CS 221 Project Proposal

Jorge Garcia and Joey Asperger

Data-driven pitch selection

The pitcher is an integral factor in the result of baseball games. He begins every play when he releases the ball from his hand, and his choice of pitch and its effectiveness determine whether his opponent will be able to hit the ball. The selection of pitch type and location is a venue primed to be studied through artificial intelligence. Starting in the 2006 Major League Baseball playoffs, advanced cameras were installed throughout the league parks. These cameras, as part of the PitchF/x system, track numerous characteristics about pitches, including pitch type, starting and ending position, velocities, spin speed, direction, break speed and direction, and other characteristics all to a very high degree of precision. All the data captured by this system for every Major League Baseball game since 2006 is freely available to download. Through our project we hope use this data to examine the effectiveness of pitch types, location, and sequences. In our project we wish to build a machine learning based system that will suggest effective pitches for select situations. To build this system we will incorporate information like previous pitch, the count, handedness, among others. From these possible inputs, we would be able to get outputs like most effective pitch type and location.

A possible baseline implementation would be a linear classifier that would pick among possible pitch types using a feature vector that would include handedness, count, and previous pitch type. An oracle implementation would be one where the actual outcome of the situation is provided as a feature so that the pitch would be thrown given a positive outcome otherwise another pitch is suggested. The gap that exists is the percentage of positive outcomes for the actual situation, which would be between 60-70 percent. To measure the success of our algorithm, we will divide up our data into training data and test data. After training on the training data, we will go through the test data and predict the most effective pitch for the pitcher to throw in each situation. Then, we will look at what pitch the pitcher throws next and determine whether it matched with the pitch we predicted. If the effectiveness when the pitcher throws what we predicted is significantly higher than when the pitcher does not, then our algorithm is correctly predicting effective pitches. Baseball is a complex environment and there are many variables that are potentially useful information. One of the challenges that the project presents include making wise feature vector selection so that the model does not overfit for the data that it was taught on, but instead generalizes for other baseball data. Machine Learning and decision making modeling would appear to the be the topics most likely to address our system. Related work that might be of use would include work by John Guttag presented at the MIT Sloan Sports Analytics Conference: *Predicting the Next Pitch* and *A Data-driven Method for In-game Decision Making in MLB* along with a freelance piece by Isaac Laughlin: *The Data Science Behind Baseball Pitching Strategy*. All three of the pieces use machine learning to accomplish different sports related tasks, but nevertheless should be useful for insight on how to guide our project.