GS-EECS6327A Project 2 Proposal

Yunge Hao 214361760

Topic

SVM and Neural Networks for classifying the successfulness of movies from IMDb

Abstract

Classifying how successful a movie yet to be released is of high interest to movie companies and entertainment industry in general. I plan to apply some machine learning techniques doing the classification, by their expected revenue and future rating. The outcome will be in one of "Excellent", "Good", "Average" and "Low" categories. The revenue and rating could be predicted in advance based on several features of that movie already known to us.

Data and Feature

The data sources are IMDb and Rotten Tomatoes.

The feature set can be divided into 2 categories, direct, numeric features, such as budget, gross revenue, overall rating, and secondly, indirect, text features, such as the company that produces this film, its director, cast and so on. The grouping of feature sets in also differs in how I am going to learn each of them. For the numeric features, they are properties closely attached to this movie, and I can directly use them in computation. So numeric features are also direct features. The data from text-based features has little direct computation value, if not at all. But they provide presumably more significant insights of training a predictor. For example, given the name of an actor/actress, or director, we can learn from their past movies, their ratings and revenues, and make use of them in embedding the potential impact of these names on this new movie we are learning. Therefore, the classification is then partially based on its "pre-history", the data of movies produced by or cast a similar crew.

I intend to constrain the data within movies produced in the states and in English, the main reason is for the huge amount of movies American industry produces each year which means large set of data. And I only collect movies of recent 10 years.

The movies are labeled by several genres, e.g., history, romantic, action, and each genre will contain 1000 movies, with at least 100 movies for each year per genre.

Model

The references that I found applied linear regression, logistic regression and SVM. But given the additional idea of a movie having a history linked to other older movies through its production crew, I think NN is also a good choice. The results from my references are not very strong, so I would like to see if NN can do better.

In my model I plan to include SVM and NN only. Besides analyzing the accuracy of movie classification, I would also like to have a performance comparison with SVM and NN on purely numeric features, a comparison with SVM and NN on purely text-based features, and together, an overall comparison. How each model behaves on various genres of movies will also be tested.