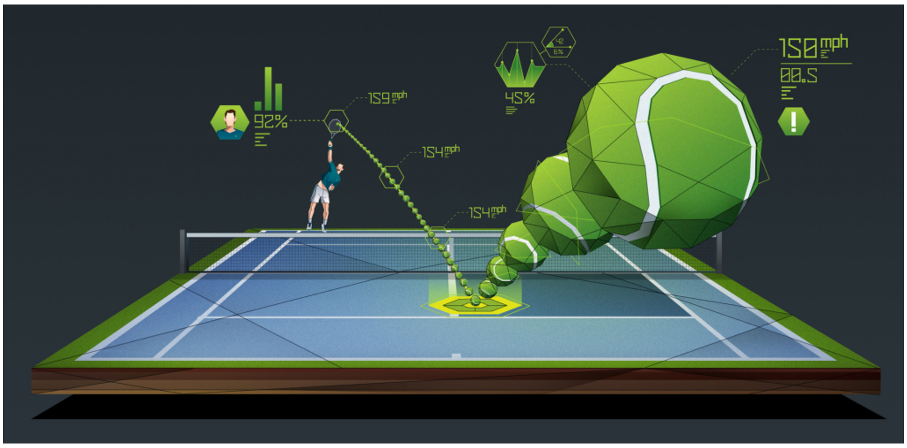
Machine learning models for sports prediction and betting

Proposer: Lorenzo Amato (Università degli studi di Firenze & Universidad Complutense de Madrid)



Introduction

Between all mainstream sports tennis is, for two reasons, probably the most suitable for using machine learning models to predict results. The first reason is that there is no tie in tennis: the match goes on until the detection of the winner. That means that our problem is a binary classification problem. The second reason is that being played between only two players, the number of variables that intervene is relatively low (compared, for example, to the 22 players who influence a football match). Predictions of tennis results using mathematical models began in the 2000s. The first models were logistic regressions that used only the two-player ranking as a feature. In subsequent papers there were developed features as historical averages of particular aspects of the game of each player, such as the average number of *aces* made per game.
In the Modelling Week we aim to develop machine learning models that can improve the predictions made by bookmakers and, secondly, use these predictions to generate a profitable betting strategy.

Work plan

The work for predicting the result of a match passes through 4 phases:
1. **Data extraction.** On the website https://github.com/JeffSackmann/tennis_atp we can find the statistics of the games played in the last 20 years. On the site http://www.tennis-data.co.uk/alldata.php we can find the odds of some bookmakers for the same games.
2. **Feature Engineering**. The features must be created on each player for each moment in time. At this stage, great care must be taken not to include the data of the game itself in the features that we will use for the prediction.

3. **Train machine learning models**. The main metric used in the literature is the accuracy (the number of times our model makes correct prediction over the total number of attempts). The bookmakers achieve accuracy between 63% and 68%. Accuracy of above 68% is a very good result. Accuracy above 72% can be a sign of improper use of future data.
4. **Betting strategy.** Ultimately, we can develop a betting strategy and evaluate our model by betting against a bookmaker. In this case, the main metric used to evaluate the models in the literature is the Return on Investments.

Recommended literature

We recommend reading the following 3 articles. The first, being a thesis, explains step by step how the modeling is done, but it uses a dataset not available for free that contains the player statistics already calculated and for this reason it does not contain the construction part of the dataset. The second one is particularly useful for taking a cue on the features to be built. The last one provides a comparison between many models.

1. Sipko, M., & Knottenbelt, W. (2015). Machine learning for the prediction of professional tennis matches. *MEng computing-final year project, Imperial College London*.
2. Gu, W., & Saaty, T. L. (2019). Predicting the outcome of a tennis tournament: Based on both data and judgments. *Journal of Systems Science and Systems Engineering*, *28*(3), 317-343.
3. Wilkens, S. (2020). Sports Prediction and Betting Models in the Machine Learning Age: The Case of Tennis. *Available at SSRN 3506302*.

Pre-requisites

Basic statistics knowledge on variance and correlation between variables. Manipulation of data by using any program language. Machine learning models: theory of at least 2 models and application with any program language.