Project Proposal

Predicting result of NCAA football games

Machine Learning and Computational Statistics
Spring 2018
DS-GA 1003

Team Members:

Diwakar Paliwal
Kumar Mehta
Tejal Lotlikar

Advisor:

David Frohardt-Lane

Abstract:

The project will make prediction on the winning team in a NCAA football match. The initial model will address the problem statement by implementing a baseline algorithm, over which performance improvements for consecutive models will be observed. The project will initially predict the winning team with winning probability. Going ahead it will predict other winning statistics such as score of the game and average margin victory points for both teams.

Data Set:

We have found a large and complex data-set compiled by cbstats.com, which has information on College Football Statistics from 2005-2013, recovered via wayback link. The data has statistics about approximately 720 games played between 240 teams, for every season from 2005-2013. Of these 240 teams, about half of them are FBS(Football Bowl Subdivision) teams and the other half are FCS(Football Championship Subdivision) teams. FBS teams have high number of matches (approximately 10-12 games) per season. FCS team are mentioned only when the match happened against a FBS team and therefore the data is very limited for these teams(0-2 games per team). Hence, our prediction will be focused for FBS teams only and we will not consider games that includes FCS teams.

Each years data has various fields and details such as:

- Game statistics: Home and visiting teams, points scored, detailed information on every rush, punt, reception, kickoff, pass, etc.
- Teams: Information of team composition, combined statistics for the team in games like rush yards, number of passes completed, kickoff return yard, field goals made, tackles, fumbles, etc.
- Players: Information about player positions, statistics of each player in every game like number of points scored, number of tackles, number of tackle assists, number of fumbles, etc.
- Environment: Information on ground-surface, total attendance for the game, stadium statistics, total time duration of the game, etc.

Link to the dataset:
https://old.datahub.io/dataset/college-football-statistics-2005-2013

Performance Measures:

We will measure the performance of our algorithms to predict each team's winning probability using log-loss that takes into consideration the uncertainty of prediction based on how much it varies from the actual label. In the next phase, when we are predicting the margin of victory points for the teams, we will use mean squared error as the performance measure to compare algorithms for predicting the score of game. We will compare these parameters for various algorithms that we implement and select models with least log-loss and mean-squared errors.

Baseline algorithms:

Algorithm 1:
We will use only one feature from the dataset for our first baseline algorithm. We predict the probability of the winning team with a better win ratio (defined as number of games won over total games played), among all games played before the current game in the same season. The probability will be calculated as the ratio of win ratios for both the teams. In case of both teams with zero win ratio (a corner case), by definition both teams will have equal probability to win the current match since both of them have same win ratio.

Algorithm 2:
We will add few more features to our first baseline algorithm like the total point difference(difference of the number of points scored and the number of points conceded in all matches) for both teams, which is the home team and time of possession in the previous games, and finally apply logistic regression on these features to predict the winning probabilities for both teams in the current game.

Methods:

We will divide the dataset into 3 parts- training set, validation set, test set by dividing the dataset into 60%, 20%, 20% for the respective sets. For every season, we will skip all the games in the first 2 weeks and predict the result of the remaining matches in the season based on collective features of all previous games of that particular season. For the baseline algorithms, we will assume that the result of the current game does not depend on the statistics of previous seasons. For the remaining models, we will give weights to results and statistics of previous seasons. Since the data has a lot of features, we will initially assume the results do not depend on some features like the team composition(players) for every season. We plan to compare performance of different models such as Logistic regression, Neural networks and Support Vector Machine.

Approximate Timeline:

● Feature Extraction	-	April 4, 2018
● Baseline Algorithms Implementation	-	April 14, 2018
● Implementation of first algorithm(Regression)	-	April 20, 2018
● Implement other models(Neural networks, SVM)	-	April 30, 2018
● Fine tuning the hyperparameters and identifying the best model	-	May 5, 2018