EECS 349 – Machine Learning

Final Project Proposal

Tim Herrmann - trh701
Mikhail Todes - mbt810
Nathan Corwin - nca318

Task: Our goal is to predict a baseball player's season long batting average based on statistics from previous years. Baseball is a multi-billion dollar sport in the US with an incredible amount of collected statistics. General Managers are regularly making million dollar decisions based gut feelings and observational data. Statistical analysis through machine learning provides an opportunity to make better decisions based on hard data.

Additionally, this model could be useful in a gambling setting, as there exist prop bets based on player performance.

Data Source: We found a comprehensive list of baseball statistics dating back over 100 years at http://www.seanlahman.com/baseball-archive/statistics/ . This data is free to use and covered under a creative commons license. We plan to use data from the last 10 years only. We will randomly select samples to set aside for test data.

Features: We plan to use the following previous year features to create our model:

Batting Average
Salary
Age
Strike Outs
Walks
Hits
At-Bats

We may add more features at a later time.

Approach:

The first step will be to discretize the numerical data. Then, we will develop a distance function based on all 7 of the features and train the model using a nearest neighbor method. Additionally, we will attempt a 10-fold cross-validation using decision trees. We will compare the two models and proceed accordingly. We will also consider additional models as we learn more in class. We will evaluate model performance using held out test data.