4P76 Project Proposal

Nick Seelert & Preston Engstrom

Purpose: Comparison of artificial neural networks with linear regression to predict likelihood of UCL injury leading to Tommy John Surgery.

Background: UCL injury has become an increasing concern in professional baseball. The prevalence of the surgery has exploded in the last decade; recent estimates are that close to 25% of MLB pitchers have undergone the surgery. As such it has become a hot topic in the sports analytics world trying to determine the causes that have lead to this situation. However, as of today, there is still no clear indication of exactly what factors are responsible. However, as more studies are published it appears that there are some factors that may be more involved than others.

Study Approach: The status quo in sports analytics is linear regression. As such we would like to compare the use of neural networks to linear regression to try and predict the likelihood of UCL injury. As of today, it doesn't appear that any studies have specifically looked at this set of inputs. This study will require a sound statistical approach, so each attribute will need to be tested for power a priori to ensure our sample size is large enough for the regression. For the neural network we will try to build as large a data set as possible including and even match of players who have had the surgery and those that have not. It will be important to ensure injury players and non-injury players are matched as equally as possible on other variables to limit the influence of confounding factors (ex. Age, position, height, weight, etc.)

Inputs of Interest:

- Pitch velocity
- # of pitches thrown
- # of innings pitched
- % fastballs thrown
- Strikes thrown
- Days on DL for elbow injury in same year before UCL injury
- Average days rest between appearances
- Advanced pitcher stats like FIP or SIERA

Machine Learning Details:

- K-folds cross-validation
- Neural network with back propagation
- ***Need to research how to use neural network for probability or confidence output***
- Will use a combination of R libraries and python scikit-learn to carry out experiments and analysis as well as neural network from assignment 1

Data:
Data is being acquired from 2 sources. PitchF/x data holds all the performance statistics of professional baseball players and is publically available (http://www.brooksbaseball.net/). The injury data is being accessed through a private database (http://www.baseballic.com/). Access to this database has been granted to Nick via a 3rd party industry partner who is asking for this work to be provided to them in return.

Roles:
Considering this is a new area of research the project will require a heavy literature review as well as a machine learning experiment. Overall, work will be split evenly between Nick and Preston. Nick will likely handle more of the domain knowledge requirements (ie. literature review, experimental design, data acquisition, pre-processing) and Preston will likely handle more of the experiment and analysis coding requirements. Both Nick and Preston will analyze results and contribute to the written report based on their specific roles in the project. These are rough guidelines for the work split and could change depending on the demands of the project.