Homework 4

Due Thursday, April 26

Setup

You are a machine learning specialist for a company that electronically delivers movies to users. The company has a major new initiative, to build a recommendation system that will help users identify content that might interest them. Your part in the process is to employ a machine learning algorithm that will make sense of users in terms of the genres of movies they watch and like. As you know you have several options of algorithms for such analysis, including clustering, PCA etc. Your boss trusts you to pick a good one.

But as a ML specialist, you know that the hard part isn't applying the code. The raw data is rarely usable out of the box to answer a particular question of interest. Instead, there are several things that need addressing. For instance, in this case, the data doesn't contain a measurement of the user's preference of genre. Instead it contains user ratings of movies. And anyways, how does one measure preference for genre- is it, say, a user's viewing distribution over genres? Is it some statistic of the ratings they give in a certain genre? A combination? You know you will have to think about this carefully- perhaps trying a few different approaches and analyzing the results.

Files

1. 'movie_genreID_genre.Rdata' – read in with load(). Contains a tibble called movie_genre which is in long format. The columns should be self explanatory.
2. 'movie_rating_data.csv' – read_csv(). A dataframe with userId, movieId (which aligns with the movieId in the previous file), rating, and timestamp (which you can ignore).

Tips

1. Use things like inner_join() to merge the two datasets
2. Keep the data in LONG format for as long as possible and use pipelines to obtain genre information for each user (of course you have to think about what that genre information will encompass (see second paragraph in "Setup")

Other bits of info

1. I will be uploading a few other files as well that you will not need for your homework, but if you get bored you might try out a few variations.
2. The data is from Kaggle. Feel free to find it on Kaggle. However, please use the data provided, as I have removed some of the rows from the movie_rating_data.csv file. ***Please do not try to figure out what rows are missing!*** For one, it is a waste of your time. For two, I will be using them as test cases when you submit your homework.

Deliverable

A writeup of your analysis, including examples of users from the dataset and some visualizations. Write up your method and justify your choices. Writeup should be submitted as a PDF.