Data Mining Methods for Sports Prediction

Jacob Mitchell
University of Minnesota, Morris
4/15/17

Why?

- Scouting and season analysis
- Coaches and managers can use this info to find optimal lineups versus given opponents
- Sports betting

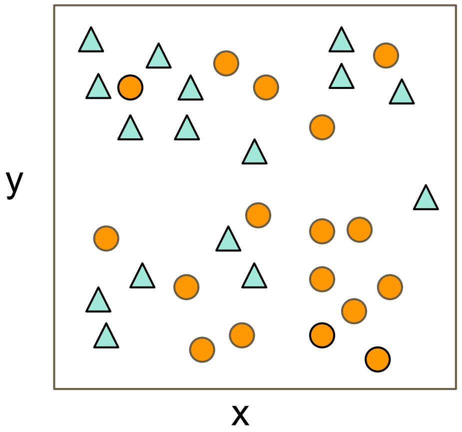
Outline

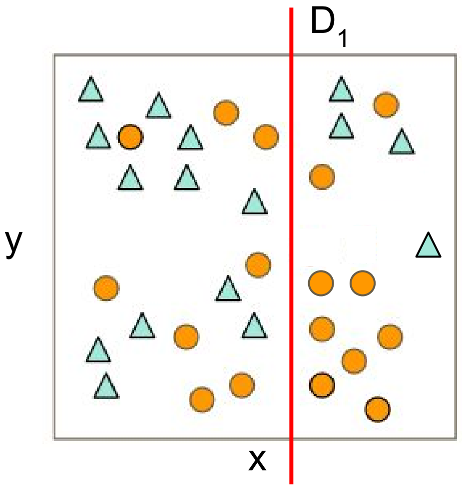
- Background
  - Random Forests
  - Neural Networks
- Trials
  - Rugby
  - English Premier League
  - Multiple Leagues
- Results and Discussion

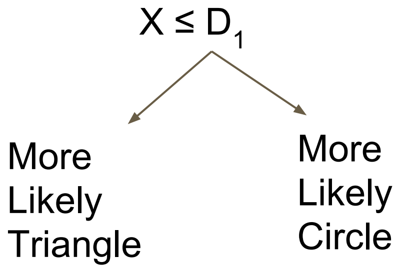
Outline

- Background
  - Random Forests
  - Neural Networks
- Trials
  - Rugby
  - English Premier League
  - Multiple Leagues
- Results and Discussion

Decision Trees

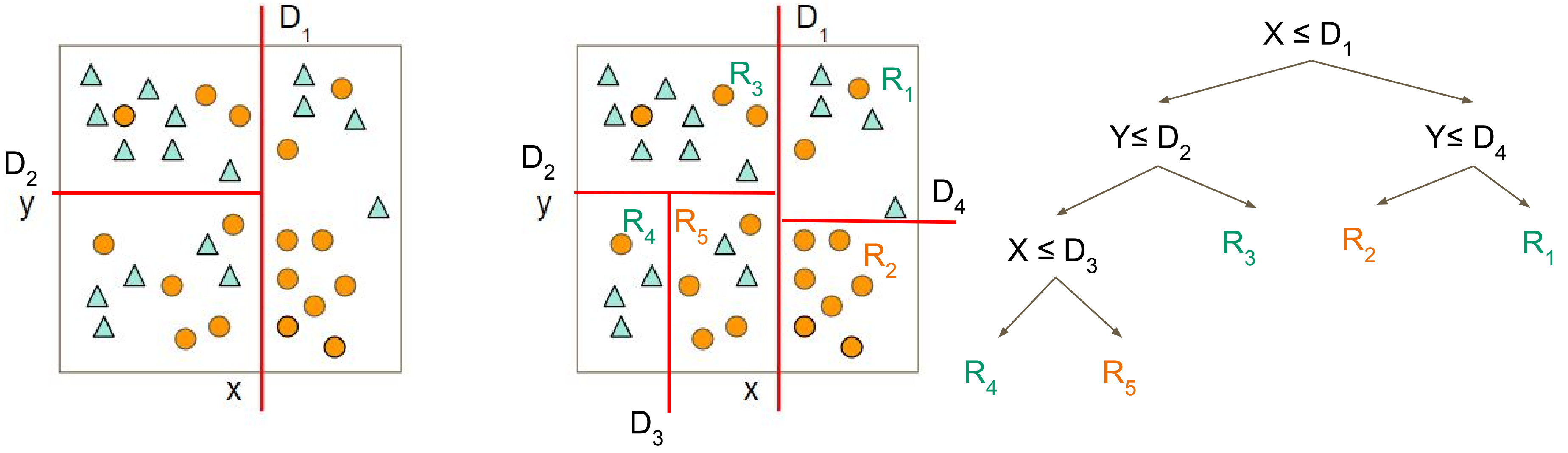






Left in tree means true, right means false

Decision Trees



Forest Building

- Takes a group of decision trees and their outputs
- Voting is done on these outputs and the majority is chosen as final output
- Randomization can come from differences in tree divisions or input data

Breiman’s Random Forest

- Breiman is generally considered the creator of random forests how we use them today
- Selects a random subset of the data for each tree- feature bagging

Outline

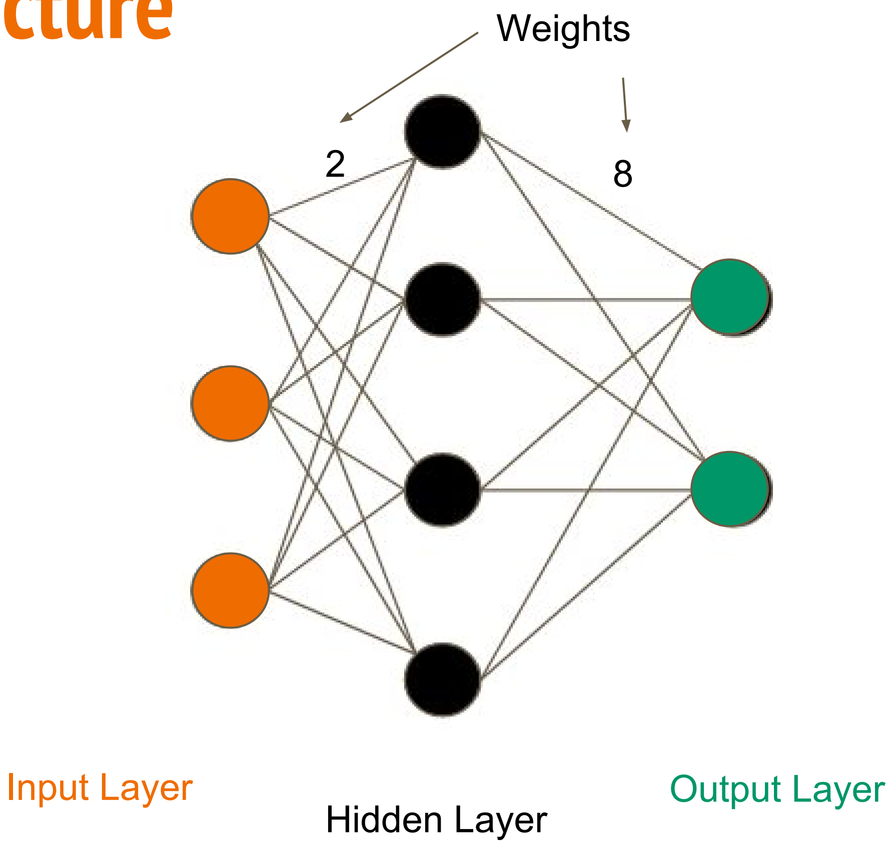
- Background
  - Random Forests
  - Neural Networks
- Trials
  - Rugby
  - English Premier League
  - Multiple Leagues
- Results and Discussion

Structure

Three Layers

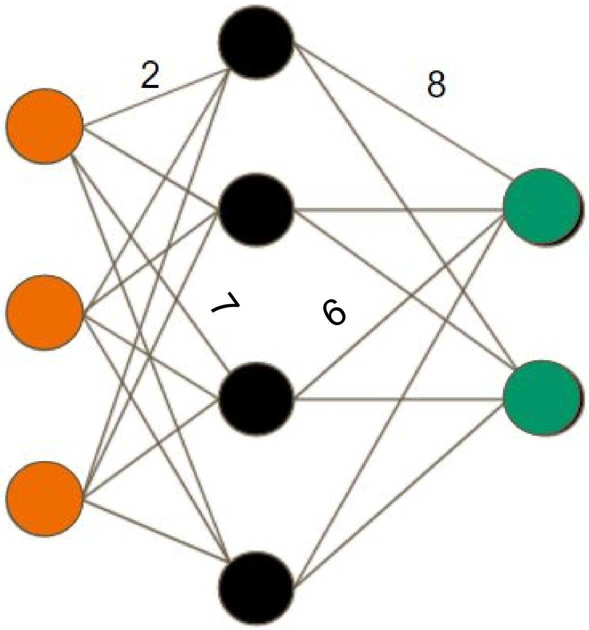
1. Input Layer
2. Hidden Layer
3. Output Layer

- Weights connect the layers and show importance of given nodes



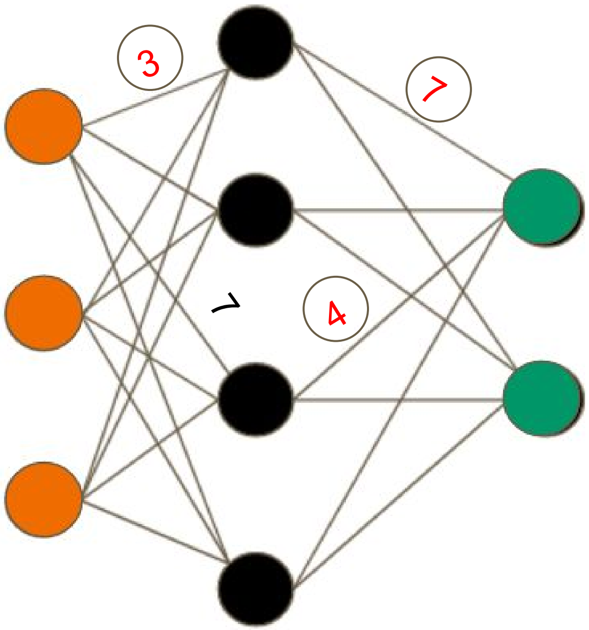
Training

- Features- data selected for training
- Involves running algorithm multiple times to produce optimal weights of the nodes
- Each run reassigns weights based on new data and adjusts accordingly



Training

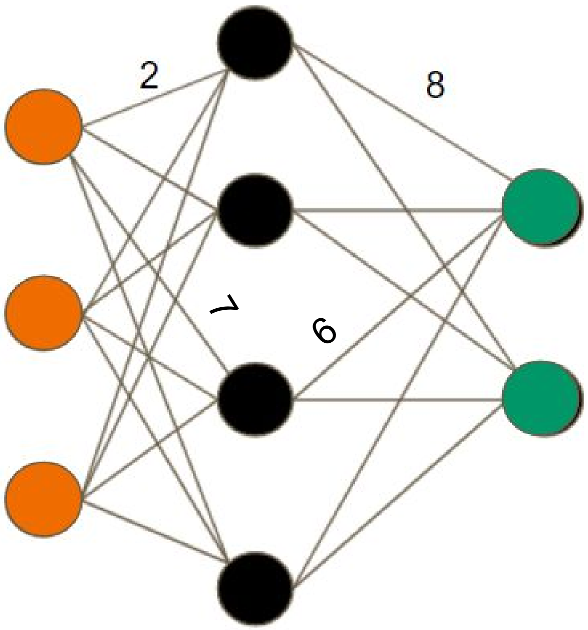
- Features- data selected for training
- Involves running algorithm multiple times to produce optimal weights of the nodes
- Each run reassigns weights based on new data and adjusts accordingly



Backpropagation and MLPs

- Used for training the network
- Repeats a 2 part cycle: propagation and weight update
- Propagation:
    - Input is shuffled through the network to the output layer
    - Output is compared to desired result
    - Error value is calculated for each node in output layer
    - Error values are propagated backwards until each node has an associated error value
- Weight Update
    - The error value at each node is used to update the weight between nodes
- MLP (Multilayer Perceptron)
    - Each layer of the network is connected fully to the next

Backpropagation Example



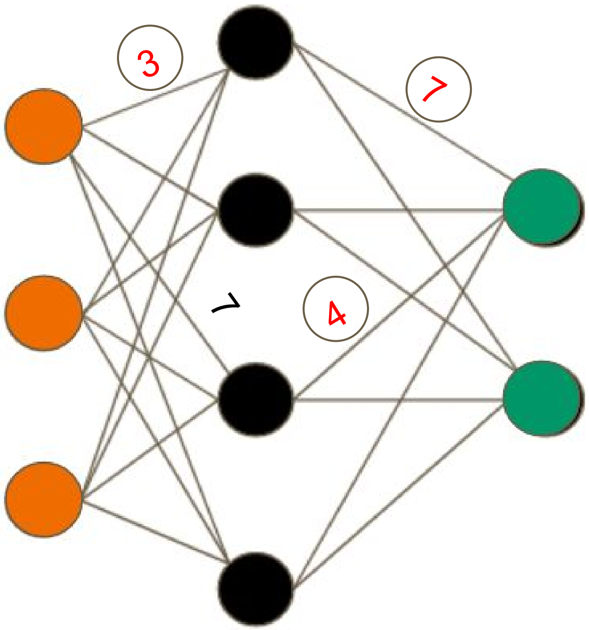
- Compares initial output to desired output

- Error value is assigned to each output node

- Error = x - y
  - x is desired output
  - y is actual output

Backpropagation Example

- Error values are propagated back through the network
- Weights are updated to account for the error values assigned to each node

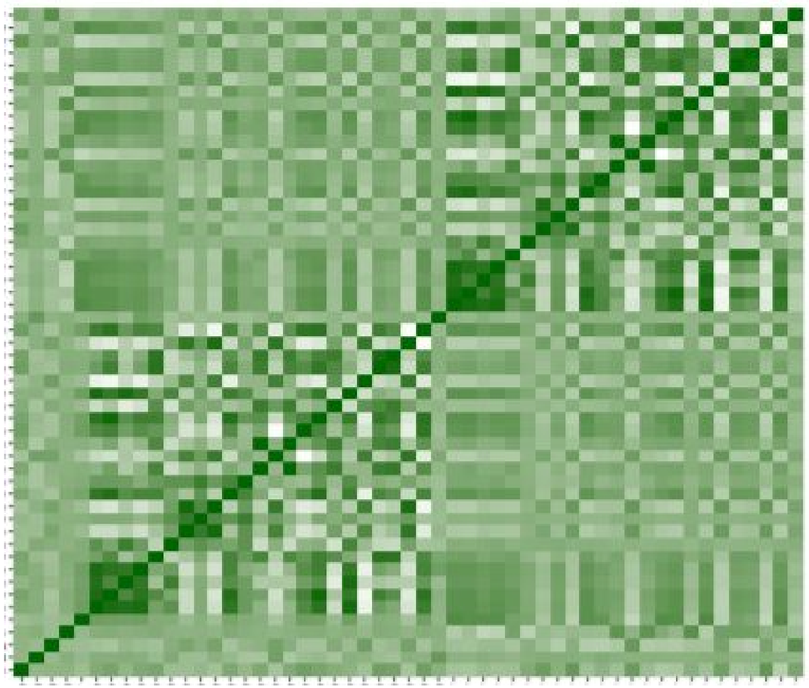


Outline

- Background
    - Random Forests
    - Neural Networks
- Trials
    - Rugby
    - English Premier League
    - Multiple Leagues
- Results and Discussion

Data

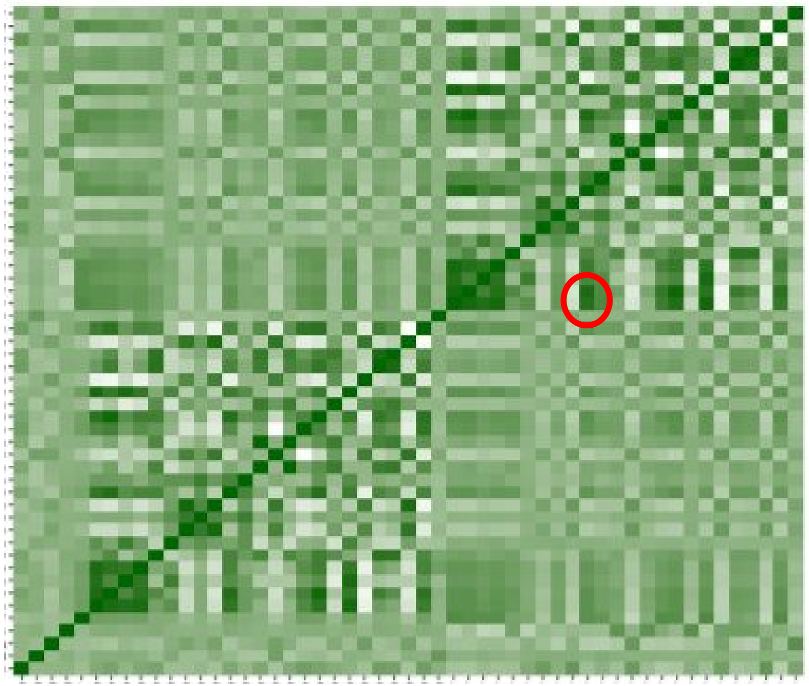
- Pretorius and Parry (2016) tested on every 2nd match from the start of 2015 to start of the tournament (late 2015)
- Examples of features: outcome, home/away, rank-home, rank-away
- x and y axis are identical lists of the features



Pretorius and Parry (2016)

Data

- Pretorius and Parry (2016) tested on every 2nd match from the start of 2015 to start of the tournament (late 2015)
- Examples of features: outcome, home/away, year, month, largest points scored home/away, rank-home, rank-away
- Ex: games drawn away, games won away



Method

- Breiman's Random Forest RI (Random Input)
    - Uses orthogonal splits of the variable space
- Chosen on fast training time (14.32 secs) and low test error (19.05%)
- Ensemble size 200
- Input data was updated after each completed match

Results

- Authors prediction- human methods (SuperBru and OddsPortal) would be superior (null hypothesis)
- Conclusion- evidence showed random forests were at least as accurate

Approach	Correct	Accuracy
Breiman Forest-RI	43/48	89.58%
OddsPortal	41/48	85.42%
SuperBru	41/48	85.42%

Outline

- Background
  - Random Forests
  - Neural Networks
- Trials
  - Rugby
  - English Premier League
  - Multiple Leagues
- Results and Discussion

Data

- Kyriakides, Talattinis, and George use aggregations of data from www.football-data.co.uk (2014)
- Targets for machine learning approaches were sum of goals and matches up to the current game
- Training set was always number of matches played in current season

Methods

- Breiman’s Random Forest
  - Uses random subset (around 66%) to train each tree
- Neural network: multilayer perceptron using backpropagation for learning
  - Starts with random weights for each weight and updates based on the delta rule
- Predicted win, loss, or draw

Results

- Random forests were far superior in hindsight prediction
- Neural networks were better at foresight especially when focused on profit
- Both methods at least slightly more accurate than linear algebra methods also tested

Hindsight

Season	RF	NN
2010/2011	94.74%	51.32%
2011/2012	96.32%	50.53%
2012/2013	95.79%	45.79%

Foresight

Season	RF	NN
2010/2011	41.58%	46.32%
2011/2012	37.89%	46.84%
2012/2013	48.42%	50.53%

Outline

- Background
  - Random Forests
  - Neural Networks
- Trials
  - Rugby
  - English Premier League
  - Multiple Leagues
- Results and Discussion

Australian Football League

English Premier League

National Rugby League

International League

Data and Methods

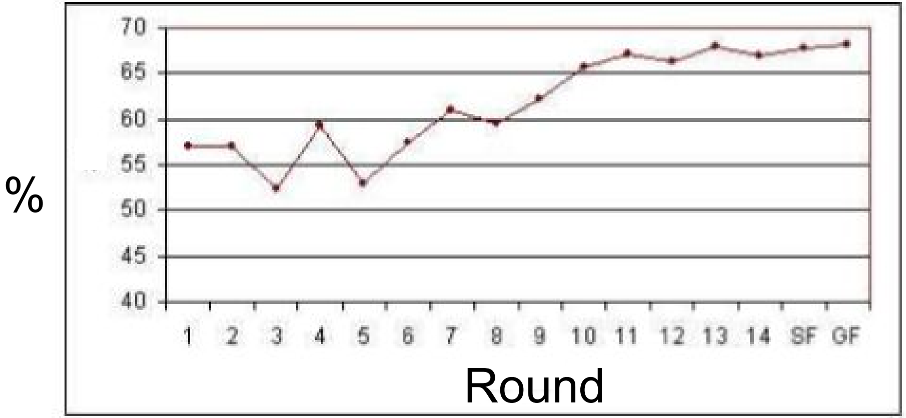
- McCabe and Trevathan (2008)
- Multilayer perceptron using backpropagation
- Features:
  - Points-for
  - Points-against
  - Win-loss record
  - Home-away record
  - Previous game result
  - Previous *n* game performance
  - Team ranking
  - Points-for and against in previous *n* games
  - Location
  - Player availability

Results

McCabe and Trevathan (2008)

- Showed expected growth of course of season- early rounds show how random weights affect predictions
- Super Rugby- 2 new teams introduced- algorithm adjusted quickly

League	Best	Worst	Average
AFL	68.1%	58.9%	65.1%
NRL	67.2%	52.2%	63.2%
Super Rugby	75.4%	58.0%	67.5%
EPL	58.9%	51.8%	54.6%



2006 Super 12

Outline

- Background
  - Random Forests
  - Neural Networks
- Trials
  - Rugby
  - English Premier League
  - Multiple Leagues
- Results and Discussion

Comparisons

- Neural Networks- accurate in foresight prediction and profitable in betting
- Random Forests- hindsight prediction accuracy, showed in some cases to be profitable in betting
- Both: were at least slightly superior to both human and linear algebraic methods at predicting results

Applications

- Random Forests- scouting, season analysis, possible betting profitability

- Neural Networks- lineup optimizations, seems to be a definite possibility as a betting tool

Questions?

Works Cited:

G. Kyriakides, K. Talattinis, and S. George. Rating systems vs machine learning on the context of sports. 2014.

A. McCabe and J. Trevathan. Artificial intelligence in sports prediction. 2008.

A. Pretorius and D. A. Parry. Human decision making and artificial intelligence: A comparison in the domain of sports prediction. 2016. ACM.