CS 221 PROJECT PROPOSAL

STUART SY AND YUSHI HOMMA

1. Introduction of Task

ESPN fantasy baseball is a common pastime for many Americans, which, coincidentally, defines a problem whose solution could potentially be predicted by artificial intelligence. The particular ESPN fantasy baseball game that we will analyze in this project is the ESPN Baseball Challenge. The basic task for these fantasy baseball participants is to predict which players will have successful games based on a standard scoring system and optimize the total score of their fantasy team under a budget. The fantasy teams consist of 9 players (one for each fielding position plus DH) and an entire pitching staff. This scoring system is provided by ESPN.com, which we define below in Section 5.

2. Literature Review

Upon searching for literature on similar topics, I came across two related former Stanford CS 229 project papers. "Predicting the Major League Baseball Season", written by R. Jia, C. Wong, and D. Zeng, regarded predicting wins and losses of teams in a particular season. Another that I came across, "Applying Machine Learning to MLB Prediction & Analysis", similarly tried to predict games to be wins or losses based on previous statistics.

3. Input/Output

The input into our system is the previous season's statistics. The system will then use this input to output the optimal team of 9 players of each position and a pitching staff with total budget of $50 million.

4. Dataset

The dataset will include season statistics for every player in the MLB since 1871 matched with the scores of the corresponding players in the next season. We obtain this data from the baseball statistics archive accumulated by Sean Lahman in his collection of statistics found at http://www.seanlahman.com/baseball-archive/statistics/. The statistics for batters consist of the year played, their team, position, age, numbers of games, at-bats, runs, hits, 2B, 3B, home runs, runs batted in, stolen bases, attempts caught stealing, walks, and strikeouts. For pitchers, the dataset includes year played, team, age, salary, number of games, wins, games started, complete games, shutouts, saves, innings pitched (in outs), shits, earned runs, home runs, walks and strikeouts.

5. Evaluation

The scoring system based on individual stats for the hitters and stats for the entire pitching staff of a team. The score for a single hitter is defined by the following formula:

$$\text{Score}_{\text{hitter}} = \text{R} + \text{TB} + \text{RBI} + \text{BB} + \text{SB}, \tag{1}$$

where $\text{R} = (\text{number of runs})$, $\text{TB} = (\text{total number of bases})$, $\text{BB} = (\text{number of walks})$, $\text{RBI} = (\text{number of runs batted in})$, and $\text{SB} = (\text{number of stolen bases})$.

The score for an entire pitching staff for a season is defined by:

$$3 \cdot \text{IP} - 3 \cdot \text{ER} - \text{H} - \text{BB} + \text{K} + 5 \cdot \text{W}, \tag{2}$$

Date: October 20, 2015.

where IP is the total number of innings piched, ER is the number of earned runs, H is the total number of hits that the pitching staff gave up, $\mathrm{BB} = (\text{number of walkspitched})$, $\mathrm{K} = (\text{number of strikeouts pitched})$, and $W = (\text{number of wins})$.

We will evaluate our system's predictions with the scores on the leaderboard of the ESPN Baseball Challenge listed for the past 4 years. These will be our test datasets.

6. Infrastructure

Using the pandas data science library, we extracted relevant parts of the CSV data and processed it into usable python objects. We separated the data by season because each season is one sample of data. We also determined what position each player is categorized as based on the metric given by ESPN.com (having played 20+ games at that position in the previous season). We also matched up the playing statistics in Batting.csv and Pitching.csv with the calculated ages from Master.csv.

7. Challenges and Approach

This project poses two main challenges: to predict the scores of each player/pitching staff for the next season using the previous season's statistics and to find the maximum total team score within the constraints of the budget and player positions.

Since the flow of a baseball game has natural breaks to it, and normally players act individually rather than performing in clusters, the sport lends itself to easy record-keeping and statistics. Ever since the explosion of fantasy sports, the analysis of player/team performance with statistics has become ever more popular/important.

As stated before, our task is to construct the best possible performing team (defined with a custom fantasy baseball score) under a budget constraint. To model this task, we want to model predicting the price/performance of individual players as a linear regression problem, and to model picking the optimal team as either a search or constraint satisfaction problem.

One challenge that may be difficult to completely remedy is that our statistics only include season statistics instead of game-by-game statistics. This means that once we choose a team for a season, we have no additional information to change the members of the team during the season. However, in the ESPN Baseball Challenge, the fantasy baseball participants can change their rosters throughout the season, which could give them an advantage over our system due to fluctuations of players' performance throughout the season. This is why our oracle might not outperform the top participants of the ESPN Baseball Challenge, but it will be the best that we can aim toward with our algorithm.

8. Baseline and Oracle

We will perform our baseline and oracle on the year 2014. We defined our baseline to be just dividing the budget into equal parts to budget each individual position, and then obtaining the top-scoring player at each position from the previous year. We used 3/10 of the budget on the pitching staff because that is roughly how much MLB teams pay their pitchers. This leaves 3.9 million for each hitter and 15 million for the pitching staff. This resulted in a score of 4583. We defined our oracle to be the top participant in the ESPN Baseball Challenge in the year 2014. The oracle was GEORGETHEBOSS1 who got a score of 8362. Thus, there is room for much improvement.