Capstone Proposal

Proposal

**Domain Background**

**In this section, provide brief details on the background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited in this section, including why that research is relevant. Additionally, a discussion of your personal motivation for investigating a particular problem in the domain is encouraged but not required.**

For my Capstone project I will be applying several different unsupervised learning algorithms to Major League Baseball data. Statistical analysis is in the DNA of the game of baseball, as game data has been recorded for well over a century. Because of this, baseball data lends itself very well to machine learning, particularly with the recent implementation of Statcast. Statcast “is a state-of-the-art tracking technology, capable of measuring previously unquantifiable aspects of the game. Set up in all 30 Major League ballparks, Statcast collects data using a series of high-resolution optical cameras along with radar equipment. The technology precisely tracks the location and movements of the ball and every player on the field, resulting in an unparalleled amount of information covering everything from the pitcher to the batter to baserunners and defensive players” (What is Statcast?, 2016). Machine learning has been applied to baseball data both publicly among the baseball community for pleasure and privately among professional teams with the goal of gaining a competitive advantage. An example of applying machine learning algorithms on a Statcast data set would be using a Random Forrest classifier to predict whether a batted ball will be a hit or an out (Petti, 2016).

There are several reasons I chose the domain of baseball for my final project. First and foremost, I am a big fan of baseball, particularly the analytical and quantitative side of the game. It was this interest that lead me to the field of machine learning in the first place, and even helped me learn some machine learning and statistical software (Marchi & Albert, 2013). Secondly, Statcast is an extremely new technology, and can offer insights into the game that were previous unavailable. Finally, the datasets are both large and free to use by anyone, making it very conducive toward statistical analysis.

**Problem Statement**

**In this section, clearly describe the problem that is to be solved. The problem described should be well defined and should have at least one relevant potential solution. Additionally, describe the problem thoroughly such that it is clear that the problem is quantifiable (the problem can be expressed in mathematical or logical terms), measurable (the problem can be measured by some metric and clearly observed), and replicable (the problem can be reproduced and occurs more than once).**

The problem I will be attempting to solve is “can unsupervised Machine Learning algorithms correctly classify pitches thrown by Major League Pitchers into balls and strikes?” Using Statcast data, consisting of perceived velocity, spin rate, release point and pitch type, I will apply unsupervised machine learning algorithms in my effort to accomplish this. Statcast tracks the above features for every pitch that is thrown in major league baseball, as well as the result of each pitch and numerous other metrics. The metric I will be attempting to identify is the resulting classification of the pitch, which is ball or strike.

**Datasets and Inputs**

**In this section, the dataset(s) and/or input(s) being considered for the project should be thoroughly described, such as how they relate to the problem and why they should be used. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included with relevant references and citations as necessary It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.**

Statcast tracks many different features of each pitch thrown in a major league baseball game, with all data being free to use and publicly available via [www.BaseballSavant.com](http://www.BaseballSavant.com). The impetus for the project came from watching the game - it is the job a hitter to identify whether each pitch thrown is going to be a ball or a strike before the pitch cross the plate, and then in turn decide whether to swing at the pitch. This identification is based on some combination of the velocity of the pitch, the spin direction and rate, and the release point, as well as some randomness (it is likely guesswork is involved to some degree though how much is unknown. To replicate the decision-making process, pitch classification is conducive to using unsupervised algorithms because it is a form of binary classification with many features with which to use.

**Solution Statement**

**In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).**

The feature we are trying to predict is ball or strike, which is identified along with each other feature that Statcast tracks. I will be using accuracy score and f1-score to determine how well the algorithms perform at classifying each pitch.

There are two possible ways pitches can be classified. The first would be to use the call made by the umpire, and the second would be to determine where the ball crosses the plate and whether that position is in the defined strike zone. There are pros and cons to each method, which I will detail below.

The first method to classify pitches is to use the call made by the umpire. The advantage of this method is that is requires no processing on the part of the user – the call is marked in the data set already – and that it potentially plays into the decision made by the hitter. The strike zone of an umpire may not match that of the rulebook exactly, yet the umpire is the final authority. This means that a hitter may take into consideration the strike specific to that umpire in determining whether a pitch will be a ball or a strike.

The second option is to use the rulebook definition of the strike zone in conjunction with the coordinates of where the ball cross the plate. The benefits of this are a consistent strike zone for every pitch in the dataset, which would be more conducive toward accurate classification. The downside is additional processing however.

**Benchmark Model**

**In this section, provide the details for a benchmark model or result that relates to the domain, problem statement, and intended solution. Ideally, the benchmark model or result contextualizes existing methods or known information in the domain and problem given, which could then be objectively compared to the solution. Describe how the benchmark model or result is measurable (can be measured by some metric and clearly observed) with thorough detail.**

This one I’m unsure about – would benchmark model be actual results? I.e. whether pitch was classified as a ball or a strike, and the effectiveness of the pitcher.

In this case, the benchmark model is that actual classification of the pitch.

**Evaluation Metrics**

**In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).**

Accuracy score computes the accuracy of the prediction, presented as a fraction of the correct predictions out of the total number of predictions (3.3. Model evaluation: quantifying the quality of predictions, 2010-2016). F1-score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal (3.3. Model evaluation: quantifying the quality of predictions, 2010-2016).

**Project Design**

**In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.**

Approach for solution is to download data, process/clean data, test several different algorithms, identify which works best by accuracy/f-score.

**Before submitting your proposal, ask yourself. . .**

**Does the proposal you have written follow a well-organized structure similar to that of the project template?**

**Is each section (particularly Solution Statement and Project Design) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?**

**Would the intended audience of your project be able to understand your proposal?**

**Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?**

**Are all the resources used for this project correctly cited and referenced?**