Generalized Linear Regression- Tennis Matchup

Vasco Villas-Boas

September 5, 2022

MODEL CONSTRUCTION

Suppose that all tennis players i at time t are endowed with characteristics a_{itS}, a_{itR} that measure the strength of i as a server and the strength of i as a returner, respectively. In a match between A and B at time t, we say that the probability that A wins a serve on B is

$$s(a_{AtS}, a_{BtR}) := \sigma(b_0 + b_1 a_{AtS} + b_2 a_{BtR}) = \frac{1}{1 + exp(-(b_0 + b_1 a_{AtS} + b_2 a_{BtR}))} \tag{1}$$

where σ is the sigmoid function. We want $s(a_{AtS}, a_{BtR})$ to be increasing in A's serving ability and decreasing in B's return ability. Given data on a bunch of matches, we want to estimate $b_0, b_1, b_2, a_{AtS}, a_{AtR}, a_{BtS}, a_{BtR}$.

ESTIMATION APPROACH

For a match between A and B at time t, again, restrict the dataset to all matches in the immediate year prior to t to produce the set of matches A_t where $M := |A_t|$. Also, let N be the number of distinct players in matches in A_t.

Let $a := [a_{1tS}, ..., a_{itS}, ..., a_{MtS}, a_{1tR}, ..., a_{itR}, ..., a_{NtR}]^T$. For each match m in A_t between i and j construct $x_{m1} := [0, ...0, 1, 0, ...0, 1, 0, ..., 0]^T$ where x_{m1} equals one at the index where a has a_{itS} and a_{jtR} and else 0. Also construct $x_{m2} := [0, ...0, 1, 0, ...0, 1, 0, ..., 0]^T$ where x_{m2} equals one at the index where a has a_{jtS} and a_{itR} and else 0. Then construct the design matrix $X = [x_{11}x_{12}...x_{m1}x_{m2}...x_{M1}x_{M2}]^T$.

Next, let y_{m1} be the realized serve win rate by player i on player j in match m and y_{m2} be the realized serve win rate by player j on player i in match m. Construct $y := [y_{11}, y_{12}, ..., y_{m1}, y_{m2}, ..., y_{M1}, y_{M2}]^T$.

To understand the estimation approach, let $b_1, b_2 = 1$ in equation 1. Under this restriction and our model, for any $x \in X$ defining a serve direction i on j in a match m, it is also true that $s(a_{itS}, a_{jtR}) = \sigma(b_0 + a^T x)$. I can in turn, try to estimate a as a generalized linear model where X is the design matrix and y is the target. The values b_1, b_2 are necessarily fixed for a given generalized linear model regression but I can find their optimal quantities using cross-validation.

I suspect that there are better estimation strategies for this model that take advantage of the fact that X has lots of columns and is very sparse. I need to read up on this. As one idea, there may exist sets of players that only played against each other. In that case, their serve and return strength values can be estimated separately from the serve and return strength values of the other players. As another idea, maybe we can make the targets of the generalized linear model regression to be James-Stein shrunken serve win rates for the match. As a final note, there may be better objectives (error functions) to try in estimation.

RESULTS

Using this model of serve point win rates, plugging these rates into the Markov model for the tennis match, I achieve a log-loss of 0.64799 (which implies an average implied prediction probability of 0.52310) on predicting tennis matches in the set of best-of-3 matches from 2015 (where both players had played at least two matches in the year prior). I suspect my estimation strategy may have not perfectly behaved due to the number of columns of X- I believe that a better log-loss

is achievable with this model.

I did try another objective function for my estimation: I set up a weighted generalized linear regression where weights were equal to the inverse of the variance of the data point. In other words, the more serves in a match from A to B, the more confident we were that the serve win rate (from A to B) was closer to the truth, and so the greater we weighted that data point in the loss function for estimating serve and return strengths. This approach for estimating serve and return strengths yielded a worse log-loss when plugging in the implied serve win probabilities into the Markov Model for tennis matches (log-loss: 0.6775, implied average prediction probability: 0.5079).

Currently, other extensions of this model are out of scope.