Ordinal classification via Support Vector Machines applied to Soccer outcomes and assessed by gambling strategies

Ravi Kalia

Department of Statistics

University of Oxford

Lady Margaret Hall

A thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2011

This work is dedicated to Professor Brian David Ripley, a man of unbounded wisdom and patience.

Acknowledgements

I will always be indebted to my supervisor Professor Brian Ripley, without whose patience and dedication in helping me this work would not have been possible. I would also like to thank Mr Alex Kuznetsov, Dr Bernhard Seiser and Dr Mehmet Gokcedag for help in technical aspects of writing and thinking about many problems as an economist and programmer. I am also indebted to Professor Gesine Reinert, Mr James Trinder and Mr Nitin Aggarwal for emotional support.

My eternal thanks to Dr Wendy Hawes, Mr Steven Clark and Mr James Orr who nudged me to complete this work with tactful redirections. Thanks to Mr Sol Campbell (formerly of Arsenal FC) and Mr John Terry (Chelsea FC) for inspiring me to look at football mathematically from my high school days. I am grateful to Mr Patrick Mackenzie-Charrington for insight in the professional gambling industry.

Thanks to affiliates of Queens Park Rangers FC and the Mittal family for encouragement and discussions relating to the organization of professional football. Finally I am indebted to my siblings for their endless encouragement and support.

This work would not be possible without the excellent programming environment provided by the R foundation and versatile packages from R core team members.

I am grateful to the Engineering and Physical Sciences Research Council and Lady Margaret Hall, for their financial support in completing this work.

Abstract

Ordinal classification via Support Vector Machines applied to Soccer outcomes and assessed by gambling strategies

Gambling on handicapped sports events lends itself naturally to being modelled as an ordinal classification task. In this work we show how to use ordinal classifiers to generate accurate estimates of predictive class probabilities for an artificial example and European Soccer data.

A review of statistical and machine learning strategies for ordinal classification is used to build an algorithm which explicitly accounts for non-linearity via Support Vector Machines, RankingSVM. We note that a sophisticated loss function is needed to properly deal with ordinal data modelling.

Using RankingSVM and Bayesian POLR models we demonstrate that a complex dataset of Soccer-related variables can be used to estimate the odds of betting related events which are consistent with market odds.

To transform this into a profitable strategy we investigate gambling strategies based on stochastic programming. These strategies have led us to develop a non-parametric comparison and visualizations for the predictive ability of multinomial and by extension ordinal classification models.

Performance against information in the market is measured via an investment returns back-test in the Bayesian case. The kernel based approach is used with artificial data and a small subset of data for the real world sports gambling task.

Contents

List of Figures

Chapter 1

Introduction

Gambling on professional sports is a leisure pursuit that has been revolutionized by modern computational power and communications technology. Although banned in some countries, it is a permitted (usually condoned) activity in others and has the peculiar benefit of offering tax-free returns in some jurisdictions. This has caused a 24 hour sports betting market to develop.

The onset of exchanges allows gamblers to offer back and lay odds, albeit for a small administration fee. To back an event the gambler can offer to receive payment at the rate of the odds they choose for the occurrence of the event or lose one unit if the event does not occur. For a lay bet the gambler receives, on a unitized basis, one unit of money if the event does not occur or agrees to pay out the odds they choose if the event does occur.

A legitimate fear is that the explosion in sports betting, and particularly laying bets, encourages cheating. Although match-fixing does occur (Hill [2006]) there are some major league sports where it is not economical for the competing teams, players and officials to cheat. As such in the vast majority of sporting events we can assume that the result is stochastically determined.

With online bookmakers and betting exchanges, the implied probability offered fluctuates to reflect the dynamics of market expectations, largely driven by changes in newsflow.

The objective of the exchanges and bookmakers is to increase the amount of money gambled, and in the case of the bookmaker to set odds accurately enough that they have positive expected returns on a wide variety of weakly correlated sporting events. Sports gamblers, as a general rule, have an unusual utility for their wealth which makes them interested in wagering only with the odds are relatively close to even. Their world view is shaped by a need to get the better of it - that is taking the positive expectation side of a bet valued at or close to even. Bookmakers are also advantaged from constructing portfolios of bets close to even. These bets can be repackaged and sold out to gamblers as a more complicated and higher profit margin product.

In many cases, some events can seem overly one-sided, discouraging the gambler from entering the market. For those events, bookmakers seek to handicap, by a margin, the stronger team's final score so that the odds are closer to even. Setting the margin creates a synthetic product which is in demand.

What this means is to set, before the game, a margin by which the stronger team's score will be reduced. This reduced score is used to determine the outcome of wagers. For example in a recent encounter between Manchester United and West Ham, Manchester United were handicapped at -2. Betting on Manchester United to win at this handicap, the wager will payout only if Manchester United win by 3 goals or more (if they win by 2 goals, the stake is returned; if they win by 1, draw or lose the stake is lost). Optimally (in a stochastic profit-seeking sense discussed in later chapters) determining the handicap line before a game starts is a non-trivial task.

Given that most market participants are not necessarily rational it follows that an objective model, taking account of non-linearities, can estimate probabilities more accurately. This, combined with the handicap line, allows for the possibility that some bets can be placed such that there is a positive expectation of return.

Often the quantity of interest is the difference between the scores of two competing teams or individuals. In theory this difference can have support on the space of all integers. In practice the difference tends to be restricted to a smaller space.

Nonetheless models have been suggested that treat the score difference as a distribution which fall on the integer state space (Dixon and Coles [1997]; Karlis and Ntzoufras [2003]). This can be problematic for two reasons: firstly, extreme results have strong influence on the parameters used in the predictive distribution; secondly, it may make more sense to define the state space as unequally spaced bins, this is particularly true for Baseball and Basketball. These facts expose the weaknesses of relying on an integer state space to model the score difference. It might be argued that these concerns are not important given that the sign of the score determines the outcome of the event, not the margin of victory. However, when gambling with handicaps, these issues are very important in setting probabilities accurately.

In this thesis we propose a novel approach to analyze score difference which more realistically captures its nature: non-linear ordinal classification. Splitting the score differences according to suitable cuts for a given sport and league allows a unified approach to be taken across all sports and mitigates the weaknesses outlined above. Although the game outcomes of win, lose and draw have been modelled using proportional odds logistic regression (POLR) - an ordinal classifier (Goddard [2005]; Brillinger [2007]) partitioning the score difference using an ordinal model is to our knowledge an innovation. An expert-informed partition could lead to more accurate probability estima tes and handicap setting.

There has been significant interest in ordered data in the machine learning community, motivated in part by Page *et al.* [1999]. We use approaches from machine learning in this thesis.

The statistical problem is one of finding the best ordinal classification, mapping non-linear behaviour from the covariate space to the probability space. To do so, we take inspiration from a technology developed in the machine learning community - kernel or SVM methods. Designed for binary classification, modifications have been suggested for multinomial and ordinal problems. We seek to place this technology on a firmer footing by suggesting changes that explicitly model ordinality using kernel methods.

One drawback to the kernel method approach is that it does not produce probabilistic forecasts after training. A remedy is to use post-processing; train a mapping from the kernel output to probabilities (Platt [2000]). Therefore there is still merit in pursuing the kernel method line of enquiry.

1.1 Types of odds

Probability theory and statistics have roots in studying games of chance and their associated odds. Indeed de Finetti [1974] explained coherence using an odds gambling thought experiment.

There are several senses in which odds can be defined, for each there is a fair value interpretation:

- Statistical odds

  The statistical odds of an event is just the ratio of the probability that the event will happen to probability that it will not happen.

- Asian fixed odds

  The payoff (reward and return of unit wager staked) offered by the bookmaker for one unit of money wagered if an event occurs. It is usually quoted as a decimal, hence also referred to as decimal odds. The fair value of the payoff is the recriprocal of the probability, allowing for commission. (The term is sometimes confused with the Asian handicap; a soccer handicap at which Asian prices are often quoted.)

- UK fixed odds

  Also referred to as fractional odds these are quoted as integer pairs in the ratio, reward : stake. They represent the reward that will be given if the event occurs, if the stake is risked to be lost to the gambler if the event does not occur. Note that

  $$\frac{reward + stake}{stake} = \frac{1}{P(Event)}$$

  which is the Asian odds of the event at fair value.

  The reward and stake pair offer a fair value (expected profit of 0) to the bet if when the stake is 1, the reward is

$$\frac{1 - P(Event)}{P(Event)}$$

which is the reciprocal of the statistical odds of the event, or equivalently the statistical odds of the complement of the event.

- North American odds

  Also referred to as moneyline odds, they are quoted as a dollar amount, usually integer, either positive or negative.

  If positive, it indicates the amount that will be returned to the gambler (payoff) for a 100 dollars wager if the event occurs; the 100 dollars are lost to the gambler if the event doesn't occur..

  If negative, it indicates the amount that must be wagered to return 100 dollars to the gambler.

  The moneyline odds are fair value if, when the moneyline is positive

  $$\frac{payoff}{100} = \frac{1}{P(Event)}$$

  which is the Asian odds of the event at fair value.

  When the moneyline is negative, if

  $$\frac{payoff}{100} = \frac{1}{1 - P(Event)}$$

  then the moneyline is at fair value. This is the Asian odds of the complement of the event at fair value.

Unless specified otherwise, in this work the reference to odds is to the statistical odds of the event in question.

1.2 Aim of the thesis

Modern computational techniques have outgrown many well established methods used in professional gambling circles. The objective of this thesis is to suggest a new methodology which might profitably be applied to the sports betting market. Further, we wish to illustrate that often it is better to think of the state space differently, in this case ordinally, as a robust way to model facets of the data accurately.

To do so, we need to show that our chosen predictive model has forecasting ability which is superior to that implied by the market. (It is worth noting that the market can imply different probabilities in two ways; either bookmakers offer different odds on the same event or different bets can be used to triangulate different odds on the same event with a single bookmaker. Markets should be efficient - locking out such arbitrages - but rarely are.)

Although standard statistical tools exist to assess the predictive performance of established models, a gambling wealth approach yields directly comparable results. Developing this methodology into a framework that allows one to assess such performance, is an aside in this piece of work.

This work is completed with R (R Development Core Team [2010]), a tool which has revolutionized statistical programming and made writing this thesis possible.

An artificial dataset (outlined below) is also used to help contrast performance against the real world sports betting example.

1.3 Thesis outline

The structure of the thesis intermingles our approach of ordinal classification with analysis of the problem domain in sports betting.

- In Chapter 2 we highlight well-established ordinal classifiers. These classifiers have a rich pedigree, having been used in the statistics community for at least three decades.
- In Chapter 3 we develop our support vector machine inspired ordinal classification methodology.
- Chapter 4 considers well known approaches to modelling score differences.
- In Chapter 5 we discuss tools to assess the performance of classifiers, using utility theory and log wealth optimizers.
- Chapter 6 reviews the progress of our SVM approach and Bayesian POLR methods in generating good predictive probabilities.
- Chapter 7 assesses the investment performance of our models.
- Chapter 8 concludes with final remarks.

1.4 Description of datasets

We use two datasets in our assessment of models. The first is artificial data taken from a well mined example in the machine learning community. The second is a major league sports example taken from the professional soccer leagues in Europe. Bookmaker odds are available for this data, thus we are able to back-test the performance of our model against the market.

1.4.1 Artificial data

A bi-linear function, given by Herbrich *et al.* [2000], which ranks patterns $\mathbf{x} = (x_1, x_2)$ to a class y is

$$y = i \iff 10\,(x_1 - 0.5)\,(x_2 - 0.5) + \epsilon \in (\theta(r_i), \theta(r_{i+1})) \tag{1.1}$$

where ϵ is normally distributed, $\epsilon \sim N(0, 0.125)$, $i = 1, \ldots, 5$ and

$\theta = (-\infty, -1, -0.1, 0.25, 1, +\infty)$

We simulate $n = 10,000$ examples in the unit square, anchored at the origin, contained in the positive quadrant according to a uniform distribution.

1.1 shows a random sample in the bivariate space, where colour indicates observed class membership and level set curves correspond to the rank boundaries.

Our objective is to learn this non-linear mapping in the presence of noise. That is, we wish to learn the non-linear mapping, in the presence of the normally distributed error that is affecting the rank class of the response.

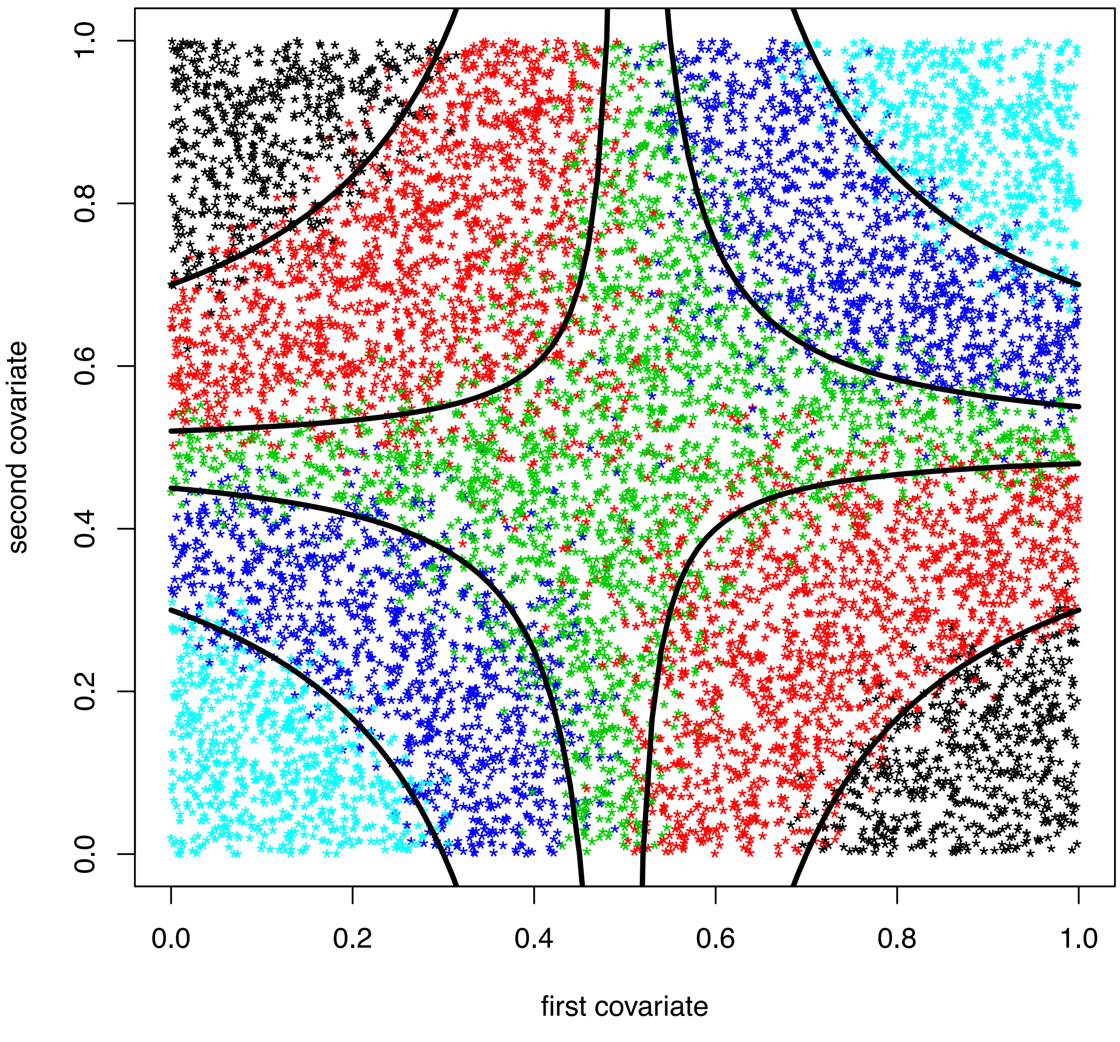


Figure 1.1: Non-linear mapping showing observed ranks and latent boundaries. Colours correspond to observed ranks: 1 to black, 2 to red, 3 to green, 4 to blue, 5 to cyan.

1.4.2 Soccer data

Data scrapped and cleaned from websites has been collected from major European football leagues, augmented by proprietary factors covering the period 1993 to 2011. Our interest lies in finding a predictive distribution for the full time goal difference. Some of the publicly available variables (cf. www.soccerbase.com) are:

- Div = League Division
- Date = Match Date (dd/mm/yyyy)
- HomeTeam = Home Team
- AwayTeam = Away Team
- FTHG = Full Time Home Team Goals
- FTAG = Full Time Away Team Goals
- FTR = Full Time Result (H = Home Win, D = Draw, A = Away Win)
- HTHG = Half Time Home Team Goals
- HTAG = Half Time Away Team Goals
- HTR = Half Time Result (H = Home Win, D = Draw, A = Away Win)
- 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
- HF = Home Team Fouls Committed
- AF = Away Team Fouls Committed
- HO = Home Team Offsides
- AO = Away Team Offsides
- HY = Home Team Yellow Cards
- AY = Away Team Yellow Cards
- HR = Home Team Red Cards
- AR = Away Team Red Cards
- HBP = Home Team Bookings Points (10 = yellow, 25 = red)
- ABP = Away Team Bookings Points (10 = yellow, 25 = red)

Asian odds data (on home win, draw and away win events) from bookmakers clusters in a constrained odds space; 1.2, so that the odds implied probability sums to near one. It is not exactly one because of the commission, or over-round, so that the bookmaker is guaranteed a profit if equal amounts are placed on each outcome. (Note that we focussed on the odds region of 1.5 to 5, so outliers don't make visualization of the chart difficult.)

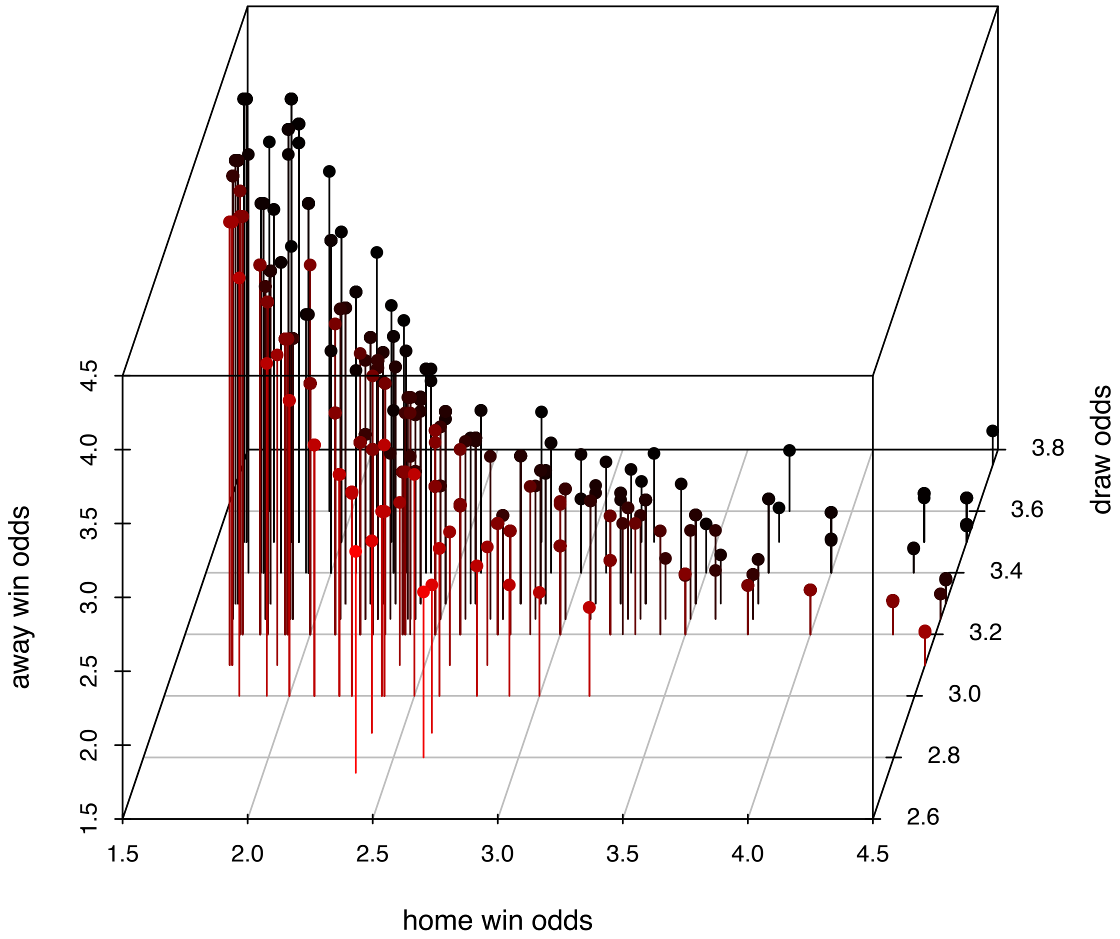


Figure 1.2: Scatterplot of bookmaker odds on home, draw and away win events

Chapter 2

Established ordinal classification methods

This chapter provides some of the background decision theory and overview of statistical methods for ordinal classifiers.

2.1 Decision theory

In evaluating the quality of a classifier we need to consider the penalty for misclassification. Suppose that we classify a pattern as $c\left(x\right)$, when it is actually of class y. Then the loss for misclassification is described by $L\left(c\left(x\right),y\right)$. It is usual to define a classification rule which minimizes the expected loss, with respect to the posterior probability of classification given the data.

There are three popular loss functions, each being well suited to nominal classification, regression, and feasibly for ordinal regression: the absolute distance, quadratic, and $0-1$ loss functions. These loss functions correspond respectively to the mode, mean and median of the posterior class probability given the data.

The absolute distance loss function is often favoured for use in ordinal classification problems. It will penalize misclassification and increase the penalty by the number of classes that the misclassification is out by.

We note that the selected loss function depends on the application, even though the default choice is the absolute distance loss function.

The absolute distance loss function suffices but it is a symmetric function, so that over-estimating, is treated the same as under-estimating. There are many practical scenarios where there would be a different loss incurred for under-estimating compared to over-estimating. For example, in classifying the condition - *Imminent failure, Weak, Good, Full strength* - of a critical component used for a hazardous task, the loss from classifying a true state *Weak* as *Good* will be higher than classifying *Imminent failure.* The absolute distance loss function cannot capture this difference. In handicapping games, the loss is affected by the direction, as well as the size, of the misclassification.

Yet, there are more deficiencies to using one of these three loss functions. By its very nature ordinal data have a finite number of classes. The loss function needs to incorporate the number of classes as well as the level of error.

In many respects the loss function needs to be chosen specifically for the classification problem at hand, as might be the case in a betting context. Below we consider a special example loss function, addressing some of our concerns.

Suppose we are interested in a penalizing over-estimation more than under-estimation and also that we wish to penalize less as the number of classes increases. Then one possible loss function would be

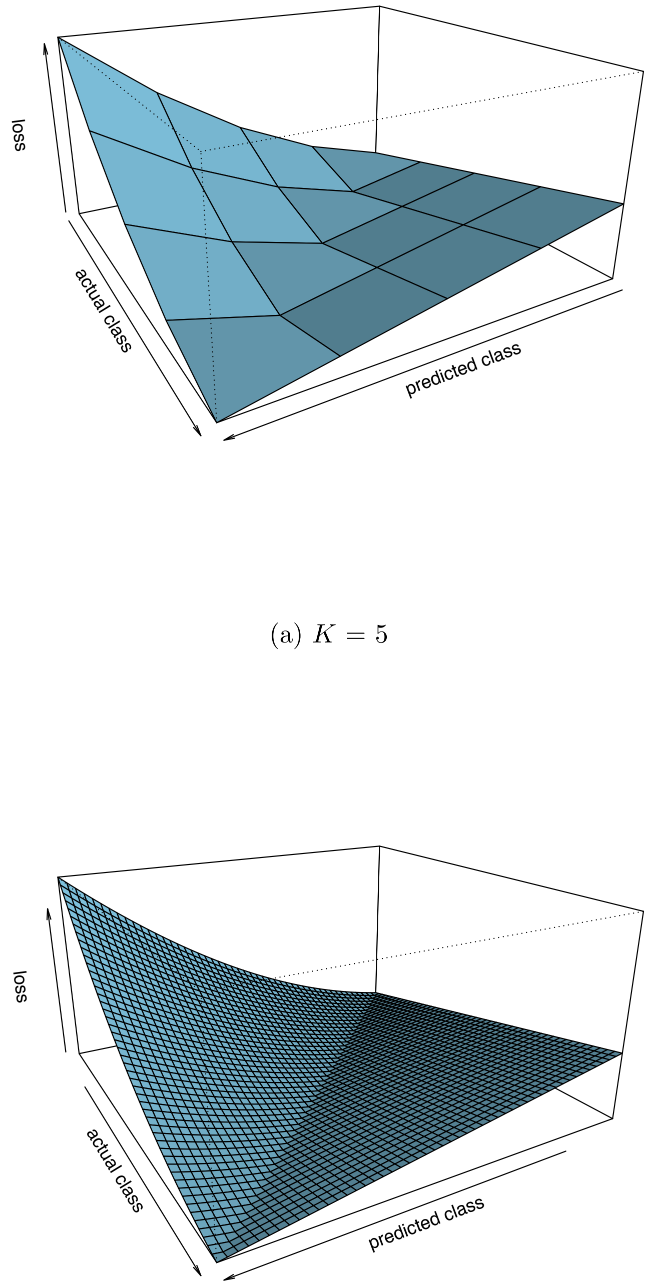
$$L\left(c\left(x\right),y\right)=\frac{\left|c\left(x\right)-y\right|+\max[0,c\left(x\right)-y]^{2}}{K} \tag{2.1}$$

where K is the number of classes. 2.1a and 2.1b show this loss function with $K=5$ and $K=50$ classes. Notice the markedly different curvature for underestimating (smaller penalty) and overestimating the ordinal class (quadratic penalty).

2.2 Established models

A naive approach to modelling ordinal variables is to perform regression, either linear or non-linear (Likert [1932]; Moody and Utans [1995]) on the ordinal labels, say 1 to K suitably, re-coding the original labels where needed. This ignores the problem that there is no metric between the ordered classes. It will respect the order relations, but the lack of a true distance will lead to results that have no understanding of the probabilities involved and the model does not guarantee any optimality in the rule chosen. Mathieson [1998] has suggested that the integer labels can be replaced with numerical scores and these could also be estimated in the optimization process.

McCullagh [1980] provides an excellent account of ordinal regression strategies, perhaps inspired by Cox [1972]. Here, there is an assumption that the ordinal state space can be modelled on a continuous underlying scale using a latent variable. The generalized linear model approach is modified in that the link



(a) $K = 5$

(b) $K = 50$

Figure 2.1: Asymmetric loss functions with K classes

function is applied to the cumulative distribution function of the ordinal variable. It is a mechanism to relate the latent utility variable, or linear predictor to the probability space.

$$LINK\left[P\left(Y \leq y|\mathbf{x}\right)\right] = \eta\left(y, \mathbf{x}\right) = \theta_y - \beta^{\mathbf{T}}\mathbf{x} \tag{2.2}$$

and the added constraint that $\infty = \theta_K \geq \theta_{K-1} \ldots \geq \theta_1 \geq \theta_0 = -\infty$ which are cut-points of the continuous scale.

We can think of the latent variable as being related to the ordinal state space such that

$$Y = j|\mathbf{x} \quad \text{if} \quad \theta_{j-1} < Y^*|\mathbf{x} \leq \theta_j$$

Then with $Y^* = \beta^{\mathbf{T}}\mathbf{x} + \varepsilon$ (Agresti [2010]) the inverse of the link function determines the distribution of ε, which we may think of as an error.

The logit function is $\log\left(\frac{\theta}{1-\theta}\right)$ and when used as a link function the distribution of the error, ε, is the logistic distribution.

The logit link leads to the proportional odds model. The odds of $\{Y \leq j|\mathbf{x}\}$ are

$$ODDS\left[\{Y \leq j|\mathbf{x}\}\right] = \frac{P(Y \leq j|\mathbf{x})}{1 - P(Y \leq j|\mathbf{x})} = K(j, x) = K_j e^{-\beta^{\mathbf{T}}\mathbf{x}} \tag{2.3}$$

Then the odds ratio for two different treatments for the event $Y \leq j$ is

$$\frac{ODDS(A)}{ODDS(B)} = e^{\beta^{\mathbf{T}}(\mathbf{x_2}-\mathbf{x_1})}$$

where $A = \{Y \leq j|\mathbf{x_1}\}$ and $B = \{Y \leq j|\mathbf{x_2}\}$

Therefore the odds are proportional to each other. Applying the logit model and after re-parameterizations the model is as above with $\theta_j = \log[K_j]$.

Under a different model, the proportional hazards model, the survival function is related to the regressors as follows

$$S(j) = P(Y > j|x) = e^{-\Lambda(j)} = e^{-\Lambda_0(j)e^{-\beta^{\mathbf{T}}\mathbf{x}}} \tag{2.4}$$

where $\Lambda(j) = \int_0^j \lambda(u)du$ and $\lambda(j) = \lambda_0(j)e^{-\beta^{\mathbf{T}}\mathbf{x}}$

It is easily verified that the hazard rate and integrated hazard functions are proportional for a given threshold and different treatments.

Applying the complementary log-log link function to the cumulative distribution function of the ordinal variable gives

$$log\left(-log\left(1 - P\left(Y \leq j|\mathbf{x}\right)\right)\right) = \eta\left(y, x\right) = \theta_y - \beta^{\mathbf{T}}\mathbf{x} \tag{2.5}$$

This link function corresponds to the proportional hazards model. It also implies that the underlying distribution of ε is an extreme value distribution.

A probit link function implies that the distribution of ε is the standard normal. Other link functions can be used, however it is not clear how one should interpret the parameters. Practitioners, and our own experience, suggests that there is little benefit to varying the link function in most contexts.

Agresti [2010] discussed an adjacent categories logit model. These are an alternative to using cumulative probability models. This model performs a regression on the logit of $P\left(Y = y | Y = y \quad \text{or} \quad Y = y+1\right)$, which simplifies to

$$logit\left(P(Y = y | Y = y \quad \text{or} \quad Y = y+1)\right)$$

$$= \log\left(\frac{P\left(Y = y\right)}{P\left(Y = y+1\right)}\right)$$

$$= \log\left(\frac{\pi_y(x)}{\pi_{y+1}(x)}\right) = \theta_j + \beta^{\mathbf{T}}\mathbf{x} \tag{2.6}$$

A convenient property of adjacent categories logits is that after re-parametrization they can be expressed as baseline category logits. This model can be used to identify those explanatory variables which have differing variables across adjacent categories.

Anderson [1984] advocated that the ordinal classification task should be split into two separate cases depending on whether there is some knowledge of the underlying scale from which the ordinal classes have been constructed. For example, for grading on a scale of (*Poor, Satisfactory, Good, Very Good, Excellent*), knowing that the scale is made from quintiles of percentage scores should be treated differently from the case when there is no knowledge of the underlying scale.

He also argues for a *stereotype* model for ordinal regression. In it, the coefficients of regression are allowed to vary for different thresholds of the ordinal scale. However he constrains the coefficients of regression to be parallel, so that they only differ by a scaling. The novelty of this approach is that scalars are used to determine how to order the classes. Indeed, the model can decide if ordering the classes is an important feature at all.

Given that usually there are a small number of classes, potentially all orderings of the classes could be tested to decide on a best fit. That would be needed if there is some doubt about the ordering.

Mean response models, where a weighted sum of the underlying probabilities is used as a response to fit a linear model have been examined by Grizzle *et al.* [1969] and Williams and Grizzle [1972]. The optimization is non-trivial as determining the probabilities is treated as a separate exercise using a multinomial likelihood function. This model benefits from providing a simple explanation of the effects of the regressors on the category. Also as K increases the model becomes closer to OLS regression model. Treating ordinal variables as real numbers will necessarily lead to higher prediction errors than models which work with the ordinality of the data directly.

2.2.1 Inference

Estimating the parameters can be done using maximum likelihood and Bayesian inference. In order to do this the likelihood needs to be written down explicitly in terms of the parameters.

The maximum likelihood estimator of the parameters will not possess a closed form expression, but can be found using numerical algorithms.

Taking logarithms of the likelihood, simplifies the objective function while exploiting the monotonic property of the logarithmic operator. Although there are a number of optimization algorithms, the optimizers can be found using Fisher scoring for multivariate generalized linear models (McCullagh [1980]; Thompson and Baker [1981] and Walker and Duncan [1967]) . It has been shown that the log-likelihood is concave and well-behaved for many link functions, so that there is a single optimum which numerical methods converge to rapidly.

Taking a Bayesian perspective, the above models can be viewed as having uncertain regression parameters. The specification of a informative prior distribution is desirable if there are reasons to believe that particular parameter values are more likely than others. The quantity of interest is the predictive distribution of the class given a pattern. We average the estimate of class probabilities over the parameter space, using the prior to weight the class probabilities. This produces final estimates that are smaller than the MLE point estimate. However, it is often difficult to evaluate the integrals analytically to obtain the predictive distributions (Johnson and Albert [2001]). Historically, the use of series expansion methods to approximate the integral provided a crude solution. More recently using Markov Chain Monte Carlo (MCMC) methods (Cowles [1996] ; Johnson and Albert [2001]) to approximate integrals has become feasible using commonly available computing resources.

2.2.2 Non-linear extensions

Hastie and Tibshirani [1987] extended ordinal regression beyond the generalized linear model paradigm with additive functions. Rather than use a model with a linear relationship being fitted to the logit transform of the cumulative distribution function of an ordinal variable, they used additive non-parametric functions (Hastie and Tibshirani [1986] ; Wood [2006]) to model the relationship between the logit transform and the covariates. That is

$$log\left(\frac{ODDS(A)}{ODDS(B)}\right) = \sum_j f_j\left(\mathbf{x}\right) \tag{2.7}$$

where A and B are defined as above and the non-parametric functions f_j are estimated using scatterplot smoothers, although other non-parametric estimation procedures could also be used. The approach is to fit the functions f_j locally so as to provide better predictive accuracy. A problem is then interpreting the model, since non-parametric procedures are difficult to interpret.

Although referred to as non-parametric functions there are usually many tuning parameters which have to be chosen to make the model fit well. We next discuss neural networks applied to ordinal classification.

Neural networks offer a flexible approach to infer a function from observations. The approach is inspired by the operation of the animal brain. A function is constructed additively from other functions which themselves may be constructed from other functions additively. Going back far enough the process relies on the input data or pattern. The weights taking each function to the next are determined as the optimizers of a loss function which scores the accuracy of the neural network. Good introductions to neural networks are to be found in Bishop [1995] and Ripley [1996].

Neural networks have been studied extensively for multinomial classification and regression problems. By comparison, the use of neural networks for ordinal classification has not been studied as intensively.

Mathieson [1996] uses the approach of modelling the latent variable underlying an ordinal variable using a neural network. He investigated the logit transform to regress cut points θ_y and a non-linear function $\psi\left(\mathbf{x}\right)$, that is

$$logit\left[P(Y \leq y|\mathbf{x})\right] = \theta_y - \psi\left(\mathbf{x}\right) \tag{2.8}$$

He found that for a suitable number of hidden units, neural networks outperformed ordinal regression using a linear predictor with the plug-in maximum likelihood estimator.

In Mathieson [1998], he approximated the predictive distribution classification given a pattern using series approximation to integrate out the parameters weighted by the prior probability of the parameters. Performing MCMC was not possible computationally - due to the multi-modal nature of the likelihood.

da Costa and Cardoso [2005] have developed a different approach to apply neural networks to ordinal regression. Instead of relying on cut-points placed on the real number line to classify ordinal variables they use the binomial distribution with a single probability parameter.

The inputs are run through a neural network to produce a single probability parameter, $p\left(\mathbf{x}\right)$. This is then used to calculate the Bernoulli outcomes K outputs, which are the probabilities that a particular input pattern belongs to that ordinal classification.

The probability that a pattern $\mathbf{x}$, belongs to a class y is given by

$$P\left(Y=y|\mathbf{x}\right)=\frac{K!}{y!\left(K-y\right)!}p\left(\mathbf{x}\right)^{y}\left(1-p\left(\mathbf{x}\right)\right)^{K-y} \tag{2.9}$$

Thus the ordinal classification variable follows a binomial distribution. It benefits from the fact that by construction there will be only one mode; at worst two which will be contiguous so that we have exactly the same probability for two classes.

In essence the model is a single output neural network whose output, a probability estimate, is processed to reach the final outputs. The training uses a $0-1$ loss function which penalizes mis-allocation according to the maximum posterior probability criteria.

Chapter 3

A new model: RankingSVM

We start by introducing support vector machines, and the kernel trick, using it to develop a model for ordinal response based on kernel methods. A fitting procedure is outlined for the RankingSVM model.

3.1 Support vector machines

Kernel methods are a useful way to convert extend model of linear relationships to non-linear models in a controlled manner. In particular, Support Vector Machines (SVMs) have been applied to many fields with success. Good introductions to the methodology are available in Burges [1998]; Duda *et al.* [2007] and Cristianini and Shawe-Taylor [2000].

The original formulation was to create a non-linear model for binary classification. Since then there have been extensions to non-linear regression, multinomial classification, and also to ordinal regression by Duan and Keerthi [2005]. However there is not a sound theoretical basis to multinomial and ordinal classification methods using SVMs.

Our approach is to model a non-linear vector function that maps a pattern to a set of scores for each class. This is done in such a way so as to minimize the distance between the examples and a set of hyperplanes which separate the classifications of the data.

An algorithm is developed which calculates the parameters of the non-linear vector function using an Iteratively Reweighted Least Squares (IRWLS) scheme. We suggest that the minimal scoring class is used to classify patterns from the test data, or if a predictive distribution is necessary, it be constructed from the vector function returned by our procedure.

What follows is a step by step introduction to SVMs and how they inspire us to construct an ordinal classifier from the technology.

3.1.1 Binary classification of linearly separable patterns

SVMs have attracted much attention because they demonstrate as great, and sometimes better generalizations ability than other machine learning methods, Keerthi *et al.* [2001].

The simplest problem, from which support vector machines are developed is binary classification of linearly separable data. Suppose n observations are collected such that

$$\{\mathbf{x}_i, y_i\}_{i=1}^n \quad \text{where} \quad \mathbf{x}_i \in \mathbb{R}^p \quad \text{and} \quad y_i \in \{-1, 1\} \tag{3.1}$$

For now we assume that the data are separable using a hyperplane. We then want to find the best hyperplane. Since the task is to classify existing patterns and generalize for unseen patterns, a plane which is the maximum distance from the nearest patterns of either class is desirable. Thus the problem does not need all the data, just the convex hull - or support vectors - of the patterns for each class. Let the hyperplane be defined by the equation

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{3.2}$$

We wish to determine $\mathbf{w}$ subject to maximizing the distance between the support vectors and the plane. Using vector calculus and Lagrange constraints, the optimization problem can be stated as

$$L_p = \frac{1}{2} \parallel \mathbf{w} \parallel^2 - \sum_{i=1}^{n} \alpha_i y_i \left(\mathbf{w} \cdot \mathbf{x}_i + b \right) + \sum_{i=1}^{n} \alpha_i \tag{3.3}$$

We use the notation L_p as this is the optimization function for the primal problem, whose dual is actually solved to find the optimizers. When this function is minimized with respect to $\mathbf{w}, \mathbf{b}$ and the Lagrange multipliers α_i subject to the constraint that $y_i \cdot (\mathbf{w} \cdot \mathbf{x}_i + b) > 0, \quad \alpha_i > 0$ while the derivative of L_p with respect to the α_i must vanish. The optimizers determine the hyperplane. That is $\mathbf{w}_{opt}, \alpha_{1,opt}, \ldots, \alpha_{n,opt}$ such that

$$L_p(\mathbf{w}_{opt}, \alpha_{1,opt}, \ldots, \alpha_{n,opt}) \leq L_p(\mathbf{w}, \alpha_1, \ldots, \alpha_n) \tag{3.4}$$

$\forall \quad \mathbf{w}, \boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_n} : \nabla L_p(\alpha_1, \ldots, \alpha_n) = \mathbf{0}$

This problem is one of quadratic programming for which well defined methods exist - faster optimization algorithms will be discussed later. Due to the geometry of the surface a single optima is guaranteed. For convenience we understand $\mathbf{w}$ to be the optimizer $\mathbf{w}_{opt}$ in our discussion, other than the prose relating to optimization.

The result is a decision rule which allocates between classes $\{-1, 1\}$ according to

$$c\left(\mathbf{x}_i\right) = sign(\mathbf{w} \cdot \mathbf{x}_i + b) \tag{3.5}$$

This is a simplified case. The practical problems below need to be dealt with (so that we may have a useful method):

- Non-linearly separable data
- Misclassifications in the training set
- Fast calibration of parameters
- Multinomial classification

Of course of paramount interest to us is applying these methods to ordinal classification. All of these issues will be discussed in building up to the use of support vector machines for ordered responses.

3.1.2 Binary kernel trick

As mentioned the technique above needs the data to be linearly separable by class in the covariate space. This restricts its usefulness. Boser *et al.* [1992] proposed using the *kernel trick*; a way to adapt linear methods to fit non-linear data.

In order to do this we first need to note that the optimization problem can be restated to its dual. The optimization problem is now to maximize L_p with respect to α where the derivative constraints of L_p are with respect to $\mathbf{w}, b$. That is $\mathbf{w}_{opt}, \alpha_{1,opt}, \dots, \alpha_{n,opt}$ such that

$$L_p(\mathbf{w}_{opt}, \alpha_{1,opt}, \dots, \alpha_{n,opt}) \geq L_p(\mathbf{w}, \alpha_1, \dots, \alpha_n) \tag{3.6}$$

$\forall \quad \mathbf{w}, \alpha : \nabla L_p(\mathbf{w}, \alpha_1, \dots, \alpha_n) = \mathbf{0}$

This is called the Wolfe dual of the problem. The optimizers are the same as the original problem. This representation allows for the kernel trick to be applied.

The kernel trick is to transform the data usually to a higher dimensional, possibly infinite, space such that even though $\mathbf{x}$ maps non-linearly to the ordinal response, by applying a vector function ϕ to the pattern $\mathbf{x}$. For this

$$\Phi((x) = (\phi_1(\mathbf{x}), \dots, \phi_M(\mathbf{x})) \tag{3.7}$$

can be modelled as linearly separable by class.

Due to Mercer's condition, these transformations can be characterized uniquely through a dot product.

For a symmetric continuous function, $K(\cdot, \cdot)$

$(K : [a, b]^n \times [a, b]^n \longrightarrow \mathbb{R}, \qquad [a, b] \in \mathbb{R}\ )$ which obeys

$$\int \int K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0 \tag{3.8}$$

known as square-integrablity condition, the following is true

$$K(x, y) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})^T \tag{3.9}$$

for some, possibly infinite, length vector of non-linear functions of the inputs.

Often $K(\cdot, \cdot)$ is referred to as a kernel. Because of this theorem, optimization problems with dot products in the objective function can be replaced by a suitable kernel. As a result a non-linear problem in a pattern space can be solved using linear methods in the feature space induced by the kernel function. It is therefore not necessary to know what the non-linear functions $\phi_i(\mathbf{x})$ are. Thus our Lagrangian can be used as before, just replacing the dot product by a kernel function.

The advantage of using the kernel trick for support vector machines is that the non-linear functions do not have to be estimated. Also the optimization is not complicated since there is a single optima to search for.

However, choosing the kernel and its parameters is done without a theoretical grounding. Duan *et al.* [2003] call these variables hyper-parameters and note that estimation via cross-validation and leave- one-out-cross-validation calibration exercises are commonly used. They argue that to reduce computational cost, crude analytical approximations should be used. A simple suggestion has been to perform a grid search over the parameters (Karatzoglou *et al.* [2006]) until the best generalization performance is achieved.

3.1.2.1 Binary misclassification in the training set

Because of the ability of support vector machines to separate out classes non-linearly, we are wary of noisy data significantly undermining the true nature of the patterns in the training set.

Suppose that some of the data in the training set has been mislabelled. Then it would be prudent to penalize the contribution of those observations that are outliers relative to the rest of the sample. The mathematical formalism for this is introduced via a slack variable ξ_i. We now require that the hyperplane $\mathbf{w} \cdot \mathbf{x}_i + b = 0$ is such that the following hold

$\mathbf{w} \cdot \mathbf{x}_i + b \geq 1 - \xi_i$ and $\mathbf{w} \cdot \mathbf{x}_i + b \geq -1 + \xi_i$

subject to $\xi_i \geq 0, \ \forall i$

Ideally we wish to minimize the slack variables as much as possible. We also associate a cost with this slack variable on the objective function (Cortes and Vapnik [1995]).

As before the dot product can be replaced by a kernel function. The problem is still a quadratic programming one with unique optima, only now we are factoring misclassification explicitly into the model. Finding the optimizers can be time consuming using standard quadratic programming problems. Specialized methods have been developed to lower the computational time in performing the optimization.

3.1.2.2 Binary optimization

As seen above in order to extract the decision rule it is necessary to solve a quadratic programming problem. Vapnik [1982] showed that the task can be broken down into smaller quadratic programming problems by removing examples that have a Lagrange multiplier which is zero. These smaller problems can then be solved using numerical methods. Details on how to reduce the time taken to optimize can be found in Platt [2000] and Keerthi *et al.* [2001].

Next we look at how this binary classification algorithm can be modified for the multinomial classification task.

3.1.3 Multinomial classification with SVMs

Several models have been proposed to make SVMs perform classification for multinomial data. The unifying theme is to perform binary SVM through redefining classes and combining the results into a classification rule.

Duan and Keerthi [2005] investigated three popular methods in tests against benchmark datasets:

- One against all

  Under this scheme each class is fitted against all others in the training set with a support vector machine. Then for each new pattern an output is associated with every class. The class with the best output is the prediction label of the pattern (Hsu and Lin [2002]).

- One against one

  More support vector machines are needed using this method. Each pair of classes have a support vector machine fitted. Then each new pattern is tested on each of the $\frac{K(K-1)}{2}$. The class that has collected the most classification decision *votes* is where the final decision pattern is assigned to.

- Pairwise coupling

  Assuming that the outputs of support vector machines can be treated as posterior probability parameters Hastie and Tibshirani [1998] constructed an algorithm to couple the probabilities and estimate the posterior probability of each class. The class with the highest probability is where each pattern is allocated to.

Benchmark tests have shown that each method has merit depending on the nature of the data. For example when training data is sparse the pairwise coupling approach gives high classification accuracy (Duan and Keerthi [2005]).

There are other methods which can be used to adapt SVMs to a multinomial classifier. Those mentioned above are some of the most popular implementations for computer algorithms and are readily available.

3.1.4 SVMs for ordinal data

Having built an understanding of SVMs, we now turn to looking at how they can be used for ordinal classification. The development of such methods to deal with ordinal classification has only recently received attention in the literature. All rely on modelling a latent variable and cut-points which depend on the pattern **x**. The assumption is that realizations of the latent variable are allocated to bins defined by the cut-points and this determines the value that the ordinal variable takes.

Herbrich *et al.* [2000] have a theoretical model for the empirical risk involved with ordinal classification and provide an upper bound for this in terms of a distribution independent risk functional. They begin with the assumption that classification rule must be asymmetric and transitive: enforced by arguing for the existence of a linear utility function. A mapping from a partition of intervals from the range of the utility function to the ranks in the data yields the classification rule.

The optimization procedure is modified to incorporate these features of the model. It is still a quadratic programming problem and estimation of cut-points is carried out after the weight parameters have been estimated. Herbrich *et al.* [2000] report that their algorithm outperforms using multinomial classification SVM methods for ranking information retrieval data.

Crammer and Singer [2005] work with parallel hyperplanes. Their idea is to have parallel hyperplanes with differing intercepts. Parameter estimation is by minimizing a loss function iteratively. The decision rule classifies patterns to ranks depending on which hyperplane is closest to the pattern. They also have an online approach, so that as the true rank of the test case is revealed the parameters of the decision rule are updated if there has been a misclassification. They note that it is possible to combine this approach with the kernel trick to model non-linear data.

Shashua and Levin [2002] introduce two approaches to determine the optimal hyperplanes under differing optimization criteria. Under the first, they maximize the distance between each pair of consecutive ranks. Under the second criteria the sum of squared distances between hyperplanes is maximized. For both cases the optimization problem is one of quadratic programming which has a unique optima.

Chu and Keerthi [2007] demonstrate that Shashua and Levin failed to implement a constraint in their approach, which is not consistent with the ranking of the data. They offer a way to implement the constraint and improve the time it takes to run the optimization by modifying the SMO algorithm. They also suggest a new approach that maximizes the distance between the hyperplanes by using training data from all ranks. The algorithm does not require the order constraint to be made explicit.

What we do next is introduce our contribution in the context of vector regression output using SVMs. We modify an algorithm known as of MIMO (Multiple Input, Multiple Output) regression to model ordinal data.

3.2 A modification of the MIMO algorithm

3.2.1 Ordinal data as a multivariate response

The data observed can be viewed as a pattern vector $\mathbf{x} = (x_1, ..., x_n)$ and a response y. It is usual that $\mathbf{x} \in S_{\mathbf{x}} \in \mathbb{R}^n$. For ordinal data $y \in T_y$ where a binary preference relation exists such that $T_y = \{y(1), ..., y(k)\}$, $y(i)$ is preferred to $y(j)$ if and only if $i < j$. Without loss of generality we assume that the ordinal responses are y are from the integers $1, \ldots, k$ corresponding to the ordered classes.

The ordinal response can be coded as a vector **z** whose r^{th} component, for response y_i is

$$z_i^r = z(i,r) = I_{i \geq r} - I_{i<r}, \qquad r = 1, \ldots k \tag{3.10}$$

We can then view all the responses through a response matrix Z. The rows of the matrix are the response vectors corresponding to each observed pattern. Each column records the response of all observations to that ordinal class.

Other codings are possible, but we use this to keep the optimization arithmetic tidy and close to the original binary classification problem of SVMs.

3.2.2 The decision rule

First we sketch the original formulation of the SVM. The data observed is from a pattern **x** and is associated with a binary response. The objective is to find the best hyperplane which separates the data in some pre-defined *feature* space.

(The feature space is a vector whose components are non-linear functions of the underlying pattern. It is chosen so that the responses can be linearly separated.) The criteria applied, to find the optimal hyperplane, is that hyperplane which separates the data and has the maximum distance between it and the nearest data points in the feature space - referred to as the maximum-margin principle. This is equivalent to minimizing the sum of squares of the weights defining the optimal hyperplane, while maintaining separation of the two classes.

Our argument for ordinal classification proceeds much in line with the SVM formulation of Boser *et al.* [1992].

Suppose that we consider the responses to a particular class r. For each observation we see a pattern $\mathbf{x}_i$ and the response will be $z(y_i, r)$. The data cloud will have points labelled with 1's and -1's which are linearly separable in a 'suitable' feature space. We are interested in finding the optimal hyperplane which separates the classes. Borrowing from Vapnik's methods we apply the maximum-margin principle. Let the hyperplane be defined as

$$f^r(\mathbf{x}) = \mathbf{w}^r \cdot \Phi(\mathbf{x})^T + b^r = 0 \tag{3.11}$$

such that, for all i $z_i^r f^r(\mathbf{x}_i) \geq 1$

[These conditions are requiring that the model chosen separates the data cloud into positive (data with labels 1) and negative examples (data with labels -1) in the feature space via a hyperplane.]

As per maximum-margin principle the optimization criteria is

$$\min_{\mathbf{w}^r, \mathbf{b}^r} \quad \| \mathbf{w}^r \|^2 \tag{3.12}$$

subject to $z_i^r f^r(\mathbf{x}_i) \geq 1$

which can be optimized via the SMO algorithm (Platt [2000]) or IRWLS (Perez Cruz *et al.* [2000]).

The optimizers are the weights and intercept of the hyperplane. Using the representer theorem we can avoid estimation of the non-linear functions, instead having kernels.

The above setup is just a vanilla application of SVMs to the pattern and response vector z_i^r for class r. To allow for multivariate response we repeat this for every class r. Each class is modelled independently of the previous class specific hyperplane parameters. To link the classes together we insist that the same feature space, of non-linear mappings, is used for all ordinal classes.

Therefore once the feature space has been chosen, via a good kernel, $K(\cdot, cdot)$ say, all we require is to solve the optimization below for the hyperplane weights

$$\min_{\mathbf{w},\mathbf{b}} \quad \sum_{r=1}^{r=k} \parallel \mathbf{w}^r \parallel^2 \tag{3.13}$$

In an ideal situation what should happen is that once trained the output of the all k hyperplanes are consistent - they predict a single class, the correct one. It may however be possible for this to be violated.

If the model is inconsistent, we can add a regularizing loss function to the optimization to encourage more *consistency*. How to choose a suitable loss function is a problem that we have not encountered in the literature. What it needs to achieve is a distance measure between the training example responses and the fitted responses.

A crude loss function, in the spirit of Sanchez-Fernandez *et al.* [2004], is

$$L_\epsilon(u_i) = \max(0, u_i - \epsilon)^2, \qquad \epsilon > 0 \tag{3.14}$$

with

$u_i = \sqrt(\mathbf{e}_i^T \mathbf{e}_i)$

$\mathbf{e}_i = \mathbf{z}_i - sign(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + \mathbf{b}^r)$

Once the model has been trained the decision rule can be formed as a function from the output vector generated by the model. One possibility is that we classify the ordinal class as k, such that all the entries of the output vector upto the k^{th} component are positive.

3.2.3 Optimization

If we do not have a penalty function for misclassification then the optimization can be performed via either the SMO or IRWLS algorithm.

When we incorporate a loss function the problem is more complicated. The optimization that we outline is taken wholesale from Sanchez-Fernandez *et al.* [2004]. It is an IRWLS procedure, which requires the solution of an intermediate least squares problem. The algorithm is based on a line search with backtracking algorithm.

We begin by defining the optimization function as

$$L_p(\mathbf{W}, \mathbf{b}) = \frac{1}{2} \sum_{j=1}^{Q} \|\mathbf{w}^j\|^2 + C \sum_{i=1}^{n} L(u_i) \tag{3.15}$$

where

$\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^k]$

$\mathbf{b} = [b^1, \ldots, b^k]$

$L(u_i) = L_\epsilon(u_i) = \max(0, u_i - \epsilon)^2$

and note that u_i is a function of $\mathbf{W}$ and $\mathbf{b}$.

Next we represent the function differently from above. Suppose (for some iteration d), we have estimates $\mathbf{W}^d, \mathbf{b}^d$ of the optimal parameters $\mathbf{W}^{opt}$ and $\mathbf{b}^{opt}$. They are described by u_i^d for all i. Then expanding about these estimates at iteration d

$$L_p^*(\mathbf{W}^d, \mathbf{b}^d, \mathbf{W}, \mathbf{b}) = L_p^*\left(u_1^d\left(\mathbf{W}^d, \mathbf{b}^d\right), ..., u_n^d\left(\mathbf{W}^d, \mathbf{b}^d\right), \mathbf{W}, \mathbf{b}\right) =$$

$$\frac{1}{2}\sum_{j=1}^{k} \|\mathbf{w}^j\|^2 + C\sum_{i=1}^{n} L\left(u_i + u_i^d - u_i^d\right) \tag{3.16}$$

where

$u_i^d = u_i\left(\mathbf{W}^d, \mathbf{b}^d\right)$

$\mathbf{b}^d = ||\mathbf{e}_i^d||$

$$\left(\mathbf{e}_i^d\right)^T = z_i^T - \phi(\mathbf{x}_i)^T\mathbf{W}^d - \left(\mathbf{b}^d\right)^T$$

We just introduced an increment about which approximations can be taken. After taking a linear approximation and then a quadratic approximation of the linear approximation, we have a modified optimization problem which can be solved via WLS. The solution of this is used to estimate a solution to the original problem.

The newly estimated parameters are put into a modified optimization and this is solved, and so on until some terminal condition is reached.

The details are presented here:

As we are approximating a convex real-valued function using Taylor series, the approximation will always be less than or equal to the original problem. Indeed when the increment is zero the two functions are the same. This ensures that the optimizer of the approximation is the optimizer of the original problem, in all but pathological cases.

Our approximating function is

$$L_p^{''}(\mathbf{W}, \mathbf{b}) = \frac{1}{2}\sum_{j=1}^{k} ||\mathbf{w}^j||^2 + \frac{1}{2}\sum_{i=1}^{n} a_i u_i^2 + Constants(\mathbf{W}, \mathbf{b}) \tag{3.17}$$

where

$a_i^d = Max\left[0, \frac{2C(u_i^d)-\epsilon}{u_i^d}\right]$

Although a_i^d depends on $\mathbf{W}^d$ and $\mathbf{b}^d$ if we consider it fixed then we can solve for $\mathbf{W}^d$ and $\mathbf{b}^d$, use these to estimate a_i^d and so on until there is little change in the parameters, when the optimal point has been reached.

The IRWLS algorithm is summarized as:

1) Set $d = 0$, $W^0 = 0$, and compute u_i^d and a_i^d.

2) Compute the solution to (3.17), which is a weighted least squares problem, and label it $\mathbf{W}_{SOL}^d$ and $\mathbf{b}_{SOL}^d$. (The $\mathbf{W}^d$ parameters are never explicitly known in general, only implied by the constants β^d, defined below.)

3) Define the descending direction as

$$\mathbf{P}^d = \begin{pmatrix} \mathbf{W}^d_{SOL} - \mathbf{W}^d \\ (\mathbf{b}^d_{SOL} - \mathbf{b}^d)^T \end{pmatrix}$$

4) Obtain the next step solution as

$$\begin{pmatrix} \mathbf{W}^{d+1} \\ (\mathbf{b}^{d+1})^T \end{pmatrix} = \begin{pmatrix} \mathbf{W}^d \\ (\mathbf{b}^d)^T \end{pmatrix} + \eta^d P^d$$

computing the step size using the back-tracking algorithm, again (the W's are not known, only implied).

5) To calculate η^d

a)initialize $l = 0$, $\eta^d_l = 1$

b) Calculate $L_p(\mathbf{W}^{d+1}, \mathbf{b}^{d+1})$ check if it is greater than or equal to $L_p(\mathbf{W}^{d+1}, \mathbf{b}^{d+1})$.

c)If so $\eta^d_{l+1} = \frac{1}{2}\eta^d_l$ and set $l = l + 1$ and return to b)

d)If not $\eta^d = \eta^d_l$ and terminate loop

6) Compute u_i^{d+1} and a_i^{d+1}, set $d = d+1$ and go back to step 2 until convergence.

Once the optimization has converged, the parameters are used to define the set of hyperplanes for the ordinal model. This can be used to find the fitted values to compare against the observed response.

Although it is easy to solve for W_{SOL} and b_{SOL} using WLS we need a different representation as the non-linear functions in the RKHS are not explicitly known. Therefore we solve for weight parameters of a kernel linear sum which can be used to evaluate all the functions of $\mathbf{W}_{SOL}$ that are needed. This is making use of the representer theorem that

$$\mathbf{w}^j = \sum_{i=1}^{n} \phi(\mathbf{x}_i)\beta_i^j = \Phi^T \beta^j \tag{3.18}$$

The linear system to solve then becomes

$$\begin{pmatrix} \mathbf{K} + Diag(\mathbf{a}^d)^{-1} & \mathbf{1} \\ (\mathbf{a}^d)^T\mathbf{K} & \mathbf{1}^T\mathbf{a}^d \end{pmatrix} \begin{pmatrix} \beta^j \\ b^j \end{pmatrix} = \begin{pmatrix} \mathbf{z}^j \\ (\mathbf{a}^d)^T\mathbf{z}^j \end{pmatrix}$$

with

$j = 1, ..., k$

$\mathbf{z}^j$ the j^{th} column of the coded response matrix Z.

3.2.4 Predictive distribution

Platt [2000] has discussed the lack of a predictive distribution with SVM models and suggested that polynomial regression using the output of SVM classification as a predictor of a probabilistic classifier. This is a natural way to model the output of RankingSVM as a predictive distribution, although the neatness of the underlying large margin approach is lost. We have implemented this method using R code and seen reasonable performance with the data we analysed. The code was checked and optimised using heuristics and the tools provided by the R community.

Nonetheless, it is non-trivial process to take the output of this approach to probabilities. Deciding the number of degrees in a polynomial fit can be achieved via cross-validation over a range of possible models. A weakness with this approach is that it is difficult to assess how the model relates to the inputs in any interpretable manner. Given the two stage nature of the fitted model, it is a problem that has no easy resolution.

Chapter 4

Existing methodologies for score difference modelling

Sports modelling in statistics has been researched from many perspectives. A good survey is Ziemba and Hausch [2008] while Brillinger [2010] provides a decent discussion on football score modelling. A related task in extracting economic benefit is determining the stake of one's bankroll to be invested (Poundstone [2005]) in a bet (which is discussed later in this chapter and related to the next chapter on assessing classification via gambling strategies).

Although our primary interest is score difference modelling, it is worth noting that among the first researchers to apply gambling methodologies commercially was Benter [1994] - to horse racing. He used a generalized additive multinomial model with available odds as a predictor along with historical measures of horse performance. By using this methodology he was able to beat the Hong Kong Jockey Club for much of the 1990s using algorithms developed in FORTRAN with data manually entered into SQL databases. A difficult part of the procedure was determining the stake since the size of the position would have impact on price.

Our focus is score difference modelling. A bivariate variable describing the final score is, (S_a, S_b), where S_q is the number of points/goals by team q, for $q = a, b$ and the score difference is given by $GD = S_a - S_b$.

4.1 Poisson models

A reasonable starting point is to assume that the score pair is a bivariate Poisson process with parameters that are possibly stochastic and inter-related.

Modelling score differences with a double independent Poisson distribution was done (Maher [1982]) as below

$$P(S_a = x, S_b = y) = \frac{\lambda^x \mu^y e^{-(\lambda+\mu)}}{x!y!} \tag{4.1}$$

where $\log(\lambda)$ and $\log(\mu)$ are the canonical parameters of a GLM model to be fitted.

The team coefficient parameters in the Poisson regression linear predictor were attack strength and defensive weakness for each team on a home and away basis. So that the they would be identifiable, Maher placed a sum to zero constraint on the latent team attack and defence weakness parameters.

The use of nominal predictors, attack and defence parameters with levels corresponding to the names of the teams has been used in all subsequent models using the Poisson approach to forecasting. The sum to zero constraint allows models to be identifiable and gives a reference point for the attack and defensive strength parameters against an average of zero across the whole league. In estimating the parameters, they introduced a temporal parameter to penalize the likelihood which is determined by cross validation.

Dixon and Coles [1997] noted that assuming independent Poisson processes was unrealistic as the bivariate final score variable usually presents non-zero correlation, which can have significant impact on the probability distribution. As a partial remedy they applied a dependence structure to the independent probability mass functions at low scoring games. Their formulation is

$$P\left(S_a = x, S_b = y\right) = \frac{\tau\left(x, y, \lambda, \mu\right)\lambda^x \mu^y e^{-(\lambda+\mu)}}{x!y!} \tag{4.2}$$

where:

- a = home team
- b = away team
- $\lambda = \alpha_a \beta_b \gamma$
- $\mu = \alpha_b \beta_a$
- α = defensive strength
- β = attack strength
- γ = league wide home effect

and

$$\tau\left(x, y, \lambda, \mu\right) = \begin{cases} 1 - \lambda\mu\rho & \text{if } \ x = y = 0, \\ 1 + \lambda\rho & \text{if } \ x = 0, y = 1, \\ 1 + \mu\rho & \text{if } \ x = 1, y = 0, \\ 1 - \rho & \text{if } \ x = 1, y = 1, \\ 1 & \text{otherwise.} \end{cases}$$

The parameter ρ is suitably constrained to ensure that τ is non-negative.

They found results which lead to economically beneficial forecasting against bookmakers on the basis of certain filtering rules. It was also noted that their model was weak at forecasting draws and away wins compared to their home win predictive probabilities. Although ad-hoc, their dependence structure , the τ function multiplying the independent Poisson distribution, helped to model the prevalence of low-scoring games.

Crowder *et al.* [2002] put the Dixon and Coles approach on a stronger footing by use of a hidden Markov model for a dependent bivariate stochastic process. To estimate the parameters they used MCMC on a approximation of their model. By use of rolling predictive likelihood statistic they found superior performance as against the Dixon and Coles Poisson model in the forecasting of English Football Association leagues in the years 1992-1997.

Rue and Salvesen [2000] used a Bayesian approach, allowing for short runs of form in team strengths by employing a Brownian bridge as a covariate on the logarithm of the Poisson rate parameters, with dependence between them. Computationally more involved, their model provided good forecast probabilities as compared to Dixon and Coles Poisson model using Premier league data only.

Karlis and Ntzoufras [2003] introduced Skellam distributions (Skellam [1946]) as a way to model the goal difference explicitly. Skellam distributions are expressible as a convolution from the difference of two Poisson distributions. Using a Bessel function of the first kind, the density function is

$$Pr\left(GD=k\right)=e^{-(\mu_1+\mu_2)}\left(\frac{\mu_1}{\mu_2}\right)^{k/2}I_{|k|}(2\sqrt{\mu_1\mu_2}) \tag{4.3}$$

They performed fitting using the expectation maximization (EM) algorithm once a suitable latent variable is introduced in the likelihood function.

They used Skellam regression on team names and home effects as factors for the final scores of Italian Serie A data. It is worth noting the recursive relation for computing probability densities is

$$Pr\left(GD=k+1\right)=Pr\left(GD=k\right)\left(\frac{\mu_1}{\mu_2}\right)^{1/2}\frac{I_{|k+1|}(2\sqrt{\mu_1\mu_2})}{I_{|k|}(2\sqrt{\mu_1\mu_2})} \tag{4.4}$$

The parameters μ_1 and μ_2 are linear predictors for the Skellam regression used to model the impact of covariates on the outcome of the competition.

This poses a potential problem in that it fixes the rate of decay of probabilities between adjacent states, controlling the latent distance between scores if viewed from a utility theory point of view. From our experience of watching sports and interviews with those around major league sports, the difference in probabilities for score differences is not governed by such a law and can vary strongly depending on the league and changes in style of football over time. Therefore, this is a weakness of using Poisson models.

A benefit is that there is no need to assume that the final score follows a particular joint Poisson distribution with the Skellam approach. Although a Skellam distribution can be described as a difference of two statistically independent Poisson random variables, as long as each random variable is Poisson, their joint structure can be of any dependent form. The difference of the dependent variables (which are marginally Poisson) must follow a Skellam distribution. In fact, Karlis and Ntzoufras suggested that a latent variable should be introduced to account for the excess of draws observed.

Their performance measures, diagnostic tests and deviance residuals, indicated a superior fit by inflating the Skellam score difference distribution at zero. This may in part reflect the nature of the data they handled; the Serie A Italian league is noticeable for its paucity of goals.

In a follow up paper, Karlis and Ntzoufras [2009] investigated the use of Bayesian methods with non-informative priors to alleviate some of the problems associated with extreme outliers to model the joint score with an explicit dependence structure. This avoided the need for the EM approach as a Gibbs sampler could be employed to generate parameter samples from the posterior. The parameter samples are then averaged to construct a smoothed probability estimate from all the model parameters.

Karlis and Ntzoufras specified zero mean large variance normal distribution priors for the parameters of the Skellam distribution and a uniform prior on the probability parameter, while maintaining the sum to zero constraint via restriction on the attack and defence parameter of one team, say the first. That is

$$A_1 = -\sum_{j=2}^{K} A_j$$

$$D_1 = -\sum_{j=2}^{K} D_j$$

where $A.$ and $D.$ are the latent parameters describing the attack and defensive strengths for the K teams being observed.

They used a sampling augmentation scheme which relied on proposals via the Metropolis-Hastings algorithm to generate dependent samples from the posterior distribution. The posterior samples lead to predictive distributions; averaging them provides a fully Bayesian prediction model, which was noted to have superior performance compared to the frequentist approach adopted in an earlier paper (Karlis and Ntzoufras [2003]).

One interesting property of Skellam distributions is that for any bivariate distribution on the non-negative integers, its difference can be represented by an appropriately chosen Skellam distribution. As a consequence, the score difference from a Dixon-Coles random vector could be recast as a Skellam distribution. However, the time decay scaling might prevent this from being true.

4.2 Ordinal models

An alternative approach from the Poisson setting, empirical or Bayesian, is to use ordinal models. This has been done in the setting of Win, Draw and Lose, whereas our innovation is in introducing ordinal models for the score difference state space.

One might also consider a multinominal classifier, where a wider set of procedures become available, however the censored score difference is naturally an ordinal variable and modelling this facet of the data will lead to better predictive performance.

Goddard [2005] investigated using many lagged predictors to determine the probabilities of Win, Draw and Lose using a discrete choice ordinal model. This is just a proportional odds logistic regression model. Using t-tests he selected statistically significant variables in the final model.

Goddard noted that it was not possible to construct a profitable strategy from the predictive probabilities from his selected models. Interestingly, he fitted a discrete choice model to results in cup as well as league competitions. This is a contentious issue, as given the revenues attached to league and European games, cup games are often neglected by teams, with second or third choice players being used to form a cup team in a significant number of games.

In the statistics community, Brillinger [2007] made use of a small number of predictors to assess Norwegian football. Visualizations suggested that recent games and distance between teams impacted the performance of teams in this division. No back-test against market odds was provided. The paper was focussed on explaining the fit of the model rather than assessing predictive performance. Similarly, McHale and Davies [2007] used ordinal models to assess the effectiveness of FIFA world rankings of international teams, illustrating how non-league matches can be modelled.

We believe that Bayesian ordinal models on the state space offer a similar approach to that of the RankingSVM methodology developed. Although not explicitly non-linear, by use of suitable priors they offer flexibility and are able to mitigate over-fitting by implicitly model averaging over parameters in a non-linear way.

Cowles [1996] provides a framework in which a Bayesian analysis of ordinal models can be performed using a multivariate normal prior on parameters. Albert and Chib [2001] placed priors on the cut-point and covariate multipliers of the proportional odds logistic regression model, assuming independence between them. In their formulation, the cut-points would have a diffuse prior and an additional latent variable. These can be used to sequentially simulate MCMC chains across all parameters.

An interesting proposition would be to recover the joint distribution bivariate score given the score difference distribution. Here we could take inspiration from McHale and Scarf [2007], using Copula functions to map say the home team's score random variable with the score difference to get to the joint distribution. This would let one then model the distribution of the total goals, for example. This may seem like an unnatural way to arrive at the joint distribution, but the score difference drives the game and we might reasonably expect that home teams tendency to score is stable enough for estimation.

4.2.1 Surrogate Skellam

We noted that Skellam distributions and ordinal models are two ways to solve closely related problems. There is a parallel between this and surrogate Poisson models (where a Poisson regression is fit to multinomial data). As a consequence we are inspired in a later chapter to use a surrogate Skellam distribution to fit to ordinal data. One natural question that arises is how to choose the shift and whether there should be padding parameters to separate out the integer state space, so as to best map the ordinal classes. This is an interesting line of enquiry, which merits further investigation by experts in the statistics community.

4.3 In-play models

There is limited literature in the area of in-play modelling, with the only substantive reference being Dixon and Robinson [1998], where a continuous time Markov jump process is used to reconcile an in-play model with the full time Poisson approach. Assuming an exponential waiting time to the next home or away goal, with the current score being (x, y), the likelihood of particular game k, may be written as

$$L\left(t_k, J_k\right) = e^{-\Lambda[0,1]} * e^{-Y[0,1]} \prod_{l=1}^{l=m_k} \lambda_k(t_{l,l+1})^{1-J_{k,l}} \mu_k\left(t_{l,l+1}\right)^{J_{k,l}} \tag{4.5}$$

where

$\Lambda[t_1, t_2] = \int_{t_1}^{t_2} \lambda_k(t)dt$

$Y[t_1, t_2] = \int_{t_1}^{t_2} \mu_k(t)dt$

$\lambda_k(t_{l,l+1})$ home team's tendency to score in game k between the l^{th} and $(l+1)^{th}$ goals of the game.

$\mu_k(t_{l,l+1})$ home team's tendency to score in game k between the l^{th} and $(l+1)^{th}$ goals of the game.

t_i are the times of change in the score difference.

$J_{k,l}$ is an indicator that is 0 for a home goal and 1 for an away goal, at the l^{th} goal of the k^{th} game.

We have not been able to devise ordinal time-dependent models that reconcile with the pre-play approach. It seems plausible that the cut points would shift over time as the score difference dynamics change, however quite how to do so - estimate parameters from time specific score difference data - is not clear at this time.

An alternative solution would be to have a continuous time Markov process that jumped through states - corresponding to a censored score difference – in an ordinal manner and estimated the parameters using a likelihood or Bayesian approach. The estimation process would involve estimating the generator matrix from time specific score difference data.

A Bayesian approach to ordinal time-dependent score differences, would need the development of a specialized approximating Markov chain so that an MCMC procedure could be used to reach posterior samples for the parameters of the model. Computationally, it is likely that the estimation would be expensive if real-time updates of the in-play probabilities are needed. Having said that, it may be possible for low level language and multiple processor implementation to yield results at sufficient speed to be commercially relevant.

4.4 Handicapped bets

Handicapping the stronger team so that the odds are closer to even (like asian odds of 2), (and the probability that either team wins is closer to equality), generates interest among gamblers in an otherwise less attractive, one-sided, gambling event.

The more likely or unlikely an event is, the harsher the penalty in mis-setting the odds over a large number of games. Beating the counter-party in setting the line excites gamblers. If the handicap is an integer then there will be three payoffs associated with the game (a positive return corresponding to the odds offered, zero loss or full loss of wager).

When the handicap is fractional, there are several payoffs which can occur, designed, in theory at least, to slice up a score difference distribution that does not separate into equal probabilities via the integer handicap. The objective of the handicapper is to determine the handicap - or line - that makes fair value of asian odds close to 2,equivalent to making the statistical odds closer to even. Even statistical odds indicate even chance of the event occuring, which encourages gamblers to take both sides of the handicapped bet. Bookmakers and other market participants will then nibble away at the odds and shift the line, or handicap, to reflect the flow of money.

The fair value of such a handicapped bet can be calculated once the probability distribution has been determined with a basis that is superior to that used by the market to set odds. In particular a shifted score difference random variable, with line h is

$$S(h) = S_a - S_b + (-1)^{\alpha} h = S + (-1)^{\alpha} h \tag{4.6}$$

where

$\alpha = 1$ if team $P(S_a > S_b) > P(S_a < S_b)$ and 0 otherwise.

The handicapped bet has an expected return, which is

$$ER(S(h)) = \sum_{s_h} CF\left(s_h\right) P\left(S(h) = s_h\right) \tag{4.7}$$

where $CF\left(s_h\right)$ is the net cash-flow to the bettor on the handicapped bet if the handicapped score difference is s_h.

In the simple case where h is an integer, the expected return is zero if the notional odds, payoff (profit and return of unit risked) to the gambler on the shifted score random variable exceeding 0, of the bet is

$$\frac{P\left(S(h) > 0\right)}{P\left(S(h) < 0\right)} + 1 = \frac{P\left(S(h) \neq 0\right)}{P\left(S(h) < 0\right)} \tag{4.8}$$

Which models should we prefer to handicap bets? Statistically we can determine better fitting models from those considered. However, there are four reasons that make ordinal score difference modelling particularly attractive:

- Although Poisson models may seem like a natural approach, they do not reflect the dynamics of many sports - in particular football. Teams tend to *chase* the score difference rather than the final score. For example, it is natural in football that a team will change its style of play depending on the extent to which it is behind or ahead in a particular game. This impacts the final score, but in fact the score difference is the driver.

- It is restrictive to assign parametric constraints on the score difference distribution's latent distance between classes. Even if one uses a Skellam distribution the flexibility is not as great as that provided by ordinal classification models. This is because ordinal models have more free parameters to explicitly model the distance between score differences as classes via cut-points on a latent utility variable.

- The impact of extreme scores is controlled by suitably chosen end-point bins that capture the tails of the distribution.
- Handicap bets have payoffs that depend on the score difference at the end of a game, so determining the score difference probabilities accurately leads to correct pricing and staking for these bets.

In total, our view is that score difference modelling via ordinal classifiers is one way to make more accurate probability predictions. This rationale applies more so when considering bets on games that are in-play; before the final score has been determined.

Chapter 5

Assessing classification performance by gambling strategies

This chapter defines a tool to assess predictive ability that the author noted in his analysis of learner performance in classification tasks for test data. The idea arose one night while playing card games, which abstracted into a thought experiment concerning wagers against a unlikely drunken yet infinitely wealthy bookmaker. It generalizes with ease to the case of multiple classification and with some assumptions to ordinal classification. It implies that model and variable selection may be done without nested models and that the likelihood ratio test is redundant.

5.1 Introduction

Consider a classification task. Following analysis and model fitting we set out to assess the generalization performance of our forecasting tool to unseen cases from the test set. There are scenarios where we are interested in assessing the accuracy of estimated predictive probabilities rather than matching the classes of the test cases. For example consider the case where we have two models A and B for a trinomial classification task, where A and B have the same modal class for the pattern given by a test case, but Model B has a less peaked distribution, 5.1.

For a case such as that above, the maximum a posterior (MAP) estimates of the class label are the same - illustrating the bluntness of MAP as an instrument in assessing superiority between models. Less subtle differences in predictive distributions are still hard to assess using test cases. (Although there exist stochastic dominance theorems relating utility functions and preferred distributions (Bawa [1975]) it is different from using test data to assess models.) As most models tend to have some predictive capacity, this is a common problem faced by practitioners of predictive modelling for classification tasks. For example those competing with implied distributions from speculative markets which are pricing risk-taking contracts using supply and demand based on a large number of confounding predictors, at least some of which can be subjectively interpreted. In such an environment, minuscule improvements in estimating the predictive distribution can have significant pecuniary implications.

Confusion matrices and MAP error rates are of limited use in settings such as those above. Strategies such as prequential analysis (Dawid [1992]), averaging over observations in the covariate space and cross-validation provide some possible

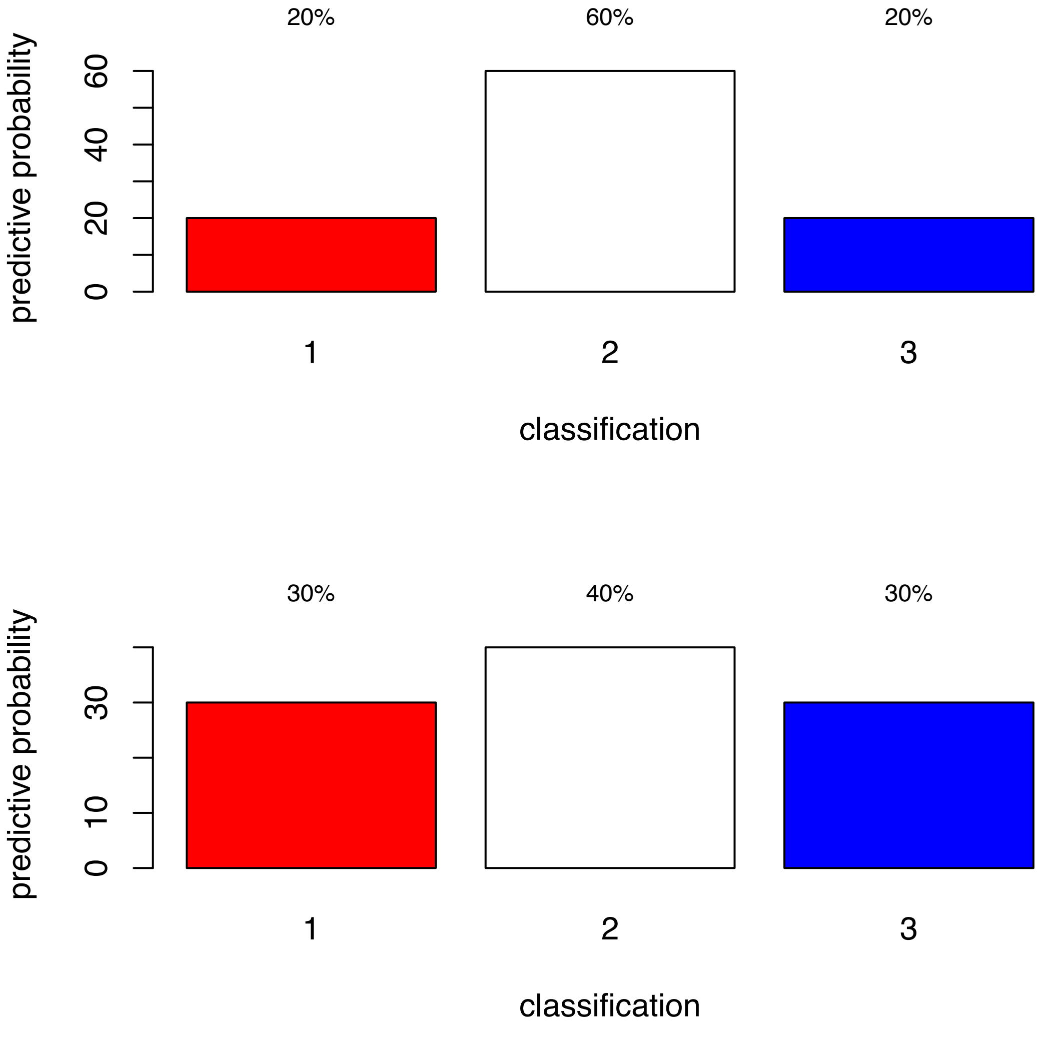


Figure 5.1: Model A (top) and Model B (bottom) have different predictive distributions for a given test case. They assign to the same classification and their loss is the same under the maximum a posterior decision rule.

remedies, (Geisser [1993]). Other measures are Brier and logarithmic scoring (Brier [1950]; Good [1983]). Ripley [1996] and Adams and Hand [2000] note that the best measure of classification performance depends on the problem, providing a detailed discussion with examples. A non-parametric tool to compare model performance across the range of predictive probabilities is a different way to look at the problem.

Following the work of Shannon [1948], Kelly [1956] and Breiman [1961] we propose a thought experiment - initially restricted to the realm of binary classification tasks. Imagine that n test cases are serially offered as wagers to us by a drunken bookmaker, she of infinite wealth. In principal she offers Asian odds of O_d to 1 on the gamble that the test case is true, where odds $O_d(\geq 1)$ are chosen at random, (in order to visualize effects in simulations we will restrict attention to the odds space 1.3 to 4). The profit or loss on each test case is either $O_d - 1$ or -1 for one unit of money notional wager depending on whether the test case is true or not. She is also very accommodating, in allowing us to take the other side of the wager at the odds she offers. In other words, we may choose instead to receive, if we wish, either a profit of 1 if the test case is false or else suffer a loss of $1 - O_d$ if the test case is true, for one unit of money notional wager.

Since we are playing against an ill-informed counter-party and believe the test cases to be Bernoulli realizations from the model predictive probabilities, it follows that we can choose which side of each wager to take, so that the expected payoff for us is always positive. Let us assume that we begin with an arbitrary wealth level, W_0. Should we wager at one time only, for the outcome of a single predictive case, our profit maximizing strategy is bet all of our wealth on the side of the wager that offers us the chance to buy bookmaker implied probability at a discount to the test case fair value model implied odds. In most cases we expect to survive many periods so wealth annihilation - on a single or short run of games, is an unacceptable possibility, whose probability we must ensure is zero under all outcomes of our selected stratagem.

We seek to use our information - the predictive probability of the test case - to stake a fraction of our current wealth on the wager and keep the remaining amount in a risk-free, zero interest earning, asset. This would allow us to benefit from the advantage we have over our bookmaker friend, while resting assured that we can survive with some positive wealth level to wager again if the favourable event is not realized. Intuitively, it follows that the fraction of our wealth staked should depend on the comparison between odds offered and predictive probability estimates. We can specify the nature of the stake differential with respect to increments in odds offered and predictive probability estimates. For example the fraction of wealth risked should increase as:

- The probability of the favourable event increases.
- The difference between the probability and bookmaker implied probability increases.
- the bookmaker odds increases.

Symbolically we can state these requirements as

$$\frac{df}{dp} > 0$$

$$\frac{df}{d\left[p - O_d^{-1}\right]} > 0$$

$$\frac{df}{dO_d} > 0$$

where f, p and O_d are the fractional investment, probability of success and odds on offer for a particular classification. We could also add boundary conditions, such as $f\left(p, p^{-1}\right) = 0$ – no fraction of notional wealth is to be invested when the odds are the reciprocal of the probability of the favourable event.

Rather than construct a closed form expression, solve partial differential inequalities or imply one from simulations – for the optimal fraction – we are aided by the work of Kelly [1956] and Breiman [1961] in taking maximum advantage of our ill-informed bookmaker – informally known as the Kelly criterion (in the two payoff context), which is one of many scenarios that occurs in using an expected logarithmic loss function.

By using an informed proportional wagering strategy we can not only take advantage of the bookmaker, but we construct a measure of how good our predictive probabilities are. In fact by simulating the odds offered by many drunken bookmakers we can see how well our model probabilities forecast over a large test set. Generalizations based on re-ordering the test cases and one-period wagering on the outcome of multiple test cases offer interesting possibilities. Because misspecification of predictive probabilities leads to poor investment performance, we are provided with a ready method of assessing the accuracy of predictive probabilities on test sets.

It is worth noting that such matters are often analyzed through gambler ruin problems, often using Martingales (Doob [1990]). Yet, they tend to assume that the gambler will always stake an equal amount on each wager - a fixed fraction of her initial wealth, with simple analytic wealth metric expressions existing for the case when the odds and probabilities are fixed from trial to trial. This would not aid us in assessing predictive ability. We can change the constant wagering assumption and stake a variable fraction of our current wealth at the time of each game.

In the next section we suggest a particular Markov process W_t, $t = 1, ..., n$ to aid our analysis and call it a Breiman-Ripley wealth process.

5.2 Allocation decisions by maximizing expected logarithmic wealth

In this section there are two types of games, or lotteries, which can be used for Breiman-Ripley processes. The first is where a single premium is paid to enter the game and all payoffs are functions of this sole position. The second is where multiple positions are taken which relate to the outcome payoff of a game. These positions cannot be factored into a single premium which is a function of the payoffs under all scenarios. It is effectively defined by statistically dependent bets on the outcome of a single game.

5.2.1 Sole position payoff games

In a general context, as Breiman [1961] one can start with a initial wealth level, W_0 and play classification games, or wagers, against an ill informed bookmaker (in the sense that the each game has positive expected payoff for the player) sequentially so that the relative change in wealth between games is

$$\frac{W_t}{W_{t-1}} = 1 + f\,(t)\,X_t = S\,(f, t) \tag{5.1}$$

where $-\infty \leq f(t) \leq 1$ is the notional fraction of one's wealth invested in the game. X_t is the payoff on playing the t^{th} game when risking 1 unit of money. X_t has a discrete distribution such that $P(X_t = x_{t,j}) = p_{t,j}$

The support of X will usually have one negative number, 0 and several positive real numbers. The distribution underlying X will have been estimated by some learning algorithm applied to the training data.

Choosing f so that the expected logarithm of wealth, $\log\left(S\left(f, t\right)\right)$, is maximized offers tantalizing statistical properties; Breiman [1961] showed that using f_{opt} leads one to maximize the median of $S(f)$, the relative wealth change, and also that the expected time to reach any given level of wealth is minimized by using such a strategy asymptotically. Also, the probability of ruin is 0 under such a wagering scheme – although the wealth level may get infinitesimally small. We refer to the stochastic process $W_t : t \in \{1 : n\}$ with optimal $f(t)$, $t \in \{1, ..., n\}$ as a Breiman-Ripley wealth process. It is a Markov process and can be modified into the discrete-time Martingale M_t.

$$M_t = \frac{W_t}{\prod_{t=1}^{t=n} \left(1 + f_{opt}\left(t\right) \sum_j p_{t,j} \cdot x_{t,j}\right)} \tag{5.2}$$

Using the martingale M_t, we believe that Breiman's asymptotic results can be arrived at, and possibly extended via Doob's martingale convergence theorems, but such matters are best left to experts in the field.

Note that for a collection of optimized wealth fractions $f_{opt}\left(t\right)$, $t \in \{1, ..., n\}$

$$W_n = \prod_{t=1}^{t=n} \left(1 + f_{opt}\left(t\right) X_t\right) \tag{5.3}$$

The distribution of W_n, the Breiman-Ripley wealth process at time n, is unaffected by rotating the order of the games. In the context of test cases, the final value of W_n is invariant to permutating the order in which the test cases are presented as games, once the pattern and outcomes of the test cases have been revealed. However, varying the ordering of the test cases changes the path taken by the wealth process.

In analysis, plots of $\log\left(W_t\right), S(f,t)$ and $\log\left(S(f,t)\right)$ will be useful in understanding the adherence of the model to the data generating process.

5.2.2 Binary payoff games

For the case of binary classification, with a traditional odds wager on the outcome that the test case Y_t ($Y_t \in \{1, 0\}$) is true

$$P(X_t = O_t - 1) = p_t = 1 - P(X_t = -1) = P(Y_t = 1) \tag{5.4}$$

we note the following.

$$f(t; opt) = \frac{(O_t p_t - 1)}{(O_t - 1)}$$

$$\mathbb{E}[X_t] = O_t p_t - 1 = \mu_{X_t}$$

$$\mathbb{V}ar[X_t] = (1 - p_t)(O_t^2 p_t - 1) = \sigma_{X_t}^2$$

$$\mathbb{E}[\log(S(f,t))] = \log(1 + f(t)O_t)p_t + \log(1 - f(t))(1 - p_t) = \mu_{\log(S(f,t))}$$

$$\mathbb{V}ar[\log(S(f,t))] = (\log(1 + f(t)O_t))^2 p_t + (\log(1 - f(t)))^2 (1 - p_t) - \left(\mu_{\log(S(f,t))}\right)^2$$

A plot of the back and lay fractional investment surface is provided, 5.2. It is easier to see the behaviour of the back only surface, which is a capped version of the back and lay fractional investment surface (negative investment bets are set to zero), 5.3. Notice how it increases the fractional investment as the odds and probability increase, while remaining non-interested when the expected returns are zero.

For large n, by the central limit theorem, $\frac{W_n}{W_0}$ tends to the log normal distribution with parameters

$\mathbb{E}\left[\log\left(\frac{W_n}{W_0}\right)\right] = \sum_{t=1}^{t=n} \mu_{\log(S(f(t),t))}$

$\mathbb{V}ar\left[\log\left(\frac{W_n}{W_0}\right)\right] = \sum_{t=1}^{t=n} \sigma^2_{\log(S(f(t),t))}$

For small n, the distribution of $\frac{W_n}{W_0}$ can be arrived at exactly by iterating through all outcomes on the joint state space of X_t. That is, list all outcomes; calculate their chance using independence between the games; work out each sample path

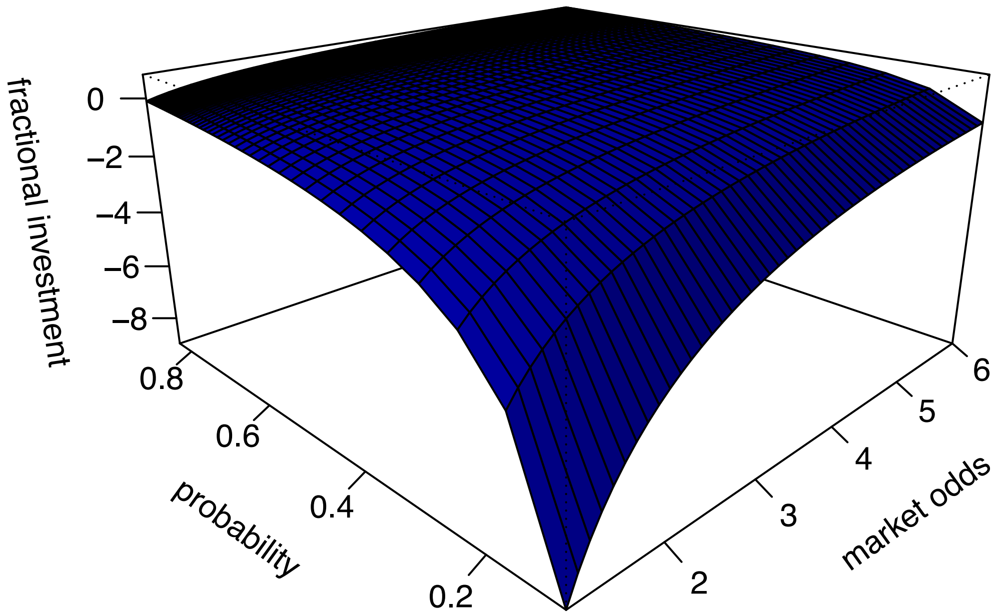


Figure 5.2: Notional fractional investment as a function of the market odds and underlying probabilities.

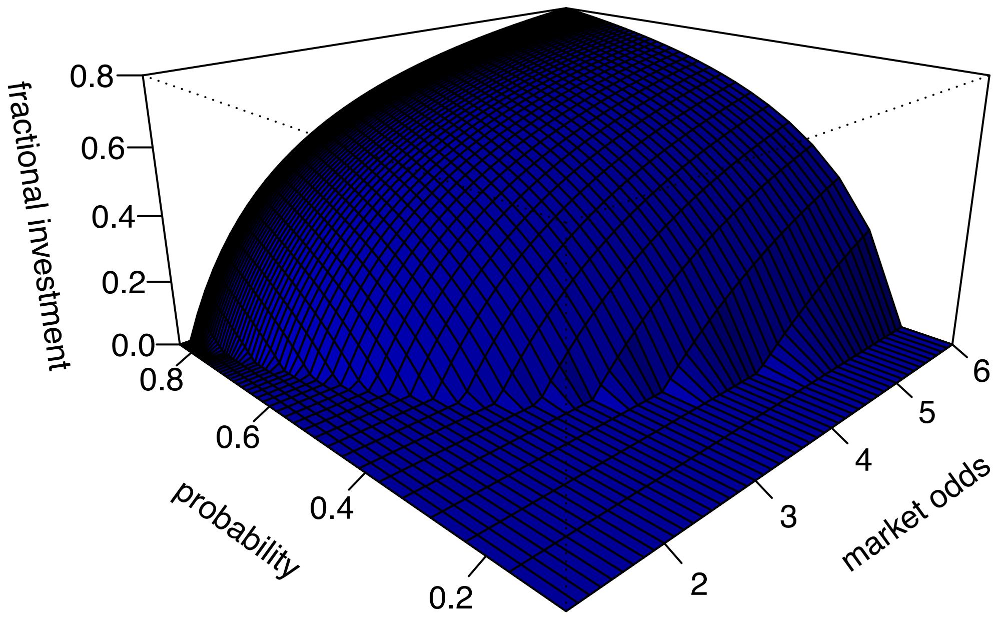


Figure 5.3: Notional fractional investment as a function of the market odds and underlying probabilities, with a floor function that lay bets are changed to a zero investment of wealth.

and the terminal value of $\frac{W_n}{W_0}$. By summing the probabilities when outcomes lead to the same terminal value, we would have discovered the exact distribution of $\frac{W_n}{W_0}$.

Using $f(t) = f_{opt}(t)$ we achieve Breiman's mentioned asymptotic statistical properties for the wealth process. This optimal fraction also satisfies the heuristic differential properties (outlined above) which make sense for a rational gambler in the game.

The next section provides some examples of model comparison via Breiman-Ripley wealth processes.We set $W_0 = 1$ for all the examples in this chapter, without loss of generality. As we place no ordering on the test cases used for constructing the wealth processes, we assume that *time* progresses as the indexing of the test cases in the examples below - the test cases are randomly ordered. Our analysis uses confusion matrices, MAP estimates and plots of Breiman-Ripley wealth processes and returns on their natural and log scales.

5.2.2.1 Examples of binary Breiman-Ripley wealth process

In this section we work with simulated and real datasets. In each, assessing predictive probability estimates is the objective. Two models, estimated with different technologies and covariates are trained. Each is sequentially offered to wager any proportion of its current wealth on the outcome of the test case, based on its estimate of the probability and knowing the implied odds (inverse of predictive probability) offered by the other model.

For the first dataset we use simulated data. The data are vectors (x, y) of length $20,000$. The y's are Bernoulli realizations with success probability given by:

$$p(x) = \frac{exp\,(-\eta(x))}{1 + exp\,(-\eta(x))}$$

where

$$\eta(x) = x + \epsilon$$

and the x and ϵ are sampled uniformly from the interval $(-1.09, 1.09)$ and $(-0.218, 0.218)$ respectively. These sample ranges of x and ϵ are chosen so that $p(x)$ corresponds to reasonable odds.

Half of the sample is used as the training set and the other half as the test set. The superior model's estimated success probabilities for the test set of $10,000$ cases are generated by a glm model with formula given by $y \sim x$ and trained on $n = 10,000$ observations.

Our inferior model is a drunken bookmaker, offering odds corresponding to random choice of predictive probabilities (in the range $0.15 - 0.85$ because extremely likely or unlikely events distort the process with unreasonable jumps that make it difficult to visualize). We also provide the MAP error rate and confusion matrices for each model on the test set. For the superior model:

Simulated data: superior predictive model

$$\begin{array}{lll} & & \qquad\qquad\quad \textit{Predicted}\,\% \\ \textit{Actual}\,\% & \begin{array}{l} \\ \textit{True} \\ \textit{False} \\ \textit{Column total} \end{array} & \begin{pmatrix} \textit{True} & \textit{False} & \textit{Row total} \\ 22.4 & 22.8 & 45.2 \\ 42.1 & 12.7 & 54.8 \\ 64.5 & 35.5 & 100 \end{pmatrix} \end{array}$$

MAP classifier error rate = 64.9%

Assessing our superior model with MAP criteria we are disappointed to find that we do worse than random in getting an error rate of 64.9%. The confusion matrix confirms that making absolute predictions indicates poor performance. However, our predictive probabilities are only small perturbations of the true probabilities for the data generating process, making the results surprising.

Using just randomly selected predictive probabilities we have the following confusion matrix:

Simulated data: randomly chosen predictive probabilities

$$\begin{array}{lll} & & \qquad\qquad\quad \textit{Predicted}\,\% \\ \textit{Actual}\,\% & \begin{array}{l} \\ \textit{True} \\ \textit{False} \\ \textit{Column total} \end{array} & \begin{pmatrix} \textit{True} & \textit{False} & \textit{Row total} \\ 21.3 & 23.9 & 45.2 \\ 26.9 & 27.9 & 54.8 \\ 48.2 & 51.8 & 100 \end{pmatrix} \end{array}$$

MAP classifier error rate = 50.8%.

Random predictive probabilities give better predictive performance than using predictive probabilities from a superior model, when measured using MAP and confusion matrices. MAP is not always the best criteria when assessing predictive performance.

Individual returns and sample wealth paths along with expected returns are provided on Breiman-Ripley compounded and cumulative logarithmic Breiman-Ripley compounded basis using the simulated data, 5.4.

From the charts, the Breiman-Ripley wealth process approach makes it clear that predictive probabilities from the model are better than randomly choosing predictive probabilities (as a drunken bookmaker may set odds). That the logarithm path is largely conforming to expected returns we are encouraged that the predictive model is a good fit to the data generating mechanism, relative to the random bookmaker odds. In other words, most heterogeneity has been accounted for by the predictive model. In addition a bivariate plot of the predictive probabilities from the two models, 5.5, is provided. It serves to indicate that there is real variation in the two models which is being picked by the Breiman-Ripley wealth process in a non-parametric way.

As expected, playing against a less informed model as bookmaker, the superior model leads to exponential growth in wealth, with the logarithm of wealth performing inline with expectations. This shows that the model has most of the information present in the data-generating process, although how to quantify this adherence to expectations is not obvious.

Our second example uses real data from Hosmer and Lemeshow [1989]. The dataset concerns 189 births at a US Hospital. As Venables and Ripley [2002] we model the low birth weight binary variable. Two different models are con-

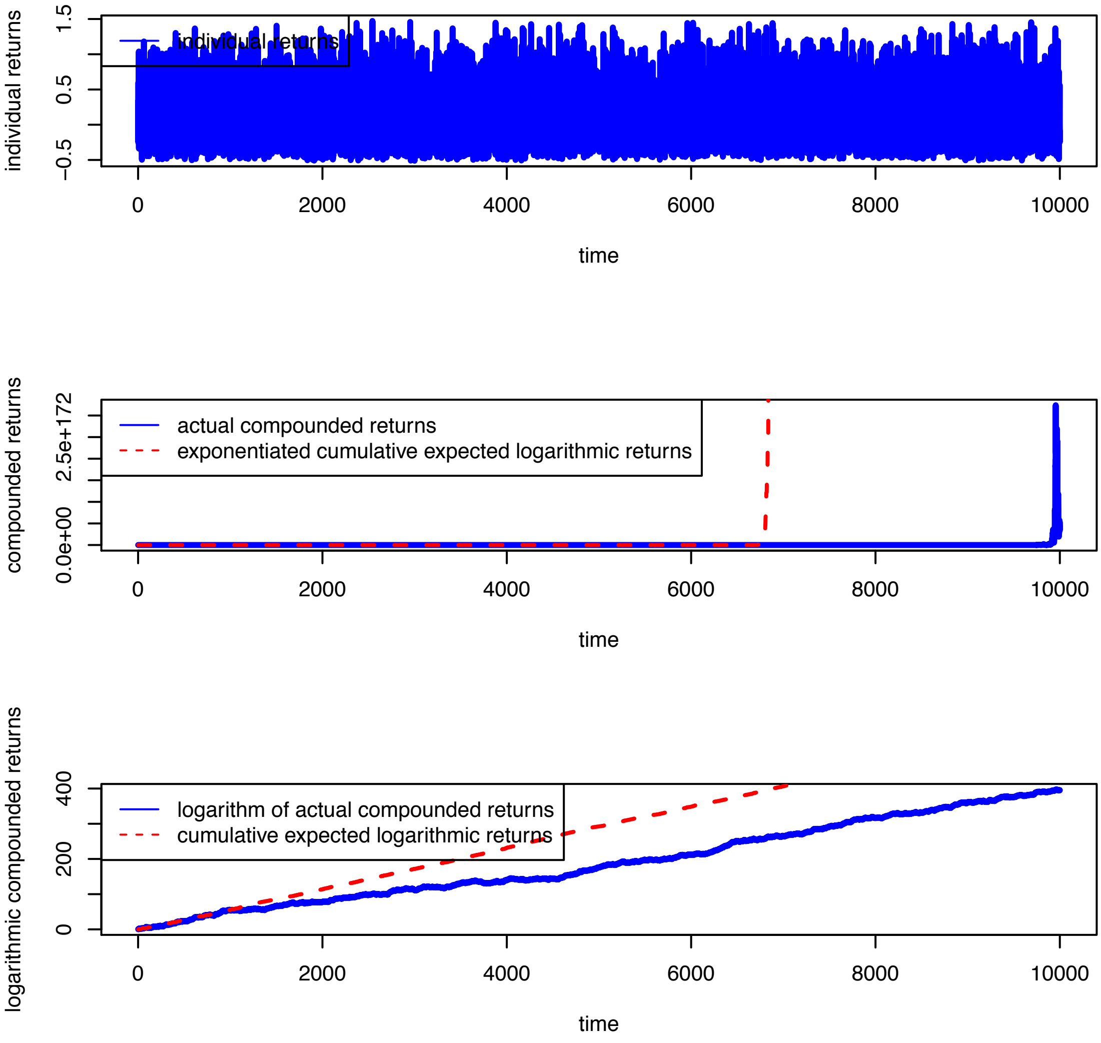


Figure 5.4: Simulated data: plots of individual returns, Breiman-Ripley compounded and logarithmic Breiman-Ripley wealth paths generated by betting for and against binary classification games.

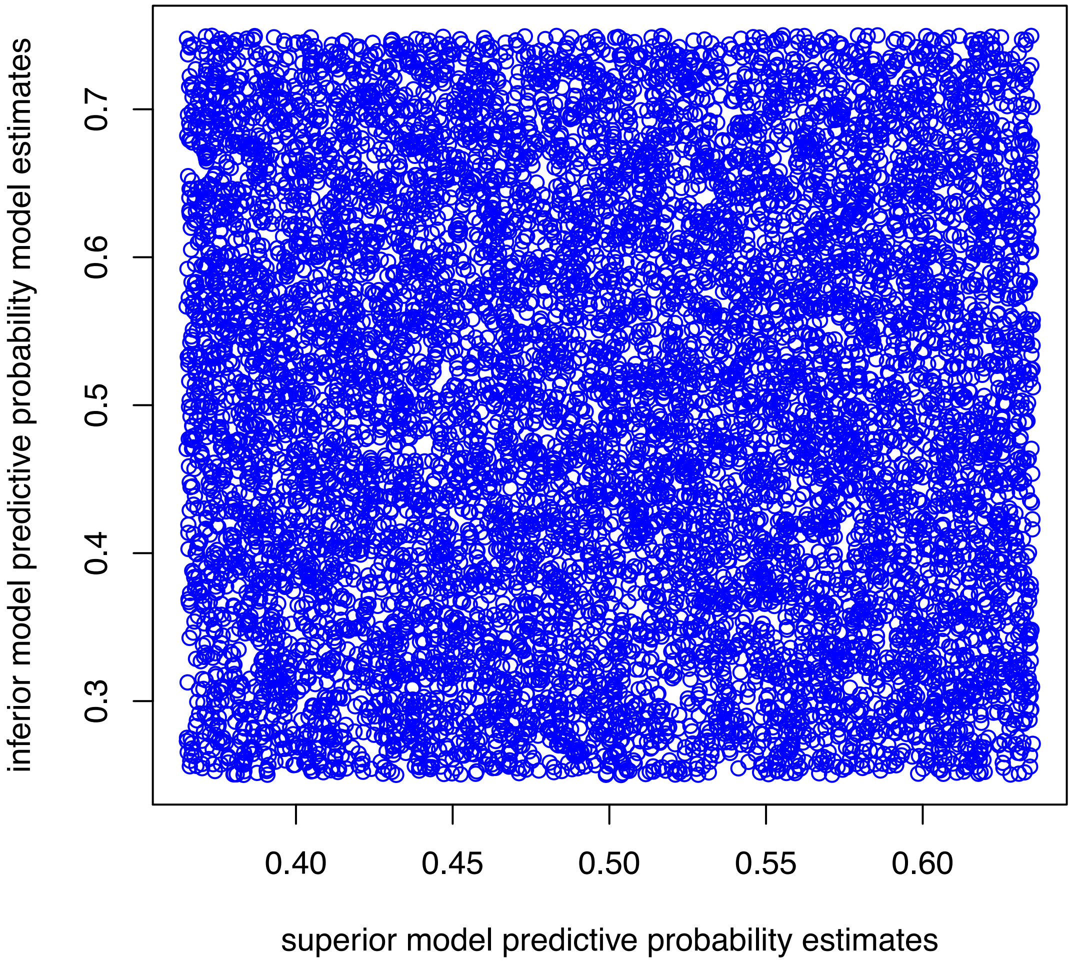


Figure 5.5: Simulated data: Comparison of probability estimates from two models for all test cases.

sidered. Firstly a logistic regression using race and smoker status of the mother as covariates, we will call this the small glm model. Secondly a stepwise model selected from the set of covariates modelled in the analysis of Venables and Ripley [2002]. These covariates are race, smoker status, uterine irritability, history of hypertension, number of physician visits in the first trimester, weight of mother at last menstrual period(in lbs) and whether the mother had previous premature labours. This is referred to as the stepAIC model. The predictive probabilities from the two models are plotted, 5.8. To make the most of the small dataset we use leave-one-out-cross-validation in estimating predictive probabilities for each model. A Breiman-Ripley wealth path, using randomly ordered test cases, logarithm of Breiman-Ripley wealth path and actual returns from wagering the two models against each other is given, 5.6 and 5.7. In addition we provide the MAP error rate and percentage confusion matrices for both models.

Infant birth-weight data: Small covariate glm model

$$
\begin{array}{ccc}
 & & \textit{Predicted}\,\% \\
\textit{Actual}\,\% & \begin{array}{l} \\ \textit{True} \\ \textit{False} \\ \textit{Column total} \end{array} & \left(\begin{array}{ccc} \textit{True} & \textit{False} & \textit{Row total} \\ 3.7 & 65.1 & 68.8 \\ 2.6 & 28.6 & 31.2 \\ 6.3 & 93.7 & 100 \end{array}\right)
\end{array}
$$

MAP classifier error rate = 67.7%

Infant birth-weight data: stepAIC model

$$\begin{array}{lll} & & \qquad\qquad Predicted\,\% \\ Actual\,\% & \begin{array}{l} \\ True \\ False \\ Column\ total \end{array} & \begin{pmatrix} True & False & Row\ total \\ 6.9 & 61.9 & 68.8 \\ 12.1 & 19.1 & 31.2 \\ 19.0 & 81.0 & 100 \end{pmatrix} \end{array}$$

$$\text{MAP classifier error rate} = 74\%$$

Although the MAP measure would suggest that the small glm model with formula: low birth weight $\sim$ smoke + race (in Rogers-Wilkinson notation, Wilkinson and Rogers [1973]) is better, the Breiman-Ripley wealth path makes it clear that the stepAIC model is superior at estimating predictive probabilities more accurately.

The next section deals with games on multinomial state spaces, which encapsulate treatment of the ordinal case.

5.2.3 Multiple position payoff games

Suppose that we have, as in our real-world problem, multiple states that a random variable may take and a ready market where odds are offered on each state. As before for the n^{th} game – each game is played with statistical independence of others and only one game per time period – let the distribution be defined by $P(X_n = x_{n,j}) = p_{n,j}$, where the support is defined as one of K classes, denoted by

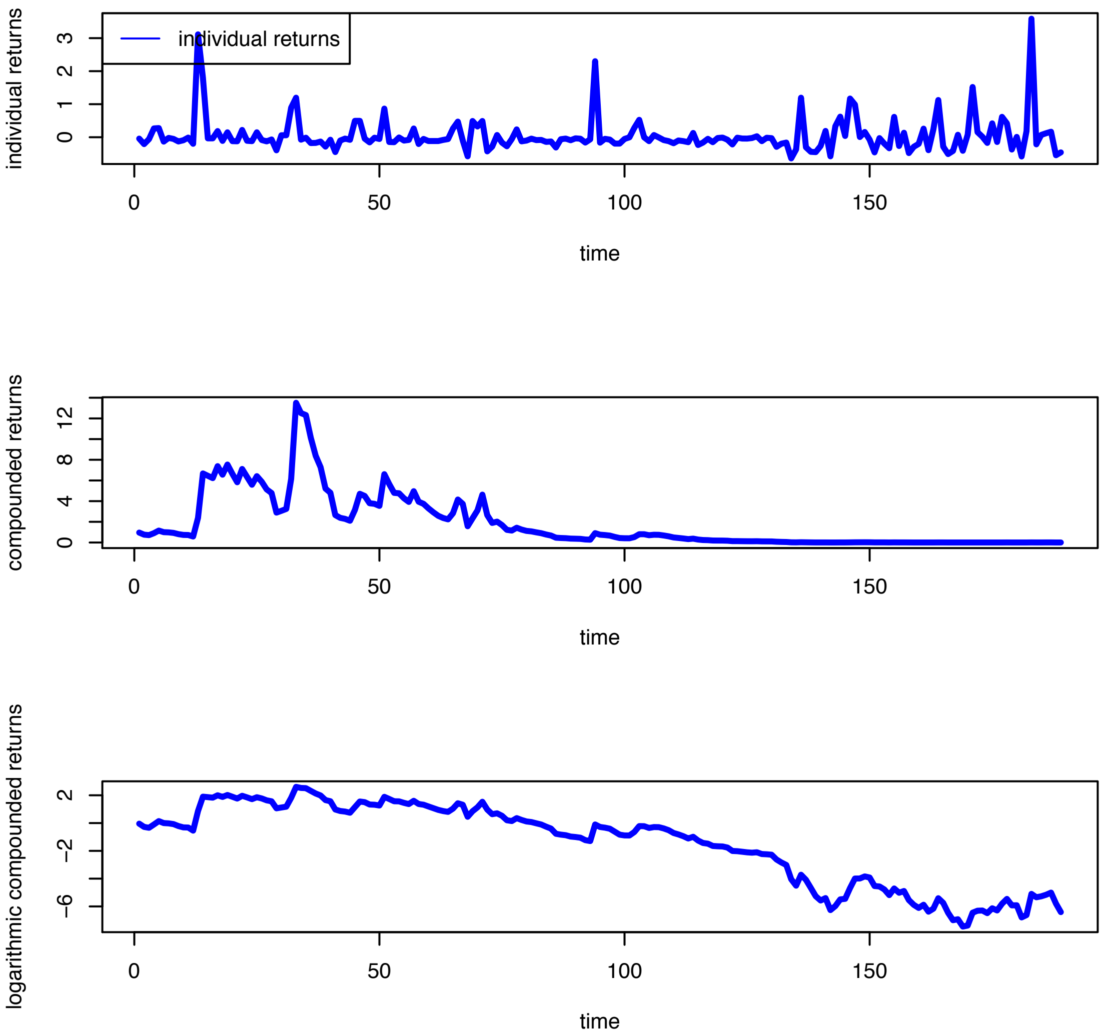


Figure 5.6: Infant birth weights: plots of individual returns, Breiman-Ripley compounded and logarithmic Breiman-Ripley wealth paths generated by betting for and against binary classification games with the stepAIC model being the bookmaker and small covariate glm being the gambler.

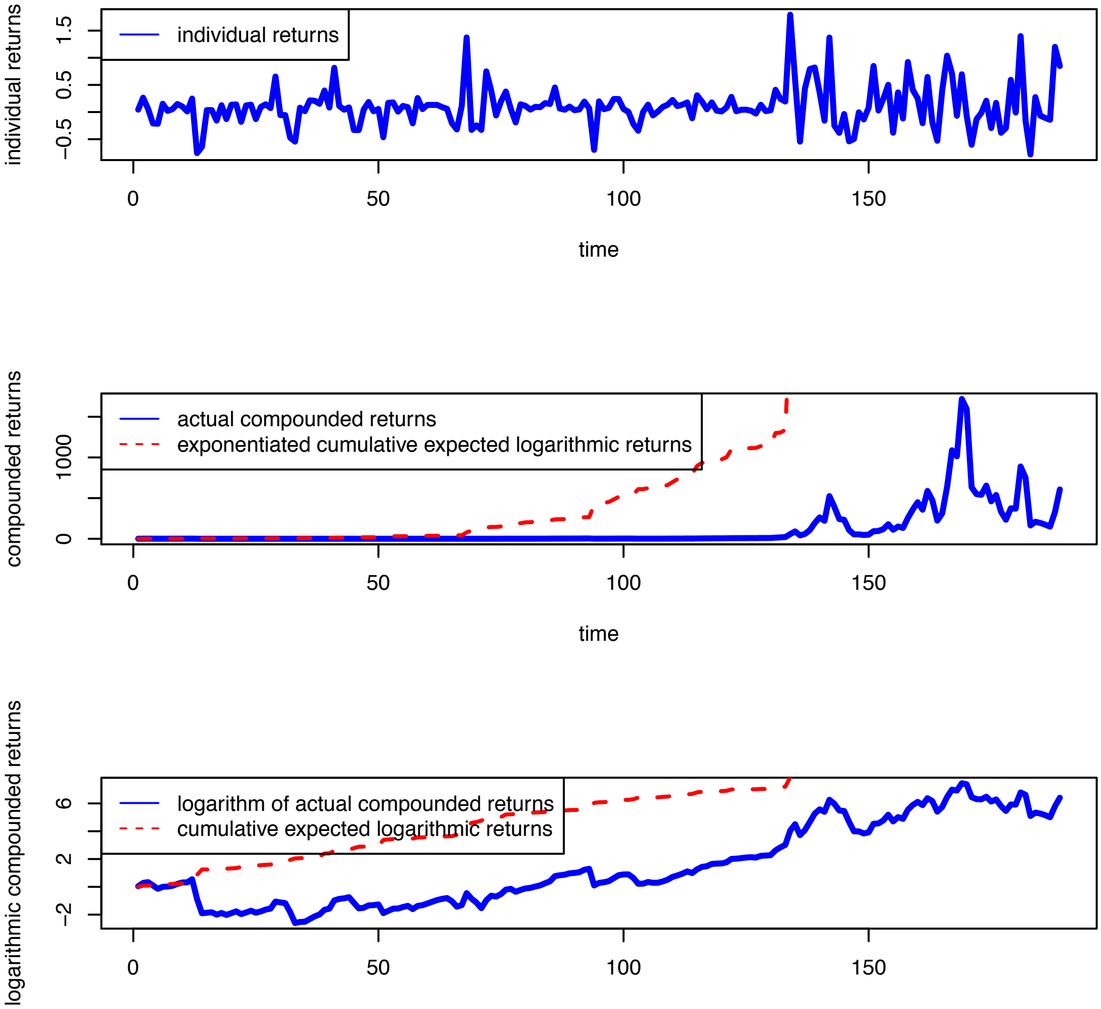


Figure 5.7: Infant birth weights: plots of individual returns, Breiman-Ripley compounded and logarithmic Breiman-Ripley wealth paths generated by betting for and against binary classification games with the small covariate glm being the bookmaker and stepAIC model being the gambler.

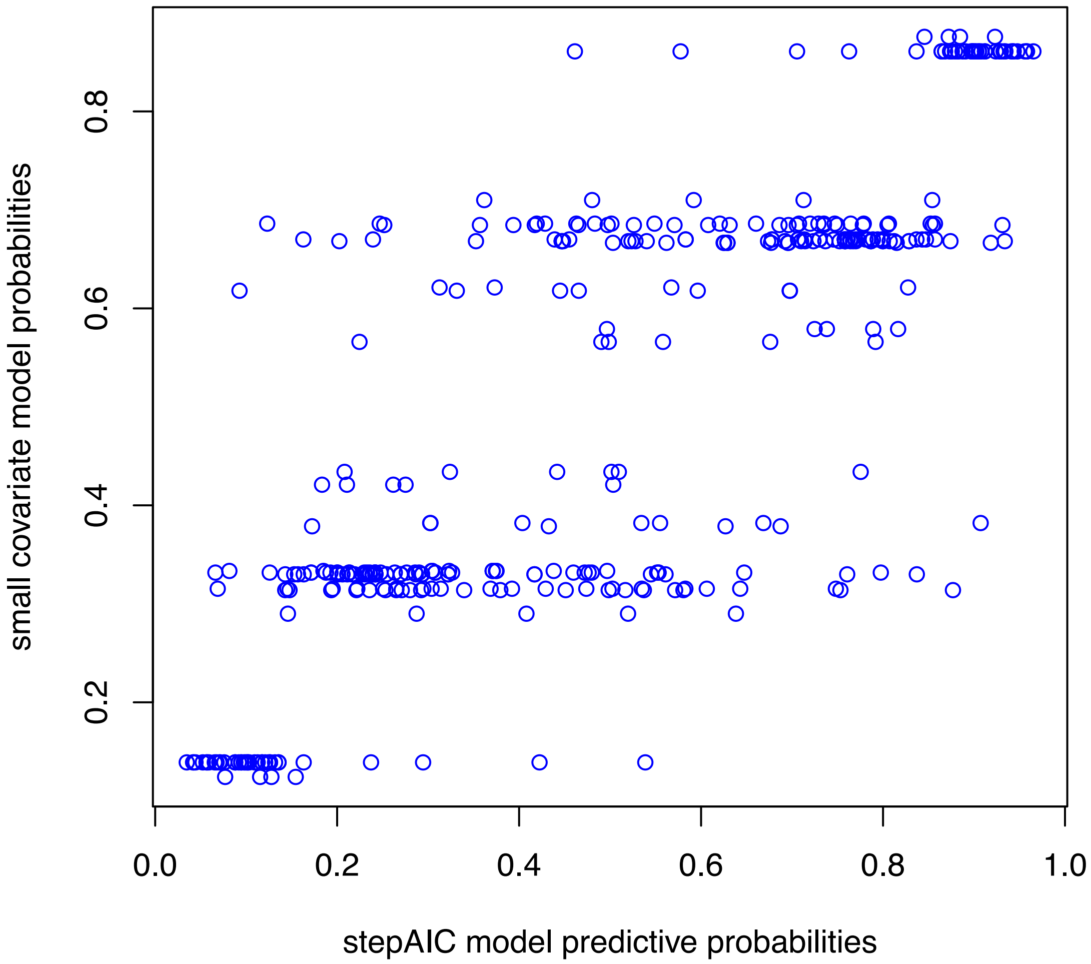


Figure 5.8: Infant birth weights: comparison of probability estimates from two models for all test cases.

$$x_{n,1}, ..., x_{n,K}$$

Further the odds associated with these states are given by

$$O_{n,1}, ..., O_{n,K}$$

In the absence of strong theoretical foundations, we believe that the case of ordered classes should be treated as that of multinomial classification games. (Any additional information from the ordinal nature of such games is assumed to be captured by the market odds setting process.)

Assume that the game is positive expectation at the odds offered; the market is irrational and one is able to wager on all outcomes in any proportion, but cannot keep any wealth in the riskless asset. Breiman [1961] showed that the best allocation of one's assets in such a game is to place one's wealth on each outcome in proportion to the probability that it occurs. That is, the best allocation is achieved by ignoring the odds on offer and staking the entire wealth available according to the probability distribution underlying the game. This counter-intuitive result is easily verified with Lagrange multipliers and the monotonic property of the logarithm operator.

This sets us up well for creating multiple position Breiman-Ripley wealth processes for multinomial classification games.

5.2.3.1 Example: Danish housing data

Danish housing data was analyzed by Cox and Snell [1981]. As done by Venables and Ripley [2002] we fit a multinomial and POLR model to the three level (*High, Medium, Low*) ordered response satisfaction with respondent housing conditions. The covariates are type of housing, contact and influence,being modelled additively. Using leave-one-out-cross-validation we generate predictive distributions for test cases of both models. As with the binary classification case, we make the models play wagering games against each other. The odds offered for each class being correct are those implied by the probability distribution of each model. We assume all information relating to the ordinal nature of the classification game have been incorporated into the odds given. Using a multiple position Breiman-Ripley wagering strategy each model chooses to place its wealth on every test case in proportion to its estimation of the predictive probabilities (which are generated from leave-one-out-cross-validation). The corresponding multiple position Breiman-Ripley wealth plots, individual returns and difference in predictive probabilities are provided, 5.9, 5.10 , 5.11.

In line with Venables and Ripley [2002] the wealth path approach shows that the POLR model is superior to the multinomial approach. Note that we didn't need to resort to asymptotic statistical tests. However, the paths do not exhibit as strong a growth trajectory as the simulation example. In part, this is explained by the small sample, increased number of classes and likely influenced by heterogeneity not accounted for by the data.

In comparison confusion matrices and MAP error estimates for the two models are not as useful in comparing models, suggesting only a slight benefit in using the POLR model. Also the confusion matrices show that the POLR model never predicts the middle class. Intuitively, this may seem problematic, however never forecasting a particular class is not in itself a problem. It is feasible that a reasonable model will have a distribution that never forecasts some class over all test cases using the MAP criteria. By contrast, the Breiman-Ripley wealth approach will assess the model across its whole distribution, whether it be playing against another model or a random bookmaker.

Danish housing data: multinomial model

$$\begin{array}{lll} & & \qquad\qquad\qquad \textit{Predicted}\,\% \\ \textit{Actual}\,\% & \begin{array}{l} \\ \textit{Low} \\ \textit{Medium} \\ \textit{High} \\ \textit{Column total} \end{array} & \begin{pmatrix} \textit{Low} & \textit{Medium} & \textit{High} & \textit{Row total} \\ 12.5 & 1.4 & 19.4 & 33.3 \\ 12.5 & 1.4 & 19.4 & 33.3 \\ 12.5 & 1.4 & 19.4 & 33.3 \\ 37.5 & 4.2 & 58.3 & 100 \end{pmatrix} \end{array}$$

MAP classifier error rate = 68.1%

Danish housing data: POLR model

$$\begin{array}{lll} & & \qquad\qquad\qquad \textit{Predicted}\,\% \\ \textit{Actual}\,\% & \begin{array}{l} \\ \textit{Low} \\ \textit{Medium} \\ \textit{High} \\ \textit{Column total} \end{array} & \begin{pmatrix} \textit{Low} & \textit{Medium} & \textit{High} & \textit{Row total} \\ 12.5 & 0 & 20.8 & 33.3 \\ 12.5 & 0 & 20.8 & 33.3 \\ 12.5 & 0 & 20.8 & 33.3 \\ 37.5 & 0 & 62.5 & 100 \end{pmatrix} \end{array}$$

MAP classifier error rate = 66.7%

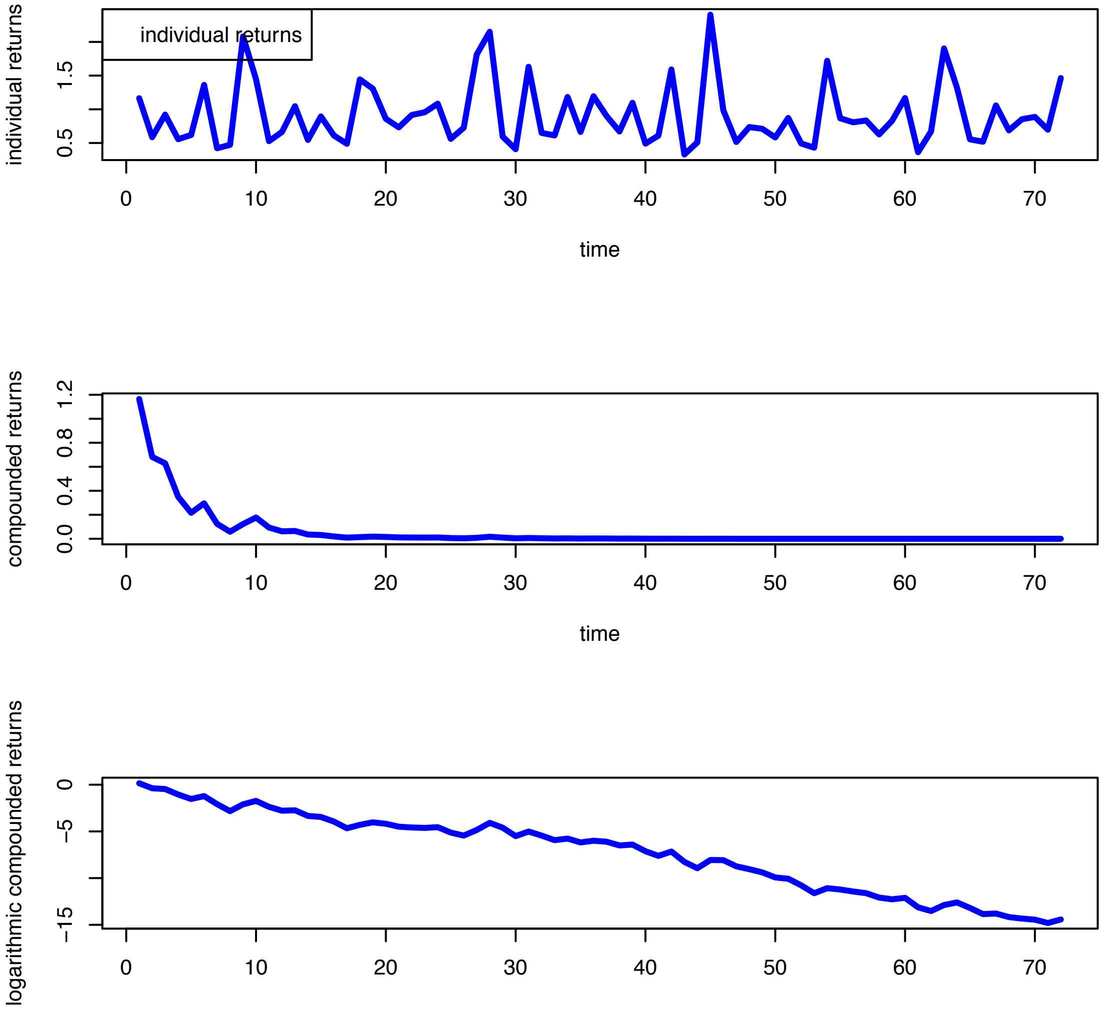


Figure 5.9: Danish housing data: plots of individual returns, Breiman-Ripley compounded and logarithmic Breiman-Ripley wealth paths generated by betting for and against classification games with the POLR model being the bookmaker and multinomial model being the gambler.

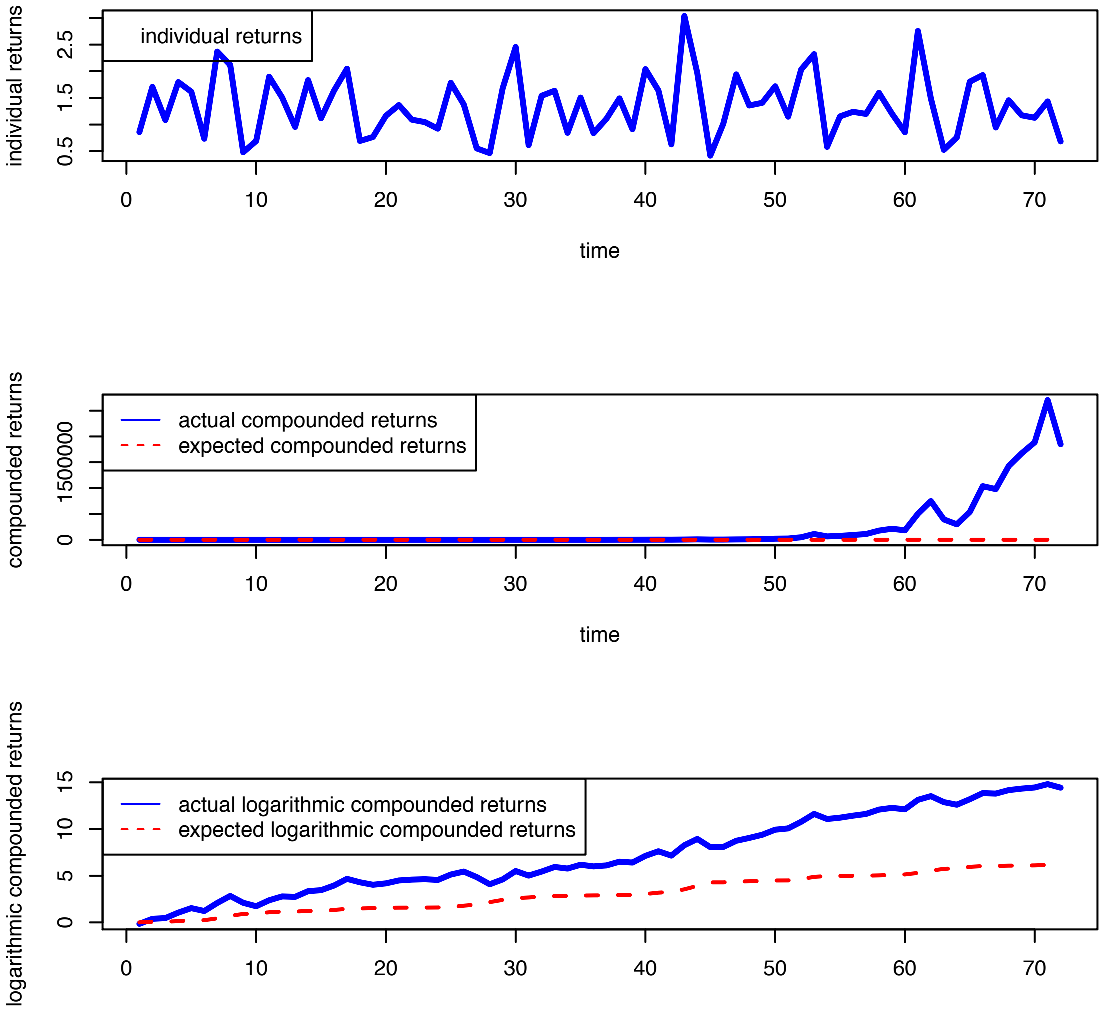


Figure 5.10: Danish housing data: plots of individual returns, Breiman-Ripley compounded and logarithmic Breiman-Ripley wealth paths generated by betting for and against classification games with the multinomial model being the bookmaker and the POLR model being the gambler.

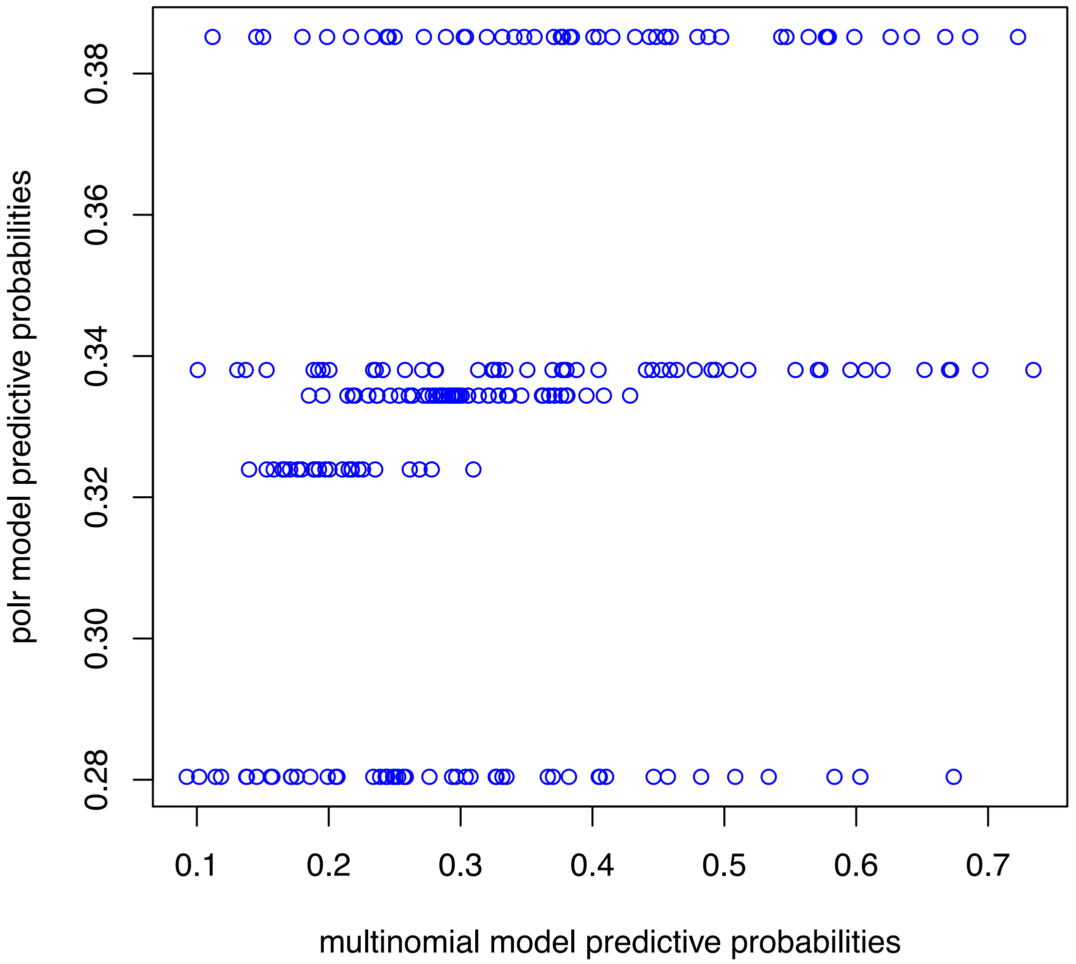


Figure 5.11: Danish housing data: comparison of probability estimates from two models for all test cases.

5.3 Wagering with imperfect models

In multinomial applications, ordered and unordered, one usually meets an obstacle that prevents this elegant strategy from being applied. Determining the predictive probability distribution accurately across all (outcome) classes is seldom possible. There tends to be material uncertainty in the class probabilities for at least a pair of classes, with adjacency between the problematic classes being common for ordinal data. (Tomas [2008] suggests strategies that might be employed to combine many classifiers so that one improves overall classification performance.) Faced with this reality, it would be fool-hardy to wager on all outcomes. We should, instead, restrict ourselves to considering wagers on classes where we have confidence in the accuracy of the model at setting predictive probabilities.

This restricted optimization problem changes the optimal fractions to be invested in considered classes (those we have faith in determining predictive probabilities accurately). If $cc = (C_1, ..., C_q)$ is the set of q considered classes, those which we are happy to gamble on, then the wealth increment random variable after allocating $\mathbf{f}_{cc} = \left(f_1^{cc}, ..., f_q^{cc}\right)$ to the considered classes, is given by

$$\frac{W_n}{W_{n-1}} = 1 + \sum_{u=1}^{u=q} f_u^{cc,n}(O_{n,C_u} I\{X_n = x_{n,u}\} - 1) = S\left(\mathbf{f}_{cc}, n\right) \tag{5.5}$$

with the constraint that $\mathbf{f}_{cc}$ lies in the positive quadrant simplex that is constrained to sum to at most one.

The optimal vector $\mathbf{f}_{cc}$ is determined as that which maximizes the expected value of $\log\left[S\left(\mathbf{f}_{cc}, n\right)\right]$; this can be done using the Nelder-Mead, or amoeba Olsson and Nelson [1975] method with suitable error handling, or else the geometry of the surface may be used via a multinomial logistic transform of the parameters. The expected value and variance quantities for the optimal wealth path are similar to those highlighted above for the binary classification case.

A matter not touched on is how to allocate one's wealth over multiple games being played in the same time period. This is a problem that explodes exponentially in the number of terms, for the number of games and number of outcome classes. It is not relevant to assessing classification performance since we may assume that games are played sequentially even if this is not the case in reality.

However, for gambling applications it is important to determine the complete wealth allocation over simultaneous games. The expression to be optimized is the expected value of the logarithm of

$$\frac{W_n}{W_{n-1}} = 1 + \sum_{JGO} \sum_{t=1}^{t=m_n} \left(\gamma\left(JGO, WFIG\left(t\right)\right)\right) = S\left(\mathbf{f}_{cc}, n\right) \tag{5.6}$$

where:

- JGO = joint game outcomes, is the set of game outcomes that describe $X_{t,n}$, for $t = 1, ..., m_n$.
- $X_{t,n}$ is the outcome of the t^{th} game played at time n.
- $WFIG(t)$ = vector of wealth fractions in game t
- $\gamma(JGO, WFIG) = \sum_{v=1}^{v=q} \left[WFIG(t)\right]_v \left(I_{(X_{t,n}=x_{t,n,v})} * O_{t,n,v} - 1\right)$

- I_A is the logical indicator.

- m_n is the number of games being played simultaneously at time n.

Computationally, once more than 15 games are considered simultaneously, it becomes unfeasible to perform the optimization with the technology currently available due to the number of terms in the expression being more than the memory capacity of the computing facilities available. In other words, the function to be optimized becomes too large in the number of terms for even a reasonable number of games and outcomes.

Thorp [1997] solved the problem for the case of two simultaneous wagers (with both games being given at odds of two) for different estimated probabilities. From his experience, 5–10 binary classification games could be played simultaneously with a scaled fractional investment of the single game case. We believe that scaling the individual investment fraction is a reasonable (but conservative) strategy. Approximations to solving this optimization problem have been suggested by use of first and second moment approximations, or using genetic algorithms to simulate the largest components of the derivatives of the function (Whitrow [2007]).

A joint game Breiman-Ripley process may lead to more stable returns by reducing the variability of the process and concentrating the distribution about smaller range than a single game Breiman-Ripley process.

5.4 Assessing model predictive performance

Our ideas on assessing model performance using maximizing expected logarithmic wealth are:

- The sample wealth path for a model with predictive ability should grow exponentially. Significant unexpected movements can be seen as indicative of problems in the fit.

- The actual wealth path can be compared with that expected.

- The test set can be re-ordered and performance of wealth paths to expected compared.

- The performance of the model can be compared against many drunken bookmakers. Summary statistics could then be compared to those of the hypothesized wealth distribution.

- Two different models can be compared, as playing against the same drunken bookmaker. Differences in the wealth path can be used to indicate strengths and weaknesses in the predictive model.

- Playing multiple games at the same time which allocates our wealth across multiple wagers at each time. This would introduce stability to the wealth process, providing a different vista. It would also be an implicit check on the independence of the test cases.

- One might focus on a region of the predictive probability range, which is of particular interest. In that region, strategies as above could be employed to find a locally optimal model.

- The process allows for variable selection within the same nested family of models as well as insight into models from different families.

- Just as model search can be done so that the MAP error is minimized, there are likely strategies which allow using one model to leverage into improving another. This would be along the lines of selecting those parameters which maximize the Breiman-Ripley terminal wealth for a training set, when the model to be optimized is played against another model, say a naive Bayes classifier.

Chapter 6

Results of model fitting

The model fitting process has been performed using the RankingSVM algorithm and also a non-informative Bayesian proportional odds logistic regression (POLR) model for the artificial data; Bayesian POLR for the soccer data with all covariates, along with RankingSVM using a smaller set of covatiates for the soccer data.

Our assessment tools are various types of calibration plot, Breiman-Ripley wealth paths and confusion matrices. For the POLR models we check the assumption of proportional odds.

6.1 Artificial data

6.1.1 POLR models

We tried the Cowles [1996] Bayesian POLR (with non-informative Gaussian priors) model with the two covariates for the data, but the results were disappointing.

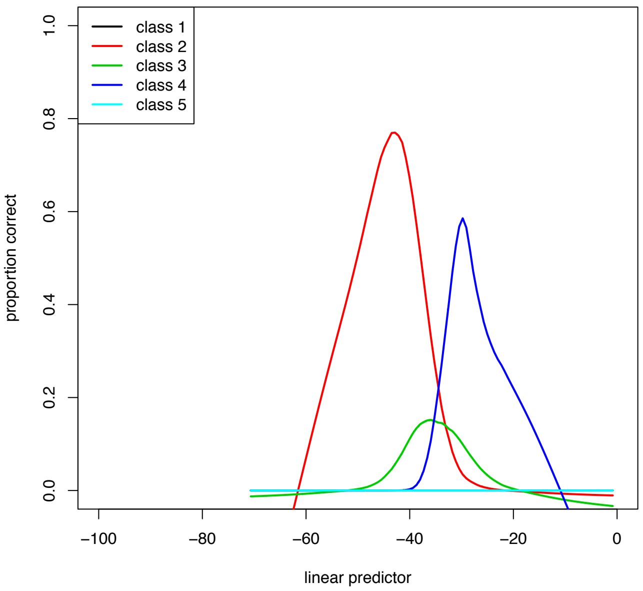


Figure 6.1: Empirical class proportions fitted by lowess models. Class 1 is coincident with class 5.

The model was only able to forecast to a single class. The weakness of POLR, and other parametric models is that they are unable to pick up interesting patterns within the covariate space unless their formulation is sufficiently flexible. However, we found that allowing for interaction between the covariates produced a better predictive model, by MAP criteria.

It is reasonable to check the proportional odds, or parallel logits, assumption in assessing the model. Since there is no natural partitioning of the covariate space, we plot the proportion correct for each class, 6.1 and empirical logits against the linear predictor (shifted by relevant cut-points), 6.2. A parallel empirical logits assumption seems plausible.

There is however a problem with checking the parallel logits assumption with a linear predictor estimate for POLR. That is the cut points of the POLR model.

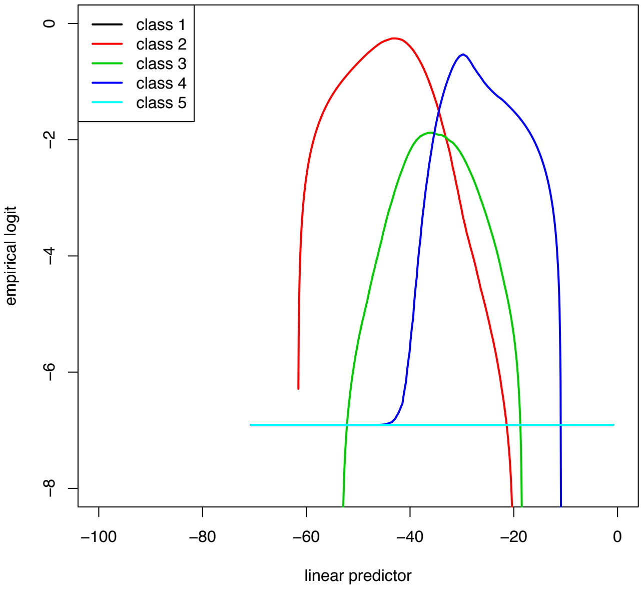


Figure 6.2: Empirical logit for each class. Class 1 is coincident with class 5.

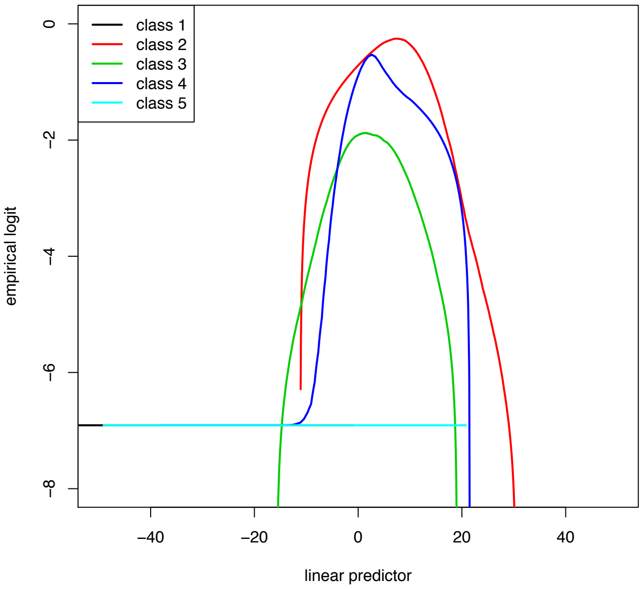


Figure 6.3: Empirical logit shifted by cut-point estimates. Class 1 is co-incident with class 5.

In order to look on the same scale, we need to shift the linear predictor by the cut points. This introduces additional uncertainty in the estimates of the shifted linear predictor. Once we apply the cut-point shifts, we see more clearly that a parallel empirical logits assumption is reasonable, 6.3.

A confusion matrix and MAP error rate are provided, both of which perform rather well. The MAP error rate of the POLR model is substantially better than a randomly selecting classes without knowledge of the covariates, which is encouraging.

POLR model with interaction between covariates

$$\begin{array}{ll} & \begin{array}{cccccc} \;1\; & \;2\; & \;3\; & \;4\; & \;5\; & \textit{Row total} \end{array} \\ \begin{array}{l} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \textit{Column total} \end{array} & \left(\begin{array}{cccccc} 11.2 & 1.3 & 0.0 & 0.0 & 0.0 & 12.5 \\ 0.8 & 25.6 & 4.2 & 0.0 & 0.0 & 30.6 \\ 0.0 & 2.6 & 18.8 & 1.7 & 0.0 & 23.1 \\ 0.0 & 0.0 & 3.0 & 18.5 & 0.6 & 22.1 \\ 0.0 & 0.0 & 0.0 & 0.6 & 11.1 & 11.7 \\ 12.0 & 29.5 & 26.0 & 20.8 & 11.7 & 100.0 \end{array} \right) \end{array}$$

MAP classifier error rate = 14.8%

We can also look at a calibration plot across all class probabilities, 6.4. It shows that the model is good at accurately forecasting predictive probabilities, agreeing with the results of Herbrich *et al.* [2000].

Using a Breiman-Ripley wealth process, when compared with a multinomial model - fit without the interaction term, the interaction term POLR model shows it is superior at estimating the true odds across the predictive distribution, 6.5.

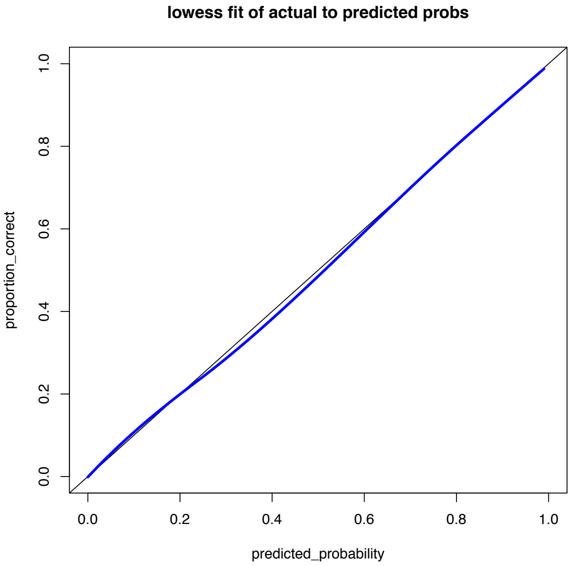


Figure 6.4: Calibration plot for the POLR model.

6.1.2 Surrogate Skellam model

As discussed in Chapter 4, we tried a surrogate Skellam fit to the artificial data. Although it gets the relativities correct, the model underestimates class probabilities. Looking at the confusion matrix, we note that the model never selects classes 3 and 4, by a MAP criteria. The MAP error rate is also substantially higher, reflecting an inability to discriminate classes 3 and 4 of the model.

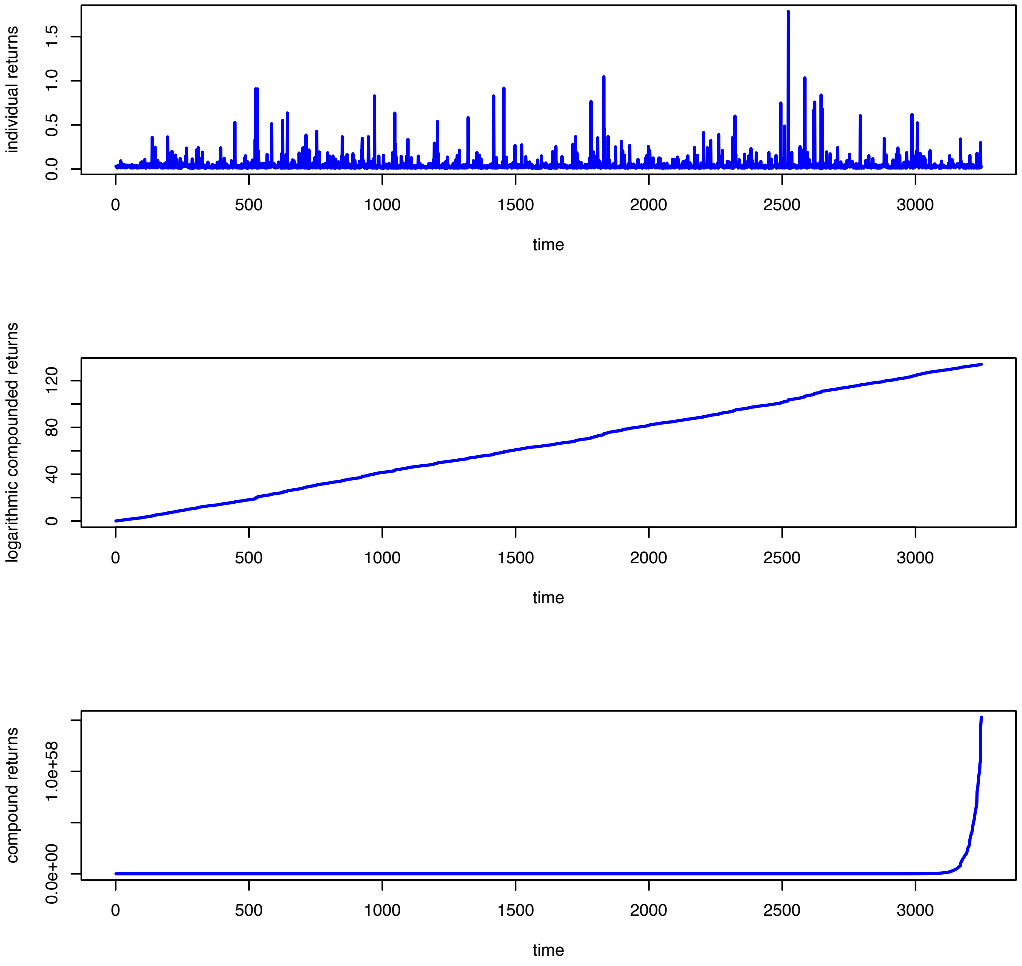


Figure 6.5: Breiman-Ripley wealth process for a multinomial model as bookmaker and POLR model as gambler.

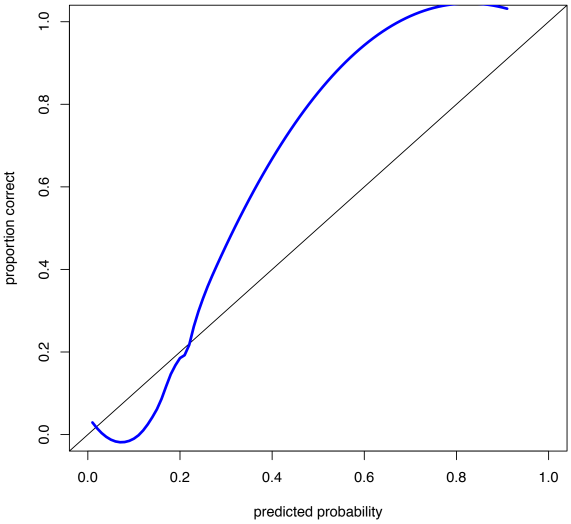


Figure 6.6: Surrogate Skellam calibration plot for artificial data.

$$\begin{array}{lcccccc}
 & 1 & 2 & 3 & 4 & 5 & \textit{Row total} \\
1 & 11.5 & 0.0 & 0.0 & 0.0 & 1.0 & 12.5 \\
2 & 8.0 & 22.6 & 0.0 & 0.0 & 0.0 & 30.6 \\
3 & 0.0 & 2.6 & 0.0 & 0.0 & 3.9 & 23.2 \\
4 & 0.0 & 0.0 & 0.0 & 0.0 & 20.4 & 22.1 \\
5 & 0.0 & 0.0 & 0.0 & 0.0 & 11.7 & 11.7 \\
\textit{Column total} & 19.5 & 43.5 & 0.0 & 0.0 & 37.0 & 100.0
\end{array}$$

MAP classifier error rate = 61.9%

As might be expected, the Breiman-Ripley wealth process performed badly for the Skellam fit against the POLR model, where the leave-one-out-cross-validation procedure has been used.

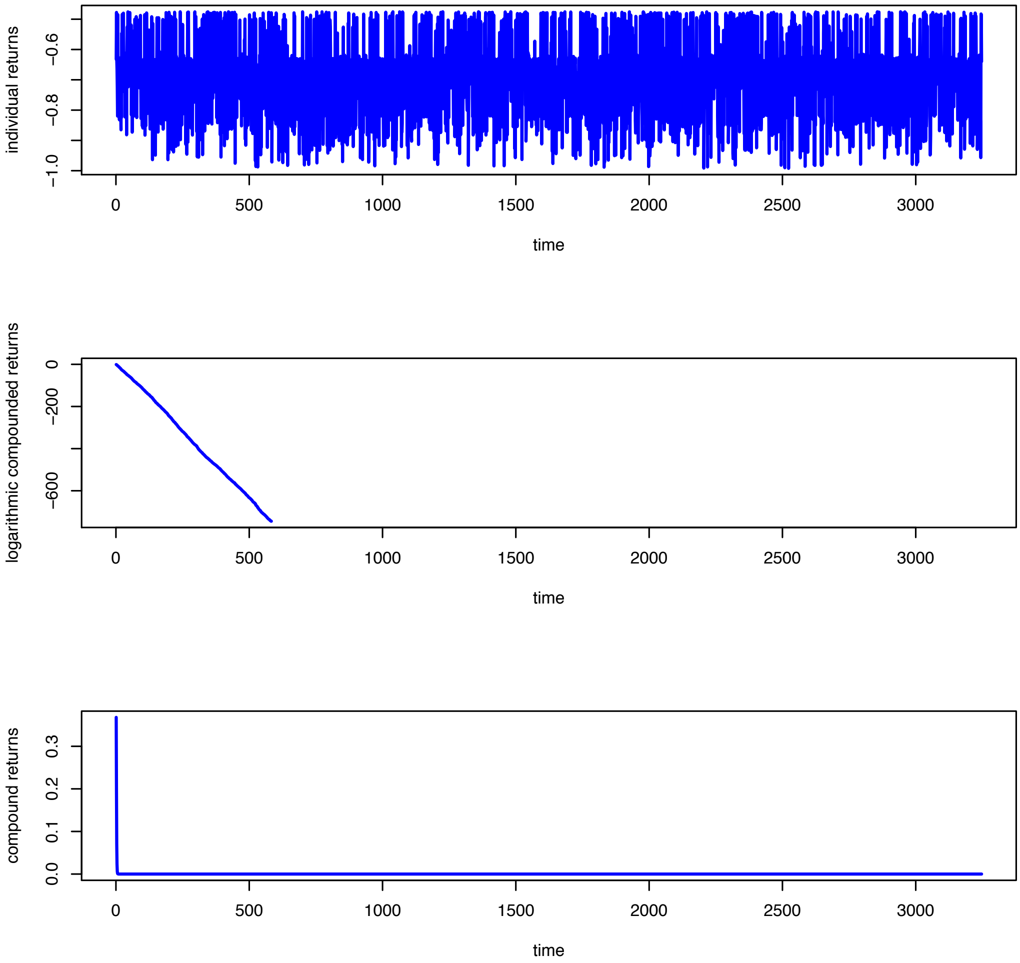


Figure 6.7: Breiman-Ripley wealth process generated by POLR model as bookmaker playing against Skellam model as gambler.

6.1.3 RankingSVM model

Using the RankingSVM algorithm with a MAP classifier based on post-processing the RankingSVM output with a Bayespolr model, we achieve the following confusion matrix

$$\begin{array}{l} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \textit{Column total} \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & \textit{Row total} \end{array} \\ \left(\begin{array}{cccccc} 10.9 & 1.3 & 0 & 0.0 & 0.0 & 12.2 \\ 0.5 & 26.3 & 4.8 & 0.0 & 0.0 & 31.6 \\ 0.0 & 2.3 & 18 & 1.9 & 0.0 & 22.2 \\ 0.0 & 0.0 & 3.1 & 19.2 & 0.0 & 22.3 \\ 0.0 & 0.0 & 0.0 & 11.7 & 0.0 & 11.7 \\ 11.4 & 29.9 & 25.9 & 32.8 & 0.0 & 100.0 \end{array}\right) \end{array}$$

Post-processed RankingSVM MAP classifier error rate = 25.6 %

The error rate is respectable, however not as good as Bayesian POLR which uses interactions between covariates. Note that the classifier is unable to predict any test cases as class 5. The purpose of the post processing predictive fitting is to allow us to create a Breiman-Ripley wealth process and compare the quality of the model across the probability curve. When playing against the Skellam distribution bookmaker the performance is rather good. However, when playing against the Bayesian POLR model 6.8, the RankingSVM model does not perform as well as might be expected, 6.9.

We can also assess for the fitting via a calibration plot, 6.10. This shows that the post processed probabilities of the RankingSVM model are accurate at estimating class probabilities, across the range of predictive probabilities.

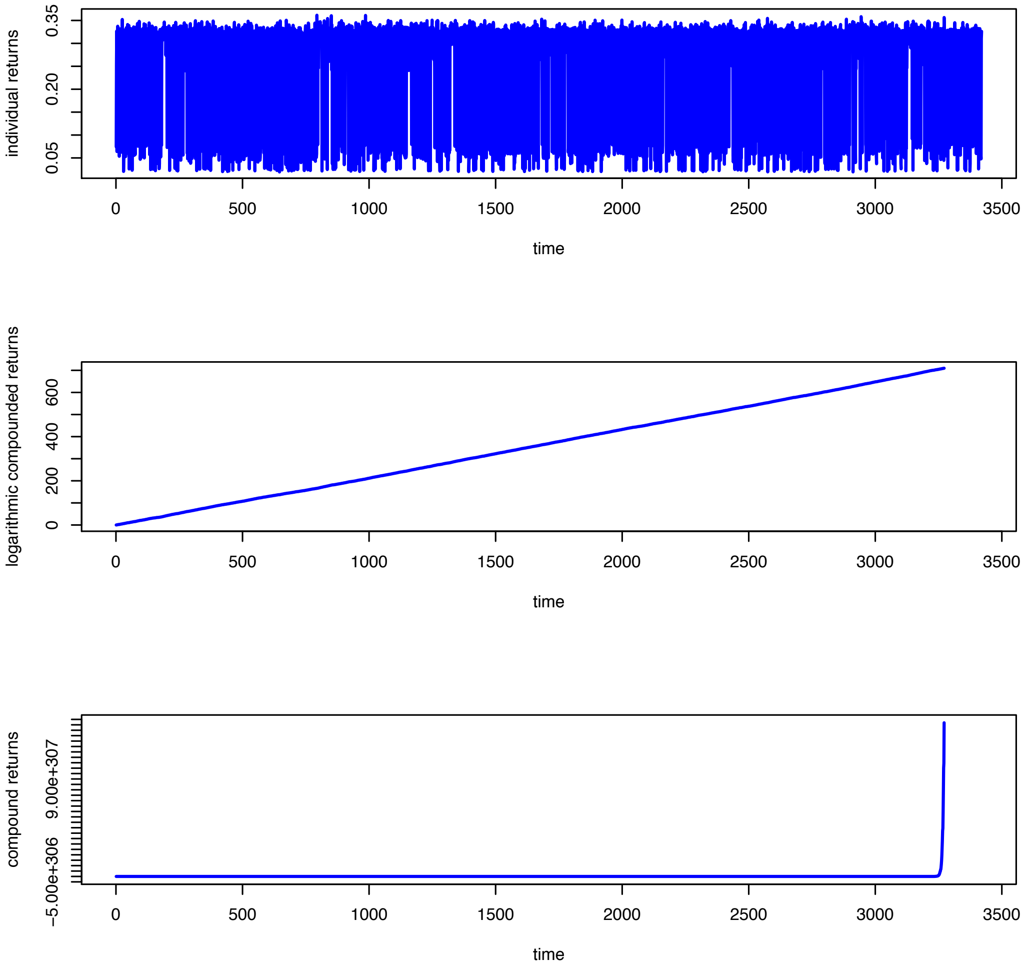


Figure 6.8: Breiman-Ripley wealth process generated by Skellam model as bookmaker playing against RankingSVM model as gambler.

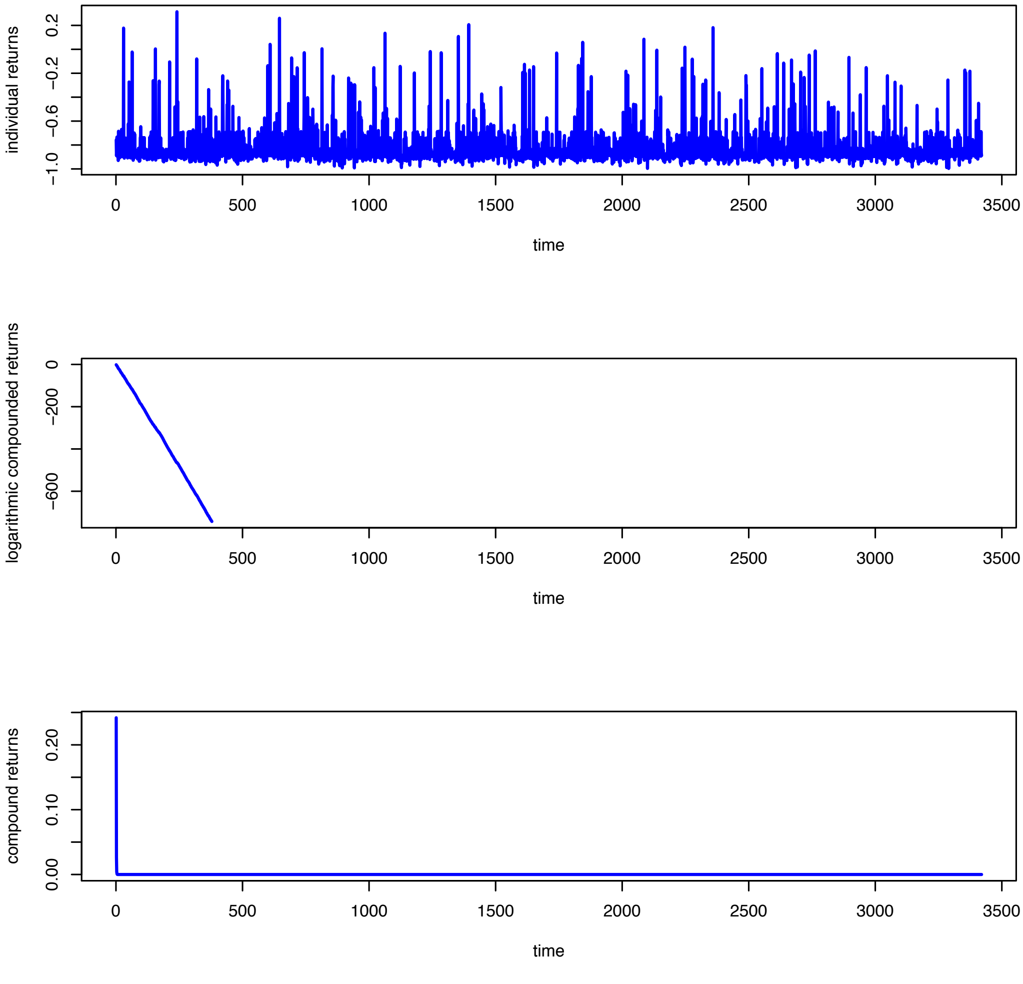


Figure 6.9: Breiman-Ripley wealth process generated by POLR model as bookmaker playing against RankingSVM model as gambler.

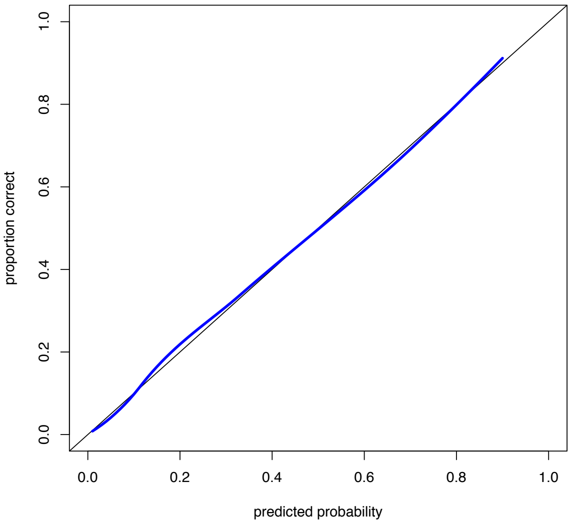


Figure 6.10: Calibration plot of post processed RankingSVM probabilities.

Our overall assessment is that the RankingSVM model produces reasonable predictive probability estimates, but they are inferior to those produced by a Bayesian POLR model with interaction terms for the covariates.

6.2 Soccer data

6.2.1 POLR model

For the POLR model thegoal difference response data were binned (in all competitions and countries) using expert opinion from practictioners in the gambling industry. The classes correspond to the goal difference being one of $' < -3',' -3',' -2',' -1',' 0',' 1',' 2',' 3',' 4',' 5'$ or > 5.

Predictive probabilities were generated out of sample by using all data up to but not including the date of the prediction case. Predictions were made for games across Europe for the 1999-2010 seasons. Using cross validation, a look-back period of two years (from the date of prediction) of historical data was selected for training the model. The output of the model can be observed as a barplot on the score difference space. For example, the Arsenal vs Wigan premiership game on the 22nd of January 2011 had a goal difference distribution as 6.11. (As we write the score is Arsenal $1-0$ Wigan at half-time.) From the score difference distribution we can cacluate the fair value asian odds and predictive probabilities for home win, draw and away win.

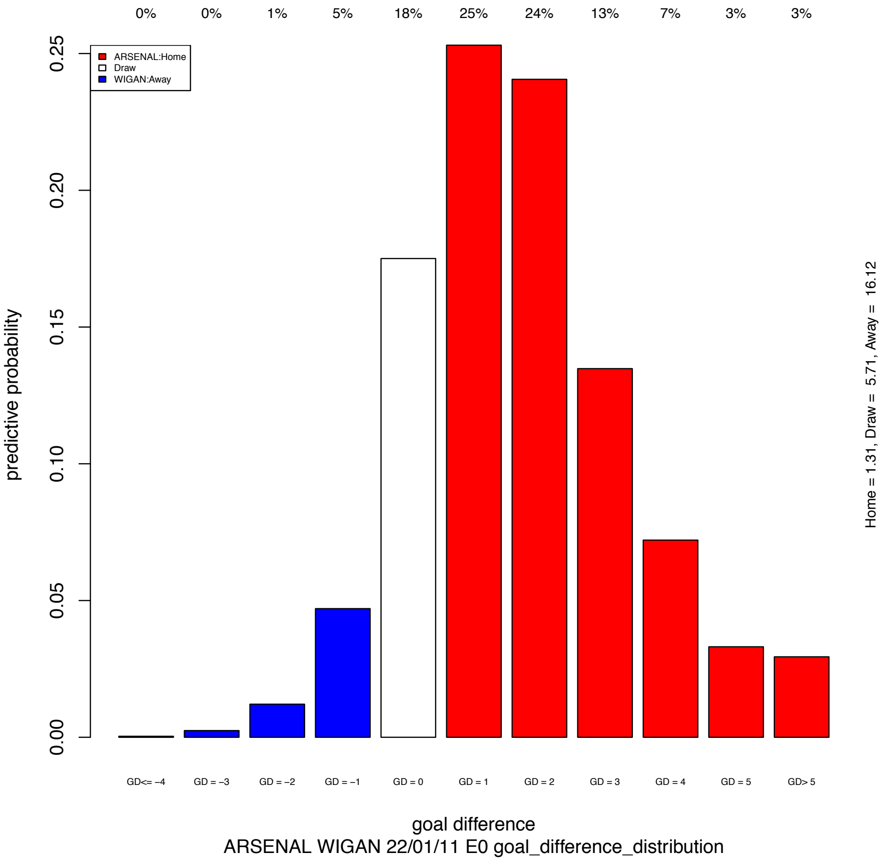


Figure 6.11: Arsenal vs. Wigan @ Emirates on 22 January 2011. The percentages across the top of the chart correspond to class probabilities.

Pricing handicapped bets is straightforward once the goal difference distribution is known.

Assessing model fit for the Bayesian POLR model is done via a calibration plot - Venables and Ripley [2002] and Dawid [1992]. Although the goal difference is being modelled by the algorithm, there is also real world interest in the home win, away win and draw outcomes of the games.

We note that the calibration plot on the goal difference performs very well in the region of $2 - 30$ % class probabilities. At higher class probabilities, there are few cases and it is difficult to assess model fit. However, we are not able to put confidence bounds on the prequential analysis, since the calibration plot has been formed from a dependent sample of classifications, which do not impact the expected proportion, but do affect the second and higher moments of the lowess proportion estimator.

Although 6.12 is promising, we also consider plots of home win, draw and away win using lowess smoothing.

We note that the home win predictive probabilities fits data rather well. The model over-fits at predictive probabilities less than one half and under-fits above one half predictive probabilities, but the direction is aligned with the predictive probabilities.

Draws probabilities fit the data reasonably, for the narrow range that predictive probabilities produced by the model, 6.14. The kink observed is because of the concentration of probability estimates around the $1/3$ and above range for draws.

Away win predictive probabilities are less accurate, although the direction is correct 6.15. The model fits at predictive probabilities less than one 30 per cent

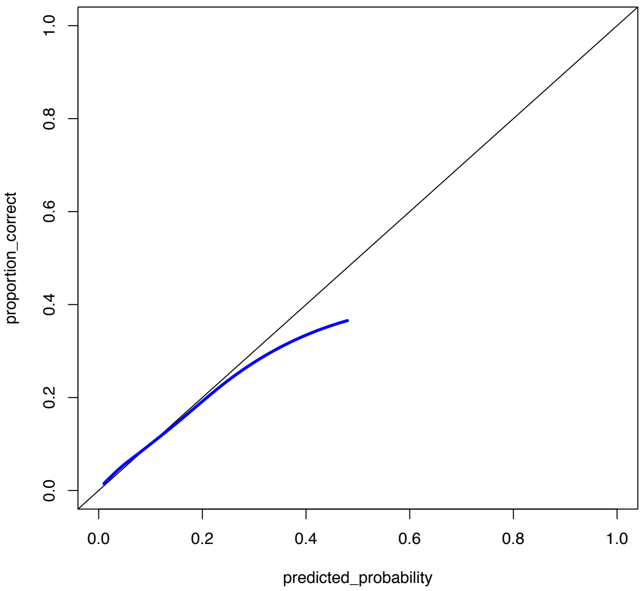


Figure 6.12: Goal difference calibration plot using Bayesian POLR

very well and under-fits above that threshold.

6.2.2 RankingSVM model

We had convergence issues with using the full set of covariates for RankingSVM data. We opted for a smaller set of covariates, namely the home, draw, away and asian odds; the home team, away team and goal difference in the previous 2 years. Since the RankingSVM algorithm produces vector output which are not probabilities, we used them as covariates for a Bayesian POLR on the goal difference. This produced a meta probability model based on pre-processing the covariates using the RankingSVM algorithm.

The associated calibration plots for goal difference, home win, draw and away win for RankingSVM are similar to those of the Bayesian POLR model. This is

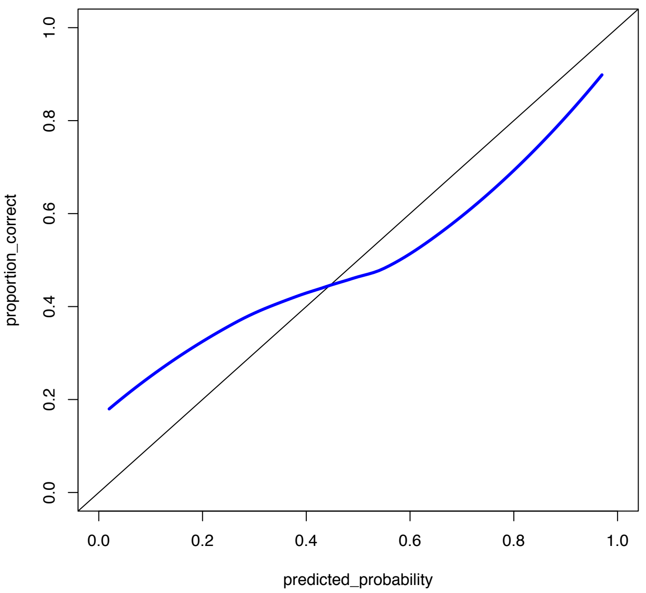


Figure 6.13: Home win probability lowess calibration plot using Bayesian POLR

not surprising as the predictive probabilities of the two models are similar. By visual inspection, we note that across all events the calibration plots seem to produces slightly better estimates than the Bayesian POLR model on its own. It is difficult to know which model is better, given the subtlety of differences. In such a scenario, the Breiman-Ripley wealth process can help to judge between the quality of models.

6.2.3 Discussion

It appears that the RankingSVM algorithm pre-processed for Bayesian POLR performs well on the small example artificial data when compared with the Bayesian POLR; in contrast it is unable to computationally manage a large (many covariate) analysis with the current resources available. However judiciously cho-

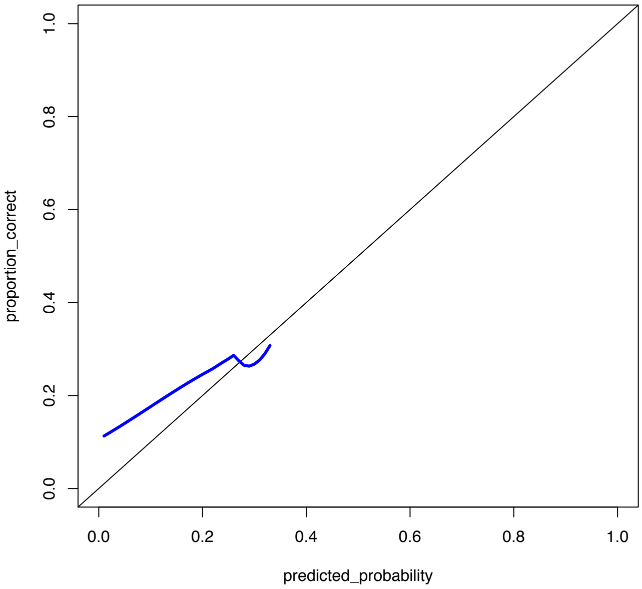


Figure 6.14: Draw probability lowess calibration plot using Bayesian POLR

sen covariates for the soccer data perform well for the RankingSVM algorithm. The surrogate Skellam model fit to the artificial data showed weaker estimation of predictive probabilities than the RankingSVM or Bayesian POLR approach.

The Bayesian POLR model performs less well on the artificial data as compared to the RankingSVM algorithm. We believe that additional heterogeneity from unobserved covariates accounts for the unexplained variability in the prequential analysis above. We also note that although the score (or goal) difference fits rather well, when considering home win, draw and away win events, misspecified probabilities seem to aggregate, weakening predictive performance. Our thoughts are that a remedy may be re-training the output of the Bayesian classifier with a either a binary classifier or POLR model with log class predictive probabilities as covariate(s). However, since the output of the Bayesian classifier sums to one,

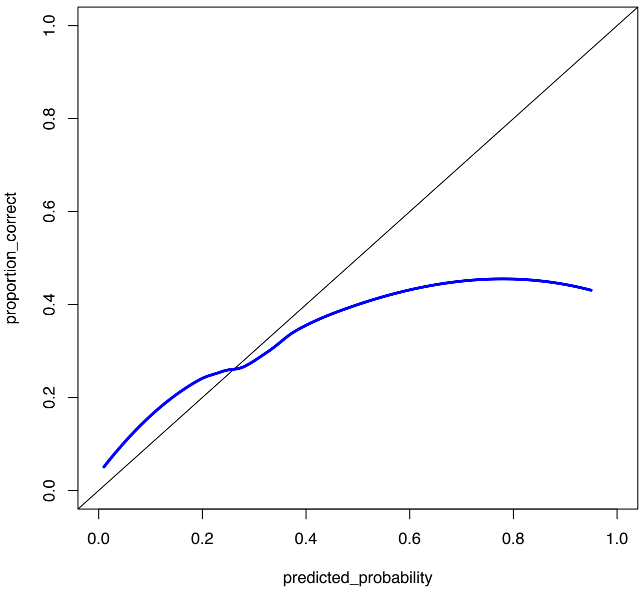


Figure 6.15: Away probability lowess calibration plot using Bayesian POLR

some thoughts would be needed on how to model multi-collinearity in these new covariates.

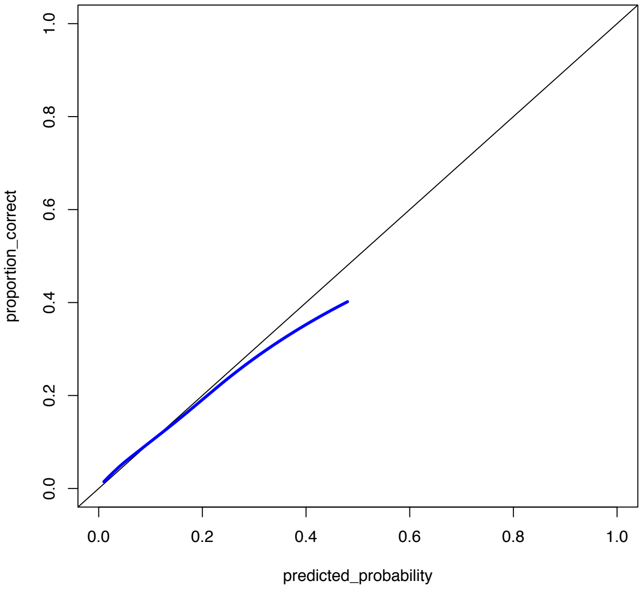


Figure 6.16: Goal difference calibration plot using RankingSVM

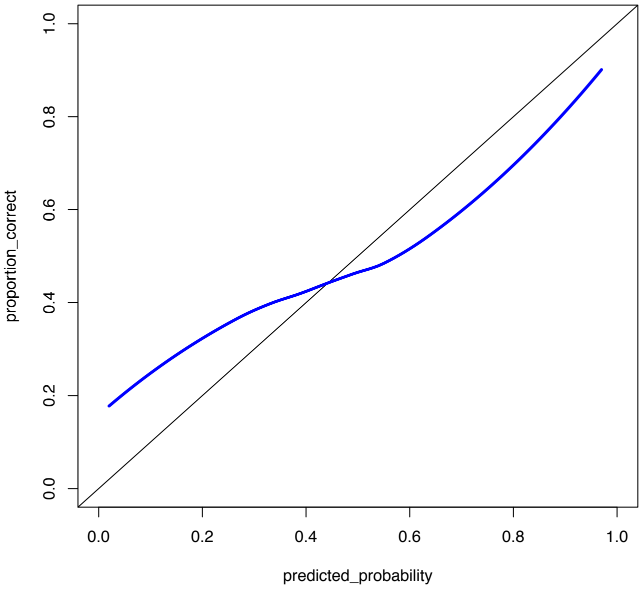


Figure 6.17: Home win probability lowess calibration plot using RankingSVM

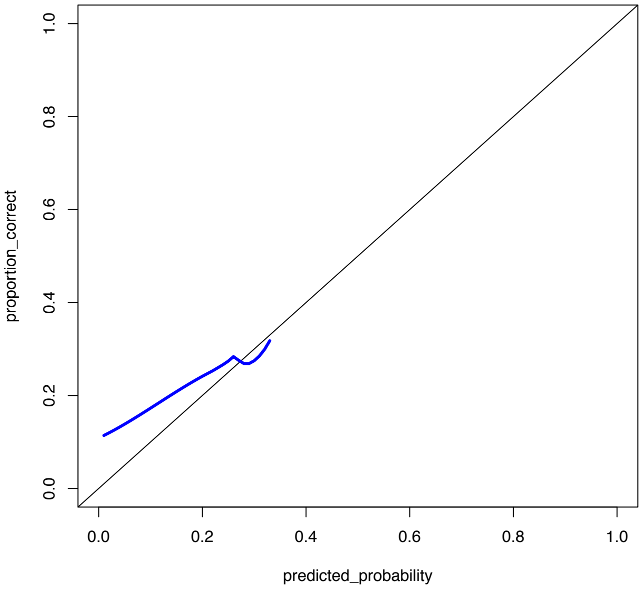


Figure 6.18: Draw probability lowess calibration plot using RankingSVM

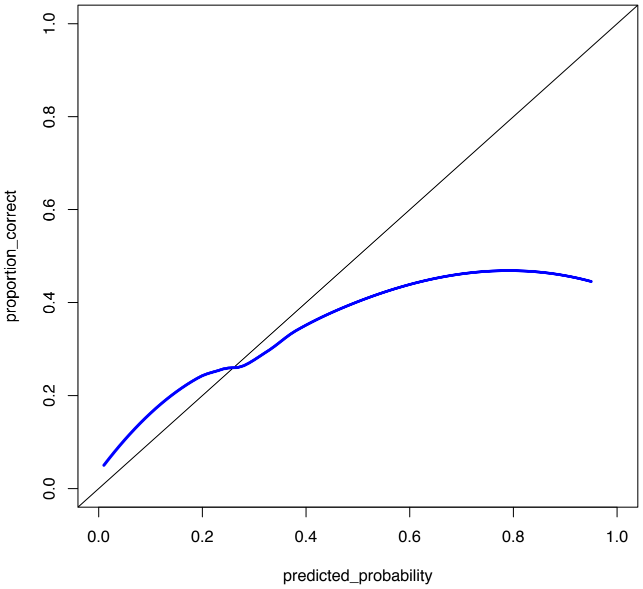


Figure 6.19: Away probability lowess calibration plot using RankingSVM

Chapter 7

Investment performance of sports betting

The purpose of this chapter is to demonstrate the effectiveness of predictive probability estimates to take advantage of the market, and yield superior investment performance. We play against the best odds posted online pre-game from the 2007-2010 football seasons across Europe.

Of course, there is an unrealistic assumption that the best online bookmaker odds would be available for the large size wagers throughout the duration of the test. (Most bookmakers refuse to accept sizeable bets from long run winning customers at advertised rates, forcing the winning patrons to resort to subterfuge to get their bets accepted.) We also considered that the games take place at different times. As a consequence, it is possible to break up individual days into multiple trading sessions. However, for issues to do with poor data availability regarding time-zones, extra-time and game delay considerations, we decided not to do so.

We consider wealth processes, returns and risk metrics in the strategies offered by using Bayesian POLR and RankingSVM pre-processed Bayesian POLR probability engines.(At each match day the predictive probabilities were generated with a lookback period of 2 years and goal difference classes as in the previous chapter.)

As a comparison to traditional investments we provide benchmarks against the Standard and Poors closing index of the 500 biggest large-cap stocks actively traded in the United States, the S&P 500. The S&P 500 is viewed as a standard benchmark of diversified investment performance at any given time, so provides a good comparison to soccer betting across countries. To marry the irregular time series on different days we take the nearest financial trading day to the match day being considered.

The investment strategy is to make one of four exclusive decisions regarding selection of investment: home win, draw, away win or cash for each game. If cash is chosen for a game then no capital is at risk for the outcome of that game, but there is also no reward. We assume that cash earns no interest. As we can only allocate the capital available each morning, we allocate to bets, ordered by the size to be invested in each bet, until all bets or the available capital is exhausted. At that point, no further bets are made. The size of capital to each bet is decided by a logarithmic wagering strategy, scaled down by a factor of ten. Experience showed this to be a reasonable approach. That is, each bet is individually sized by maximising the expected log wealth, then the portfolio is made by scaling to one-tenth the investment for each bet in the portfolio.

However, we note that it is a sub-optimal allocation (Thorp [1997]). There is a diversification benefit because it is extremely unlikely that all bets will fail, thus

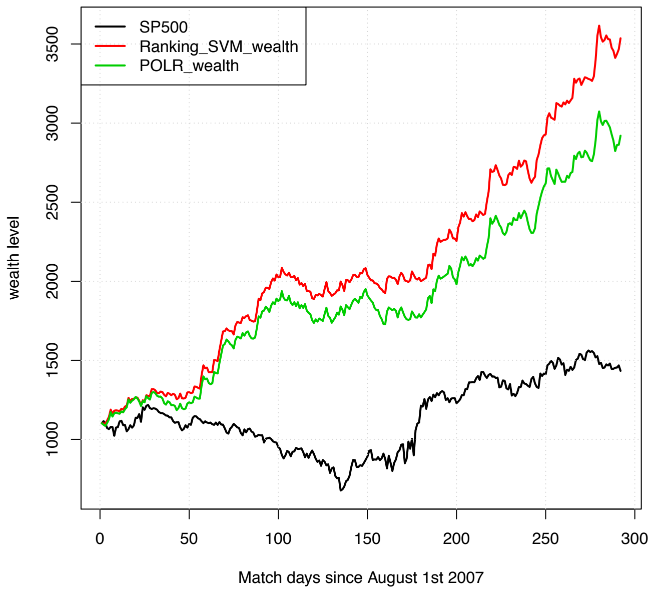


Figure 7.1: Wealth level from each strategy over time.

we could wager more than the scaled down bet sizes if the optimisation could be performed faster. As outlined in our discussion of joint-game Breiman-Ripley processes, when there are more than a few games, the expected wealth has more terms than can be computed in reasonable time for optimisation tasks.

7.1 Wealth levels

The wealth (or price level) and logarithmic wealth plots are shown in 7.1 and 7.2 for each strategy and a comparison to the SP500. We see that a pre-processed Bayesian POLR probability engine outperforms the SP500 and Bayesian POLR probability engine.

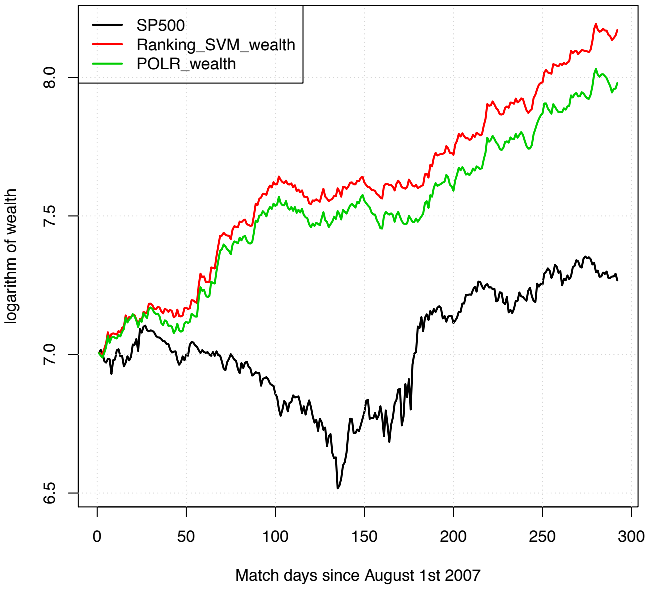


Figure 7.2: Logarithm of wealth level from each strategy over time.

7.2 Returns

Individual returns are plotted using histograms, time series plots and qqplots. We also consider a plot of the square returns, as in most financial time series it is usual that the returns are uncorrelated, but the squares of returns are correlated (Bollerslev [1986]). We can see a benefit of sports betting is the uncorrelated nature of the returns, both within themselves and with other assets. We note however that the RankingSVM model approach bets more heavily on winning days than the Bayesian POLR, in line with our estimates.

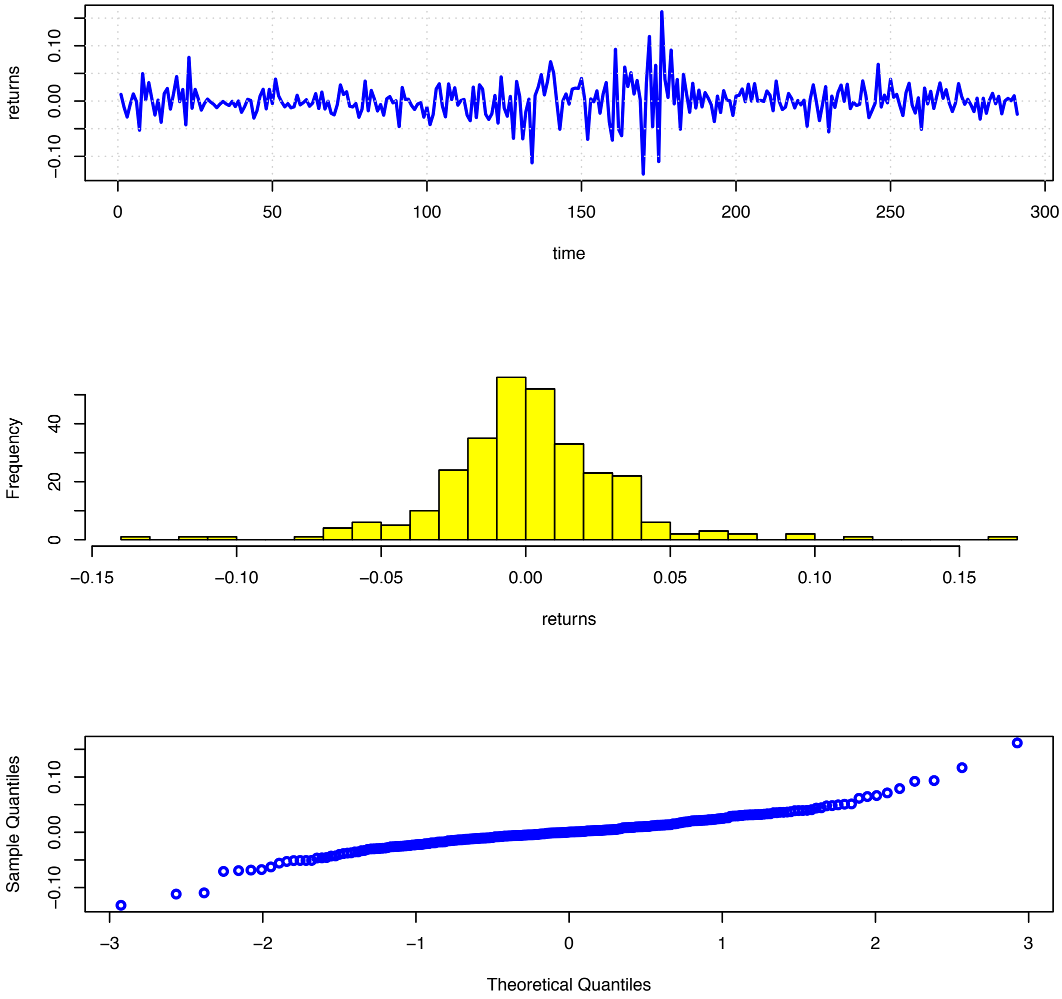


Figure 7.3: Returns of SP500: histogram, qqplot and time series.

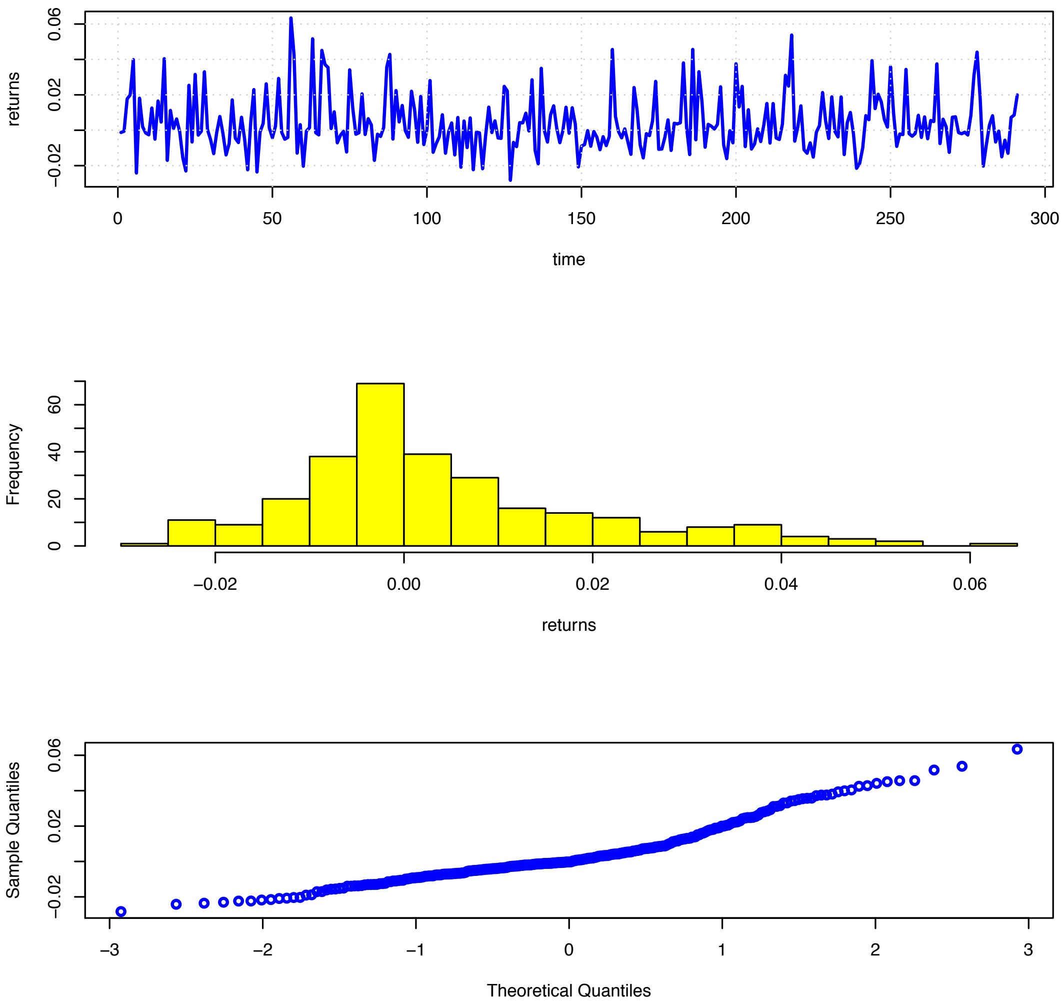


Figure 7.4: Returns of RankingSVM pre-processed Bayesian POLR: histogram, qqplot and time series.

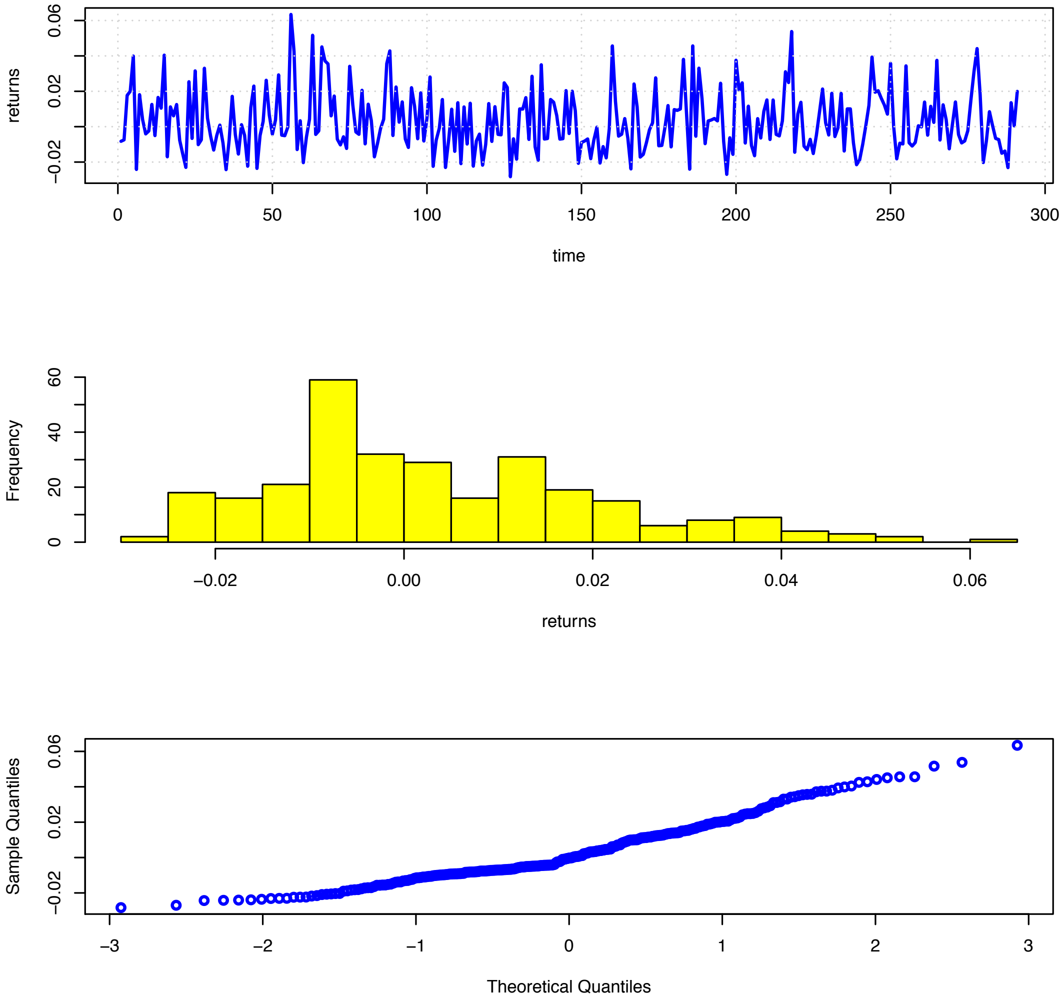


Figure 7.5: Returns of Bayesian POLR probability model: histogram, qqplot and time series.

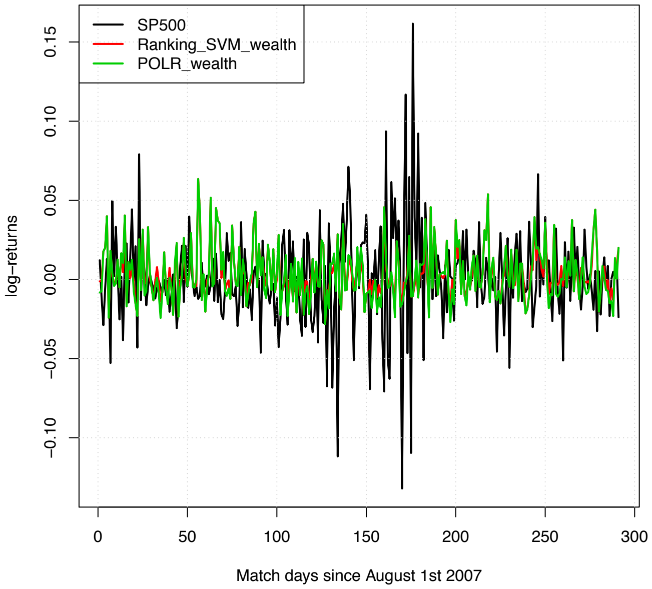


Figure 7.6: Logarithm of strategy returns

7.3 Risk metrics

In order to assess returns we need to measure them against the level of risk being taken. There has extensive research in this field, with a number of rolling ratios (we used a 50 period return) being popular (Pzier [2008]; Sharpe [1968]). The Sharpe ratio for a returns series

$r_1, ..., r_n$, is given by

$$\frac{\bar{r} - r_{free}}{\sqrt{(n-1)^{-1} \sum_{i=1}^{n} (r_i - \bar{r})^2}} = \frac{\bar{r} - r_{free}}{\hat{\sigma}^2} \tag{7.1}$$

(We use $r_t = \log(W_t/W_{t-1})$, but $r_t = W_t - W_{t-1}$ is also used and often close to the logarithm of the ratio of wealth levels. r_{free} is the risk free rate of interest, currently at 2 per cent for US treasury yields.)

It is a measure of the return per unit risk, if risk is measured using standard deviation, also known as the *reward-to-variability* ratio. In our case with zero risk free interest rate, it is the reciprocal of the coefficient of variation.

The Sharpe ratio varies without any obvious systematic component dependent on time, suggesting that the return per unit risk does not vary over time, 7.11.

The Sortino ratio is a a risk adjusted measure of return for a returns series. It is a modification of the Sharpe ratio, in that only returns falling below some threshold contribute to the volatility measure. It is usual to set a return threshold of 0, so that the expression becomes excess return over downside censored standard deviation.

$$\frac{\bar{r} - r_{free}}{\sqrt{(n-1)^{-1} \sum_{i=1}^{n} \min\left(r_i - \bar{r}, 0\right)^2}} \tag{7.2}$$

On a rolling basis, the Sortino ratio, 7.9, demonstrates no systematic variability, indicating that the return per downside risk is constant.

Another approach is to measure the drawdown and maximum drawdown on a rolling basis, important metrics when attracting funds from investors. For a prices, or wealth, series $W_1, ..., W_n$ as above the drawdown $D(T)$ and maximum drawdown $MDD(T)$, at time T are given by

$$D\left(T\right) = \max\left[0, \max_{0 \leq t \leq T} W_t - W_T\right] \tag{7.3}$$

$$MDD\left(T\right) = \max_{0 \leq \tau \leq T}\left[\max_{0 \leq t \leq \tau} W_t - W_\tau\right] \tag{7.4}$$

The drawdown at some time of interest, is measure of the decline from the historical peak of the wealth level at the time of interest. If the wealth level is at an all-time high, then the drawdown at the current time is 0. The maximum drawdown is the maximum peak to trough decline of the wealth level, since measurements started.

The rolling measures for D(T) and MDD(T), 7.7 and 7.8, indicate that the drawdowns experienced scale with the size of the accumulated wealth level, which is not too disturbing.

The CALMAR (California Managed Accounts Report) ratio is a moving average return measure that divides the compounded returns by the maximum drawdown over a period. It is seen as a way to compare performance amongst highly volatile assets. For our strategies, the CALMAR ratio, 7.10, shows the type of volatility that might be expected over a 50 period basis.

7.4 Value@Risk

Since we take the events to be independent realizations of random variables whose predictive distribution is know it is possible to construct a profit and loss distribution for our bet portfolio on any given day. For a portfolio of wagers, the profit at the end of the day, a sum of scaled Bernoullis, is

$$P = \sum_{i=1}^{n} (O_i Y_i - 1) f_i \tag{7.5}$$

where

- n is the number of bets in the portfolio
- O_i is the Asian odds offered on bet i
- f_i is the fraction of wealth invested in bet i
- Y_i is 1 if the bet wins and 0 otherwise.

There are 2^n possible outcomes for the portfolio of bets. When n is greater than 15, there are too many combinations, so that useful calculations become impractical. In such cases, simulations help to approximate the distribution of the profit random variable.

We provide an example from European weekend fixtures, 7.1, and construct the profit distribution, 7.12, arrived at by simulating scenarios, from singular Kelly betting on backing the home outcome only over the 25 bets in the table.

Our fractional investment in each bet affects the type of distribution that is constructed for the portfolio. In general, an expected logarithmic wealth optimal wagering strategy has the effect of maximizing the median of the distribution. For computational efficiency we rely on approximating the optimization problem, but this should have small effect on the quantiles of the distribution. To reach the distribution, for a small number of cases we can iterate through all outcome scenarios, generate the impact on the aggregate profit and loss, and calculate the likelihood of that scenario. Once this has been done, we can order the profit and loss amounts and suitably bin them to construct a distribution. We had to simulate and smooth to estimate the distribution as with 25 bets we cannot analytically calculate the exact distribution due to the computational cost of generating and iterating through all scenarios.

Note how the distribution is right skewed, indicating that gambler will profit more often than not; the median of the distribution is positive. This allocation is not the optimal one though. By running a genetic algorithm to simulate derivatives, Whitrow [2007], we could arrive at a better distribution, which would be closer to the optimal one. (Although we know how to reach optimal distribution, computationally achieving it takes more time and resources than a gambler would consider worthwhile.)

Date	Div	Home Team	Away Team	P(Home)	Odds
05/03/2011	SC2	AIRDRIE UTD	AYR	0.517	2.75
06/03/2011	N1	AJAX	AZ ALKMAAR	0.630	1.62
04/03/2011	B1	ANDERLECHT	GENK	0.670	1.80
05/03/2011	E0	ARSENAL	SUNDERLAND	0.760	1.44
05/03/2011	SP1	ATH MADRID	VILLARREAL	0.587	2.10
07/03/2011	D2	AUGSBURG	F. DUSSELDORF	0.660	1.73
05/03/2011	SP1	BARCELONA	ZARAGOZA	0.922	1.10
05/03/2011	SC3	BERWICK	ELGIN	0.618	2.25
05/03/2011	E0	BIRMINGHAM	WEST BROM	0.641	2.38
05/03/2011	E2	BOURNEMOUTH	OLDHAM	0.575	1.83
05/03/2011	E2	BRISTOL RVS	DAG & RED	0.485	2.40
06/03/2011	T1	BURSASPOR	BUYUKSEHYR	0.645	1.62
05/03/2011	E3	CREWE	BURTON	0.481	2.10
05/03/2011	E1	DERBY	BARNSLEY	0.509	2.00
05/03/2011	SC2	DUMBARTON	LIVINGSTON	0.302	4.00
05/03/2011	D1	E. FRANKFURT	K'LAUTERN	0.488	2.30
05/03/2011	I2	EMPOLI	MODENA	0.667	1.85
06/03/2011	D2	ERZGEBIRGE AUE	U. BERLIN	0.688	2.10
06/03/2011	T1	E'SPOR	BUCASPOR	0.737	1.80
05/03/2011	N1	EXCELSIOR	PSV EINDHOVEN	0.095	11.00
05/03/2011	E0	FULHAM	BLACKBURN	0.647	1.73
05/03/2011	T1	GALATASARAY	KARABUKSPOR	0.734	1.57
05/03/2011	T1	G'TEPSPOR	SIVASSPOR	0.559	1.80
06/03/2011	N1	GRONINGEN	HERACLES	0.626	1.60
06/03/2011	D1	HAMBURG	MAINZ	0.567	1.91

Table 7.1: Backing home team to win: betting data for a particular weekend of fixtures in Europe.

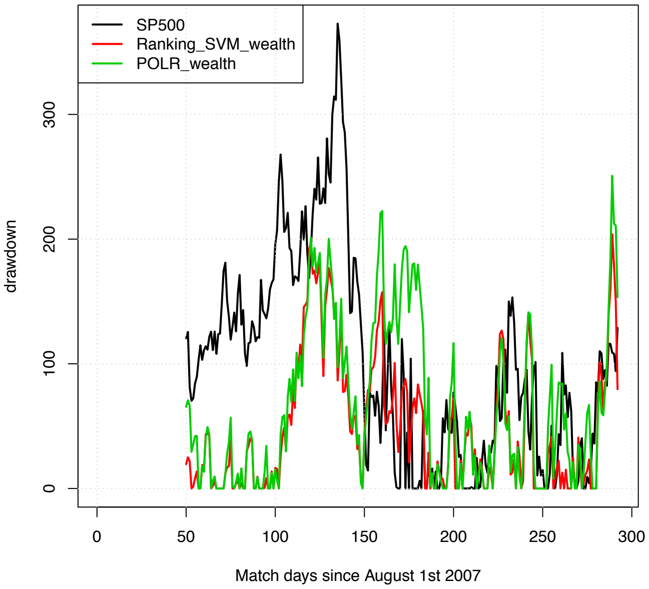


Figure 7.7: Rolling drawdown of strategies

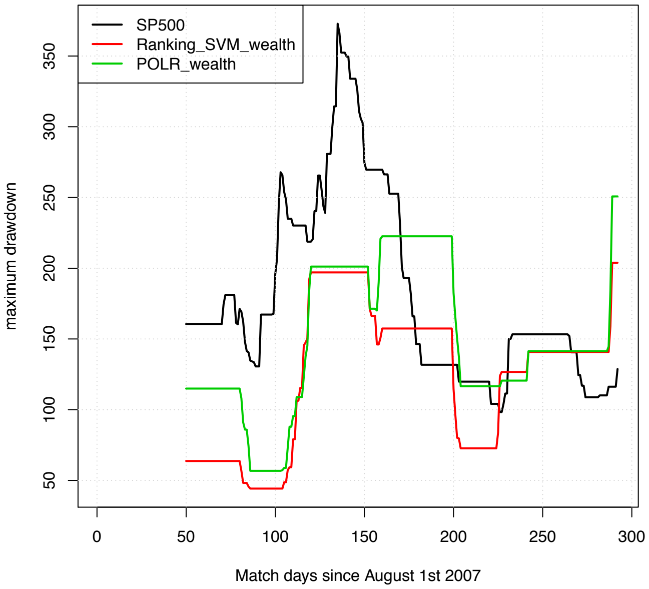


Figure 7.8: Rolling maximum drawdown of strategies

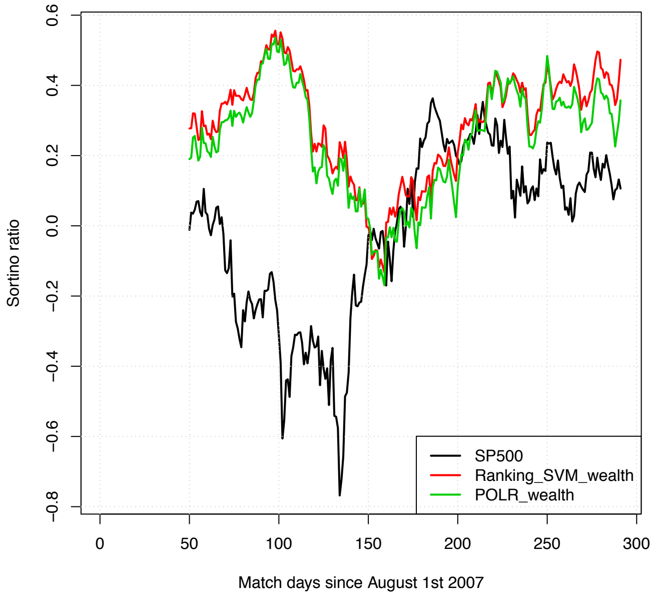


Figure 7.9: Rolling Sortino ratio of strategies.

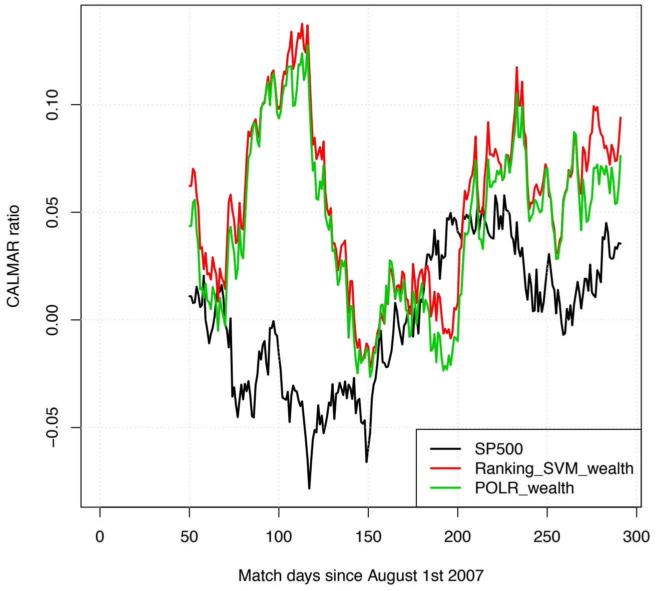


Figure 7.10: Rolling CALMAR ratio of strategies.

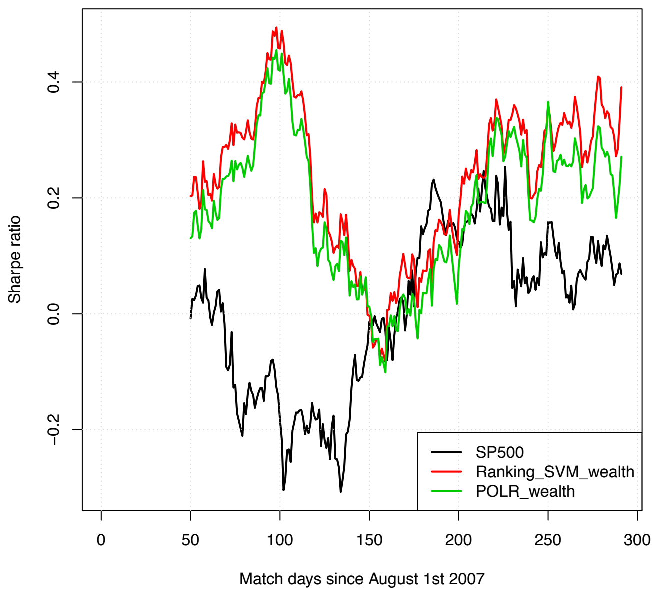


Figure 7.11: Rolling Sharpe ratio of strategies.

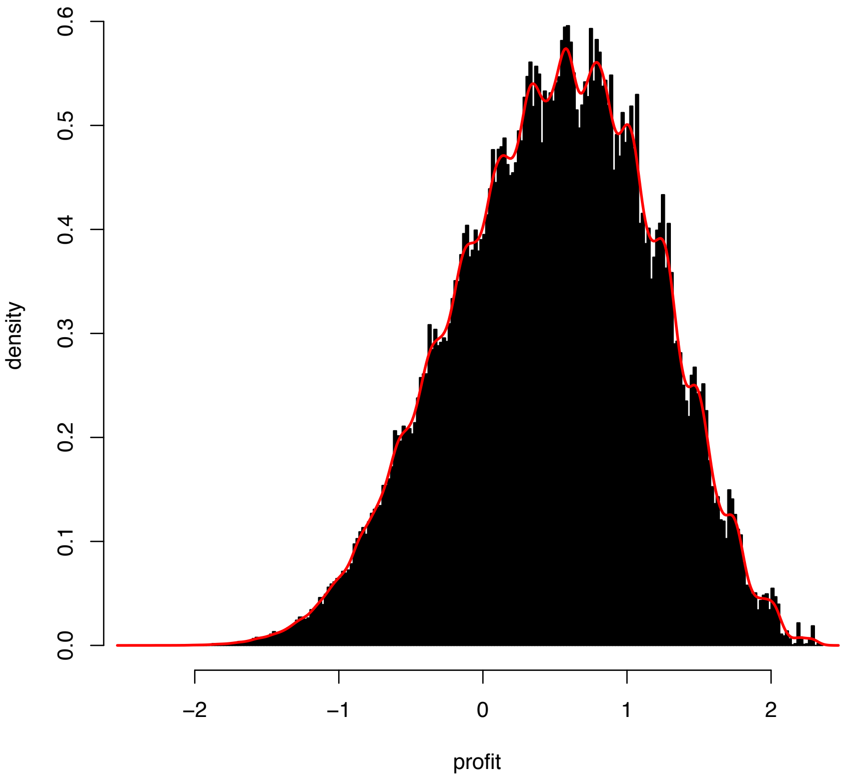


Figure 7.12: Simulated profit distribution from Kelly wagering on Bayesian POLR model predictions for fixtures in the table using Friedman-Diaconis breaks to bin the data. The red line is a smoothed density estimate given that exact distribution is computationally infeasible.

Chapter 8

Conclusions

The outcome of this research has been two-fold. We have developed a theoretical underpinning for working with ordinal data using kernel methods and found a predictive model for score differences. The computational resources needed to apply the RankingSVM algorithm to a large soccer dataset has been too great, but results with a smaller subsample are promising. We were able to introduce a new approach to modelling the outcome of soccer games by using a Bayesian POLR model. This model has demonstrated good predictive performance, by estimating predictive probabilities close to market expectations, which we know to be accurate.

The real world test of this approach has been to back-test investment performance against the market. Here we have found good investment performance which outperforms the SP500 according to many standard financial metrics. Although encouraging, it would be difficult to generate this performance as finding counterparties to take on such volume would be very difficult, unless non-legitimate sources were used.

In our live trading we noted that at times the model produces odds that are not plausible when compared to the market. When the *distance* between market and implied odds is implausible, we would like to discard such predictions as untrustworthy. This is based on our knowledge that the market seldom sets the odds drastically wrong from what they should be, cf the calibration plots in chapter 6. This seems related to the Kullback-Leibler divergence (Kullback and Leibler [1951]) but we were unable to formalize an automated algorithm for this.

8.1 Outlook for the future

There are risks to the application of supervised learning models for generating profitable strategies in the long term. These are based on the changing nature of the betting market and soccer competition.

For now, the major gambling syndicates rely on being shrouded from their counterparties via agents. These agents are a legacy business, which has come about because of the need of Asian bookmakers to out-source their credit risk and customer service to local representatives. The bookmakers have been happy to operate a franchise business which pays a commission on volume to the agents so long as the agents have a pool of bettors that consistently loses. As internet technologies reach Asia, it is only a matter of time before agents are bypassed by bookmakers to directly transact with their customers. When this happens, like British bookmakers, Asian bookmakers will no longer accept a portfolio of bets from individual customers, particularly those that keep winning. Already some bookmakers in the West Indies have become alert to professional gamblers, but use them to help determine efficient prices by allowing small volumes through, but stopping larger bets being placed.

In addition, the development of online betting exchanges will lead to prices moving to efficiency, as specialist in assessing these risks can access the market directly, providing capital to purchase and sell risks that are mispriced. In the long run, only superior information and proprietary models can sustain profitable strategies. With the onset of Web 2.0 technologies and the relative cheapness of high performance computing, controlling this technology is not practical. It seems unlikely that super-linear returns will be sustainable, as the market will become efficient and only a small number of players will be able to sustain economic benefit from professional gambling.

There is a risk to the effectiveness of the predictive probability estimates due to the game evolving. Barclay [2006] noted that Jose Mourinho changes his team tactics depending on who the competition is and the form of his own team. When a team alters its behavior, with the objective of altering the score difference distribution, it becomes problematic for forecasters. Although this happens to some extent already, we are able to capture it algorithmically. Should there be more managers that use specific strategies to counteract covariate and team specific effects, then it would be difficult extract economic benefit from the market.

One final trend is to do with the pool of losers that professionals drink from. The government of the People's Republic of China has severely controlled gambling. They are making strong efforts to limit gambling activity and have set up a specialist unit to capture illegal gambling activities. There have been reports of IP addresses being blocked in greater China by authorities keen to stop online gambling. This poses a threat to the sustainability of professional gambling activities worldwide as a large proportion of losing bettors are in China.

All of the above points being true, we still believe that substantial returns will be plausible for at least a decade from now.

8.2 Developing models further

Advanced models for sports betting need to take account of many covariates, some interacting and some not relevant. There is a need for an automatic approach which can lead to a best model. Bayesian model averaging might be one way to do this, (Draper [1995]; Hoeting *et al.* [1999]). The details of how this might be achieved for ordinal models is worthy of further study.

An area where we have seen limited literature is in-play modelling. We believe that this is where the highest returns can be achieved and the highest barriers to entry for competitors will exist. One way to do this is to have a time parameter in a modified POLR model. There should be a stochastic process to model in-play performance responding to covariates and state of the score difference with time decay pushing the distribution towards its current state. It might be achieved via a vector Markov process linked to the POLR model. Optimizing such a model and finding fair tests for its performance is an area where we are still developing our thoughts.

Once applied to soccer data, these models can easily be extended to other team sports, such as North American sports and horse-racing. This is because ordinal models can be applied in the same way once the underlying variable has been suitably binned into classification categories.

The level of sophistication in the market gleaned from players indicates that there are substantial returns to be generated by applying modern machine learning methods to price sports wagering risks on betting exchanges, or that bookmakers might consider purchasing risk cheaply rather than setting odds to balance the flow of money on their books.

The citation page numbers at the end of each reference are given in blue.

References

Adams, N. M. and Hand, D. J. (2000) Improving the practice of classifier performance assessment. *Neural Computation*, **12**(2), 305–311. 63

Agresti, A. (2010) *Analysis of Ordinal Categorical Data*. Wiley, New York, Second edition. 18, 20

Albert, J. H. and Chib, S. (2001) Sequential ordinal modeling with applications to survival data. *Biometrics*, **57**(3), pp. 829–836. 54

Anderson, J. A. (1984) Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **46**(1), pp. 1–30. 20

Barclay, P. (2006) *Mourinho: Anatomy Of A Winner*. Orion. 133

Bawa, V. S. (1975) Optimal rules for ordering uncertain prospects. *Journal of Financial Economics*, **2**(1), 95–121. 61

Benter, W. (1994) Computer based horse race handicapping and wagering systems:A Report. In *Efficiency of Racetrack Betting Markets* (Eds V. S. Y. L. William T. Ziemba and D. J. Hausch), pp. 183–193. Elsevier North Holland. 46

Bishop, C. M. (1995) *Neural Networks for Pattern Recognition.* Oxford: Oxford University Press. 24

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**(3), 307–327. 117

Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992) A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory.* 30, 38

Breiman, L. (1961) Optimal gambling systems for favorable games. *Fourth Berkeley Symposium on Probability and Statistics*, **1**, 65–78. 63, 65, 67, 83

Brier, G. W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**(1), 1–3. 63

Brillinger, D. (2007) Modelling some Norwegian soccer data. In *Advances in statistical modelling and inference: essays in honor of Kjell A. Doksum* (Ed. V. Nair), pp. 3–20. World Scientific, Singapore. 4, 53

Brillinger, D. (2010) Soccer/world football. University of California, Berkeley. http://www.stat.berkeley.edu/tech-reports/777.pdf. 46

Burges, C. J. C. (1998) A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, **2**(2), 121–167–167. 26

Chu, W. and Keerthi, S. S. (2007) Support vector ordinal regression. *Neural Computation*, **19**, 2007. 36

Cortes, C. and Vapnik, V. (1995) Support-vector networks. In *Machine Learning*, pp. 273–297. 32

da Costa, J. F. P. and Cardoso, J. S. (2005) *Machine Learning: ECML 2005, 16th European Conference on Machine Learning*, chapter Classification of ordinal data using neural networks, p. 690 697. Springer. 24

Cowles, M. (1996) Accelerating Monte Carlo Markov Chain Convergence for Cumulative link Generalized Linear Models. *Statistics and Computing*, **6**, 101 –111. 22, 54, 94

Cox, D. R. (1972) The analysis of multivariate binary data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **21**(2), pp. 113–120. 16

Cox, D. R. and Snell, E. J. (1981) *Applied Statistics; Principles and Examples.* London: Chapman and Hall. 84

Crammer, K. and Singer, Y. (2005) Online ranking by projecting. *Neural Computing*, **17**, 145–175. 35

Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, first edition. 26

Crowder, M., Dixon, M., Ledford, A. and Robinson, M. (2002) Dynamic modelling and prediction of English football league matches for betting. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **51**(2), pp. 157–168. 49

Dawid, A. P. (1992) Prequential data analysis. *Lecture Notes- IMS Monograph Series*, **17**, pp. 113–126. 61, 107

de Finetti, B. (1974) *Theory of Probability.* London: Wiley. Italian original 1970. 4

Dixon, M. J. and Coles, S. G. (1997) Modelling Association Football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **46**(2), pp. 265–280. 3, 48

Dixon, M. J. and Robinson, M. E. (1998) A birth process model for association football matches. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **47**(3), pp. 523–538. 55

Doob, J. L. (1990) *Stochastic Processes (Wiley Classics Library).* Wiley-Interscience. 66

Draper, D. (1995) Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), pp. 45–97. 134

Duan, K., Keerthi, S. S. and Poo, A. N. (2003) Evaluation of simple performance measures for tuning svm hyperparameters. *Neurocomputing*, **51**, 41–59. 31

Duan, K.-B. and Keerthi, S. S. (2005) Which is the best multiclass SVM method? an empirical study. In *Multiple Classifier Systems* (Eds N. C. Oza, R. Polikar, J. Kittler and F. Roli), volume 3541 of *Lecture Notes in Computer Science*, pp. 278–285. Springer Berlin / Heidelberg. 26, 33, 34

Duda, R., Hart, P., and Stork, D. (2007) Pattern classification. *Journal of Classification*, **24**(2), 305–307. 26

Geisser, S. (1993) *Predictive Inference: An Introduction.* Number 3. Chapman and Hall. 63

Goddard, J. (2005) Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, **21**(2), 331 – 340. 4, 53

Good, I. J. (1983) *Good Thinking - The Foundations of Probability and its Applications*. Minneapolis: University of Minnesota Press. 63

Grizzle, J. E., Starmer, C. F. and Koch, G. G. (1969) Analysis of categorical data by linear models. *Biometrics*, **25**(3), pp. 489–504. 21

Hastie, T. and Tibshirani, R. (1986) Generalized additive models. *Statistical Science*, **1**(3), pp. 297–310. 23

Hastie, T. and Tibshirani, R. (1987) Generalized additive models: Some applications. *Journal of the American Statistical Association*, **82**(398), pp. 371–386. 23

Hastie, T. and Tibshirani, R. (1998) Classification by pairwise coupling. *The Annals of Statistics*, **26**(2), pp. 451–471. 34

Herbrich, R., Graepel, T. and Obermayer, K. (2000) *Large Margin Rank Boundaries for Ordinal Regression*. MIT Press. 9, 35, 97

Hill, D. (2006) *The Fix: Soccer and Organized Crime*. McCelland and Stewart. 1

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: A tutorial. *Statistical Science*, **14**(4), pp. 382–401. 134

Hosmer, D. J. and Lemeshow, S. (1989) *Applied Logistic Regression*. New York: John Wiley and Sons. 75

Hsu, C.-W. and Lin, C.-J. (2002) A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, **13**(2), 415–425. 33

Johnson, V. E. and Albert, J. H. (2001) *Ordinal Data Modeling (Statistics for Social and Behavioral Sciences)*. Springer. 22

Karatzoglou, A., Meyer, D. and Hornik, K. (2006) Support vector machines in R. *Journal of Statistical Software, Volume 15*, pp. 1–28. 32

Karlis, D. and Ntzoufras, I. (2003) Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society*, **Series D (The Statistician) 52**(3), pp. 381–393. 3, 49, 52

Karlis, D. and Ntzoufras, I. (2009) Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, **20**(2), 133–145. 51

Keerthi, S., Shirish Shevade, C. B. and Murthy, K. K. (2001) Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*. 27, 33

Kelly, J. L. (1956) A new interpretation of information rate. *Bell System Technical Journal*. 63, 65

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics*, **22**(1), pp. 79–86. 132

Likert, R. (1932) A technique for the measurement of attitudes. *Archives of Psychology*. 16

Maher, M. J. (1982) Modelling Association Football scores. *Statistica Neerlandica*, **36**. 47

Mathieson, M. J. (1996) *Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets*, pp. 523–536. World Scientific. 24

Mathieson, M. J. (1998) Ordinal models and predictive methods in pattern recognition. D.Phil Thesis, The University of Oxford. 16, 24

McCullagh, P. (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **42**(2), pp. 109–142. 16, 22

McHale, I. and Davies, S. (2007) *Statistical thinking in sport*, chapter Statistical analysis of the FIFA world rankings, pp. 77–90. Chapman and Hall. 53

McHale, I. and Scarf, P. (2007) Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica*, **61**(4), 432–445. 54

Moody, J. and Utans, J. (1995) *Neural Networks in the Capital Markets*, chapter Architecture selection strategies for neural networks: application to corporate bond rating prediction. Chichester : Wiley. 16

Olsson, D. M. and Nelson, L. S. (1975) The Nelder-Mead simplex procedure for function minimization. *Technometrics*, **17**(1), pp. 45–51. 90

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999) The PageRank Citation Ranking: Bringing Order to the Web. http://www.cis.upenn.edu/∼mkearns/teaching/NetworkedLife/pagerank.pdf. 4

Perez Cruz, F., Navia-Vazquez, A., Alarcon-Diana, P. and Artes-Rodriguez, A. (2000) An IRWLS procedure for SVR. *European Signal Processing Conference (EUSIPCO), Tampere (Finland)*. 38

Platt, J. C. (2000) *Advances in Large Margin Classifiers*, chapter Probabilistic outputs for support vector machines and comparison to regularized likelihood methods, p. 6174. MIT Press. 4, 33, 38, 44

Poundstone, W. (2005) *Fortune's Formula: The Untold Story of the Scientific Betting System That Beat the Casinos and Wall Street*. Hill and Wang. 46

Pzier, J. (2008) Maximum certain equivalent excess returns and equivalent preference criteria. ICMA centre discussion papers in finance, Henley Business School, Reading University. 121

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 7

Ripley, B. D. (1996) *Neural Networks and Pattern Recognition*. Cambridge University Press. 24, 63

Rue, H. and Salvesen, O. (2000) Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **49**(3), pp. 399–418. 49

Sanchez-Fernandez, M., de-Prado-Cumplido, M., Arenas-Garcia, J. and Perez-Cruz, F. (2004) SVM Multiregression for Nonlinear Channel Estimation in Multiple-Input Multiple-Output Systems. *IEEE Transactions on Signal Processing*, **52**, 2298–2307. 39, 40

Shannon, C. E. (1948) A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, **5**(1), 3–55. 63

Sharpe, W. F. (1968) [mutual fund performance and the theory of capital asset pricing]: Reply. *Journal of Business*, **41**. 121

Shashua, A. and Levin, A. (2002) *Advances in Neural Information Processing Systems*, chapter Taxonomy of Large Margin Principle Algorithms for Ordinal Regression Problems, pp. 937–944. MIT Press. 35

Skellam, J. G. (1946) The frequency distribution of the difference between two poisson variates belonging to different populations. *Journal of the Royal Statistical Society*, **109**(3), p. 296. 49

Thompson, R. and Baker, R. J. (1981) Composite link functions in generalized linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **30**(2), pp. 125–131. 22

Thorp, E. O. (1997) The Kelly criterion in Blackjack, Sports betting, and the stock market. The 10th International Conference on Gambling and Risk Taking. Montreal. http://www.bjmath.com/bjmath/thorp/paper.htm. 91, 115

Tomas, A. (2008) A dynamic logistic model for combining classifier outputs. D.Phil. Thesis, The University of Oxford. 89

Vapnik, V. N. (1982) *Estimation of Dependences based on Empirical Data*. New York: Springer – Verlag. 33

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S. Fourth Edition*. New York: Springer. ISBN 0-387-95457-0. 75, 78, 84, 107

Walker, S. H. and Duncan, D. B. (1967) Estimation of the probability of an event as a function of several independent variables. *Biometrika*, **54**, 167–179. 22

Whitrow, C. (2007) Algorithms for optimal allocation of bets on many simultaneous events. *Journal Of The Royal Statistical Society Series C*, **56**(5), 607–623. 91, 125

Wilkinson, G. N. and Rogers, C. E. (1973) Symbolic description of factorial models for analysis of variance. *Applied Statistics*, **22**, 392–399. 79

Williams, O. D. and Grizzle, J. E. (1972) Analysis of contingency tables having ordered response categories. *Journal of the American Statistical Association*, **67 (337)**, 55–63. 21

Wood, S. N. (2006) *Generalized Additive Models: An Introduction with R.* Boca Raton, FL: Chapman & Hall/CRC. ISBN 1-584-88474-6. 23

Ziemba, W. T. and Hausch, D. J. (2008) *Handbook of Sports and Lottery Markets.* Elsevier North Holland. 46