Assessing classification performance by gambling strategies

Ravi Kalia
Department of Statistics
University of Oxford
1 South Parks Road
Oxford, OX1 3TG , England
ravkalia@gmail.com

November 4, 2011

Abstract

This paper 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.

Keywords: Breiman-Ripley stochastic process, Kelly betting, classification performance, MAP, model selection, variable selection, model fitting, ordinal, multinomial, binary, martingale, entropy and gambling.

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, 1.

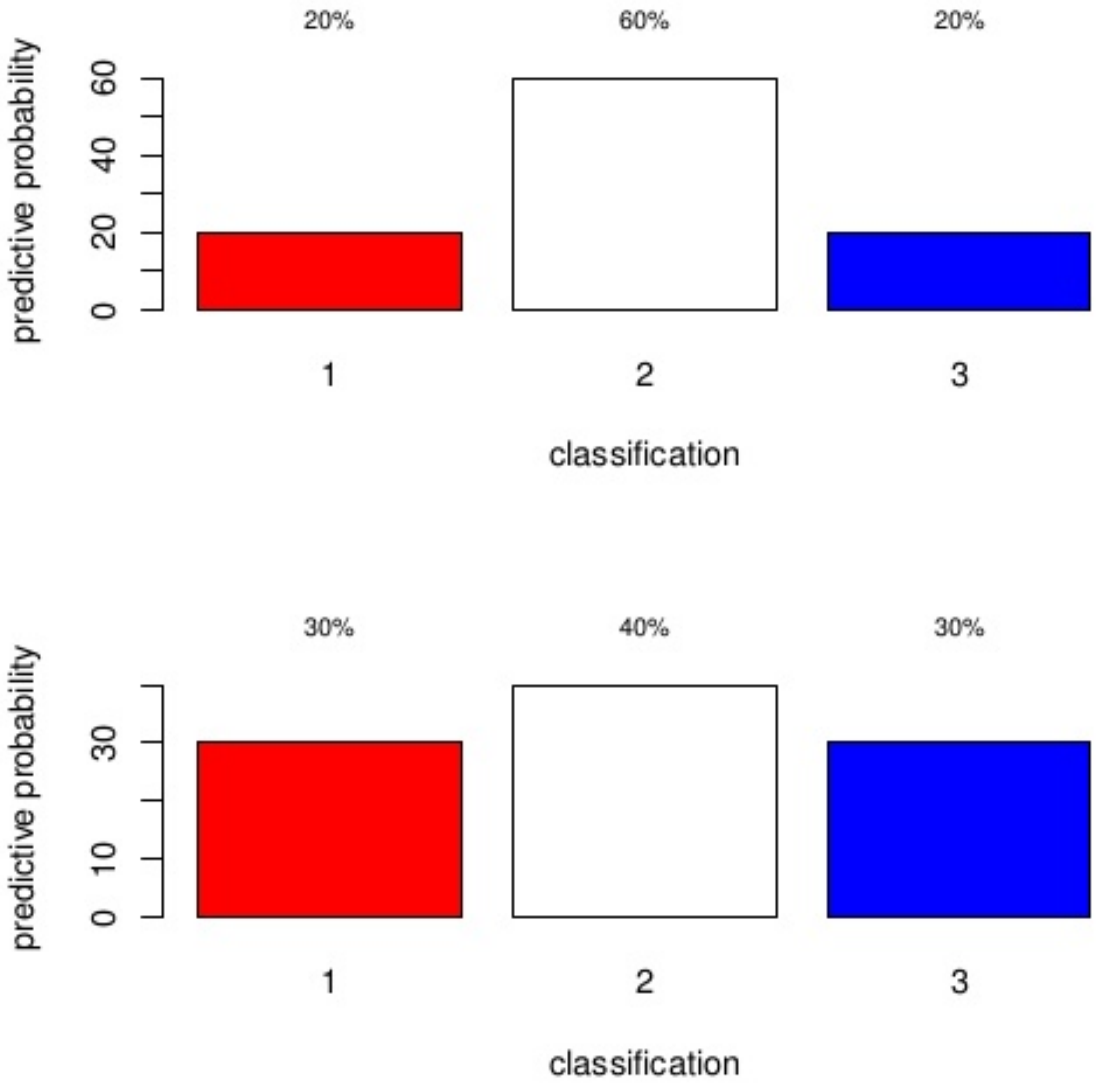

Figure 1: Model A (top) and Model B (bottom) have different predictive distributions for a given test case. They assign to the same classification and their loss is the same under the maximum a posterior decision rule.

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 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 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 mis-specification 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.

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.

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

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

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.

$$M_t = \frac{W_t}{\prod_{t=1}^{t=n}\left(1 + f_{opt}\left(t\right)\sum_j p_{t,j} \cdot x_{t,j}\right)} \tag{2}$$

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 $f_{opt}\left(t\right), t \in \{1,...,n\}$

$$W_n = \prod_{t=1}^{t=n}\left(1 + f_{opt}\left(t\right)X_t\right) \tag{3}$$

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.

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

$$P(X_t = O_t - 1) = p_t = 1 - P(X_t = -1) = P(Y_t = 1) \tag{4}$$

we note the following.

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

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

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

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

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

A plot of the back and lay fractional investment surface is provided, 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), 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

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

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

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 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

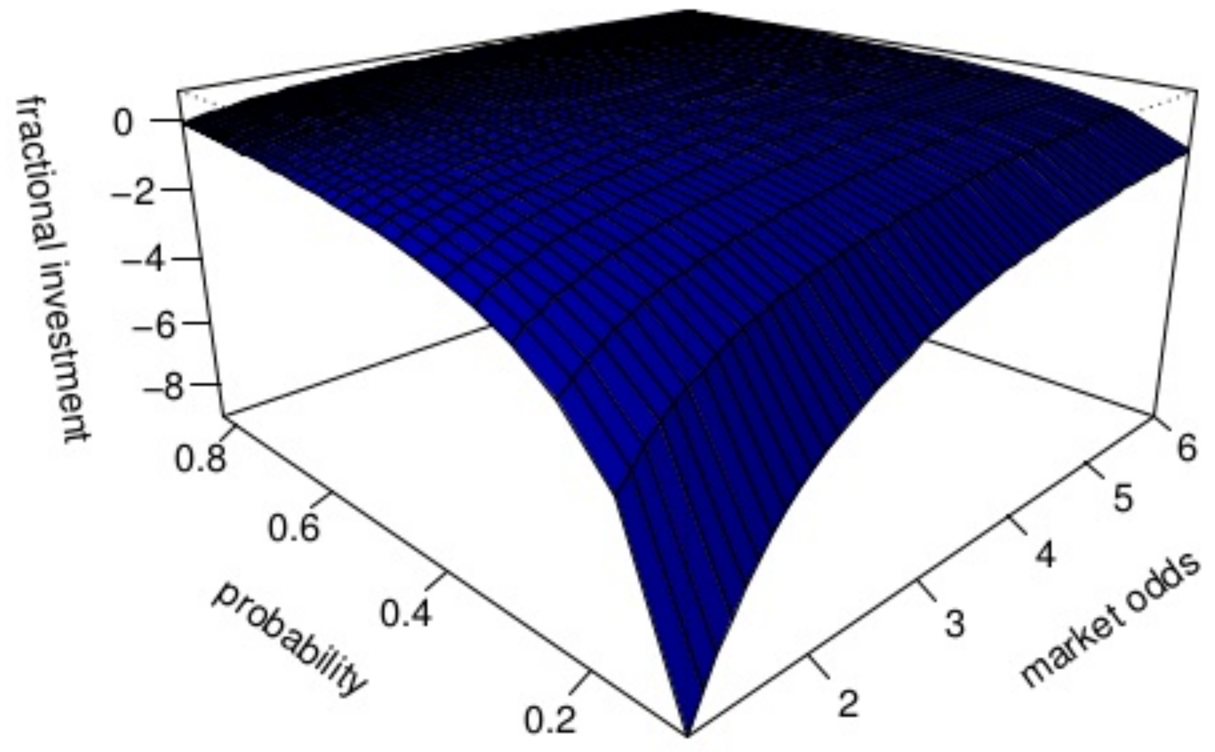

Figure 2: Notional fractional investment as a function of the market odds and underlying probabilities.

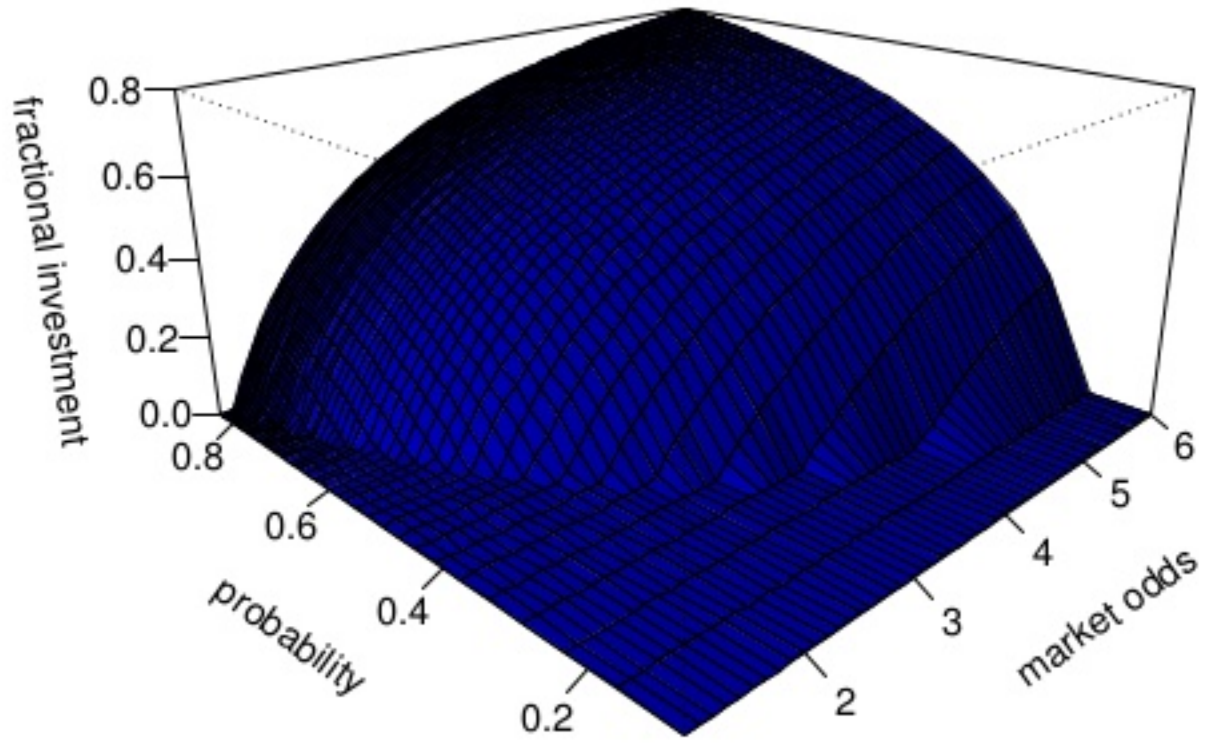

Figure 3: Notional fractional investment as a function of the market odds and underlying probabilities, with a floor function that lay bets are changed to a zero investment of wealth.

and plots of Breiman-Ripley wealth processes and returns on their natural and log scales.

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

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

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

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

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, 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

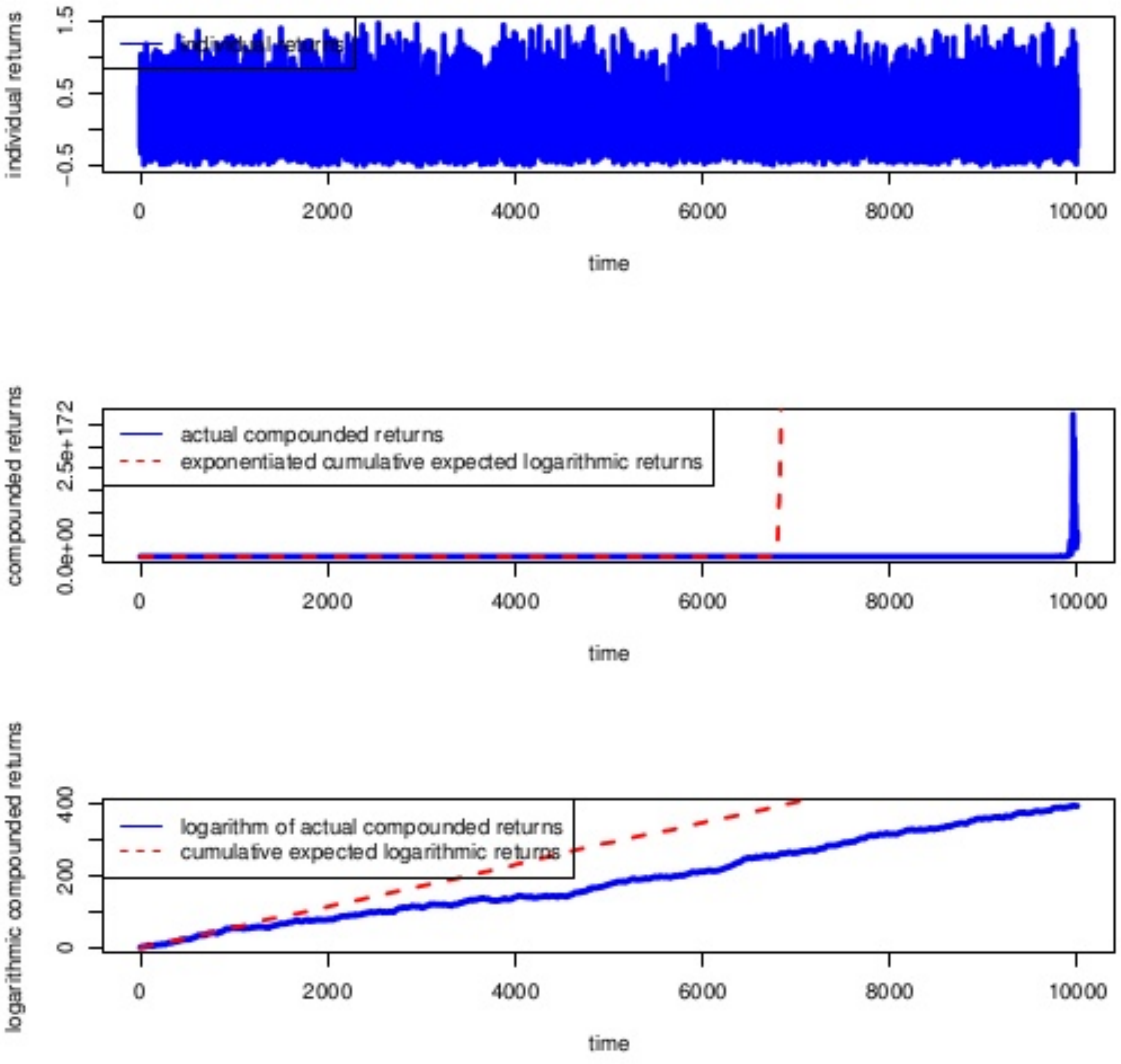

Figure 4: Simulated data: plots of individual returns, Breiman-Ripley compounded and logarithmic Breiman-Ripley wealth paths generated by betting for and against binary classification games.

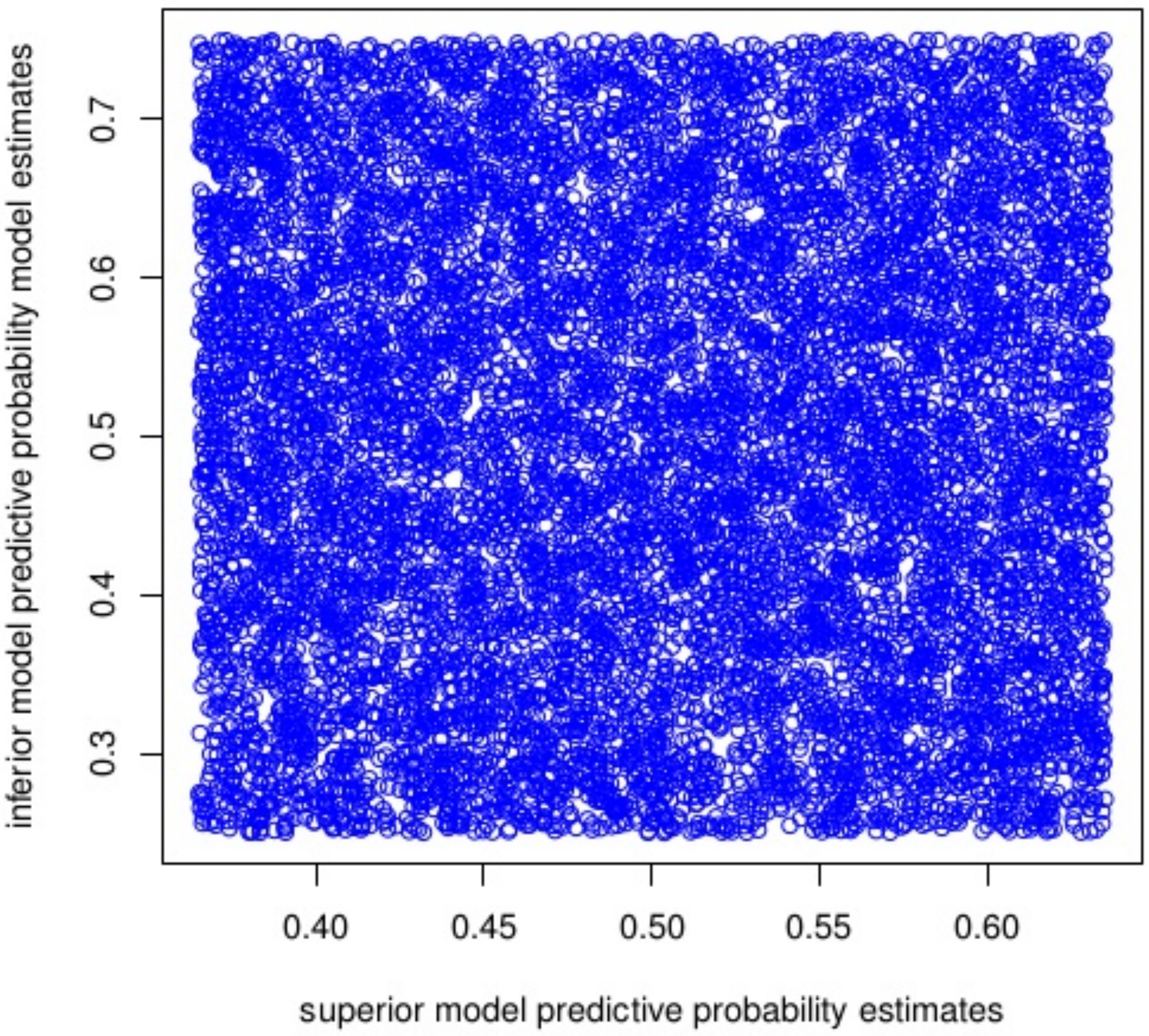

Figure 5: Simulated data: Comparison of probability estimates from two models for all test cases.

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, 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 considered. 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, 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, 6 and 7. In addition we provide the MAP error rate and percentage confusion matrices for both models.

Infant birth-weight data: Small covariate glm model

$$\begin{array}{lll} & & \text{Predicted } \% \\ \text{Actual } \% & \begin{array}{l} \\ \text{True} \\ \text{False} \\ \text{Column total} \end{array} & \begin{pmatrix} \text{True} & \text{False} & \text{Row total} \\ 3.7 & 65.1 & 68.8 \\ 2.6 & 28.6 & 31.2 \\ 6.3 & 93.7 & 100 \end{pmatrix} \end{array}$$

MAP classifier error rate = 67.7%

Infant birth-weight data: stepAIC model

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

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.

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

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

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.

□ □

Like this slideshow? Why not share!

- Share
- Email
-
-

- The AI Rush The AI Rush by Jean-Baptiste Dumont 1209335 views
- AI and Machine Learning Demystified... AI and Machine Learning Demystified... by Carol Smith 3669912 views
- 10 facts about jobs in the future 10 facts about jobs in the future by Pew Research Cent... 688247 views
- 2017 holiday survey: An annual anal... 2017 holiday survey: An annual anal... by Deloitte United S... 1106962 views
- Harry Surden - Artificial Intellige... Harry Surden - Artificial Intellige... by Harry Surden 650782 views
- Inside Google's Numbers in 2017 Inside Google's Numbers in 2017 by Rand Fishkin 1231839 views