Machine Learning for Recommender Systems

1. Recommender systems are a hot topic. Recommendation systems can be formulated as a task of matrix completion in machine learning. Recommender systems aim to predict the rating that a user will give for an item (e.g., a restaurant, a movie, a product).

2. Download the movie rating dataset from: https://www.kaggle.com/rounakbanik/themovies-dataset. These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

3. Building a small recommender system with the matrix data: "ratings.csv". You can use the recommender system library: Surprise (http://surpriselib.com), use other recommender system libraries, or implement from scratches.

a. Read data from "ratings.csv" with line format: 'userID movieID rating timestamp'.

b. MAE and RMSE are two famous metrics for evaluating the performances of a recommender system. The definition of MAE can be found via: https://en.wikipedia.org/wiki/Mean_absolute_error. The definition of RMSE can be found via: https://en.wikipedia.org/wiki/Root-mean-square_deviation.

c. Compute the average MAE and RMSE of the Probabilistic Matrix Factorization (PMF), User based Collaborative Filtering, Item based Collaborative Filtering, under the 5-folds cross-validation

GITHUB LINK: https://github.com/CamachoBry/CAP5610-ML/blob/main/Homeworks/HW5/HW5.ipynb

d. Compare the average (mean) performances of User-based collaborative filtering, item-based collaborative filtering, PMF with respect to RMSE and MAE. Which ML model is the best in the movie rating data?
The PMF had the lowest RMSE and MAE so the PMF would be the best in the movie rating data.

e. Examine how the cosine, MSD (Mean Squared Difference), and Pearson similarities impact the performances of User based Collaborative Filtering and Item based Collaborative Filtering. Plot your results. Is the impact of the three metrics on User based Collaborative Filtering consistent with the impact of the three metrics on Item based Collaborative Filtering?

The graphs are not fine-grain on the y-label but from an overview, it does not seem to cause a significant impact in this case.

f. Examine how the number of neighbors impacts the performances of User based Collaborative Filtering and Item based Collaborative Filtering? Plot your results.

g. Identify the best number of neighbor (denoted by K) for User/Item based collaborative filtering in terms of RMSE. Is the best K of User based collaborative filtering the same with the best K of Item based collaborative filtering?

For User-based, the lowest RMSE score stemmed from 45 neighbors and for Item-based, the lowest score was from 50 neighbors. They are not the same but slightly similar.

Additional Questions:
    • Approximately how many hours did you spend on this assignment?
        About 2 hours.
    • Which aspects of this assignment did you find most challenging? Were there any significant stumbling blocks?
The most challenging part for me was the graphing which I always struggle with to get the graph I would like to see.
    • Which aspects of this assignment did you like? Is there anything you would have changed?
I actually liked the graphing which allowed me to get the habit of looping through various parameters and graphing to easily see experimental results.

GITHUB LINK: https://github.com/CamachoBry/CAP5610-ML/blob/main/Homeworks/HW5/HW5.ipynb

    Please submit a PDF report. In your report, please answer each question with your explanations, plots, results in brief. DO NOT paste your code or snapshot into the PDF. At the end of your PDF, please include a website address (e.g., Github, Dropbox, OneDrive, GoogleDrive) that can allow the TA to read your code.