Seong-Eun Cho
Christopher McClellan
Evan Hilton

Group Project Progress Report

For our Machine Learning Group Project we are implementing a model using backpropagation to learn what influences the final result of a movie's IMDb score. IMDb is a website with an authoritative movie database. Once a movie has been released (meaning that it has been shown publicly at least once), users of www.imdb.com give their own score from 1 to 10. These scores are averaged and this is displayed as the movie's official score on the website. We are aiming to find a correlation between certain features and the final IMDb score. If the training can produce reliable results, it might be possible to predict relatively how "good" a movie is before it is even released.

Model

For our initial attempt at solving this problem, we are going to use backpropagation with a multilayer perceptron. Since many of the features are continuous and there is a possibility of hidden relations between them, this seemed like the most reasonable approach. A decision tree would be difficult to implement for the initial run because of the continuous nature of the features. We would have to bin the data to get it to work properly. Also, we will be working with a large number of features, so a k-nearest neighbor approach would probably be hindered due to the high dimensionality. With a multilayer perceptron, inconsequential features will eventually be weighted to have little to no impact on the problem, which will help narrow down our features in the future.

Data

Data for this is available on www.kaggle.com. However, there are some missing values and a lot of extra features. The data would need to be consolidated and cleaned up for training. Most of the proposed features are numerical, but there are some categorical features in the data that could be used if needed, such as director name, actor names, language, and country of origin. Certain datasets might have additional desirable features, so those datasets would have to be union-ed together by movie title to get the best features.

There are barely more than 5000 data points in our dataset. Each data point represents a unique movie with 28 features. Some of the features of the data are the number of

facebook likes the director and actors have, the duration of the movie, when the movie was released, the budget and gross income of the movie, etc. The labels we will be attempting to predict is the IMDb score it got from the users. Below is an example of a data point for the movie Avatar, which also includes what type each feature is (nominal or real). Some features are nominal but have too many variations to lead to any reasonable relation (such as the names). These features will not be used. Of the 28, we'll use 16. All data points will be used except for those with missing values. It'll likely still be around 5000.

color	Color	Nominal
director_name	James Cameron	
num_critic_for_reviews	723	Real
duration	178	Real
director_facebook_likes	0	Real
actor_3_facebook_likes	855	Real
actor_2_name	Joel David Moore	
actor_1_facebook_likes	1000	Real
gross	7.60506e+08	Real
genres	Action|Adventure|Fantasy|Sci-Fi	Nominal
actor_1_name	CCH Pounder	
movie_title	Avatar	
num_voted_users	886204	Real
cast_total_facebook_likes	4834	Real
actor_3_name	Wes Studi	
facenumber_in_poster	0	Real
plot_keywords	avatar|future|marine|native|paraplegic	
movie_imdb_link	http://www.imdb.com/title/tt0499549/...	
num_user_for_reviews	3054	Real
language	English	Nominal
country	USA	Nominal
content_rating	PG-13	Nominal
budget	2.37e+08	Real
title_year	2009	Real
actor_2_facebook_likes	936	Real
aspect_ratio	1.78	Nominal
imdb_score	7.9	

Plan

Our next major goal is to preprocess our data and run it on a multilayer perceptron. All features' values will be scaled to be between 0 and 1. From there, we'll work on eliminating insignificant features either based on the model's weights, a wrapper function, or PCA. We plan to have this preprocessing done by November 14th.

After the preprocessing of the data, we will do our initial run on our MLP in the following week. During this time, we will play with hyperparameters and hidden layers, eliminate features, and try to get the best result possible from our MLP. Our goal will be to predict the IMDb score within +/- 1 of the actual score, but we will determine the reality of that goal during that time. Perfecting our MLP should be done by the end of the following week, on November 22nd.

Lastly, we will explore other models and see how they fare on this problem. By this point, the data should be neat and not include insignificant features, so it should be easily runnable on most models. We'll test out other models such as k-nearest neighbor, k-means, etc. and see if we can get accurate ratings close to or greater than that of our MLP. Those results will then be analyzed and our final report and model will be completed by the due date, December 5th.