Name: Jithin Jose
B Number: B00815334
Introduction to Data Mining: CS535-01

Project Report

1. Algorithm:

Used original dataset as training set. Then, generated test set which includes unrated items. Split the training and test sets. Implemented six baseline models. Model 1 predicted all ratings as a constant in range (1,5). Model 2 assigned global average of rating matrix to all items. Model 3 calculates average user rating. Model 4 calculates average item rating. Model 5 calculates the average of ratings obtained from Model 3 and Model 4. Model 6 uses LightGBM with features we got from Model 3, 4 and 5 and predicts the final rating.
A function **impute_rating(rating,pred)** then replaces the ratings of the unrated items with the predicted ratings and leaves the already existing ratings as the same.
Another function **get_rmse(y_true, y_pred)** calculates rmse for each model.

2. Cold-Start Issue:

Initially, cold-start issue comes up while generating model 2 ratings, which is solved by assigning all ratings as the global average of rating across items.
The unrated items in model 3 and 4 get assigned with the global average obtained from Model 2. This solves the cold-start issue.

3. Matrix Completion Applications:

Document-term matrix

A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. There are various schemes for determining the value that each entry in the matrix should take. One such scheme is tf-idf. They are useful in the field of natural language processing.

Computer Vision:

Matrix completion has a lot of applications in computer vision. It is used for image processing applications. One example is Depth Enhancement via Low-Rank Matrix Completion. Depth captured by consumer RGB-D cameras is often noisy and misses values at some pixels, especially around object boundaries. A depth map enhancement algorithm performs depth map completion and de-noising simultaneously.
The dimensions are x and y coordinates of pixel value. The value of the matrix gives the intensity of that pixel.

4. Existing Techniques:

Collaborative Filtering:

Collaborative filtering uses a database of user preferences to predict additional topics or products a new user might like. Collaborative Filtering provides strong predictive power for recommender systems and requires the least information at the same time. However, it has a few limitations in some particular situations. Collaborative Filtering is faced with cold start. When a new item comes in, until it has to be rated by substantial number of users, the model is not able to make any personalized recommendations. Similarly, for items from the tail that didn't get too much data, the model tends to give less weight on them and have popularity bias by recommending more popular items.

J. Breese, D.. Heckerman, and C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, *Proc. Conf. Uncertainty in Artificial Intelligence*, (UAI98) 1998

Content-Based Filtering:

Content-based filtering methods are based on a description of the item and a profile of the user's preferences. These methods are best suited to situations where there is known data on an item (name, location, description, etc.), but not on the user. Content-based recommenders treat recommendation as a user-specific classification problem and learn a classifier for the user's likes and dislikes based on an item's features.
A key issue with content-based filtering is whether the system is able to learn user preferences from users' actions regarding one content source and use them across other content types. When the system is limited to recommending content of the same type as the user is already using, the value from the recommendation system is significantly less than when other content types from other services can be recommended. For example, recommending news articles based on browsing of news is useful, but would be much more useful when music, videos, products, discussions etc. from different services can be recommended based on news browsing

Collaborative Filtering vs. Content-Based Filtering: differences and similarities by Rafael Glauber and Angelo Loula

Using Content-Based Filtering for Recommendation by Robin van Meteren and Maarten van Someren

Hybrid Recommendation System:

Most recommender systems now use a hybrid approach, combining collaborative filtering, content-based filtering, and other approaches. Hybrid approaches can be implemented in several ways: by making content-based and collaborative-based predictions separately and then combining them; by adding content-based capabilities to a collaborative-based approach or by unifying the approaches into one model. Several studies that empirically compare the performance of the hybrid with the pure collaborative and content-based methods and demonstrated that the hybrid methods can provide more accurate recommendations than pure approaches. These methods can also be used to overcome some of the common problems in recommender systems such as cold start and the sparsity problem, as well as the knowledge engineering bottleneck in knowledge-based approaches.
Netflix is a good example of the use of hybrid recommender systems. The website makes recommendations by comparing the watching and searching habits of similar users (i.e., collaborative filtering) as well as by offering movies that share characteristics with films that a user has rated highly (content-based filtering).

A Hybrid Recommendation Method Based on Feature for Offline Book Personalization

Hybrid Recommender Systems: A Systematic Literature Review Erion C and Maurizio Morisio

5. Issues:

Scalability:

It measures the ability of the system to work effectively with high performance while growing in information. Recommender system need to recommend items to users without any changes while the number of users increased, or the number of items increased too.
A solution is to use a hybrid user model by combining the item demographic information with searching for a set of neighboring users having the same interests and using a genetic algorithm to determine the weight features in the user model.

Accuracy of rating matrix:

A paper proposes a new method to generate an accurate rating matrix completion than the ordinal matrix by constructing the two-class structure of binary matrix factorization.