Predict Next Baseball Pitch Type with RNN
Yifan Pi
ypi@stanford.edu

Task
- Knowing the next pitch type can be a huge advantage for a baseball hitter.
- We try to predict whether each pitch is a fastball or not as the progress of a baseball game.

Data
- MLB Statcast dataset contains pitch-by-pitch data of every game in Major League Baseball from 2018 season. (Available in http://baseballsavant.mlb.com)
- Group pitch-by-pitch data by games to form time series:

- Numerical columns are centralized and categorical columns are converted to one-hot vectors.
- ~2500 games per-season and truncated to 128 pitches per game.

Pitcher/Hitter-2-Vec
- We encode pitcher/hitter one-hot vector into x-dimensional vector by predicting a dummy task: (P, H) -> P(Fastball)

RNN Model
- Standard RNN model to predict a sequence from another.
- Using Pitcher/Hitter Embedding vectors from PH-2-Vec.

Results
- Trained on 2014-2016 data, and validated/tested on 2017 data.
- Train/Dev/Test size (7.5k, 1.5k, 1k)

Discussions
- Bad generalization result: Too many parameters, training data size is limited, games are different every year.
- Results mainly depends on pitch count and pitcher embedding.
- Future works: explore more model settings, embed "pitch sequence" into embedding vector, change training/test split by mixing years.

References