Data-Driven Insights into Football Match Results

By Kevin Bishop ❖ *kbishop@stanford.edu*

The Goal:

To predict the outcomes of EPL matches using data available before the match. I used statistics from the previous year for each team. The data is in medium depth, including things like shots and corners for and against, but does not contain advanced statistics like chances created or possession percentage. This proves I can take fairly limited data and turn it into an accurate predictor of game outcome.

The Playing Field:

My dataset contains results from every EPL match from the 2000-2001 season through the 2012-13 season, a total of 4768 matches. The raw feature set includes full-time scoreline, halftime scoreline, and for/against stats for shots, shots on target, corners, fouls, yellow cards, and red cards. Each match is also labeled with home team, away team, year, and match result (the ground truth).

The Players:

My final feature set includes 16 computational features: the difference in previous-year averages of fulltime goals, halftime goals, shots, shots on target, corners, fouls, yellow, and red cards (for and against for each). Each is derived from the raw features, but none are the raw features themselves. What makes the difference in a match is how the strengths of one team match up against the weaknesses of the other, and vice versa, so naturally the differences between home and away team statistics in various areas, as best you know them before the match, stand to be good predictors of match outcome.

The Formations:

1. Multivariate Gaussian Naïve Bayesian Model

Parameters:

$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=y\} + 1}{m+3}$$

$$\phi_{x_j|y} = \frac{1}{\sqrt{2\pi\sigma_{j|y}^2}} exp\left(-\frac{(x_j - \mu_{j|y})^2}{2\sigma_{j|y}^2}\right)$$

Objective:

$$argmax_y \sum_{j=1}^{n} log\ p(x_j \mid y)\ log\ p(y)$$

2. Kernelized SVM using RBF kernel

Objective:

$$J(\alpha) = \frac{1}{m}\sum_{i=1}^{m} L(K^{(i)T}\alpha,\ y^{(i)})$$

$$L(z,y) = max(0,\ 1 - yz)$$

Stochastic Gradient Descent:

$$\alpha \leftarrow \alpha - \eta y^{(i)} K^{(i)} 1\{y^{(i)} K^{(i)} \alpha \le 1\}$$

Even though the SVM is a binary classifier, I solved the three-class problem using a Win vs. Draw/Loss SVM and a Win/Draw vs. Loss SVM. By running each pipeline individually, and then combining the resulting predictions using a logical 'and' function, we can produce a three-classifier of wins, draws, and losses.

The Results:

Training set size: 4180
Test set size: 208

	Training Error	Test Error
Naïve Bayes	0.5787	0.5913
SVM	0.4787	0.5240

Naïve Bayes	Actual Win	Actual Draw	Actual Loss
Predicted Win	14	11	13
Predicted Draw	13	18	24
Predicted Loss	31	31	53

SVM	Actual Win	Actual Draw	Actual Loss
Predicted Win	26	20	21
Predicted Draw	15	13	11
Predicted Loss	17	27	58

Post-Match Analysis:

The outcomes of sporting events are notoriously hard to predict; in fact the entire sports betting industry relies on the unpredictability of these outcomes. The highest accuracy I saw in my background research was around 50% for 2-classification, so I was pleased to achieve around 50% accuracy for 3-classification using the SVM. Along with the results themselves, a big insight I had is that the ingredients of victory vary a lot between leagues. I originally planned on using game data from England's top 4 leagues, but the resulting models actually turned out to be worse due to the differences in what makes a successful team in each league. Also, I could make money using my results! If I were to bet on my predictions whenever the odds for that result were better than 1:1 (which is common), I would make money in expectation.

Next Fixtures:

With 6 more months, I would like to develop a feature mapping that uses year-to-date stats rather than stats from the previous year. With bias being the primary component of error in my models, increasing the relevance of the feature set to the prediction stands a good chance of improving accuracy. I would also like to develop a team rating feature, to increase the margin of games between strong and weak teams, even when the strong team is away from home.

Backroom Staff:

- "Predicting Soccer Results in the English Premier League". Web. 21 November 2016. <https://goo.gl/Zdj87n>.
- "Football (Soccer) Data for Everyone". Web. 21 November 2016. <https://github.com/jokecamp/FootballData>.
- "Historical Football Results and Betting Odds Data". Web. 10 December 2016. <http://www.football-data.co.uk/data.php>.
- "How I Used Machine Learning to Predict Soccer Games for 24 Months Straight". Web. 16 November 2016. <http://doctorspin.me/2016/03/21/machine-learning/>.