PROJECT: DEEP LEARNING BASED MOVIE RECOMMENDER SYSTEM

Recommender system plays a pivotal role in the digital era, assisting users in discovering relevant items or content from vast datasets. As the volume of available information continues to grow, personalized recommendation engines have become integral for enhancing user's experience. Traditional recommendation methods include collaborative filtering and content-based filtering, each with its strength and limitations.

The motivation for employing deep learning techniques, particularly autoencoders, for recommenders' system stems from the need to address certain challenges associated with conventional approaches:

- Traditional methods struggle to capture intricate patterns and dependencies within user-item interactions. Deep learning models, by design, can automatically learn intricate hierarchical representations, enabling the extraction of nuanced user preferences and item characteristics.
- Recommender systems often face challenges when dealing with new users or items lacking sufficient historical data. Autoencoders, being unsupervised models, can learn latent representations even with limited data, mitigating the cold start problem to some extent.
- Deep learning models, including autoencoders, can scale effectively to large datasets and adapt to dynamic user behaviour. Their ability to continuously learn and update representations makes them suitable for evolving recommendation scenarios.

The primary objective of this project is to develop an advanced movie recommender system that leverages deep learning techniques, specifically autoencoders, to provide personalized movie recommendations to users. The system aims to enhance user satisfaction by offering accurate and relevant movie suggestions based on their preferences.

Dataset:

Link: https://files.grouplens.org/datasets/movielens/ml-1m.zip

In this project, two primary datasets are employed to build and evaluate the recommender system: the rating dataset and the movie dataset. These datasets are part of the MovieLens dataset, specifically MovieLens 1M dataset. The dataset contains 1000209 ratings from 6040 unique for 3706 unique movies.

The datasets has following columns:

- **UserID:** Unique identifier for each user.
- **MovieID:** Unique identifier for each movie.
- **Rating:** User-assigned rating for the movie on a predefined scale.
- **Timestamp:** Time at which the rating was recorded.
- **Title:** Title of the movie.
- **Genres:** categorical information about the movie's genre(s).

Data Preprocessing:

- Checked for missing values and duplicate entries in both rating dataset and movie dataset.
- The two datasets, rating and movie datasets, were integrated to create a comprehensive dataset, termed 'meta_df'

- The genre column was one-hot encoded to facilitate genre-based analysis.
- Basic statistics such as the number of unique movies, users and distribution of the ratings.
- Filtered movies only with ratings greater than 25 from the dataset.
- Extracted the movie release year from the title.
- Standardized the rating to bring them to a common scale.

Exploratory Data Analysis (EDA):

- By visualizing the distribution of number of rating per users we can see majority of users seem to have provided a relatively small number of ratings, indicating that many users might not rate a large number of movies.
- The distribution highlights the sparsity of the rating dataset. Not all users rate all movies, resulting in sparse matrix.
- In the distribution of number of rating per movie, most movies have a relatively low number of ratings, as indicated by the left side of the histogram.
- Most of the movies are rated 4, it suggests that users generally rate movies positively.
- In the average rating per movie diagram, there is a peak around 3.1, it suggests that many movies have an average rating around 3.1.
- 'Old Lady Who Walked in the Sea, The (Vieille qui marchait dans la mer, La) (1991)' is the highest rated movie and Pie in the Sky (1995) is the lowest rated movie.
- The distribution of average rating per user indicates, most users have an average rating between 3.5 and 4. This peaks at higher average indicate many users are generally satisfied with the movies they watch.
- By analysing the number of genres, most of the movies belongs to Comedy, Drama and Action.

Model Architecture:

An autoencoder is a neural network architecture designed for unsupervised learning, where the model aims to reconstruct its input data. In the context of recommender systems, an autoencoder can be leveraged to learn low-dimensional representations of users and items, capturing latent features that influences user preferences. The autoencoder architecture consist of:

1. **Input layer:** The input layer corresponds to the number of unique movies in the dataset, which is the size of the feature space for each user.
2. **Encoding Layers:** The encoding layer consist of densely connected layers that progressively reduce the dimensionality of the input data, leading to creation of bottle neck layer.
3. **Bottleneck Layer:** The bottleneck layer is the layer with the smallest number of neurons and serves as the condensed representation of the input information. It captures the most important information features of the data.
4. **Decoding Layer:** It mirrors the encoding layers in reverse order, gradually reconstructing the input data. Each decoding layer increased the dimensionality of the data until the final layer matches the size of the original input.
5. **Output Layer:** The output layer aims to reconstruct the original input data and its size matches the number of unique movies.

The autoencoder algorithm consists of three hidden layers for encoding and three hidden layers for decoding. ReLU (Rectified Linear Unit) activation function is used for the encoding and decoding layers due to its ability to handle non-linearities. The autoencoder is trained for 100 epochs, indicating that the entire training dataset is passed through the neural network 100 times.

Training and Evaluation:

The data is split into training and test sets. The training set comprises 80% of the data, while test set comprises the remaining 20%. This split helps evaluate the performance of the autoencoder on unseen data. The model is trained using Adam optimizer and Mean Squared Error (MSE) is employed as the loss function.

Evaluation:

The evaluation is performed using Mean Squared Error (MSE) on the test set. MSE is the average squared difference between the predicted value and the actual values. It provides a measure of how well an autoencoder model is able to reproduce the target variable. It expressed the average magnitude of error between predicted and actual values. Lower MSE value indicate better performance.

The Mean Squared Error (MSE) value of approximately 0.0557 on the test set indicates the average squared difference between the predicted movie ratings generated by the autoencoder and the actual ratings in the test set. It is relatively low, indicating that the autoencoder is performing well in reconstructing movie ratings.

Recommendation System:

The user embeddings represent compressed, low-dimensional representations of users' preferences. A separate model is built from the trained autoencoder, specifically its encoder portion. By feeding the user-item rating matrix through this model, we obtain embeddings that capture the essential features characterizing each user's movie preferences.

To find users with similar preferences, a Nearest Neighbors model is employed. Using cosine similarity as the distance metric, this model identifies users whose embeddings are close in the learned latent space. During training, the model learns relationships between users based on their embeddings. In the recommendation phase, it efficiently locates users in the test set whose preferences align closely with those in the training set, facilitating accurate personalized movie recommendations.

Finally, a recommendation function has been created to provide personalized movie suggestion for a target user.

```
Recommendations for User 12 are:
American Beauty (1999)
Star Wars: Episode IV - A New Hope (1977)
Raiders of the Lost Ark (1981)
Shawshank Redemption, The (1994)
Schindler's List (1993)
```