10 Statistics

10.16 The Tennis Modelling Challenge (8 units)

This project requires concepts from the Part II course Statistical Modelling. The use of R is highly recommended, although Python is a viable alternative (see Programming notes at the end of the project).

1 Introduction

In this project you shall analyse a dataset scraped from the website `tennis-data.co.uk`. The file `mensResults.csv` available from the CATAM website contains information about matches between the top male professional players in the period 2000–2016. Each row corresponds to a match and lists the following variables:

- `Winner`: The match's winner.
- `Loser`: The match's loser.
- `W1-W5`: The number of games won in sets 1-5 by the winner. If an entry is missing, the match ended before the corresponding set.
- `L1-L5`: Like `W1-W5`, but for the loser.
- `Surface`: The surface on which the match was played.
- `Series`: Name of ATP tennis series (Grand Slam, Masters, International or International Gold).
- `Tournament`: Name of tournament.
- `Round`: Round within the tournament.
- `Date`
- `B365W`: Betting odds on website Bet365 for the winner.
- `B365L`: Betting odds on website Bet365 for the loser.

2 Bradley–Terry model

The Bradley–Terry model for a tournament considers each match to be independent, and

$$\Pr(\text{Player } a \text{ wins a match against Player } b) = \frac{\exp(\beta_a - \beta_b)}{1 + \exp(\beta_a - \beta_b)}, \tag{1}$$

for a vector of parameters β of equal length as the number of players. This model was first invented by the mathematician Ernst Zermelo in 1929 as a way to rank chess players.

> **Question 1** The Bradley–Terry model is a Generalised Linear Model (GLM). Specify the exponential family distribution of the response, the link function, and the design matrix. What happens when we exchange the order of the players in Eq. (1)?

Question 2 Fixing the coefficient for the player "Agassi A." at 0, obtain the maximum likelihood estimator for the coefficients β, using the data for the period 2000–2014. Does the choice of the player "Agassi A." for the reference class affect the fitted values? Report the logistic loss (the negative log-likelihood divided by the number of samples) in the training data, as well as test data containing all the matches in 2015 and 2016.

Question 3 Obtain a 68% confidence interval for the probability that Roger Federer beats Andy Murray in a match.

It is well-known among tennis enthusiasts that certain players enjoy an advantage on specific surfaces. For example, Rafael Nadal does very well on the Roland–Garros tournament, which is played on clay, while Roger Federer has an excellent record in Wimbledon, played on grass. You suggest a new model with

$$\Pr(\text{Player } a \text{ wins a match against Player } b \text{ on surface } s) = \frac{\exp(\beta_a + \beta_{a,s} - \beta_b - \beta_{b,s})}{1 + \exp(\beta_a + \beta_{a,s} - \beta_b - \beta_{b,s})},$$

where $\beta_{a,s}$ can be interpreted as the advantage of player a on surface s compared to a baseline fitness β_a.

Question 4 Describe an appropriate constraint to make the parameters of this model identifiable. Fit the model using the data from 2000–2014 and compare to the model fit in Question 2 using the logistic loss incurred on training and test data.

Question 5 Perform a formal hypothesis test which might allow you to reject the model in Question 2 in favour of the alternative in Question 4. Can you reject the simpler model at the 1% level? Does this agree with the cross-validation comparison of the two models?

Question 6 Discuss how you might use the variables `W1-W5` and `L1-L5` as an output in a GLM.

3 Regularisation

The *lasso* estimator $\hat{\beta}^{(\lambda)}$ for a GLM without an intercept solves the problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimise}} \; -\frac{1}{n}\mathcal{L}(\beta) + \lambda \|\beta\|_1, \tag{2}$$

where $\mathcal{L}(\beta)$ is the log-likelihood, $\lambda \geqslant 0$ and $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. When $\lambda = 0$, $\hat{\beta}^{(0)}$ is the maximum likelihood estimator. When $\lambda > 0$, the second term penalises large values of the coefficients. While this biases the estimator toward 0, it can actually reduce the mean squared error, especially when there are many input variables.

This estimator is also convenient because it can be exactly 0 for a subset of the coefficients, which is a form of automatic variable selection. Increasing the parameter λ produces estimators of increasing sparsity.

The penalty term in (2) is separable, i.e. it is a sum over the coefficients in the model. The function `glmnet` in R allows us to choose a vector $w \in \mathbb{R}^p$ through the argument `penalty.factor` in order to solve the problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimise}} \; -\frac{1}{n}\mathcal{L}(\beta) + \lambda \sum_{j=1}^p w_j |\beta_j|. \tag{3}$$

Question 7 Using the same constraints applied previously, refit the model in Question 4 with a lasso penalty on the parameters $\beta_{a,s}$ for every player a and surface s, but no penalty on the parameters β_a for each player a. Fit the model to the data from 2000–2014 and use 10-fold cross validation to find the optimal value of λ. The function `glmnet` standardises the columns of the design matrix by default. Why might this make sense, in general? Does this make sense in your model? Explain why or why not and adjust the argument `standardize` accordingly.

Question 8 With the optimal value of λ, how many of the estimates for coefficients $\beta_{a,s}$ are non-zero? Compute the logistic loss on the held-out data from 2015 and 2016 and compare to the maximum likelihood estimator in Question 4.

Question 9 It is likely that the ability of each player drifts from year to year. Modify the model in Question 2 to take account of this fact and apply a lasso penalty of your choice. Plot the probability that Federer wins against Nadal as a function of time. Compare your proposal against previously considered alternatives in any reasonable way.

4 Can you outperform the betting market?

The dataset contains betting odds from the site Bet365. Each number is a ratio of the payoff, including the gambler's stake, to the stake. For example, suppose that in a match between Federer and Nadal which was won by Federer, the variable `B365W` equals 1.85 and the variable `B365L` equals 1.90. Then, betting £1 on the event that Federer won would have a payoff of £0.85 in addition to the gambler's stake, while betting £1 on the event that Nadal won would have a payoff of £0.90 in addition to the gambler's stake.

Suppose you have a budget of £1,000 to bet on k tennis matches played in 2015, and you have access to data from 2000–2014. Let $\phi_{i,W}$ and $\phi_{i,L}$ be the amounts we bet on the winner and the loser in match i, respectively*. We call the vector $\phi = (\phi_{1,W}, \phi_{1,L}, \ldots, \phi_{k,W}, \phi_{k,L})$ the *portfolio*.

A Markowitz portfolio aims to maximise the expected profits while minimising the risk, measured by the variance of the profit, under a model which specifies the probabilities of each outcome in each match. Letting $R_{i,W}, R_{i,L}$ be the profit made on the bet on the winner and loser of match i, and $R = \sum_{i=1}^{k}[R_{i,W} + R_{i,L}]$ be the total profit, we can formalise the problem as follows

$$\begin{aligned} &\underset{\phi \in \mathbb{R}^{2k}}{\text{minimise}} \; -\mathbb{E}(R) + \nu \text{Var}(R) \qquad (4) \\ &\text{subject to } \phi_{i,W} \geqslant 0, \phi_{i,L} \geqslant 0 \text{ for } i = 1, \ldots, k, \; \sum_{i=1}^{k}[\phi_{i,W} + \phi_{i,L}] = 1000. \end{aligned}$$

Question 10 Write this problem as a quadratic program with affine constraints

$$\begin{aligned} &\underset{w \in \mathbb{R}^{2k}}{\text{minimise}} \; w^\top Q w - w^\top q \\ &\text{subject to } w^\top \mathbf{1} = 1000, \; w_i \geqslant 0 \text{ for } i = 1, \ldots, 2k. \end{aligned}$$

for some positive semidefinite matrix Q and a vector q. Provide an expression for Q and q in terms of the betting odds and the predictions of a logistic regression model.

*The winner and loser are not known *a priori*, but we use these labels as a way to distinguish the two players in each match.

Question 11 Using any software library for quadratic programming, find the optimal portfolio for $\nu \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ with the maximum likelihood predictions for the model in Question 4. Plot the profits which would have been obtained with this portfolio in 2015 against the parameter ν. Comment on the results.

The variance of the profits R in the definition of the Markowitz portfolio (4) assumes that the probability of each outcome for each match is known exactly. This does not take into account the uncertainty of the model parameters, which can increase a portfolio's risk. A *Bayesian Markowitz portfolio* assumes that the probabilities predicted for each match by the model are random and distributed according to some posterior distribution. The expectation and variance in Eq. (4) take into account this source of randomness.

Question 12 Prove that a Bayesian Markowitz portfolio is still a quadratic optimisation problem with affine constraints.

In reality bets are made sequentially, as we wouldn't know which matches are played in 2015 in advance. One approach to sequential betting is the *Kelly* criterion. Before match i, we bet fractions $\phi_{i,W}$ and $\phi_{i,L}$ of our current bankroll on the winner and loser of the match, which maximise the expectation of the logarithm of the bankroll after the bet. Here, $\phi_{i,W}, \phi_{i,L}$ are constrained to be non-negative and $\phi_{i,W} + \phi_{i,L} \leqslant 1$. The *fractional Kelly* strategy further reduces the risk by staking fractions $\rho\phi_{i,W}$ and $\rho\phi_{i,L}$ of the bankroll on each player, where ρ is a constant smaller than 1.

Question 13 Write a function to compute the Kelly fractions numerically using a grid search. Evaluate the fractional Kelly strategy with $\rho = 0.1$ on the data from 2015 using the predictions of the model in Question 4. Plot the bankroll as a function of the date in 2015.

The results of this section should be rather surprising, as the betting market pools the opinions of many experts and the betting odds ensure that the bookmaker turns a small profit. It is possible to improve the statistical model significantly, using any of the variables provided for the period 2000–2014. **One way to obtain excellence marks in this project, but not the only way, is to try to improve upon the models outlined above and share your results[†].**

5 Programming notes

The use of R is strongly recommended. In particular, the package `glmnet` can be used to fit all the models in the project and may be installed through the following commands.

```
install.packages("glmnet")
library(glmnet)
```

The function `glmnet` fits the more general *elastic net* estimator, in which the penalty term in Eq. (2) is replaced by

$$\lambda \left\{ \frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right\};$$

[†]You may also guard them as a trade secret.

setting $\alpha = 1$ specialises to the lasso. The function `cv.glmnet` may be used to optimise the parameter λ through 10-fold cross validation. The documentation for these functions must be read carefully prior to the analysis. Refer to [1] for further guidance.

You may find `glmnet` useful to obtain maximum likelihood estimators in Section 2, as this function accepts design matrices which are sparse. Familiarity with the data type `sparseMatrix` and the function `sparse.model.matrix` will prove useful. In addition, the function `as.date` is useful to handle variables which indicate dates.

The function `solve.QP` in the `R` library `quadprog` may be used to solve the portfolio optimisation problem. When the matrix Q is singular, an approximate solution may be obtained by adding weight to the diagonal elements, replacing Q by $Q + 10^{-8}I$.

The Python library `Scikit-learn` also has methods to fit GLMs with a lasso penalty, and may be used together with the package `pandas` for manipulating data frames.

References

[1] James, G., Witten, D., Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*. Springer, New York, 2013. [`http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf`]

