Exploring Predictors of Team Success in Ultimate Frisbee:

An Analysis of Game Statistics for Stanford Women's Ultimate

Caitlin Go
cgo2@stanford.edu

PREDICTING

Given a list of player and opponent statistics for a game, I wanted to predict if Stanford Women's Ultimate would win the game and determine the most influential features in this prediction.

In Ultimate Frisbee, a relatively young sport, there is very little use of statistics to predict game results or to analyze factors in team success. An understanding of what features are most useful in predicting success can suggest ways to start growing a body of data for the sport and beginning to analyze it.

Results show that the classifier is still a ways away from being accurate on unseen data, but the feature selection and cluster centroids do appropriately reflect elements of Ultimate Frisbee gameplay.

DATASET

The dataset consists of player and game statistics collected by Robin Davis, head coach of Stanford Women's Ultimate, and game and team statistics published by USA Ultimate from 2011 to 2017. Rows of the first dataset represent the statistics for one player per game, while rows in the second dataset represent the statistics for one opponent per game. The ground truth is the outcome (Stanford win or lose) of each game.

FEATURES

Derived features were used in place of categorical statistics in the data and in place of individual opponents and players.

The K-Means used both raw statistics - including goals, disc touches, points in the game, and national ranking - and derived statistics such as percent of game played, field position, etc.
15 player features, 12 opponent features

Feature selection used the derived tournament number and round number. Clustered features included opponent cluster and averaged stats across players belonging to each player cluster.
227 features total

MODEL PIPELINE

K-MEANS

K-Means was used to cluster the opponent and player statistics per game into types of opponents and players, minimizing using the Euclidean norm.

$$J(c,z) = \sum_{i=1}^{m} ||x^{(i)} - z_{c^{(i)}}||_2^2$$

Clusters were initialised to random samples in the training set.

Normalization weights were stored and used to expand out the cluster centers for analysis.

LOGISTIC REGRESSION

Logistic regression was used to make the win/ lose predictions per game.

The player and opponent data were first categorized by the closest cluster centroid (using Euclidean distance). Then the data was merged using a stored game index to take the following form.

x(i) =	opponent type	tourna-ment	round	# players in game in cluster 0	avg stats for players in cluster 0	•••

Using the method discussed in class, I used stochastic gradient ascent to maximize the log-likelihood of the data using the sigmoid of the dot product of x and theta as my classifier. [1]

$$max_\theta l(\theta) = max_\theta \sum_{i=1}^{m} y^{(i)} log h_\theta(x^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))$$

FEATURE SELECTION

Using the logistic regression model, I ran wrapper or forwards feature selection as discussed in class. Since this method can be slow, I chose logistic regression as a simple and fast model.

The program institutes a feature cap at 17, which was around the number of in-game statistics the head coach is currently keeping track of.

RESULTS

Clustering Results

After testing a couple of different cluster numbers, I ended up with 8 opponent clusters and 16 player clusters. Sample included below:

Opponent Centroids

	Score diff	SU Wins	Opp Wins	Ranking	Pts in Game
Weaker opponents that are blown out	11.88	1.13	0.00	129.63	14.12
Highly ranked opponents that often beat SU	3.86	.46	1.96	4.54	22.21

Player Centroids

	Position	Throw %	A. Hucks	Assists	% Played
Main throwers with high completion rates	-.98	.92	1.22	1.91	.67
Thrower/ receiver hybrids with more risky decisions	-.33	.86	5.23	3.05	.79

Feature Selection Results

LR Model	train error	test error	train accuracy	test acccuracy
n=227 (before selection)	.0002	.2656	1.0	.72
n=17 (after selection)	.0131	.2706	1.0	.64

The train set had 103 game samples, the training dev set had 24 game samples, and the test set had 25 game samples.

Feature Results

Over 17 feature additions, error was reduced from .25 to .0131 and was still decreasing. Sample below:

F=3, Cluster=14, theta=.64
% points played by strong throwers with very high completion rate and many throws, and assists has positive weight

F=4, Cluster=8, theta=1.37
completed throws per point by players with very few touches in early season tournaments has strong positive weight

FUTURE

I would like to get more data from other teams and conduct experiments exploring the feature space of the data, such as PCA. Currently, the models are tailored to one team that has historical tendencies to win/ lose in particular ways. More insight into the data would be helpful in moving on to applying more advanced machine learning techniques.

DISCUSSION

Clustering

While the cluster centroids seem to portray several known Ultimate Frisbee player stereotypes, the mean-squared cost of the clustering, which was around 5-6, seemed to be high and very variable. This variance seems to indicate that either we don't have enough data to make stable clusters, or that the data itself does not cluster neatly.

Logistic Regression

A very low train error and a high test error suggests that the model may be overfit to the train data, and that it may require a wider variety of data and more regularization to generalize well.

Feature Selection

Here, a low train error and a high test error are indications that there is also high variance in the model, a problem that could potentially be fixed by removing the artificial cut-off at 17 features and adding in more data samples. The top 17 features and their corresponding weights were mostly expected - for example, the weights for drops by certains player types is negative while the weight for goals by certain player types is positive.

Summary

While it was nice to see that Ultimate Frisbee knowledge did translate to the clusters and features, the poor test accuracy (worse than the 74% chance of being right if you guessed that the team would win every time) was disappointing albeit unsurprising given the small amount of data and the simplicity of the model, which was chosen to ensure human readability of features.

REFERENCES

[1] A. Ng, "CS 229 Class Notes," CS 229. Stanford University, Stanford. Class Notes. [Online]. Available: http://cs229.stanford.edu/notes/cs229-notes1.pdf/. [Accessed Dec. 10, 2017]