Milestone 2

GitHub repository: https://github.com/cocochrane/CS109B_FinalProject (https://github.com/cocochrane/CS109B_FinalProject)

Discussion about Imbalanced Nature of Data Set and how we want to address it

We have selected 16 genres that we would like to predict. As we will describe below, we are planning on performing a binary classification for each of these genres in the traditional supervised learning phase, and we'll do a multi-label deep learning classification. Because we have so many genres, some of these genres are much more rare than others.

For many supervised learning techniques, we will be able to provide sample weights that will prevent the models from underperforming on the rare class (the genre we are predicting). If we believe that this approach is not effective, we can perform a stratified sampling to balance the classes. This is a more complex and challenging approach because for most observations, more than one genre tag is present, so it may not be possible to completely balance the classes through re-sampling.

In the coming lectures, we expect to learn more about how neural networks handle unbalanced data, and may need to modify our approach in order to do multi-label classification.

Description of Data

Our data comes from the IMDB and TMDb databases. We thought that there was plenty of information contained in these databases to accurately predict genre without using other sources. We have approximately 60,000 instances and 21 features (some are expanded using bag of words) in our final dataset. In addition to these numerical and categorical features, we have also downloaded the posters associated with each movie for which a poster url is available.

We only kept movies that were in both the TMDB and IMDB database and had at least one genre assigned to them.

What does our Y look like?

Our Y is a matrix. For the supervised learning we are going to try to predict each of the genres we're considering separately, whereas with the neural network we'll do a multi-label classification problem. The genres that we're considering are the 16 genres that appear in both databases, specifically:

- Action
- Adventure
- Animation
- Comedy
- Crime
- Documentary
- Drama
- Family
- Fantasy
- History
- Horror
- Mystery
- Romance
- Science Fiction
- Thriller
- War
- Western

We looked in-depth into the genres that we are not considering and determined that either these were not "movie" genres (e.g. talk show, news, music) or had no clear pair in the other database. For the later case we checked to make sure that the "bad" genre was often found with one of the genres that we're keeping so that we don't lose too many instances.

What features do we use for X and why?

From TMDB we're keeping:

1. revenue
2. overview (using bag-of-words)
3. title (bag-of-words)
4. vote_count
5. popularity
6. budget
7. vote_average
8. runtime
9. release date year
10. release date month
11. average value for red channel in movie poster
12. average value for green channel in movie poster
13. average value for blue channel in movie poster

We believe that revenue, vote_count, vote_average, popularity, and budget all give information for the popularity of the movie and popular movies tend to be from specific genres (action, comedy, drama). Overview and title (using bag-of-words) may give some indication of genre (for instance, the word 'police' might indicate a crime genre). Runtime may have an interesting interaction with movie genre (e.g. perhaps Romance movies tend to be shorter than Science Fiction ones). Release date and year may show some effects (perhaps Horror is more likely for movies released in October, for instance). The average color of the pixels may give a significant amount of information (a red average may indicate a crime movie, a blue average may be a drama movie, etc.). Alot of care goes into movie poster design and colors often evoke very specific feelings that tend to relate to the genres.

From IMDb we're keeping:

1. rating
2. votes
3. color info
4. language
5. countries
6. director
7. writer
8. cast

As above we include rating and votes because we hypothesize that certain genres are more "popular" than others. Color info is included because black and white movies may be more often in a certain genre (e.g. 'western'). There may be cultural differences in terms of what genre movies tend to be produced- language and countries get at this notion. Director, writer and cast often seem predictive of certain movie genres- often certain actors get cast for mostly action roles (so action genre), etc.

How do we sample our data, how many samples and why?

First we extracted data for all movies in the TMDB database. Then for those movies that had IMDB ids available, we extracted data from IMDB. So the resulting set contains only movies that are present in both databases.

We chose this approach because we wanted to maximize the amount of data we can get, but at the same time want to limit the number of "nulls" present in our data.

Our final dataset includes 63911 movies.