University of Liège

Big Data Project

Review 1 - Pre-analysis, literature review

Master 1 in Data Science & Engineering

Authors :
Tom Crasset
Maxime Lamborelle
Antoine Louis

Professors :
G. Louppe
P. Geurts
B. Cornelusse

Academic year 2018-2019

Introduction

Tennis has become one of the most popular sports in the world and its popularity keeps growing. Each year, a multitude of tournaments are held by the Association of Tennis Professionals (ATP) in 30 different countries. The matches attract a huge amount of spectators, both in person in the stadium and behind their screens, and enormous amounts of money are bet on either one of the competitors. This has lead to a big increase in the research focusing on tennis match prediction for a few different reasons. The decision of which player is likely to win is important because of the financial assets involved in the betting process. Also, sports managers can benefit from a model that can determine a certain weakness or strength of the opponent that might not be apparent to the naked eye. Many of this is linked to the fact that player and match data has been made increasingly (publicly) available.

The goal of this project is to investigate the possible methods to best predict the outcome of one of the most prestigious tournaments in the tennis world, the Roland Garros tournament, also called French Open. The Roland Garros tournament is held each year in June in the stadium of the same name in the iconic city of Paris.

The scoring system has a hierarchical structure, where matches are composed of sets, which are in turn composed of games and these games are also divided into individual points. Some prediction approaches rely heavily on this hierarchical structure to define expressions for the probability of each of the competitors to win a certain game. By supposing that the points are independently and identically distributed (iid), one only needs to compute the probability of each competitors winning the points on their serve. Using this and the massive amount of data that is available, the probability of a certain player winning a game ca be determined, then a set, and lastly the match.

Other more performing models are based on numerous other factors involved in the games, such as the results of historical matches, player performance indicators, and opposition information.

Pre-analysis

Type of model

The first step of the process is to understand and formalize the problem. A lot of documentation about sport result prediction is available on the web, including tennis. Unfortunately, it only concerns the result prediction of particular matches. Using some data as inputs, would it be possible to implement a model that would predict, not the winner of a single match, but the winner of an entire tournament ? After some consideration, it doesn't seem obvious at all.

Then, the obvious solution was to use a match by match model to predict our tournament's winner. That choice would imply that all the possible draws must be considered. For each one of them, each match of the first round would be analyzed and would lead to a set of predicted winners that would become the players of the second round. Then again, each match of the second round would be analyzed and predicted winners would become the players of the next round. One thing leading to another, the winner of each possible draw would be predicted and probabilities on each winner of all draws could tell which player has the most chance to win the tournament. That way of thinking could seem relevant, but it actually raises a problem. To understand it, the concept of how the players are selected must be clear.

Roland Garros draw

In the Roland Garros draw, 128 tennis players are taken into account. Among them, there are :

— The world's best 104 players according to the official ATP ranking, which is stopped by the Roland Garros' organizers six weeks before the tournament.

— 16 players ranked outside the TOP 100 and that come from the qualification draw. The qualifications begin five days before the real tournament and welcome 128 players among which 16 access the real tournament.

— 8 players ranked outside the TOP 100 that receive a "wild-card", an exclusive invitation to participate in the real tournament without having to pass the qualifications. These cards are mostly reserved for ex-champions that have suffered an injury recently and have lost a lot of points in the ranking.

From the 128 players of the tournament, the 32 best players are said to be "seeds". The goal of the seeds is to "protect" a minimum the best players in the first rounds of the tournament. Thus, each of the 32 seeds can not face another seed before the third round.

Once the selection of the players for Roland Garros is clearly understood, the problem in question seems obvious. As 128 players participate in the tournament, the number of all possible first rounds is simply a permutation of these players, divided by two because the order of two players for a match doesn't matter. It means that there are $\frac{128!}{2}$ different draws, which is way too big and impossible to consider. Another approach must be found...

By sticking with our match by match model, one way to remedy this problem could be to use some previous data to predict the first round of the tournament. Indeed, a tennis draw isn't really random. As explained above, the organizers ensure that the seeds do not meet before the third round. Furthermore, the distribution of the seeds in the draw is made so that :

— In the 3rd round, the seeds 1 to 8 face one of the seeds 25 to 32, drawn by lot, and the seeds 9 to 16 face one of the seeds 17 to 24, drawn by lot.
— In the round of 16 (4th round), the seeds 1 to 4 face one of the seeds 13 to 16, drawn by lot, and the seeds 5 to 8 face one of the seeds 9 to 12, drawn by lot.
— The seeds 1 and 2 can only meet in final and they theoretically face the seeds 3 or 4 (according to the draw) in the semifinals.

These claims are theoretical and they consider that it should normally be the best seeds that win their matches. So by knowing that a tennis draw is not completely random but arranged in order for the best players to play the furthest in the tournament, the number of possible draws could be significantly reduced by constructing them backwards, following the rules described above.

Selection of the players

The last point concerns the choice of the 128 players for which the match by match model will be trained. For that, a briefing about how the ATP ranking works must be done. The ATP Rankings are the objective merit-based method used by the Association of Tennis Professionals (ATP) for determining the qualification for entry as well as the seeding of players in all singles and doubles tournaments. The rankings are updated every Monday, and points are dropped 52 weeks after being awarded. Ranking points are awarded according to the stage of tournament reached, and the prestige of the tournament, with the four Grand Slams awarding the most points. Points awarded for the different tournaments at each stage reached are displayed in the Table 1.

Tournaments	W	F	SF	QF	R16	R32	R64	R128
Grand Slam	2000	1200	720	360	180	90	45	10
Master 1000	1000	600	360	180	90	45	10 (25)	(10)
500 Series	500	300	180	90	45	(20)	/	/
250 Series	250	150	90	45	20	(10)	/	/

TABLE 1 – Points awarded at each stage of the different tennis tournaments

After doing some research on which tournaments will happen between now and the time the organizers of Roland Garros freeze the ranking for their draw, which is, as said above, 6 weeks before the tournament (15th April 2019) takes place, it can be found that there will be 1 Grand Slam, 2 ATP Master 1000, 4 ATP 500 Series and 15 ATP 250 Series. Knowing that the ranking difference between some players can sometimes be of a few points, a lot of things can happen during these 22 tournaments and some players ranked outside the TOP 100 could easily get in. Predicting who will be the 128 players competing Roland Garros is thus the hardest problem. One solution could be to analyze the 128 players selected for previous big tournaments and to predict the ones that will be selected for Roland Garros 2019.

Data

There are a lot of statistics available on the web for tennis. At first, web scraping was a possible solution to collect these data but fortunately, some websites directly provide huge data sets such as `www.tennis-data.co.uk`. The next operation consists in pre-selecting the information needed to :

— Build the possible draws of the tournament.
— Predict the result of each match.

In Table 2, the first pre-selection of the data is presented by category. These are the data we will focus on and, if necessary, other data can be added.

Player detail
Name
Date of birth
Country of birth
ATP rating points over time
ATP rank over time

Match details
Tournament name
Tournament type (e.g., Masters, Grand Slam)
Tournament state (e.g., final, semi-final)
Surface
Location
Weather
Date
Results
Prize money
Bet Odds

Statistics for both players
First Serve percentage
Aces
Double faults
Unforced Error
Percentage of points won on first serve
Percentage of points won on second serve
Percentage of receiving points won Winners
Break points (won, total)
Net approaches (won, total)
Total points won
Fastest serve
Average first serve speed
Average second serve speed

Table 2 – Data to consider for our match by match model

To analyze all these data, a database is needed to have a specific location to store all the datasets that could be gathered, for the possible data cleaning that could be done and to select among all the data only the one needed for our machine learning model. The choice of a NoSQL database is justified because multiple data formats can be accepted without changing the layout of the data just to be able to insert it into a database. As for the specific database choice, it is MongoDB because it is easy to use, has many libraries in multiple languages to interact with, has an online service and is one of the most used in the community.

Models

Without entering into too much detail, some models could be very relevant for our tennis match prediction, especially Markov Chains and machine learning models.

Markov chain

Nowadays the game of tennis is modelled using a Markov chain because each point played is independent and identically distributed (iid). It does not depend on the previous point. We could then model each game by a Markov chain because of the hierarchy of the scoring architecture. The probability of a player winning the point is only influenced by the player who serves. This model could be quite complete but disregards a lot of details such as the surface, the age of the players, the possible injury,...

Machine Learning model

There are different types of machine learning model we can use to predict the outcome of a particular match.

— Binary classification : for instance a decision tree which has only two outputs, the game is won or lost by one player. Here there is no probability.

— Linear regression model : it consists of a vector of n match features x = (x_1, x_2,..., x_n) and a vector of n + 1 real-valued model parameters $\beta = (\beta_0, \beta_1, ..., \beta_n)$. To make a prediction using the model, we first project a point in our n-dimensional feature space to a real number :

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Then using the logistic function we can transform z into a probability between 0 and 1. To train the model consists in adjusting the parameters $\beta = (\beta_0, ..., \beta_n)$

— Artificial neural network : ANN is a system of interconnected “neurons”, inspired by biological neurons. Each neuron computes a value from its inputs, which can then be passed as an input to other neurons. Each input has its own weight. Training the model consists in adjusting these weights.

All these models have pros and cons but an artificial neural network seems to be the most accurate and powerful to model tennis games.