Abstract

XXX

Intro/background of the problem

What can the characteristics of a pitch tell us about the most likely outcome we can expect? What patterns can we detect in pitches, and is it possible to use this knowledge gained to provide insights that a pitcher could realistically take action on?

Methods

Feature Engineering

Trackman data does not provide the true spin, only the total spin. The total spin is comprised of both gyro spin and transverse spin (aka true spin), depending on its direction. The gyro spin is like that of a bullet and does not have an impact on the movement of the ball. Therefore, before analyzing the characteristics of the pitch, I separated out the true spin using calculations provided by Alan Nathan. This involved quite a bit of math, including the calculation of 30 or so other variables before I could arrive at the transverse spin.

Cluster Analysis

First I am dividing the data into widely accepted cases, which is the pitch type and the pitcher handedness. For each case (e.g. fastballs by left-handed pitchers) I am breaking all the balls of that case into clusters using k-means clustering based on movement and positional variables. After I have placed each pitch into a cluster, I am running the data on each cluster through some code I wrote to detect patterns. This code simply calculates which areas of the strike zone resulted in the highest percentage of certain desired outcomes. For example, a pitch with this type of movement (cluster 4) results in the highest % of swinging strikes when it passes through zone 3 of the strike zone.

The three desired outcomes I will be examining are the highest percentage of whiffs, soft contact balls, and strikes. This first piece of the project is using machine learning to determine the clusters (k-means), however, this is not predictive analytics in the sense that it is training a model. This happens in the following step, creating a pitch score.

Using Predictive Analytics to Create a Pitch Score

Machine learning model to predict xwOBAcon of a pitch, creating a contact quality metric:

The goal for this model is to look at the quality of contact for players based on exit velocity and launch angle. This is similar to what MLB Statcast is doing and gives you an idea of the true quality of the hitter/pitcher quality that isn't influenced by as much luck and fielders as using batting average or similar statistics.

Results

XXX

Discussion

Pitch type is something that is defined by the player in the MLB; (Note for final paper: insert interesting information from an article on how this is done using machine learning.) For this reason, the assigned pitch type of a ball is somewhat arbitrary, as the same exact pitch could be identified as different pitch types by different pitchers.

I could have chosen to run all the pitches together through a clustering algorithm. This would have allowed the machine learning algorithm to decide whether pitches are similar without the arbitrary use of a label. This does make sense to me and I may go back and try this at some point out of curiosity. However, for this project, I chose to first separate the pitches by pitch type before running each pitch type separately through a clustering algorithm. The reason is that I am trying to be mindful of how this could be used by a pitcher. When communicating with a pitcher, it makes sense to talk about the assigned clusters within each of their pitch types.

Conclusion

XXX

The future plans to continue this project:
Pitch Quality Metric:
This tells you how good each pitch is based on its physical characteristics and is unaffected by the hitter’s response to the pitch. We can use this to tell you which pitches are working best for you.

Pitch Optimization:
This tells you which pitches you should throw more/less of and how that varies based on count and handedness of the hitter.

Acknowledgments

I would like to acknowledge Vicente Iglesias for providing this data set. Alan Nathan's work in the realm of baseball physics was heavily relied on during the feature engineering portion of my project. Specifically in how to calculate transverse spin from trackman data. I used the data from API's of several websites in the calculations for the air density during each pitch:

- Historical weather:
- Elevation of the stadium:

Last but not least, I would not have put together this project in this timeline if it weren't for our instructor, Catie Williams, and her awesome course!

References

Nathan, A. (2018, Aug 31). *Determining the 3D Spin Axis from Statcast Data.* Retrieved from http://baseball.physics.illinois.edu/trackman/SpinAxis.pdf.