Javier Palomares
James Autry
Porter Perry

EE382V: Scalable Machine Learning
University of Texas at Austin – Fall 2019
Professor Alex Dimakis

Team Project Proposal

1.0 Introduction

In the sport of baseball, America's pastime, the game is won or lost based on the matchup of the pitcher and batter. A typical Major League Baseball (MLB) pitcher has numerous pitch types (like the fastball, curveball, cutter, sinker, etc.) and strike zone locations at their disposal to keep the batter guessing. For each batter, knowing the type of pitch thrown next by the pitcher would be a huge advantage as the hitter can appropriately time the setup of his swing and increase the likelihood of a statistical hit.
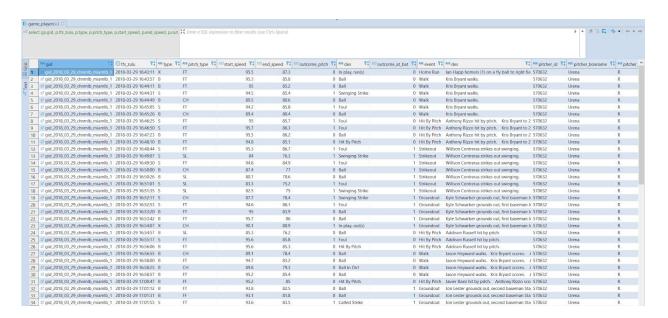
Our team proposes using machine learning techniques learned in this course to build a Neural Network model that can predict the next pitch type of a particular MLB pitcher (multi-classification), unlike related works that only predict if the next pitch is a fastball or not (binary classification). We intend to explore and ask several questions along the way: How much pitch data is required to build a decent prediction model for each pitcher? Is game situation data of the next pitch (like current pitch count, state of bases, inning, and current score) sufficient to build a decent prediction model? Which features are the most relevant to a good prediction? Would a Recurrent Neural Network (RNN) based model [like a Long Short Term Memory (LSTM) network] with memory of one or more past pitches do much better? For the latter, memory of how many previous pitches would provide a noticeable improvement?

2.0 The Dataset

We intend to use gameday statistical data published by MLB Advanced Media website at http://gd2.mlb.com/components/game/mlb/. In addition to standard baseball stats, the data includes detailed pitch tracking data collected from systems like PITCHf/x (camera) and TrackMan (radar) which provides speed and trajectory details of pitches. Standard baseball stats are available from as far back as the late 19th century. The more detailed pitch tracking data is available as far back as the 2008 MLB season.

For our project, we have collected all the available data for pitches thrown during the 2018 season from data repository. The data was collected using a python package that parses the data from the repository and writes it to a database.

The database maintains a relational model between players, innings, games, innings, at bats and pitches allowing us to easily lookup and relate pitches thrown by a specific pitcher. Our plan is to query the pitches thrown by a pitcher, in thrown order. We'll use the result set as the training and testing datasets for our model.

```
select gp.gid, p.tfs_zulu, p.type, p.pitch_type, p.start_speed, p.end_speed, p.out    Enter a SQL expression to filter results (use Ctrl+Space)
```

#	gid	tfs_zulu	type	pitch_type	start_speed	end_speed	outcome_pitch	des	outcome_at_bat	event	des	pitcher_id	pitcher_boxname	pitcher_
1	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:43:11	X	FT	95.5	87.3	0	In play, run(s)	0	Home Run	Ian Happ homers (1) on a fly ball to right fie	570632	Urena	R
2	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:43:57	B	FT	95.3	85.8	0	Ball	0	Walk	Kris Bryant walks.	570632	Urena	R
3	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:44:11	B	FT	95	85.2	0	Ball	0	Walk	Kris Bryant walks.	570632	Urena	R
4	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:44:31	S	FT	94.5	85.4	1	Swinging Strike	0	Walk	Kris Bryant walks.	570632	Urena	R
5	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:44:49	B	CH	89.5	80.6	0	Ball	0	Walk	Kris Bryant walks.	570632	Urena	R
6	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:45:05	S	FT	94.2	85.8	1	Foul	0	Walk	Kris Bryant walks.	570632	Urena	R
7	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:45:26	B	CH	89.4	80.4	0	Ball	0	Walk	Kris Bryant walks.	570632	Urena	R
8	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:46:25	S	FT	95	85.7	1	Foul	0	Hit By Pitch	Anthony Rizzo hit by pitch. Kris Bryant to 2	570632	Urena	R
9	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:46:50	S	FT	95.7	86.3	1	Foul	0	Hit By Pitch	Anthony Rizzo hit by pitch. Kris Bryant to 2	570632	Urena	R
10	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:47:23	B	FT	95.5	86.2	0	Ball	0	Hit By Pitch	Anthony Rizzo hit by pitch. Kris Bryant to 2	570632	Urena	R
11	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:48:10	B	FT	94.8	85.1	0	Hit By Pitch	0	Hit By Pitch	Anthony Rizzo hit by pitch. Kris Bryant to 2	570632	Urena	R
12	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:48:44	S	FT	95.3	86.7	1	Foul	1	Strikeout	Willson Contreras strikes out swinging.	570632	Urena	R
13	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:49:07	S	SL	84	76.3	1	Swinging Strike	1	Strikeout	Willson Contreras strikes out swinging.	570632	Urena	R
14	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:49:30	S	FT	94.8	84.9	1	Foul	1	Strikeout	Willson Contreras strikes out swinging.	570632	Urena	R
15	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:50:00	B	CH	87.4	77	0	Ball	1	Strikeout	Willson Contreras strikes out swinging.	570632	Urena	R
16	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:50:26	B	SL	80.7	70.6	0	Ball	1	Strikeout	Willson Contreras strikes out swinging.	570632	Urena	R
17	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:51:01	S	SL	83.3	75.2	1	Foul	1	Strikeout	Willson Contreras strikes out swinging.	570632	Urena	R
18	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:51:35	S	SL	82.5	75	1	Swinging Strike	1	Strikeout	Willson Contreras strikes out swinging.	570632	Urena	R
19	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:52:11	S	CH	87.7	78.4	1	Swinging Strike	1	Groundout	Kyle Schwarber grounds out, first baseman J	570632	Urena	R
20	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:52:53	S	FT	94.6	86.1	1	Foul	1	Groundout	Kyle Schwarber grounds out, first baseman J	570632	Urena	R
21	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:53:20	B	FT	95	83.9	0	Ball	1	Groundout	Kyle Schwarber grounds out, first baseman J	570632	Urena	R
22	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:53:42	B	FT	95.7	86	0	Ball	1	Groundout	Kyle Schwarber grounds out, first baseman J	570632	Urena	R
23	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:54:07	X	CH	90.1	80.9	1	In play, out(s)	1	Groundout	Kyle Schwarber grounds out, first baseman J	570632	Urena	R
24	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:54:57	B	SL	85.3	76.2	0	Ball	0	Hit By Pitch	Addison Russell hit by pitch.	570632	Urena	R
25	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:55:17	S	FT	95.6	85.8	1	Foul	0	Hit By Pitch	Addison Russell hit by pitch.	570632	Urena	R
26	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:56:06	B	FT	95.6	85.3	0	Hit By Pitch	0	Hit By Pitch	Addison Russell hit by pitch.	570632	Urena	R
27	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:56:55	B	CH	89.1	78.4	0	Ball	0	Walk	Jason Heyward walks. Kris Bryant scores. /	570632	Urena	R
28	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:58:00	B	FT	94.7	83.2	0	Ball	0	Walk	Jason Heyward walks. Kris Bryant scores. /	570632	Urena	R
29	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:58:25	B	CH	89.8	79.3	0	Ball In Dirt	0	Walk	Jason Heyward walks. Kris Bryant scores. /	570632	Urena	R
30	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 16:58:57	B	FT	95.2	85.4	0	Ball	0	Walk	Jason Heyward walks. Kris Bryant scores. /	570632	Urena	R
31	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 17:00:47	B	FF	95.2	85	0	Hit By Pitch	0	Hit By Pitch	Javier Baez hit by pitch. Anthony Rizzo sco	570632	Urena	R
32	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 17:01:12	B	FT	93.8	82.5	0	Ball	1	Groundout	Jon Lester grounds out, second baseman Sta	570632	Urena	R
33	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 17:01:31	B	FF	93.1	81.8	0	Ball	1	Groundout	Jon Lester grounds out, second baseman Sta	570632	Urena	R
34	gid_2018_03_29_chnmlb_miamlb_1	2018-03-29 17:01:55	S	FT	93.6	83.5	1	Called Strike	1	Groundout	Jon Lester grounds out, second baseman Sta	570632	Urena	R

3.0 Our Strategy

Our plan is to initially approach the problem of predicting if the next pitch is a fastball or not (binary classification) in order to become familiar with the data and get confident in out models. Note that this is a problem that has been solved as detailed in the Sloan Conference paper linked at the end of this proposal. Next, we'll move onto the more interesting problem of classifying the next pitch into one of the 17 possible pitch types and 12 location zones.

a) Start with binary classification problem - fastball or not
b) Multi-classification - what specific pitch type will be thrown next (17 classes)
c) Comparison of the two models
    i) does layering the models produce more accurate results vs a singular multi-classification model
d) Evaluate LSTM model to include pitch history in prediction
e) Add location to the model. (Adds a new dimension to predict so the are 17*12 possible pitch type and location combinations).

4.0 References

[1] Bell, Alexander. "Predicting the Next Baseball Pitch: Utilizing machine learning to gain an advantage." November 1, 2018.
https://medium.com/@alexwbell/predicting-the-next-baseball-pitch-utilizing-machine-learning-to-gain-an-advantage-cef35bbd6e26

[2] Ganeshapillai, Gartheeban , and John Guttag. "Predicting The Next Pitch." MIT Sloan Sports Analytics Conference 2012, March 2-3, 2012.
http://www.sloansportsconference.com/wp-content/uploads/2012/02/98-Predicting-the-Next-Pitch_updated.pdf

[3] PI, Yifan. "Predicting Next Baseball Pitch Type with RNN." 2018.
http://www.sloansportsconference.com/wp-content/uploads/2012/02/98-Predicting-the-Next-Pitch_updated.pdf