Project identification

Summary:

Here I'm going to cover the steps that the coding goes through to achieve optimal anime referrals. We start by loading in our rating data from Kaggle. We formant and summarize the data to be easier to read for our algorithm. Then we imported the anime names file. This matches each anime rating and stats with the proper anime name instead of ID number. We then remove unused data like producers and Japanese names.

The first thing to sort the data is by using a binary classification for each anime to see what genre it is. Then we can take the users anime they have seen and see what genres they have watched. Then a very simple algorithm will recommend the number one rated anime for each genre they like.

Now we sort the user data in a 70-30 split so we can train on the bulk and use the last 30 percent to check. We have two different search algorithms using Naive Bays with a Bernoulli distribution and Naive Bays with a Gaussian distribution. Once we have trained both programs, we can calculate the errors for both programs. In the Bernoulli distribution the overall error was low but it was very rarely 100% accurate where the Gaussian distribution has a high error but was 100% accurate about a quarter of the time. So, Bernoulli is the best system for our recommender system.

With the recommender system trained we can now input our own knowledge of anime and see what it recommends for us.

Motivation:

Our motivation is that we like anime. Anime is a good source of entertainment and Japanese culture. The data sets are also widely available and are massive, compiling hundreds of thousands of reviews. This is good for a recommender as more data will lead to a better recommender.

Data Source:

https://www.kaggle.com/datasets/hernan4444/anime-recommendation-database-2020

This dataset contains information about 17,562 anime and the preference from 325,772 different users. In particular, this dataset contain:

- The anime list per user. Include dropped, complete, plan to watch, currently watching and on hold.
- Ratings given by users to the animes that they has watched completely.
- Information about the anime like genre, stats, studio, etc.
- HTML with anime information to do data scrapping. These files contain information such as reviews, synopsis, information about the staff, anime statistics, genre, etc.