Project Proposal: Deep learning methods for premier league predictions

Arman Tadjrishi
Udacity Machine Learning Engineer Nanodegree

19/03/2019

1. Domain Background

The field of sports betting and gambling is one that has been growing steadily in the last decade**??**. Gamblers take up wager offers from the bookmaker (the house) and qualify for the winnings only when their bet wins.

In recent years, bookmakers have made use of a multiple sophisticated strategies in order to set odds at a level where the least exposure to risk (hence the most chance of steady profit) and the maximum attractiveness to clients. These strategies use massive amounts of collected data, often live feeds for in-play odds, with a host of machine learning algorithms to calculate the odds. Clients involved in betting are often layman to the underlying mathematical complexities of the odds and often have no means of accurately judging the likihoods of outcomes.

The most popular and developed sport in the world is Football, and arguably the English Premier league is the most competitive, and it is therefore the most natural choice to study. The game of football is constantly evolving, and features that may be important in one generation may become less so later - there has never been an analytic definition of the game. This is where machine learning could prove invaluable.

2. Problem Statement

The question that is brought about is whether can utilise machine learning and the depth of data available to effectively predict outcomes of football matches, and to that extent, can we build a machine learning agent that can effectively judge the amount of money to place on particular particular bets using what the agent has learned and the data available, to maximize profits?

The premise then lies, to build a machine learning framework, that can use historic data from football matches between two teams and learn how to best predict outcomes of games. The next step is to investigate whether an agent can use the learned information to make betting decisions in order to maximise profit (or minimise loss) using historic odds data.

3. The Data-set

Premier league data is sourced from football-data.co.uk**??**, individual match data is available from this website, with additional high-level statistics that describe the match. Below is an outline of the fields available from this data. The data present will need to be manipulated and target-feature sets can be built to feed through the algorithms.

Column	Meaning
Div	League Division
Date	Match Date
HomeTeam	Home Team
AwayTeam	Away Team
FTHG and HG	Full Time Home Team Goals
FTAG and AG	Full Time Away Team Goals
FTR and Res	Full Time Result
HTHG	Half Time Home Team Goals
HTAG	Half Time Away Team Goals
HTR	Half Time Result

Column	Meaning
Attendance	Crowd Attendance
Referee	Match Referee
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
HHW	Home Team Hit Woodwork
AHW	Away Team Hit Woodwork
HC	Home Team Corners
AC	Away Team Corners

Column	Meaning
HF	Home Team Fouls Committed
AF	Away Team Fouls Committed
HFKC	Home Team Free Kicks Conceded
AFKC	Away Team Free Kicks Conceded

(0)

4. Solution Statement

The aim of this project is to be able to effectively predict the outcome of a football match, with details that are only available before the match. Knowing this, we must understand what our algorithm inputs will be. Although not finalised, it is proposed that a single match will be the input into the algorithm, with features that accompany it such as 'Previous Home Result', 'Previous Away Result' and 'Seasonal Goal Difference'. There are an almost limitless number of potential features, so it will be interesting to measure which features are important.

The reason why a Deep Neural Network was chosen is because of all the potential combination/non-linear relationships between the different data fields.

5. Previous Benchmarks

It is difficult to source state-of-the-art projects such as this as they are often done by sports gaming companies with incentive to keep their results proprietary. However there are a few that we can talk about. 'xG' is an alias for models that predict expected goals of a football match or a number of games. These have become fairly popular as for late and are used by amateur data-scientists to make predictions. Mackay Analytics**??** described some these, some models have been shown to have a root-mean-square error of 0.75. soccerlogic**??** demonstrates a simple model that has 0.85 validation accuracy and correctly predicts 60 % of goals.

6. Evaluation Metrics

Our model will aim to initially classify the results of the match as a True/False output for the home team victory. Simply, the best evaluation metric will be the accuracy of the result. Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ We are less concerned with Precision and Recall, because our predictions are not based on dependencies on False positives/ True negatives. The focus is to generate a true or false result.

On top of this, we can compare accuracy to a truly random through hypothesis testing. So a potential evaluation metric would be the χ^2 test:

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k}$$

7. Project Design

The project will attempt to use a deep neural network in order to classify the results of football matches. These neural networks will have to their hyperparameters evaluated, and potentially this can be done through a confusion matrix.

The input to the DNN will be the relevant match information that can be obtained before a game, for both the home and the away sides. A data set must be made, by sampling from the database, a set of features such as the ones mentioned in section 3 will be used as input. One-hot-encoding will be required to properly quantify some of the fields, for example the score of the game should be one-hot encoded as there are not that many potential combinations.

The target data will be the result of the match, aka whether the home-side won or not. The actual design of the neural network, i.e. the number of layers, activation functions and optimizer algorithm will be the subject of experiment and will be reflected in the final report.