CS 216 - Proposal write up Duke University | March 7th, 2017
Team 14: Mary Ziemba, Melanie Krassel, Alex Deckey, David Duquette, Camila Vargas

MMM - March Madness Mania - Midterm Report

First Things First: 9.2 Quintillion

In our project write up we mentioned that there are 9.2 quintillion possible brackets to be made for the NCAA tournament. There are 63 total games in the tournament. For each of those games, two teams play, and one team wins, which means each game has two possible outcomes. Therefore, there are 2^{63} possible brackets (9.2 quintillion).

If we knew nothing about basketball, that is, if we assumed that each team in each game in the entire tournament had a 50% chance of winning, the probability that you got the entire bracket correct would be $\frac{1}{2^{62}}$. Yet, we do know some things about basketball: like that fact that the last seed has never beaten a first seed, and that every year a 12 seed upsets a 5 seed. Here's where we come in.

Progress Made So Far

By this point, we have met several times as a team and we have been working individually in different tasks. During these meetings we have made significant progress in framing our goals, questions, and next steps. We have decided our factors, and we have scraped and prepared all of our data. We also have skeleton code for a Naive Bayes Classifier, which we will be using to make our predictions.

Framing the Project's Goals

The feedback we got for our project write up mentioned that our hypothesis/predicted variables were still a little vague. Thus, we have narrowed our approach and defined our objectives.

Our goal is to construct a Naive Bayes classifier trained on a set of historical NCAA tournament data that will predict the winning team of each game in the 2017 Men's March Madness tournament with reasonable accuracy. Right now, we are hoping to achieve 80-85% prediction accuracy on the outcome of games Once we construct the Naive Bayes we want to look back at its predictions and compare them with the real outcomes for this March Madness season. We are interested to know how well our model does with predicting upsets, as well as if we have discovered factors not yet universally looked at that do affect the outcome. We now have a very solidified and diverse set of factors we are looking at.

Defining variables

Based on the data we have scraped and found available online, we have decided that the following factors will be included, at least initially, in our Naive Bayes:

- Seed difference weighted by what round the game is being played in.
    - What data do we need for this factor? Seeds of each team in each NCAA tournament since 1993. We already have this data available.
- Difference between each team's points per game (PPG) in their regular season games.
    - What data do we need for this factor? Each team's point per game average for each season since 1993. We already have this data available.
- Difference between each team's points allowed (PPG) in their regular season games.
    - What data do we need for this factor? Each team's point allowed average for each season since 1993. We already have this data available.
- Difference between each team's regular season win percentage
    - What data do we need for this factor? Each team's win percentage for each season since 1993. We already have this data available.
- Difference between each team's geographical score
    - The geographical score is a construct we have defined in the following way: a team gets 1 point for every timezone they have to cross to get to the game venue from your respective campus. The team gets -1 point if the game is played in the same state as their campus.
    - What data do we need for this factor? We need the location of every game for every NCAA tournament since 1993, as well as the state each team belongs to. We have scraped this data and have it formatted and ready to use.

Next Steps

We have divided our project in the following stages:

1. **Data explorations and simpler models**

   This stage will involve creating simpler algorithms, like K-nearest neighbors, that may provide better insights into which factors would contribute the most to the bayes algorithm. We will also work on exploratory data analysis and data visualizations of the 'raw' data in order to inform the construction of the algorithm. Exploratory Visualizations will include:
    - Winners of each tournament and visualize their characteristics, to find out which factors winning teams have in common.
    - Radar charts to visualize team characteristics

- Scatterplots (correlation matrix):
    - PPG vs. win percentage in regular season
    - PPG allowed and win percentage in regular season
    - PPG vs. PPG allowed .
- Visualization of geography score using bar graphs

2. **Constructing the Naive Bayes**

- We have a skeleton of the Naive Bayes theory that two of us have implemented in another class (CS270: Intro to Artificial Intelligence). We will most definitely need to change a significant part of it as it will be used differently/will take in very different types of information in the CS216 project. We were wondering if given this fact, we have permission to build off of/reference that assignment from CS270. We will perform cross validation on our data set.

3. **Looking back and comparing the Naive Bayes to this season's results**
4. **Creating visualizations based on our results | Explanatory Visualizations**
   We are anticipating to create at least three visualizations once we have a fully functioning Naive Bayes model in place:
    1. Prediction Bracket visualization: predicting the 2017 winners and comparing that to the actual outcome
    2. Bar graph of probabilities of winners (since ultimately, using Bayes, we will be working with probabilities of winning versus simple win/loss binary prediction
    3. Visualization of how factors compare (since we anticipate that some factors will be stronger in their predictive abilities than other factors).

Tools

-We have installed and learned how to use git so that we can work collaboratively.
-We will use the following tools to carry out the project:

- Python sklearn
- Python pandas
- Python numpy

Division of labor:

We have regular meetings at The Edge every Friday to check in as well as meeting on the weekends to further collaborate and plan the project. Nonetheless, each of us will are focusing more intensely in one of the sections:

- *Mary and Melanie* will be working together on the Naive Bayes
- *Alex* will be working on data cleaning/wrangling and new data scraping
- *David and Camila* will be working on the exploratory and explanatory data visualizations