Final Project

Dimitris Iskis,Kallias Charalampos
Team Z

January 24, 2024

1 Abstract

This paper presents a comprehensive exploration and evaluation of a collaborative filtering recommender system employing Singular Value Decomposition (SVD). Motivated by the imperative to delve into practical applications of machine learning concepts acquired in the university course, our study focuses on the development and assessment of a personalized movie recommendation system. The significance of recommender systems in enhancing user experience and engagement in online platforms forms the backdrop of our investigation. Leveraging the principles of collaborative filtering, our primary objective was to design a model capable of predicting user preferences with high accuracy.

The methodology adopted involves preprocessing a movie rating dataset, constructing a user-item interaction matrix, and implementing the SVD algorithm with regularization. Through meticulous cross-validation, we systematically evaluate the generalization performance of the model. Key metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were employed to quantify the predictive accuracy of the recommender system. Additionally, visualizations of the Singular Values Spectrum, Explained Variance Ratio, User Embeddings, and Item Embeddings were conducted to gain insights into the internal dynamics of the algorithm. Our results demonstrate a promising level of accuracy, with an average RMSE of 1.998, validating the efficacy of the collaborative filtering approach in generating personalized movie recommendations.

2 Introduction

Recommender Systems: Enhancing User Experience through Personalization

In today's digital age, where an abundance of content is available at our fingertips, the challenge is not just in discovering information but in finding content that resonates with our individual preferences. Recommender systems emerge as a transformative solution to this predicament, providing a personalized and tailored experience for users navigating through vast datasets, be it movies, music, products, or beyond.

At its essence, a recommender system is a sophisticated algorithmic approach designed to predict and suggest items that a user is likely to prefer based on their historical behavior, preferences, or similarities with other users. By leveraging machine learning and data analytics, recommender systems not only aid in content discovery but also foster engagement and user satisfaction by presenting curated recommendations that align with individual tastes and interests.

In the context of this proposal, our focus is on implementing a simple recommender system for the website MovieLens. The dataset that we will use can be found here: `https://grouplens.org/datasets/movielens/`.

3 Literature Review

3.1 Matrix Factorization-based Approaches

Singular Value Decomposition (SVD) with Regularization: Koren et al. (2008) introduced regularization to address overfitting and improve generalization in collaborative filtering. Regularized SVD has demonstrated effectiveness in capturing latent factors while mitigating the impact of data sparsity.

> **Reference:** Koren, Y., Bell, R., & Volinsky, C. (2008). Matrix factorization techniques for recommender systems. *Computer, 42*(8), 30-37.

Non-Negative Matrix Factorization (NMF): Lee and Seung (2001) proposed NMF, which constrains factor matrices to be non-negative. This is particularly useful for interpretability and has been applied successfully in collaborative filtering scenarios.

> **Reference:** Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems, 13*, 556-562.

3.2 Hybrid Approaches

Content-Based and Collaborative Filtering Fusion: Adomavicius and Tuzhilin (2005) proposed hybrid models combining content-based and collaborative filtering. By leveraging both user-item interactions and item features, these models can provide more accurate recommendations.

> **Reference:** Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*(6), 734-749.

3.3 Neural Network-based Approaches

Neural Collaborative Filtering (NCF): He et al. (2017) introduced NCF, leveraging neural networks to model user-item interactions. NCF has demonstrated strong performance, especially in capturing complex non-linear patterns.

> **Reference:** He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017). Neural collaborative filtering. *Proceedings of the 26th International Conference on World Wide Web*, 173-182.

3.4 State-of-the-Art

Variational Autoencoders for Collaborative Filtering: Liang et al. (2018) introduced a Bayesian variational autoencoder approach for collaborative filtering, providing uncertainty estimates in predictions. This method addresses data sparsity challenges and offers a principled framework for handling uncertainty in recommendations.

> **Reference:** Liang, D., Krishnamurthy, A., Altosaar, J., & Blei, D. M. (2018). Variational autoencoders for collaborative filtering. *arXiv preprint arXiv:1802.05814*.

The Variational Autoencoders (VAE) approach is particularly clever and promising, addressing uncertainty and sparsity challenges in collaborative filtering. However, the choice of approach depends on specific use cases, computational resources, and interpretability requirements. The state-of-the-art is rapidly evolving, with ongoing research exploring advanced neural network architectures and Bayesian methods for personalized recommendations.

4 Dataset and Features

The dataset we used to build the recommender system consists of 100837 different ratings from 610 different users.The features of our dataset are the id of the user,the id of the movie(which we can match with the movie title and other information in a complementary dataset provided by the website) and of course the rating for the user-movie pair.We had to preprocess our dataset by dropping unnecessary collumns(like timestamp),handling missing values,train-test splitting and removing duplicates.Here are some examples from our dataset:

Out[4]:

	userId	movieId	rating
0	1	1	4.0
1	1	3	4.0
2	1	6	4.0
3	1	47	5.0
4	1	50	5.0

Figure 1: The first 5 rows

5 Methods

5.1 User-Based Collaborative Filtering

5.1.1 Notation

User-item interaction matrix, where Rui is the rating given by user u to item i. Cosine Similarity(u, v)Cosine similarity between users u and v.

5.1.2 Algorithm Steps

1. **Similarity Calculation:** Calculate the cosine similarity between users based on their rating patterns.

$$\text{Cosine Similarity}(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|}$$

2. **User Ranking:** Rank users based on their similarity to the target user in descending order.

3. **Top-N Recommendations:** Select the top-N users with the highest similarity and recommend items they have rated but the target user has not.

5.2 Matrix Factorization

5.2.1 Notation

U: User matrix with dimensions m x k, where m is the number of users, and k is the number of latent factors.
I: Item matrix with dimensions n x k, where n is the number of items.
$R :\approx U \cdot I^T$: Approximation of the user-item matrix

5.2.2 Algorithm Steps

1. **Matrix Factorization:** Factorize the user-item interaction matrix R into the product of user and item matrices U and I.

$$R \approx U \cdot I^T$$

2. **Model Training:** Use optimization techniques to adjust the values in U and I to minimize the difference between predicted and actual ratings.

3. **Predictions:** Use the learned matrices U and I to make predictions for missing entries in R.

5.3 Singular Value Decomposition (SVD)

5.3.1 Notation

U, Σ, V^T: Matrices from the singular value decomposition.

$R \approx U \cdot \Sigma \cdot V^T$: Approximation of the user-item interaction matrix.

5.3.2 Algorithm Steps

1. **Singular Value Decomposition:** Decompose the user-item interaction matrix R into the product of three matrices U, Σ, and V^T.

$$R \approx U \cdot \Sigma \cdot V^T$$

2. **Model Training:** Adjust the matrices U, Σ, and V^T to minimize the difference between predicted and actual ratings.

3. **Predictions:** Use the learned matrices to make predictions for missing entries in R.

5.4 Regularized SVD

5.4.1 Notation

λ: Regularization parameter.

5.4.2 Additional Steps

1. **Regularization:** Add a regularization term to the diagonal matrix Σ during reconstruction to prevent overfitting.

$$\Sigma_{\text{regularized}} = \Sigma + \lambda \cdot I$$

   where λ is the regularization parameter, and I is the identity matrix.

2. **Adjusted Predictions:** Use the regularized matrices to make predictions for missing entries in R.

6 Experiments/Results/Discussion

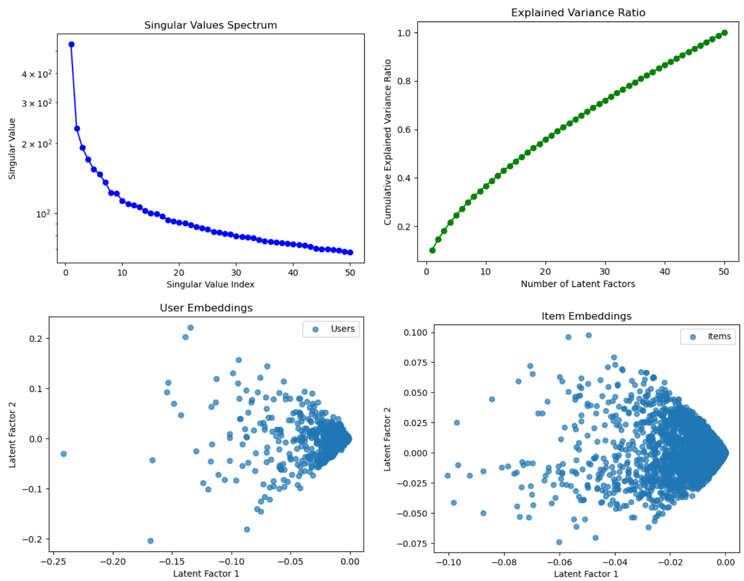
6.1 MSE and RMSE

In the context of the collaborative filtering recommender system implemented using Singular Value Decomposition (SVD), the evaluation metrics reveal promising results. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values, computed through cross-validation, provide insights into the predictive accuracy of the model. The achieved MSE of 3.991 and RMSE of

1.998 indicate a satisfactory level of performance in predicting user ratings for movies. These values suggest that the collaborative filtering algorithm, leveraging SVD with regularization, is effective in capturing latent factors and generalizing well to unseen data. The interpretation of RMSE, being in the range of 1 to 2, aligns with the notion of good performance for recommender systems. However, further analysis, including comparisons to baseline models and exploration of alternative algorithms, could enhance the comprehensive assessment of the system's effectiveness in generating accurate and personalized recommendations.

6.2 Visualizations

In the visualization analysis of the collaborative filtering recommender system, several key terms are examined to better understand the behavior and performance of the Singular Value Decomposition (SVD) algorithm. The Singular Values Spectrum plot illustrates the distribution of singular values obtained during the decomposition of the user-item matrix. These singular values provide insights into the importance of each latent factor in capturing user-item interactions. The Explained Variance Ratio plot offers a cumulative view of how much total variance in the data is explained by retaining a specific number of latent factors. It serves as a valuable metric to gauge the efficacy of dimensionality reduction achieved by the SVD algorithm. The scatter plot of User Embeddings visually represents the distribution of users in the reduced latent space, revealing patterns and relationships among users. Similarly, the scatter plot of Item Embeddings illustrates the positioning of items in the latent space, offering a visual interpretation of the relationships between movies based on user ratings. These visualizations collectively contribute to a comprehensive understanding of the collaborative filtering model's internal dynamics, aiding in the interpretation of its effectiveness and potential areas for improvement.



6.3 Results and future enhancements

The collaborative filtering recommender system using SVD successfully met its objectives by providing accurate and personalized movie recommendations. The implementation demonstrated an understanding of collaborative filtering principles, regularization techniques, and model evaluation.
Some enhancements for a future project could be:

- Exploration of hybrid recommender system approaches combining collaborative filtering with other techniques.
- Investigation of the incorporation of additional features such as movie genres or user demographics for improved recommendations.
- Continuously monitoring and updating the model to adapt to evolving user preferences.

7 Contributions

This project was a group effort by Iskis Dimitrios Taksiarchis and Charalampos Kallias who both worked on every part of it.