Jack Rogers and Kai Franke
Stat 4051
March 27, 2024

Clustering Algorithms Applied to Baseball Pitch Type Classification

Introduction

Major League Baseball pitchers use a variety of grips and spins in an attempt to get their pitches to move on the way to the plate. These pitches are commonly grouped into different categories, with subtypes within. Fastballs have high velocity and little movement. Breaking balls commonly generate movement through high spin rates. Changeups are noted for their lower velocity in an attempt to throw off hitters' timing. Pitch classification is a very important and difficult problem that often involves both supervised and unsupervised learning techniques. These systems often rely on pre-conceived ideas of the types of pitches that are possible, yet as pitchers become more creative, these tags become increasingly vague. In this project, we will examine pitch types within the specific category of breaking balls.

Pitch classification is very important for analysis as many players will call their pitches by different names, which can lead to mismatch of what specific pitch type is. Objectively classifying pitches with an algorithm takes away this subjectivity in naming pitches and allows for precise and accurate analysis.

Dataset & Plan for Analysis

The dataset that will be used for this analysis will be pitch-level data which consists of metrics such as the pitch's movement, speed, release position from the pitcher, and its spin rate. The movement is in x and y coordinates, the speed is in miles per hour, the release position is in x, y, and z coordinates, and the spin rate is in rotations per minute. Baseball Savant houses the dataset that we will be using and all of the information is collected from tracking systems at MLB stadiums. This data collection was initialized back in 2015 when tracking systems made their way into Major League parks.

Missing values will be identified and will either be imputed or thrown from the dataset entirely and it will be filtered down to just breaking balls. Once this is finished, a dissimilarity matrix will be created in order to use the Hierarchical Clustering algorithm. K-means Clustering will also be used and the two will be evaluated against each other to determine which represents the data best. After clustering the pitches, analysis will be done to see which clusters of pitches perform the best.