Modeling and Evaluation Report for Movie Recommendation System.

Introduction

This report details the development and evaluation of a movie recommendation system. The goal is to provide personalized movie recommendations to users based on their past ratings. We employ a collaborative filtering approach using Singular Value Decomposition (SVD).

Modeling

Data Preparation:

> **Data Source**: The system utilizes two datasets: 'ratings.csv' containing user-movie ratings and 'movies.csv' providing movie titles and genres.
>
> **Preprocessing**: The ratings data is split into training (80%) and testing sets (20%). We create a user-movie matrix from the training data, filling missing values with zero, indicating unrated movies.

Algorithm:

SVD: Singular Value Decomposition, a matrix factorization technique, is used to decompose the user-movie matrix. It captures latent factors underlying the interactions between users and movies.

Hyperparameter Tuning: The key hyperparameter, n_components (number of latent factors), is tuned for optimal performance. The range explored is [20, 50, 100, 200].

Training: The model is trained on the user-movie matrix from the training dataset.

Implementation:

Prediction: For a given user, the system predicts ratings for movies not rated by the user. It calculates the dot product of the user's latent vector and the movie's latent vector.

Recommendation: The top 5 movies with the highest predicted ratings, which the user has not already rated, are recommended.

Evaluation

Metrics

Root Mean Square Error (RMSE):
RMSE is used to measure the accuracy of predicted ratings against actual ratings in the test set. Lower RMSE indicates better accuracy.

Evaluation Process

RMSE Calculation: For each user in the test set, we predict ratings for movies that are common between the training and test sets. The RMSE for these predictions is calculated and averaged over all users.

Results: The average RMSE across all users in the test set provides a measure of the model's predictive accuracy.

Criteria

Model Selection: The model with the lowest RMSE in the hyperparameter tuning phase is selected as the best model.

Performance: The model's performance is deemed satisfactory if it achieves a low RMSE, indicating accurate predictions of user ratings.

Results

Best Model: The SVD model with n_components = 200 achieved the lowest RMSE.

Test Performance: On the test dataset, the model achieved an average RMSE of 1.070. This indicates a medium level of accuracy.

Future work

To further improve the performance and capabilities of the movie recommender system, several strategies can be considered. The key areas for enhancement include model refinement, data quality and diversity, and incorporating additional features.

Model Refinement

Advanced Algorithms: Beyond Singular Value Decomposition (SVD), explore more sophisticated algorithms like Non-negative Matrix Factorization (NMF), Deep Learning approaches, or Ensemble methods which might capture complex patterns more effectively.
Hyperparameter Optimization: Conduct more extensive hyperparameter tuning. Utilize techniques like Random Search or Bayesian Optimization for a more thorough exploration of the hyperparameter space.
Cross-Validation: Implement k-fold cross-validation for model evaluation instead of a single train-test split. This will provide a more robust assessment of the model's performance.
Handling Cold Start Problem: For new users or items with limited data, explore hybrid approaches that combine content-based and collaborative filtering, or utilize demographic/user information.

Data Quality and Diversity

Richer Dataset: Incorporate additional data sources, such as user demographics, temporal effects (e.g., time-based trends), and contextual information to enrich the recommendations.
Data Preprocessing: Investigate advanced preprocessing techniques like handling outliers, data normalization, or feature engineering to enhance the quality of input data.
Feedback Loop: Implement a system to capture real-time feedback from users on recommended items to refine the model iteratively.

Incorporating Additional Features

Content-Based Features: Integrate content-based features like movie genres, director, cast, and user profiles to provide more personalized recommendations.
Sentiment Analysis: Utilize sentiment analysis on movie reviews to gauge user preferences more accurately.