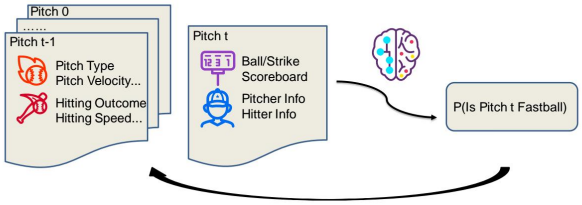
Predict Next Baseball Pitch Type with RNN

Yifan Pi
ypi@stanford.edu

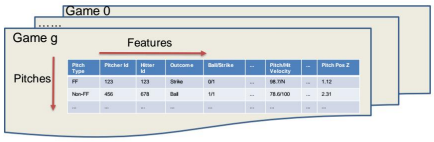
Task

- 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

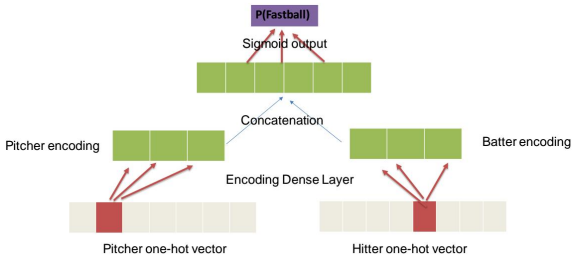
- MLB StatCast dataset contains pitch-by-pitch data of every game in Major League Baseball from 2018 season. (Available in http://baseballsavant.mlb.com)
- Group pitch-by-pitch data by games to form time series:



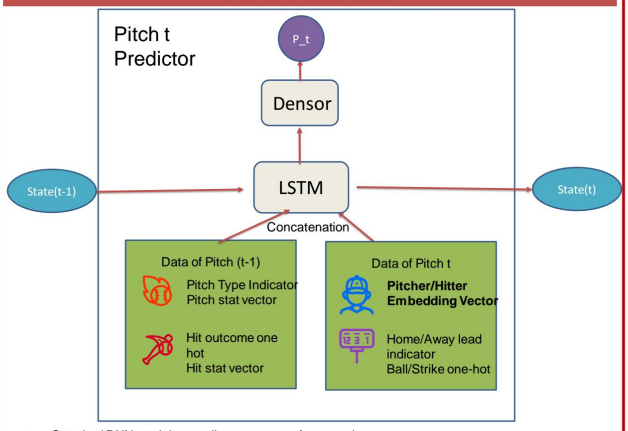
- Numerical columns are centralized and categorical columns are converted to one-hot vectors.
- ~2500 games per-season and truncated to 128 pitches per game.

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 P/H-2-Vec.

Results

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

LSTM Dim	Embed Dim	Optimizer	Learning Rate	Decay
32	8	Adam(0.9, 0.99)	0.01	0.1

Training Loss	Training Acc	Test Loss	Test Acc	Naive
32.645	73.30%	237.8	66.2%	62.8%

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

1. Michael A. Alcorn. (batter|pitcher)2vec: Statistic-free talent modeling with neural player embeddings. In MIT Sloan Sports Analytics Conference, 2018.
2.] Kaan Koseler and Matthew Stephan. Machine learning applications in baseball: A systematic literature review. Applied Artificial Intelligence, 31(9-10):745–763, 2017.
3. Gartheeban Ganeshapillaiand John Guttag. Predicting the Next Pitch. In MIT Sloan Sports Analytics Conference, 2012.

Stanford University