Predicting Pitches in Baseball

STA 531

SARAH NORMOYLE

DREW JORDAN

GONZALO BUSTOS

Motivation and Data

- Prediction rather than performance metrics is a less explored area of baseball
- Valuable for a batter to know what pitch to expect before a pitch is thrown
- Data set is available publicly on MLB.com
- 3 years, from 2013 to 2015
- 2 million observations
- 39 variables
- We focus our analysis exclusively on Los Angeles Dodgers’ pitcher Clayton Kershaw, who is considered the best pitcher in baseball.

Exploratory Data Analysis

- Tables for proportions of pitches across different covariates
- Looking at changes across the columns
- Focusing on one pitcher: Clayton Kershaw (4 different types of pitches)

	CH	CU	FF	SL
0-0	0.012	0.016	0.810	0.161
0-1	0.035	0.287	0.431	0.247
0-2	0.000	0.281	0.424	0.294
1-0	0.032	0.000	0.614	0.354
1-1	0.026	0.185	0.425	0.364
1-2	0.000	0.388	0.328	0.284
2-0	0.004	0.000	0.909	0.087

Table 1: Proportion Table for Count

	CH	CU	FF	SL
0	0.013	0.137	0.581	0.269
1	0.014	0.150	0.562	0.274
2	0.016	0.166	0.566	0.253

Table 2: Proportion Table for Pre-Outs

First: Naïve Sampling and Markov model

- Naïve sampling
  - Calculating sample probability vector
  - Predictions from sampling according to these probabilities
  - Not based on previous pitch or any covariates
- Markov model
  - Creating sample transition matrix
  - Creating sample initial probability matrix
  - Predictions from sampling according to probabilities taken from transition matrix
  - Only based on previous pitch

Hidden Markov Model

- Hidden states: not as interpretable, status of game
- Observations: different types of pitches
- Baum Welch
  - Get estimates of parameters for hidden Markov model fit to pitching sequence
  - Tried different numbers of hidden states
- Forward Algorithm
  - Can calculate probability of

$$p(x_{j+1}|x_{1:j}) = \sum_{z_j, z_j+1} p(x_{1:j}, z_j) p(z_{j+1}|z_j) p(x_{j+1}|z_{j+1})$$

Multinomial Logistic Regression

- Easy to understand as a combination of regular logistic regression
- From fit of model can calculate a probability for each category j:

$$P(y^* = j | x^*, \beta) = e^{x^* \beta_j} / \sum_{k=1}^{J} e^{x^* \beta_k}.$$

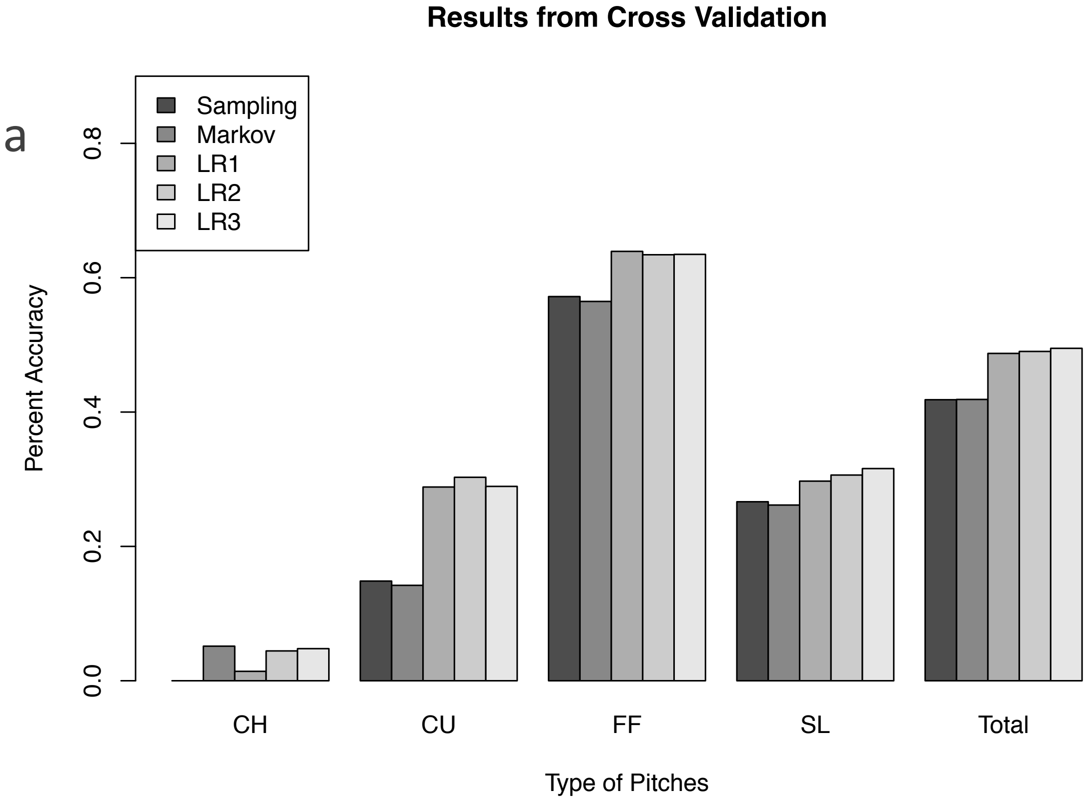
- Simple variable selection
- Reporting results for 3 models:
  - LR1: count
  - LR2: count, pre outs, inning
  - LR3: pre outs, count, pitch number, runners count, pitch count, top of inning, bat side, inning number, previous pitch type

Cross Validation Results

- 5-fold cross validation was applied to each of the techniques
- Percent accuracies of prediction were used as a comparison metric

	CH	CU	FF	SL	Total
Sampling	0.0000	0.1484	0.5719	0.2665	0.4183
Markov	0.0515	0.1420	0.5647	0.2616	0.4188
LR1	0.0140	0.2884	0.6392	0.2972	0.4873
LR2	0.0444	0.3029	0.6342	0.3062	0.4902
LR3	0.0478	0.2893	0.6347	0.3159	0.4949

Table 1: Percent Accuracies from Cross Validation



Conclusions

- Results in ability to predict baseball pitches
- Multinomial logistic regression was the most successful
- Cross validation was shown to be useful
- Clayton Kershaw’s success may partially be due to his unpredicability
- Further investigation
  - Looking into the success of simply binary logistic regression
  - Examining multiple pitchers and the differences in predicting pitches for different pitchers
  - More sophisticated techniques at variable selection