Soccer Predictor

Owen Chapman

Patrick Shao

Huu Le

Goal

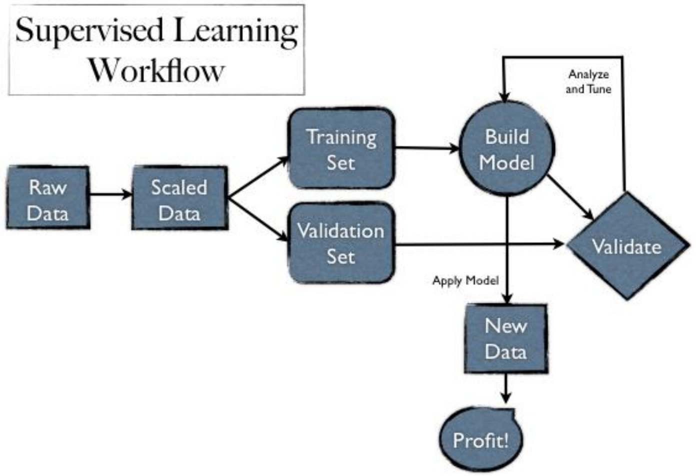
- Given multiple years of data of Soccer Premier League, can we create a classifier that can accurately predict the outcome of a match between two given teams?

Motivation

- Potential to develop an app that can accurately predict game outcomes
- Possible use for tournament organizers, sports analysts, and betting markets
- $$$

Methods

- Analyze historical data in order to generate outcome predictions for match-ups
- Supervised Learning:
  - Easy to check accuracy
  - Problem set up to have labels/features easily
- Labels: Match Outcomes
- Features: Various metrics from past games
- Implemented and tested with multiple algorithms

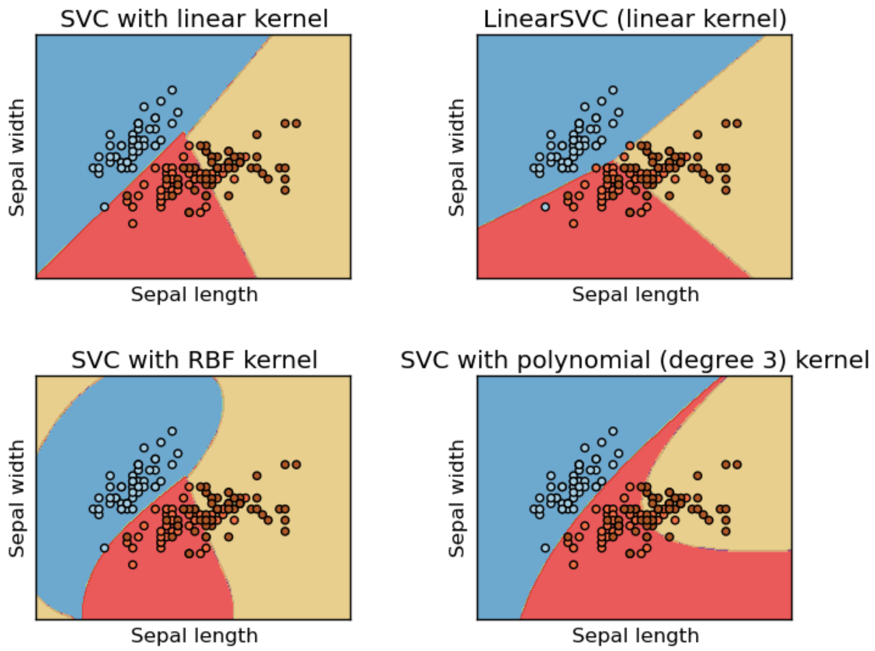


Data Parsing

- Utilized data from online databases and APIs - football-data.co.uk
- Obtained 15 years of the English Premier League and 10 years of Spanish La Liga
- Contained stats on both teams – goals scored, shots taken, fouls, etc
- Parsed data into feature vectors

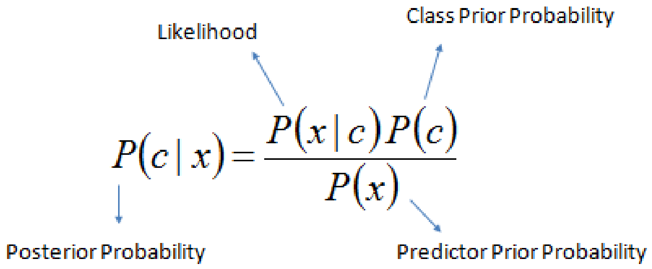
Data Sets

- Training Set: EPL seasons 2000-2009
- Validation Set: La Liga seasons 200-2004
- Testing Set: EPL seasons 2011-2014



Algorithms Used

- Gaussian Naïve Bayes
- Multinomial Naïve Bayes
- Support Vector Machine
- Perceptron
- Stochastic Gradient Descent



Likelihood

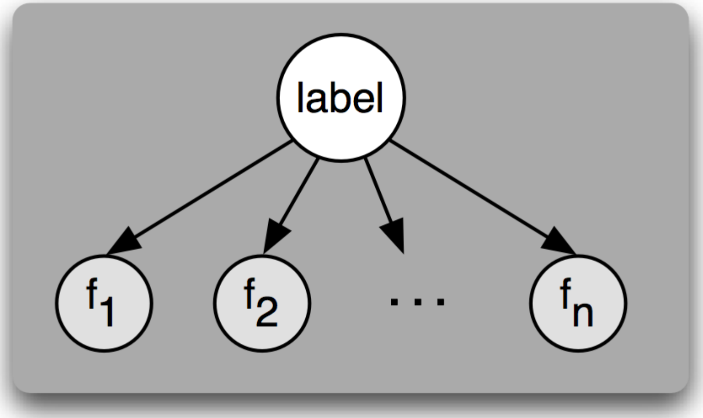
Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$



Testing Set Accuracy

Algorithm	Base Features	Shots on Target	All Features
Gaussian Bayes	42.39%	46.28%	47.99%
Mult. Bayes	42.39%	47.37%	48.39%
SVM	42.39%	42.13%	44.96%
Perceptron	37.25%	25.66%	29.07%
SGD	33.02%	25.56%	25.87%

- Base Features:
  - Head to Head, previous game's labels & scores
- Shots on Target
  - Shots made on target/opponents shots for 2 previous games with same opponents
- All Features
  - Result of previous games, goals scored, opposing goals scored

Effectiveness

- Managed to be higher than 33.33% Accuracy (Guessing)
- Use of all features proved more effective than use of select features/single feature
- Our best results managed to best our naïve (just heads to heads)
- Ideally wanted to compare to betting odds to see how our code fares against humans(but couldn’t)

Picture Credits

- http://kindersay.com/files/images/soccer-ball.png
- https://www.dudleytrophy.com/wp-content/uploads/wp-checkout/images/lil-buddy-soccer-trophy-lbr18.jpg
- http://dreamatico.com/data_images/money/money-1.jpg
- https://skitch-img.s3.amazonaws.com/20100213-djhg1re7gaj83ngygcqgj1jm2d.png
- http://www.saedsayad.com/images/Bayes_rule.png
- http://www.nltk.org/images/naive_bayes_graph.png
- http://adquadrant.com/wp-content/uploads/2014/03/Data.jpg
- http://scikit-learn.org/stable/_images/plot_iris_0012.png