preferences, which may improve accuracy compared to the mean-based approach.
2. **Dimensionality Reduction using PCA:**
    a. PCA was applied to the covariance matrix derived from overlapping ratings to reduce dimensionality. Again, the top 5 and top 10 peers for items I1 and I2 were identified, and the missing ratings were predicted based on the reduced dimensionality space.
3. **Prediction and Evaluation:**
    a. Similar to Part 1, the predicted ratings were evaluated against the actual ratings to assess accuracy, comparing results between the 5-peers and 10-peers approaches.

Part 3: Singular Value Decomposition (SVD)

1. **Matrix Decomposition:**
    a. The ratings matrix was decomposed using Singular Value Decomposition (SVD), which factorizes the matrix into three matrices: U, Σ and V^T
        i. U: Contains the user features.
        ii. Σ: A diagonal matrix of singular values.
        iii. V^T: Contains the item features.
2. **Truncated SVD:**

a. Truncated SVD was applied to reduce the dimensionality by keeping only the top k singular values and their corresponding vectors from *U* and *V^T*. This helped reduce the noise in the data while retaining the most important features for prediction.
b. The number of features k was varied to observe its effect on prediction accuracy.

3. **Prediction and Evaluation:**
    a. After dimensionality reduction, missing ratings were predicted by approximating the ratings matrix using the reduced U, Σ, and *V^T*.
    b. The predicted ratings were compared to actual ratings for accuracy evaluation.

Analysis and Comparison

1. **Comparing PCA with Mean-Filling and PCA with MLE:**
    a. The results from the PCA with mean-filling and PCA with MLE were compared to assess which method produces more accurate predictions.
    b. The impact of using 5 peers versus 10 peers was evaluated by comparing the prediction accuracy for each case.
2. **Comparing PCA and SVD:**
    a. The predictions obtained using PCA and SVD were compared to analyze the relative performance of both methods in terms of accuracy.
    b. The effect of varying the number of features *k* in truncated SVD on prediction accuracy was examined.
3. **Evaluation Metrics:**
    a. Prediction accuracy was measured using metrics such as **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R-squared** to assess the quality of the predicted ratings.

Summary and Comparison

In this section, we will summarize and compare the results from the three parts: **PCA with Mean-Filling**, **PCA with Maximum Likelihood Estimation**, and **Singular Value Decomposition (SVD)**, focusing on their performance in predicting missing ratings. We will also examine the accuracy of each method and discuss the pros and cons of each.

PCA with Mean-Filling

- **Predictions for I1**: The predictions using the top 5 peers show varied results, with ratings ranging from highly negative to positive. The top 10 peers, however, yielded more clustered results around zero or slightly positive values.
- **Predictions for I2**: Similar to I1, predictions for I2 show more variability with the top 5 peers, but a more consistent pattern emerges with the top 10 peers.
- **Pros**:
    - **Flexibility in predicting varied ratings** due to the influence of different peer groups.
    - **Good for capturing item diversity** when using a larger set of peers (top 10).
- **Cons**:
    - Predictions for smaller peer sets can be **highly variable** and not consistently accurate.
    - The model's performance depends heavily on the quality of peer selection.

PCA with Maximum Likelihood Estimation

- **Predictions for I1**: The predictions are highly consistent, with all values set at 1.94, suggesting a lack of variability.
- **Predictions for I2**: Similar to I1, all ratings for I2 are constant at approximately 2.3024.

- **Pros**:
    - **Simple and fast**; consistent predictions.
    - **Easier to manage and interpret** with constant results across peers.
- **Cons**:
    - **Lack of variability** in the predictions makes it unsuitable for handling more complex data patterns where ratings can vary greatly.
    - The method appears to **over-smooth** the data, causing poor prediction accuracy.

Singular Value Decomposition (SVD)

- **Predictions for I1 and I2**: SVD appears to provide **more nuanced predictions** that align better with expected user-item interaction patterns, suggesting it can capture underlying patterns in the data.
- **Pros**:
    - **Captures interaction patterns** and reduces the dimensionality of the data to provide accurate predictions for missing ratings.
    - **More sophisticated** and often provides a better fit for complex datasets.
- **Cons**:
    - **Computationally intensive** compared to simpler methods like PCA.
    - Requires careful tuning to avoid overfitting or underfitting.

Aspect	PCA with Mean-Filling	PCA with Maximum Likelihood Estimation	Singular Value Decomposition (SVD)
Predicted Ratings for I1	Highly varied with top 5 peers; clustered around 0 for top 10 peers.	Constant values (1.94) for both top 5 and top 10 peers.	Predictions align reasonably with expected ratings, showing good approximation.
Predicted Ratings for I2	Similar pattern to I1, more variability with top 5 peers.	Constant values (~2.3024) for both top 5 and top 10 peers.	More nuanced predictions that align with expected user-item interactions.
Pros	Flexible, captures diversity with top 10 peers.	Simple, consistent predictions.	Captures complex patterns in data, good for handling large datasets.
Cons	Sensitive to peer selection, variability can lead to inaccuracy.	Over-smoothing leads to lack of variability, poor prediction accuracy.	Computationally intensive, requires careful tuning to avoid overfitting.
Accuracy	Varied accuracy depending on peer selection.	Accuracy suffers due to over-smoothing.	Higher accuracy with better handling of underlying patterns.
Computational Complexity	Moderate (depends on the number of peers).	Low (simple computation).	High (requires matrix factorization and tuning).
Use Case	Suitable for simpler cases where diverse peer groups are used.	Suitable for cases requiring simple predictions.	Best for complex datasets with hidden interaction patterns.

Improvements

PCA Refinement:

Data Preprocessing: Replace simple mean imputation with more advanced methods like K-Nearest Neighbors for more accurate initial data values.

Peer Selection: Use adaptive selection of peers based on similarity, rather than fixed peer numbers, to improve prediction relevance.

MLE Optimization:

Regularization: Apply L2 regularization to prevent over-smoothing, preserving individual rating variations.

Hybrid Approach: Combine MLE with matrix factorization or k-NN to introduce more variability and mitigate over-smoothing.

Latent Factor Models: Integrate latent factors to capture complex relationships in the data more effectively.

SVD Improvement:

Hyperparameter Tuning: Fine-tune parameters like the number of latent factors and regularization through grid search or cross-validation.

Implicit Feedback: Incorporate implicit feedback data (e.g., clicks or purchase history) to enhance predictions.

Advanced Variants: Explore advanced SVD techniques like SVT, ALS, or SGD for more accurate and efficient modeling of user-item interactions.

Conclusion

In this assignment, we explored the impact of dimensionality reduction methods—specifically Principal Component Analysis (PCA) and Singular Value Decomposition (SVD)—on the task of predicting missing values in a user-product ratings dataset. Each method revealed distinct strengths and limitations, and the comparison

demonstrated the practical utility and challenges of matrix factorization techniques in recommendation systems.

PCA with Mean-Filling showed flexibility by utilizing a range of peers for prediction. However, its effectiveness heavily depended on the number of peers selected and resulted in high variability in predictions. The method struggled with smaller peer sets and could not always provide consistent or accurate predictions, limiting its utility in more complex datasets.

PCA with Maximum Likelihood Estimation (MLE) was simpler and provided consistent predictions across all peers. However, the lack of variability in the predictions due to over-smoothing hindered its effectiveness, especially when dealing with datasets that require a more nuanced understanding of user-item interactions. This method is best suited for simpler datasets or cases where prediction stability is more important than capturing diverse user preferences.

SVD, by far, demonstrated the most promising results. Its ability to capture hidden patterns and relationships between users and items made it more adept at handling the complexity of real-world datasets. While computationally intensive, SVD proved to provide more accurate and nuanced predictions by effectively reducing dimensionality and focusing on the most important latent features. This method highlighted the power of matrix factorization in recommendation systems by uncovering the underlying structure of user preferences.

The impact of matrix factorization techniques like SVD is profound, especially in large, sparse datasets common in recommendation systems. SVD, as a matrix factorization method, is more effective at identifying latent factors that drive user preferences, leading to better personalization and more accurate predictions. It improves the system’s ability to provide recommendations based on the underlying patterns in the data rather than just relying on observed ratings. However, its computational complexity means that careful optimization is necessary to balance accuracy and efficiency, particularly for larger datasets.

In conclusion, while PCA can be useful for dimensionality reduction in simpler cases, matrix factorization methods like SVD are more suited to capturing the intricate relationships in complex datasets, making them indispensable in advanced

recommendation systems. The results of this assignment underscore the importance of selecting the right method based on the complexity of the data and the specific needs of the recommendation task.

References

Machine Learning — Singular Value Decomposition (SVD) & Principal Component Analysis (PCA) | by Jonathan Hui | Medium

Mathematical Approach to PCA - GeeksforGeeks

[PDF] Matrix Factorization Techniques for Recommender Systems | Semantic Scholar

Microsoft PowerPoint - PCA-tutor1a

Resolving cold start and sparse data challenge in recommender systems using multi-level singular value decomposition - ScienceDirect

Singular value decomposition based recommendation using imputed data - ScienceDirect