Movie Recommendation

Zahra Sinaei

March 2021

Goals

This project has two goals:

- Movie recommendation system
- Movie rating prediction

The details of this project are available at
https://github.com/zs680/Movie-Recommendation

- Recommender systems have many applications, for instance:
  1. Targeted advertising on various website
  2. Product recommenders for online stores
  3. Content recommenders for social media platforms

Dataset

This project uses the dataset FullMovieLens dataset from Kaggle. The dataset contains

- Metadata of 45k movies released on or before July 2017, such as budget, revenue, overrview, posters, languages, genre, production company, TMBD vote counts and vote averages.
- Movie credits containing cast and crew information for all 45k movies.
- 26 million ratings from 270k users for all 45k movies.

Movie recommendation system: collaborative filtering

- Developed a collaborative filtering for rating dataset.
- Trained two recommender models using
  1. a memory based approach (KNN with means)
  2. a model based approach (Matrix factorization SVD)
- Performed hyperparameter tuning for these models.
- Obtained a winning recommender model with .87 accuracy using KNN with means.

Movie recommendation system: content-based and hybrid filtering

- Developed a content-based recommender to leverage textual data from movie overviews using NLP techniques.
- Trained a word2vec model for the overviews corpus of 45k movies.
- Extracted features by enhancing word2vec embeddings of words in movie reviews using normalized TF-IDF values as weights .
- Built a recommender model based on the cosine similarity matrix of feature vectors.
- Built a hybrid recommender algorithm by combining the above collaborative and content-based recommender models.

Movie recommendation system: future goals

- Neural Collaborative Filtering (NCF)
- Restricted Boltzmann Machines (RBM)
- Bidirectional Encoder Representations from Transformers (BERT)
- Incorporate metric to evaluate and compare various models. For instance Precision-Recall or ROC Curve.
- Use other features rather than only movie overviews to construct and train a content based recommendation model.

Movie rating prediction

- Built TMBD movie vote averages using metadata and credits data of all 45K movies.
- Performed data cleaning on metadata and credits data.
- Used the following features for regression:
  vote count, runtime, revenue, genre, popularity, budget, original language, adult, director and top 3 casts.
- Developed and trained KNN, SVR, Random Forest, and AdaBoost, and ran hyperparameter tuning using gridsearch.
- Built a final predictor, by applying a stacking model using these models as a base, which obtained results with MSE 0.39.

As a continuation I hope to improve this model to get a better MSE.