CS 4641 Team 16 Project Proposal

Nikhil Bose, Joel Joseprabu, Ji Won Lee, Logan Short, Sandeep Veludhandi

Introduction:

Every year, Americans fill out more that 70 million brackets and bet more that 2 billion dollars trying to predict the results of March Madness. So far, not a single person has managed to predict the entire tournament correctly: the closest anyone has gotten to a perfect bracket is predicting the first 49 of the 63 games correctly. We plan to use data from past college basketball seasons to predict our own bracket.

Methods:

We will be using a combination of unsupervised and supervised learning techniques to solve this problem. We will be using a BigQuery dataset of NCAA games, teams, and players. Initially, we will use a dimensionality reduction algorithm, such as principal component analysis or linear discriminant analysis, to extract useful features from past NCAA basketball game data for our problem. Then, we will use this reduced data to train a supervised learning model to predict the winner of a match up of two specific teams. By recursively matching up winning teams with each other, and predicting the outcome of these matches, we can eventually generate a full bracket.

Expected Results:

Just like in most bracket competitions, each game is awarded a set amount of points depending on which round it is in. Correctly predicting a game in the 1st round can score 1 point while a correct prediction in the semis may garner 16 points. We are trying to predict the bracket with the maximum score using the features that are most likely to predict the outcome. Using a goal score of perhaps 100 points, we can ensure that our model would predict a bracket far more accurate than most participants. This may not lead to the most correct predictions as predictions in later rounds are valued more than those in earlier rounds.

Discussion:

The ideal best outcome would be the ability to perfectly predict the outcome of future March Madness tournaments. With the odds at 1 in 9.2 quintillion, Quicken Loans and Warren Buffett have even put forth a billion dollar prize for a perfect bracket. In 2018, the average bracket entered in the NCAA challenge only scored 57 out of 192 possible points. Our goal score of 100 points would put our model well above the human average, but it would leave plenty of improvement room for future models to be refined and reach ever closer to the billion dollar 192 point mark.

References:

Balakrishnama, Suresh, and Aravind Ganapathiraju. "Linear Discriminant Analysis - A Brief Tutorial." Institute for Signal and Information Processing 18 (1998): 1-8.

Kumar, Ch Aswani. “Analysis of Unsupervised Dimensionality Reduction Techniques.” Comput. Sci. Inf. Syst. 6.2 (2009): 217-227.

“March Madness Bracket Scoring.” JellyJuke, http://www.jellyjuke.com/march-madness-bracket-scoring.html.

Shlens, Jonathon. “A Tutorial on Principal Component Analysis.” arXiv preprint arXiv:1404.1100 (2014).