Building a Movie Recommendation System

William Guo

Overview

- Recommendation algorithms are essentially foundational pieces of any online platform that serves any kind of content nowadays whether it's personalized or non-personalized
- Social media platforms push videos or photos that the algorithm deems a user might like in addition to serving advertisements that it deems relevant to the user
- Streaming services suggest movies and TV shows based on a user's viewing habits or previous content the user has consumed
- E-commerce platforms recommend products based on their recommendation algorithms
- If you are online at all, you are subject to one recommendation algorithm or another

Understanding the problem

Diversity in preferences

Consumers of films have different preferences. How do we account for these and recommend the correct content for each user?

Different algorithms

There are many different algorithmic approaches to recommendation systems, collaborative filtering, content-based filtering or a hybrid of the two. Another method is non-personalized recommendations.

Data sparsity/cold start

How to build a strong recommendation system without already having vasts amount of data?

Project objective:

Build a well-performing movie recommendation model with relatively sparse data

Data & Methodology

- MovieLens dataset from GroupLens containing 100,000 user ratings on movies
- High number of ratings but only from 610 unique users on 9742 movies
- Three modeling strategies:
    - Collaborative filtering
    - Content-based filtering
    - A hybrid approach using both of the above

Models

Collaborative Filtering	• Using the Surprise library to implement SVD • FCP: 0.68, MAE: 0.65, RMSE: 0.85
Content-based Filtering	• Using processed genre and tags as content feature matrix for cosine similarity between movies
Hybrid Model	• Assigns scores and weights to both collaborative filtering and content-based recommendations to create final recommendation

Findings

- Collaborative filtering is quite effective for movie recommendations if we have concrete ratings from users. The model was able to achieve a FCP of 0.68 which means my model is getting pairwise item ranking preferences for each user right about 68% of the time.
- Content-based filtering is a more difficult approach. Using just movie genres and tags as the content feature matrix does not seem to produce a very strong model.
- A hybrid model is theoretically probably the most robust but also the most difficult to execute well.

Limitations & Future Work

- Data sparsity is one of the limitations of this project. 100,000 entries sounds like a lot but there are only 610 users in the dataset.
- Time: Running the same algorithms on a larger dataset locally will take time and the limited timeframe for this project restricts how much I was able to do this time.
- Future work includes refining and optimizing the algorithms, particularly the content-based filtering and hybrid models in addition to using the larger 1M entries dataset.