Final Project-Proposal

Michalis Baltsiotis 2578, Nikolaos-Gerasimos Zazatis 2723, team D

November 2022

1 Introduction

In recent years, ML has a wide variety of practical applications. Popular websites and apps, such as youtube, facebook, amazon, netflix, etc, use ML to find and recommend an item to the interested user that is the most relevant to his preferences. That is called Recommendation System and it can be used for a wide range of cases, such as book recommendation, music recommendation, food recommendation, merchandise recommendation, healthcare recommendation, dating recommendation, scholarly paper recommendation and many more!

2 Data for RS

To derive a good recommendation, RS need to have some personalized information for the user. So a minimum amount of knowledge stored for some user is the items that the user has liked, reviewed and/or interacted with. Then, for every user-item pair there will be an explicit rating, where the user has rated the item withing a given range, or implicit, where the user has interacted with an item without rating it. Additional information for a given dataset can leverage the RS performance, such as the genres of a movie for a movie recommendation system, the location of the user for a food recommendation system, etc.

3 How do they work?

Generally, Recommendation Systems can use different methods to implement their functionality. The most common methods can be categorised to Collaborative Filtering and Content-Based filtering. For all these methods, an important structure is the Rating Matrix, which entries are the ratings of the users on the items. Collaborative Filtering can be further be categorised to Neighborhood Methods and Latent Factor Models. For Collaborative Filtering method, the RS finds for the user the top k similar users and averaging their rating on an item (User based) or for the item the top k similar items and averaging their rating on the user (Item based) to predict its rating. Latent Factor Models factorises the rating matrix to two new matrices with dimension $n*k$ and $m*k$, where n is the number of users, m is the number of items and k is the latent factors, which is a hyperparameter. These methods fall sort on accuracy when the ammound of data for a user is limited. That is known as cold start problem and Content-Based filtering can solve this by recommending based on additional features such as genres, timestamps, seasons etc.

4 Current Project

For this project we will implement a simple Music Recommendation System using this kaggle dataset: https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset. It contains 3 different csv files, one for books information, one for user information and one for the rating between a user and a book. The rating can be either explicit, between the range 1-10, or implicit, using 0. We have performed some basic EDA on this dataset at the notebook attached the current file. We will explore different algorithms, using simple ML models, deep learning models and hybrid models to compare their performance and computational cost. Furthermore, we are aiming to also implement a recommender that is not biased towards the most popular books, which may help some not so popular books or authors to be appreciated.