Modeling for Optimal Probability Prediction

Yong Wang YONGWANG@CS.WAIKATO.AC.NZ
Department of Computer Science, University of Waikato, New Zealand

Ian H. Witten IHW@CS.WAIKATO.AC.NZ
Department of Computer Science, University of Waikato, New Zealand

Abstract

We present a general modeling method for optimal probability prediction over future observations, in which model dimensionality is determined as a natural by-product. This new method yields several estimators, and we establish theoretically that they are optimal (either overall or under stated restrictions) when the number of free parameters is infinite. As a case study, we investigate the problem of fitting logistic models in finite-sample situations. Simulation results on both artificial and practical datasets are supportive.

1. Introduction

Recently we proposed a new approach to fitting linear models called "pace regression" (Wang, 2000). Standard techniques for this problem include *ordinary least squares* (OLS); *OLS subset selection* methods such as FPE/AIC/C_p (Akaike, 1969, 1973; Mallows, 1973), BIC/MDL (Schwarz, 1978; Rissanen, 1978), RIC (Donoho and Johnstone, 1994; Foster and George, 1994), and CIC (Tibshirani and Knight, 1999); and *shrinkage methods* such as ridge regression (Hoerl and Kennard, 1970), NN-GARROTE (Breiman, 1995), and LASSO (Tibshirani, 1996). The new approach, which adopts a methodology that resembles empirical Bayes (Robbins, 1955, 1964), was shown to always rival, and generally outperform, all these techniques in terms of predictive squared error. Moreover, it determines the dimensionality of the model as a natural by-product of the fitting process. It is theoretically established that the new approach achieves predictive optimality as the number of free parameters specified in the model approaches infinity.

The above work minimizes the squared error. The present paper extends the same ideas to the problem of predicting the probabilities of future observations. To evaluate the performance of probability prediction we use the Kullback-Leibler distance between the two distribution functions $\tilde{f}$ and f, where $\tilde{f}$ is an estimate of the probability density function (pdf) $f(x)$. This measure is $\Delta_{KL}(f,\tilde{f}) = \mathrm{E}_f \log(f/\tilde{f}) = \int \log(f/\tilde{f}) f\, dx$ (Kullback and Leibler, 1951; Kullback, 1968). This is appropriate because, almost surely,

$$\left[\prod_{i=1}^{n} \frac{\tilde{f}(x_i)}{f(x_i)}\right]^{\frac{1}{n}} \rightarrow e^{-\Delta_{KL}(f,\tilde{f})} \quad \text{as } n \rightarrow \infty, \qquad (1)$$

where $x_1, \ldots, x_n$ are iid from $f(X)$. The goal of the modeling process that is most intuitively appealing in terms of probability is to maximize the left-hand side of (1), and when n is infinite this is equivalent to minimizing the Kullback-Leibler distance $\Delta_{KL}(f,\tilde{f})$. Note that what is involved is not the expected probability but the expected log-probability (more generally, the log-density). Further, since $\Delta_{KL}(f,\tilde{f}) \geq 0$ for any f and $\tilde{f}$, the right-hand side of (1) always lies in the range $[0,1]$.

The maximum likelihood methodology is closely related to the notion of probability prediction and has found a wide range of successful applications. Nevertheless, it performs unsatisfactorily in some cases (e.g., Stuart et al., 1999, p.80). For example, it cannot reduce the dimension of the model and so on some datasets it produces an overfitted model with poor predictive power.

Our new method is based on maximum likelihood estimation. We will show that, under appropriate regularity conditions, the maximum likelihood estimator (MLE) $\hat{f}$ of the true pdf f is *not* optimal, in the sense of minimizing $\Delta_{KL}(f,\cdot)$, among all potential estimators. We will show this constructively by exhibiting better estimators than $\hat{f}$, including the optimal estimator. We will also establish the superiority of the new estimators in an asymptotic sense as the number of free parameters—and hence also the number of observations—tends to infinity.

The ideas behind the new methods are again based on the empirical Bayes methodology. We utilize the MLE's well-known asymptotic normality property to transform the original parameters into dummy ones. We form a nonparametric mixture estimate of the observed values of these dummy parameters, and finally apply an empirical Bayes analysis to minimize Δ_{KL}.

The outline of the paper is as follows. Section 2 presents the main ideas, at the core of which is the relationship between optimal probability prediction and empirical Bayes estimation. Several important issues arise, including the choice of transformation and some finite-sample considerations, and these are discussed in Section 3. Section 4 investigates fitting logistic models as a case study, and gives some actual results on both artificial and practical datasets. We end with some concluding remarks and open issues.

2. Optimal probability prediction

In this section we show how to build models that predict probability optimally in the sense of minimizing the Kullback-Leibler distance. We briefly review the empirical Bayes methodology, then extend it to estimate the mean of a multinormal distribution, then apply the idea to general MLEs, and finally exhibit a few more estimators that are only optimal under certain restrictions.

2.1 Empirical Bayes

Given independent samples $x_1, \ldots, x_k$ from distributions $F(x_i; \theta_i)$—where the θ_i may be completely different from each other—it is known that the MLE obtained from the joint distribution $F(\mathbf{x}; \boldsymbol{\theta})$ is the vector, with each entry being a univariate MLE; for example, if $F(x_i; \theta_i)$ is the normal distribution with mean θ_i, then $\widehat{\boldsymbol{\theta}} = \mathbf{x}$. (Throughout this paper we adopt the notation $\mathbf{a} = (a_1, \ldots, a_k)^T$.) The MLE estimator, however, is inferior to the empirical Bayes estimator

$$\tilde{\theta}_i^{EB} = \frac{\int \theta f(x_i; \theta)\, dG_k(\theta)}{\int f(x_i; \theta)\, dG_k(\theta)} \tag{2}$$

—(here $f(x_i; \theta_i)$ denoting the pdf corresponding to $F(x_i; \theta_i)$)—inferior in the sense that it does not minimize the expected squared error $\mathrm{E}_{f(\mathbf{x})}||\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$ with respect to the estimator $\widetilde{\boldsymbol{\theta}}(\mathbf{x})$, where $\theta_1, \ldots, \theta_k$ are iid from $G(\theta)$. Here G is the mixing distribution of the mixture $f_G(x) = \int f(x; \theta)\, dG$, and G_k in Equation (2) is a consistent estimator of G given the mixture sample $\mathbf{x}$. Robbins (1964) shows that, under weak conditions, $\widetilde{\boldsymbol{\theta}}^{EB}$ minimizes the Bayes risk as $k \to \infty$ and hence is asymptotically optimal.

We can interpret this result without taking a Bayesian perspective. Given an estimator $\widetilde{\boldsymbol{\theta}}(\mathbf{x})$ and a consistent G_k, $\mathrm{E}_{f(\mathbf{x})}\mathrm{E}_{G_k}(\tilde{\theta} - \theta)^2 \to \mathrm{E}_{f(\mathbf{x})}\mathrm{E}_G(\tilde{\theta} - \theta)^2 = \frac{1}{k}\mathrm{E}_{f(\mathbf{x})}||\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$ almost surely as $k \to \infty$. Here G is just the discrete function which jumps $\frac{1}{k}$ at each θ_i. It is not a distribution function (though it resembles one) because $\theta_1, \ldots, \theta_k$ are not randomly sampled from G: they always take the same, fixed, value in all situations. Thus G_k is only an estimate of this function. This interpretation leads to the same result as the Bayesian framework.

Many consistent estimators of an arbitrary G are available in the literature, including the MLE (e.g., Laird, 1978; Böhning et al., 1992; Lesperance & Kalbfleisch, 1992) and some minimum distance estimators (e.g. Choi & Bulgren, 1968; Deely & Kruse, 1968; Macdonald, 1971; Blum & Susarla, 1977; Wang, 2000). Here, *consistency* means that

$$\Pr(\lim_{k\to\infty} G_k(\theta) = G(\theta), \quad \theta \text{ any continuity point of } G) = 1. \tag{3}$$

2.2 Estimation for multinormal distributions

A special case of the above result occurs when $F(x_i; \theta_i)$ is the normal distribution function with mean $\mu_i \equiv \theta_i$ and common, known, variance σ^2 for all x_i's. Then the problem becomes one of estimating $\boldsymbol{\mu}$ given a single multivariate observation $\mathbf{x}$ from $N_k(\boldsymbol{\mu}, \sigma^2\mathbf{I}_k)$, where $\mathbf{I}_k$ is the identity matrix.

We now generalize this result to the problem of estimating the mean of a multinormal distribution $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the purposes of optimal probability prediction, where the covariance matrix $\boldsymbol{\Sigma}$ is assumed known. Again, only a single observation $\mathbf{x}$ is given. This problem can be transformed into the above special case by supposing that $\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{Q}^T$ for some $k \times k$ matrix $\mathbf{Q}$ (the choice of $\mathbf{Q}$ will be discussed in Section 3) and writing $\boldsymbol{\gamma} = \mathbf{Q}^{-1}\boldsymbol{\mu}$ and $\widehat{\boldsymbol{\gamma}} = \mathbf{Q}^{-1}\widehat{\boldsymbol{\mu}} (= \mathbf{Q}^{-1}\mathbf{x})$, so that $\widehat{\boldsymbol{\gamma}} \sim N_k(\boldsymbol{\gamma}, \mathbf{I}_k)$—which conforms to the previous formulation. Note that, because of the transformation, the G and G_k of Section 2.1 are now the distribution functions of the transformed parameters $\gamma_1, \ldots, \gamma_k$ instead of the original ones $\mu_1, \ldots, \mu_k$. Once $\widehat{\boldsymbol{\gamma}}$ has been upgraded to $\widetilde{\boldsymbol{\gamma}}$ (using Equation (2), or some other formula given in Section 2.4) we can obtain the upgraded estimate $\widetilde{\boldsymbol{\mu}} = \mathbf{Q}\widetilde{\boldsymbol{\gamma}}$ of $\boldsymbol{\mu}$, which is what we are ultimately interested in.

Now we justify this approach in terms of probability prediction. Denote the pdf of $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by $f(\mathbf{x})$ and the pdf of the estimate $N_k(\widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ by $\tilde{f}(\mathbf{x})$. Since, as

$k \to \infty$,

$$\begin{aligned} & \mathrm{E}_f \log \tilde{f} = C - \frac{1}{2}\mathrm{E}_f[(\mathbf{x}-\widetilde{\boldsymbol{\mu}})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\widetilde{\boldsymbol{\mu}})] \\ \to \quad & C - \frac{1}{2}\mathrm{E}_f[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})] \\ & \quad -\frac{1}{2}\mathrm{E}_f[(\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu})] \text{ (a.s.)} \\ = \quad & \mathrm{E}_f \log f - \frac{1}{2}\mathrm{E}_f ||\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}||^2, \end{aligned} \tag{4}$$

where C is a constant not depending on $\boldsymbol{\mu}$ or $\mathbf{x}$, we have $2\Delta_{KL}(f,\tilde{f}) = \mathrm{E}_f||\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}||^2$ a.s. as $k \to \infty$. This reduces the problem of optimal probability prediction to one of minimizing the expected squared errors in terms of the dummy parameter γ (not μ, as one might expect). Thus the empirical Bayes estimator $\widetilde{\boldsymbol{\gamma}}^{EB}$ obtained by (2) is asymptotically optimal in the sense of minimizing $\Delta_{KL}(f,\tilde{f})$. Some other suboptimal $\widetilde{\boldsymbol{\gamma}}$s are given in Section 2.4.

Note that the expectation E_f is over the entire sample space, including future observations sampled from the same distribution. Taking the expectation over just the training sample leads to the MLE.

2.3 Upgrading general MLEs

The above results extend naturally to the case where it is a general MLE that is being upgraded, rather than the specific one for multinormal distributions described in the previous subsections. It is well-known that, under regularity conditions, the MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ becomes multinormally distributed as $n \to \infty$, that is,

$$\widehat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \mathbf{I}^{-1}(\theta)), \tag{5}$$

where $\mathbf{I}(\theta)$ is the *Fisher information matrix*. The likelihood function reduces to

$$L(\widehat{\boldsymbol{\theta}};\boldsymbol{\theta}) \propto \exp\{-\frac{1}{2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T \mathbf{I}(\theta)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\}, \tag{6}$$

which is also the sampling pdf of the MLE. Of course, in practice $\mathbf{I}(\boldsymbol{\theta})$ is usually unknown, but it can often be accurately approximated using the consistent estimator $\mathbf{I}(\widehat{\boldsymbol{\theta}})$ (the *estimated information*), or perhaps the *observed information* $-\frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j}|_{\widehat{\boldsymbol{\theta}}}$. Then, all that is needed is to follow the derivation in Section 2.2. In terms of $\boldsymbol{\gamma}$, (5) and (6) become $\widehat{\boldsymbol{\gamma}} \sim N(\boldsymbol{\gamma}, \mathbf{I}_k)$ and $L(\widehat{\boldsymbol{\gamma}};\boldsymbol{\gamma}) \propto \exp\{-\frac{1}{2}||\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}||^2\}$ respectively.

2.4 Optimal $\widetilde{\boldsymbol{\gamma}}$s under certain restrictions

Section 2.2 gave the optimal $\widetilde{\boldsymbol{\gamma}}$ for probability prediction (i.e., $\widetilde{\boldsymbol{\gamma}}^{EB}$), under the transformation $\mathbf{Q}$. Other criteria can be applied to the dummy parameter vector $\boldsymbol{\gamma}$ in a similar way—for example, the standard techniques mentioned at the beginning of Section 1. Each standard technique implies certain restrictions, and there are analogous estimators of $\boldsymbol{\gamma}$ that are optimal under the same restrictions—and hence outperform the relevant standard technique. Of course, the advantage provided by each restriction (if any) vanishes asymptotically, and so $\widetilde{\boldsymbol{\gamma}}^{EB}$ is the overall optimum.

The following estimators, which were originally derived for fitting linear models (Wang, 2000), readily carry over to this case because optimal probability prediction is equivalent to minimizing the expected squared error of $\widetilde{\boldsymbol{\gamma}}$ (see Section 2.2). Given $G(\gamma)$ (or its consistent estimator G_k), define

$$h(t;G) = \int [\gamma^2 - (t-\gamma)^2] f(t;\gamma)\, dG(\gamma) \tag{7}$$

and

$$H(\tau;G) = \int_{-\tau}^{\tau} h(t;G)\, dt. \tag{8}$$

(For details, see Wang, 2000.)

Thresholding: Each $\hat{\gamma}_j$ is subject to a threshold τ_j for retaining that term or discarding it (i.e., setting it to zero). Standard estimators that belong to this category include OLS ($\tau = 0$), FPE/AIC/C_p ($\tau_j = \sqrt{2}$), BIC/MDL ($\tau_j = \sqrt{\log n}$), RIC ($\tau_j = \sqrt{2\log k}$) and CIC ($\tau_j = \sqrt{4\log(k/j)}$). (Note that the equivalences are asymptotic.) The optimal threshold that we found is $\tau^* = \arg\min H(\tau;G)$ (Wang, 2000). That is,

$$\tilde{\gamma}_j^{Thresh} = \begin{cases} \hat{\gamma}_j & \text{if } |\hat{\gamma}_j| > \tau^* \\ 0 & \text{if } |\hat{\gamma}_j| \le \tau^*. \end{cases} \tag{9}$$

Nested models: $\hat{\gamma}_j$s are brought into the model in a predefined sequence, say $j = 1, \ldots, k$ (without loss of generality). The optimal model of this type is

$$\tilde{\gamma}_j^{Nested} = \begin{cases} \hat{\gamma}_j & \text{if } j \le j^* \\ 0 & \text{if } j > j^*; \end{cases} \tag{10}$$

where $j^* = \arg\max_{j\in\{0,1,\ldots,k\}} \sum_{i=1}^{j} h(\hat{\gamma}_i;G)/f(\hat{\gamma}_i;G)$.

Subset models: Each $\hat{\gamma}_j$ is tested individually without specifying a threshold. In this case, the sign of $h(\hat{\gamma}_j;G)$ can be used as the criterion, that is,

$$\tilde{\gamma}_j^{Subset} = \begin{cases} \hat{\gamma}_j & \text{if } h(\hat{\gamma}_j;G) > 0 \\ 0 & \text{if } h(\hat{\gamma}_j;G) \le 0. \end{cases} \tag{11}$$

Shrinkage: Each $\hat{\gamma}_j$ is shrunk towards 0 by a given constant c_j ($-1 \le c_j \le 1$). Shrinkage estimators include ridge regression, NN-GARROTE and LASSO. We obtain

$$\tilde{\gamma}_j^{Shrink} = c_j \hat{\gamma}_j \tag{12}$$

where $c_j = \mathrm{sgn}(\frac{\tilde{\gamma}_j^{EB}}{\hat{\gamma}_j}) \min(1, |\frac{\tilde{\gamma}_j^{EB}}{\hat{\gamma}_j}|)$.

3. Discussion

Important issues that must be resolved to yield practical modelling methods include the choice of transformation matrix $\mathbf{Q}$ defined in Section 2.2, how to reduce dimensionality using the new estimators, and how to handle finite samples.

3.1 Choice of Q

The ideas presented in Section 2 involve transforming the original parameter vector $\boldsymbol{\theta}$ into the dummy parameter vector $\boldsymbol{\gamma}$, to which the empirical Bayes method is then applied. The empirical Bayes method could be applied to $\boldsymbol{\theta}$ directly, but the transformation is necessary to ensure that it is Δ_{KL}—not $\mathrm{E}_f||\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}||$, say—that is being minimized.

But the question is left open: which one should be used of the many $\mathbf{Q}$s that satisfy $\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{Q}^T$? Such $\mathbf{Q}$s include those based on eigenvalue decomposition, Cholesky decomposition, partial F-test, and indeed all orthogonal transformation of any of these. Note that the asymptotic optimality of the estimators given in Section 2 have been established under the condition that $\mathbf{Q}$ is applied to all estimators and is statistically independent of the response (i.e., the observed density/probability).

There are two kinds of choice. One is to pick some independent $\mathbf{Q}$, perhaps based on eigenvalue or Cholesky decomposition. The other is to exploit information in the data, perhaps using the partial F-test. The disadvantage of the first is that the choice may not suit the data. The disadvantage of the second is that it fails the independence condition—though hopefully only slightly. The merits of the choice are case-dependent.

The improvement of $\widetilde{\boldsymbol{\gamma}}$ upon $\widehat{\boldsymbol{\gamma}}$ depends on the diversity of the values of the γ_j. The more diverse they are, the less they can inform the adjustment of any particular $\hat{\gamma}_j$. If one $\hat{\gamma}_j$ differs radically from all others, there is little basis on which to alter its value. Therefore, $\mathbf{Q}$ should be chosen to cluster together as many of the γ_j as possible. In practice, of course the values of γ_j are unavailable, but this consideration does shed light on the choice of $\mathbf{Q}$—as the special case of dimensionality reduction shows.

3.2 Dimensionality reduction

Some of the estimators in Section 2 explicitly reduce the dimensionality of the model, in the sense that they estimate some entries of $\boldsymbol{\gamma}$ to be zero. These include $\widetilde{\boldsymbol{\gamma}}^{Thresh}$, $\widetilde{\boldsymbol{\gamma}}^{Nested}$ and $\widetilde{\boldsymbol{\gamma}}^{Subset}$, which all discard some values of $\widehat{\boldsymbol{\gamma}}$. The other estimators, $\widetilde{\boldsymbol{\gamma}}^{EB}$ and $\widetilde{\boldsymbol{\gamma}}^{Shrink}$, generally reduce dimensionality implicitly, because it invariably turns out that many of the entries of $\widetilde{\boldsymbol{\gamma}}^{EB}$ and $\widetilde{\boldsymbol{\gamma}}^{Shrink}$ are tiny and can be truncated to zero with negligible impact on Δ_{KL}.

Reducing the dimensionality of $\widetilde{\boldsymbol{\gamma}}$ as in the last paragraph is certainly not equivalent to reducing the dimensionality of $\widetilde{\boldsymbol{\theta}}$, which is usually of real interest in practice. One general way to reduce the latter is to truncate some values of $\widetilde{\boldsymbol{\theta}}$ to zero and to see whether the truncation has a negligible effect. Another way is to use partial F-test to obtain an uppertriangular $\mathbf{Q}$ (using pivoting). Then if all consecutive $\tilde{\gamma}_j$ from $j = k$ downwards are zero (or close to zero), the corresponding values of $\tilde{\theta}_j$ are zero (or close to zero) too. This case illustrates an advantage of using the partial F-test to choose $\mathbf{Q}$, because any θ_j with an effect that is zero (or close to zero) is sifted out first and the corresponding transformed values γ_j will be clustered around 0. This clustering effect will make the upgrading of these γ_js mutually self-supporting. In contrast, an arbitrarily chosen $\mathbf{Q}$ fails to exploit this information. In fact the effect is common in many practical datasets, and an arbitrarily chosen $\mathbf{Q}$ is unlikely to perform so well in these cases.

As well as improving probability prediction, dimensionality reduction often has side benefits. Reducing the number of non-zero parameters usually implies fewer measurement attributes, faster prediction, and more easily understood models. Thus it decreases the practical cost associated with the modelling operation—though this cost is difficult to quantify, and even more difficult to manipulate mathematically.

3.3 Nested model selection using the likelihood ratio statistic

The asymptotic results given above will not always work well for finite samples. Two serious problems arise. The first is that when the Fisher information matrix is unknown, its estimate—e.g., the estimated information obtained from the full model—can be inaccurate. This compromises the foundations of the theoretical analysis. The second is that the normality formulae (5) and (6) are poor approximations for finite samples—especially small ones. Hence blindly upgrading, without checking the result against the data, can yield unexpected outcomes.

Our remedy is to use the *likelihood ratio test statistic* between two consecutively nested models

$$\hat{\lambda}_j = 2l(\widehat{\boldsymbol{\theta}}_j) - 2l(\widehat{\boldsymbol{\theta}}_{j-1}), \tag{13}$$

where $\widehat{\boldsymbol{\theta}}_j$ denotes the MLE with j free parameters for $j = 0, 1, \ldots, k$, and $l(\widehat{\boldsymbol{\theta}}_j)$ its log-likelihood. This

is used only for selecting among nested models as in (10). Although the fact that $\hat{\lambda}_j$ is distributed as a non-central χ^2_1 rests on the assumption of asymptotic normality, it usually works well for finite samples. Indeed, we have $\hat{\lambda}_j = \hat{\gamma}_j^2$ asymptotically, where the $\hat{\gamma}_j$s correspond to a transformation $\mathbf{Q}$ obtained using the likelihood ratio test. The mixing distribution G is now a function of the variable λ rather than γ, and the nested model selector (10) carries over readily in terms of λ. Wang (2000) describes how to handle mixtures of non-central χ^2_1's.

There are many applications for techniques that select amongst nested models—for example, pruning tree-structured models (Breiman et al., 1984; Quinlan, 1993).

4. Case study: Fitting logistic models

Now it is time to apply the general results we have established to the special case of logistic regression models, and present simulation results.

4.1 Logistic models

Logistic models for two-class problems take the form

$$\pi(\mathbf{x};\boldsymbol{\beta}) = \frac{1}{1+\exp(-\boldsymbol{\beta}^T\mathbf{x})}, \tag{14}$$

where $\pi(\mathbf{x};\boldsymbol{\beta})$ is the probability that $y = 1$ (rather than 0) at $\mathbf{x}$. The log-likelihood function for the instances $(\mathbf{x}_1^T, y_1), \ldots, (\mathbf{x}_n^T, y_n)$ is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \big[y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\big], \tag{15}$$

where $\pi_i = \pi(\mathbf{x}_i;\boldsymbol{\beta})$. The MLE $\widehat{\boldsymbol{\beta}}$ satisfies the equation

$$\mathbf{X}^T\widehat{\mathbf{W}}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\widehat{\mathbf{W}}\hat{\mathbf{z}}, \tag{16}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, $\widehat{\mathbf{W}} = \mathrm{diag}\{\hat{\pi}_i(1-\hat{\pi}_i)\}$, and $\hat{z}_i = \widehat{\boldsymbol{\beta}}^T\mathbf{x}_i + (y_i - \hat{\pi}_i)/(\hat{\pi}_i(1-\hat{\pi}_i))$. This equation is usually solved using the iteratively re-weighted least squares method (which here is the same as the Newton-Raphson method and the Fisher scoring method). Note that the Fisher information matrix is $\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{W}\mathbf{X}$. Given $\widehat{\boldsymbol{\beta}}$, and with $\mathbf{I}(\boldsymbol{\beta})$ replaced with the estimated information $\mathbf{I}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}^T\widehat{\mathbf{W}}\mathbf{X}$, one can obtain the estimators defined in Section 2.

To evaluate the resulting models, we first resort to the Kullback-Leibler distance. To avoid numeric integration in high-dimensional spaces, our experiments used Monte Carlo integration. The Kullback-Leibler distance over a sample (superscript s) is given by

$$\Delta^s_{KL}(f,\tilde{f}) = \frac{1}{n}\sum_{i=1}^{n}\left[\pi_i \log(\frac{\pi_i}{\tilde{\pi}_i}) + (1-\pi_i)\log\frac{1-\pi_i}{1-\tilde{\pi}}\right]. \tag{17}$$

This is easy to evaluate when the true distribution function f is known—as it is in the artificial experiments. When f is unknown—as in practice—we use the negative log-likelihood over the test set:

$$-l(\widetilde{\boldsymbol{\beta}}) = -\frac{1}{n}\sum_{i=1}^{n}\big[y_i \log\tilde{\pi}_i + (1-y_i)\log(1-\tilde{\pi}_i)\big]. \tag{18}$$

This is equivalent to $\Delta^s_{KL}(f,\tilde{f})$, up to a constant.

Logistic models have a natural application to classification problems. Given an estimate $\widetilde{\boldsymbol{\beta}}$, the attribute space is split into subspaces by the linear discriminant function $\mathbf{x}^T\widetilde{\boldsymbol{\beta}} = 0$. An instance $\mathbf{x}$ is classified as 1 if $\mathbf{x}^T\widetilde{\boldsymbol{\beta}} > 0$ and 0 if $\mathbf{x}^T\widetilde{\boldsymbol{\beta}} < 0$. If $\mathbf{x}^T\widetilde{\boldsymbol{\beta}} = 0$, its class is undetermined—it belongs to each class with probability 0.5.

The *classification rate* (CR) is a natural performance yardstick for classification problems that measures the percentage of correctly classified instances. Over the whole sampling space for an estimate $\widetilde{\boldsymbol{\beta}}$, we have

$$\begin{aligned}\mathrm{CR}(\widetilde{\boldsymbol{\beta}};\boldsymbol{\beta}) = \mathrm{E}_f[&\delta(\mathbf{x}^T\widetilde{\boldsymbol{\beta}}\mathbf{x}^T\boldsymbol{\beta} < 0)\\ &+ \frac{1}{2}\delta(\mathbf{x}^T\widetilde{\boldsymbol{\beta}} = 0 \text{ or } \mathbf{x}^T\boldsymbol{\beta} = 0)],\end{aligned} \tag{19}$$

where $\delta(\cdot)$ is the indicator function and f denotes the pdf $f(y|\mathbf{x},\boldsymbol{\beta})$. There are two versions of the classification rate over a sample (as there are of Δ_{KL}), depending on whether $\boldsymbol{\beta}$ is known or unknown:

$$\begin{aligned}\mathrm{CR}^s(\widetilde{\boldsymbol{\beta}};\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[&\delta(\mathbf{x}_i^T\widetilde{\boldsymbol{\beta}}\mathbf{x}_i^T\boldsymbol{\beta} > 0)\\ &+ \frac{1}{2}\delta(\mathbf{x}_i^T\widetilde{\boldsymbol{\beta}} = 0 \text{ or } \mathbf{x}_i^T\boldsymbol{\beta} = 0)]\end{aligned} \tag{20}$$

and

$$\begin{aligned}\mathrm{CR}^s(\widetilde{\boldsymbol{\beta}},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}[&\delta(\mathbf{x}_i^T\widetilde{\boldsymbol{\beta}}(y_i - 0.5) > 0)\\ &+ \frac{1}{2}\delta(\mathbf{x}_i^T\widetilde{\boldsymbol{\beta}} = 0)].\end{aligned} \tag{21}$$

4.2 Simulation studies

Now we show the results of some experiments in fitting logistic models. In particular, we examine selection from a sequence of nested models. Each subset model is the MLE after deleting the next least significant variable, one at a time. This process builds a $\mathbf{Q}$

matrix implicitly. We investigate the following estimators: ML (maximum likelihood), AIC, BIC, RIC, CIC, and the optimal nested model selector (10) based on the likelihood ratio statistic (Section 3.3), which is denoted "New." The overall optimal estimator derived in Sections 2.1-2.3 is not included, because of the considerations explained in Section 3.3.

Three performance measures are used: the Kullback-Leibler distance over a sample, the classification rate, and the model dimension. The first two are indicators of probability prediction, while the third is of interest in dimensionality reduction. In the second experiment, where the true model is unknown, the negative log-likelihood is used as a substitute for the Kullback-Leibler distance.

We report the results of two experiments. In the first, artificial datasets are used so that accurate values of the performance measures can be obtained. This experiment shows the strengths and weaknesses of each modeling procedure in different situations. The second experiment applies the same procedures to artificial datasets, and cross-validation results are given.

Experiment 1 *Artificial datasets.* The first experiment investigates the effect of non-zero parameters in the true model using artificial datasets. Each dataset (both training and test sets) contains $n = 1000$ instances and has $k = 30$ parameters, the intercept plus 29 attribute-associated parameters. Except for the constant attribute $x_1 = 1$, $x_j \sim U(0,1)$ for $j = 2, \ldots, k$. Each parameter β_j $(j = 2, \ldots, k)$ is either 0 or β_c, with a fraction p being β_c. We set $\beta_1 = -\text{mean}(\mathbf{x}_{-1}^T \boldsymbol{\beta}_{-1})$ so that the instances fall around the point $\pi = 0.5$.

Figure 1 shows a cross-section of the results, for situations $p = \{0\%, 10\%, \ldots, 100\%\}$ and $\beta_c = \{1, 2\}$. Each data point in these graphs is the average of 50 runs under the same conditions. Figures 1(a), (b) and (c) are for $\beta_c = 1$, and correspond to $\Delta_{KL}^s(f, \tilde{f})$, $\text{CR}^s(\widetilde{\boldsymbol{\beta}}; \boldsymbol{\beta})$ and the fitted model dimension respectively. Figures 1(d), (e) and (f) show the same three measures for $\beta_c = 2$. The *New* method derived in this paper is shown by a solid line. In Figure 1 (a) and (d) it is apparent that the estimator *New* is always the best, or amongst the best, in the sense of achieving minimal values for the Kullback-Leibler distance Δ_{KL}^s—and, unlike the other methods, it performs consistently well across the whole range of p. In (b) and (e) the same effect can be observed for the classification rate CR^s, which is to be maximized. In (c) and (f) it can be seen that *New* reduces the model dimensions appropriately across the whole range too. No other estimator achieves good performance throughout the range, although each works well under its own favorable circumstances. For example, ML and CIC work well for $p = 100\%$, RIC, BIC and CIC for $p = 0\%$, and AIC for $p = 40 \sim 60\%$.

Experiment 2 *Practical datasets.* The second experiment investigates the performance of the same methods on eight practical datasets. Seven are from the Machine Learning Database Repository at UCI (Blake et al., 1998): BreastCancer (Wisconsin Breast Cancer, original), ClevelandHeart (Heart Disease, Cleveland), German (Statlog Project, German Credit), Ionosphere, Pima (Pima Indian Diabetes), Spambase, and WDBC (Winsconsin Breast Cancer, WDBC). The eighth is the Crab dataset from Agresti (1996). Some of the datasets were modified slightly: some attributes and instances were deleted to eliminate missing values, multi-class problems were transformed into binary ones, a (randomly-chosen) subset of instances were used for computational reasons. Table 2 gives the final number of instances n and attributes $(k-1)$ for each dataset in parentheses. In some cases, when the iteration of the maximum likelihood estimation does not converge, we replaced all $y_i = 1$ by 0.95 and $y_i = 0$ by 0.05. The MLEs are calculated over the slightly altered datasets.

Since the true models are unknown, cross-validation results were calculated and are shown in Table 2, in terms of the three performance measures: the negative log-likelihood $-l(\widehat{\boldsymbol{\beta}})$ $(\times 100)$, used as a substitute for the Kullback-Leibler distance; the classification rate $\text{CR}^s(\widetilde{\boldsymbol{\beta}}, \mathbf{y})$; and the model dimension. Each value is the average of twenty runs of ten-fold cross-validation. According to both the negative log-likelihood and the classification rate, the estimator *New* provides either the best or nearly the best results for six of the datasets. For the other two (Spambase and WDBC), its results are intermediate and comparable with other estimators. Along with this, it also reduces the model dimensionality, which the MLE can never do.

5. Summary and future work

We have presented a new general approach to probability prediction over future observations. Not only does it achieve optimal prediction, it also determines the appropriate dimensionality of the model as a natural by-product of optimality. We have exhibited several new estimators that are either optimal overall, or optimal when certain restrictions are enforced. In all cases, optimality is achieved as the number of free parameters approaches infinity. This suggests that performance will tend to improve as the number of parameters increases, and so the new methods are appropriate for

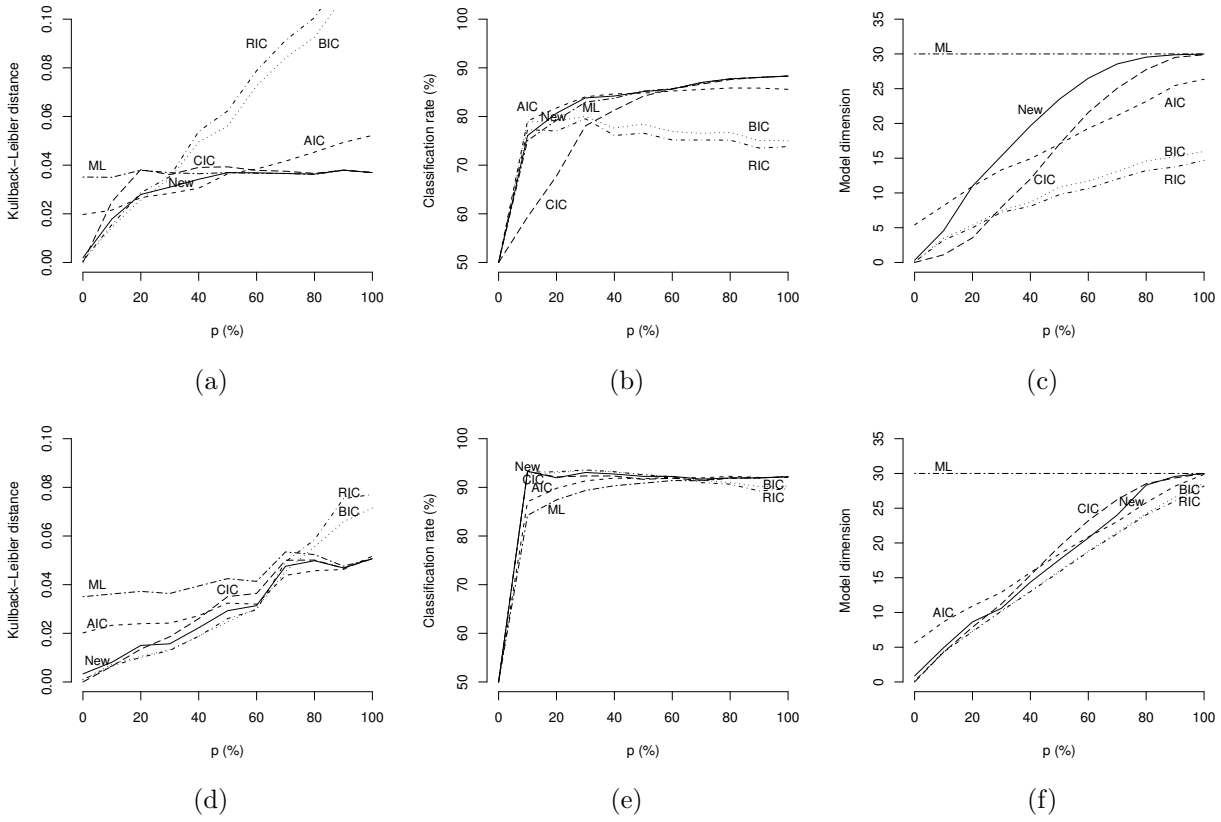

Figure 1. Experiment 1. The upper row shows results for $\beta_c = 1$ and the lower row for $\beta_c = 2$. The three graphs in each row show $\Delta^s_{KL}(f, \tilde{f})$, $\mathrm{CR}^s(\widetilde{\boldsymbol{\beta}}; \boldsymbol{\beta})$ and the model dimension respectively.

Procedure	BreastCancer (683/9)			ClevelandHeart (294/13)			Crab (173/4)			German (1000/24)		
	$-l(\widetilde{\boldsymbol{\beta}})$	$\mathrm{CR}^s(\%)$	Dim	$-l(\widetilde{\boldsymbol{\beta}})$	$\mathrm{CR}^s(\%)$	Dim	$-l(\widetilde{\boldsymbol{\beta}})$	$\mathrm{CR}^s(\%)$	Dim	$-l(\widetilde{\boldsymbol{\beta}})$	$\mathrm{CR}^s(\%)$	Dim
ML	12.5	96.7	10.0	39.5	83.3	14.0	56.9	70.4	5.0	49.4	76.9	25.0
AIC	13.1	96.4	6.8	42.4	82.1	9.4	58.4	69.0	3.0	49.8	76.5	14.2
BIC	14.1	96.1	4.9	45.1	81.4	5.3	59.5	67.7	2.1	51.2	74.5	6.1
CIC	12.6	96.7	9.7	39.9	83.1	13.6	57.4	67.0	4.7	50.1	76.2	13.9
RIC	13.9	96.0	5.4	44.8	81.3	5.4	59.4	68.1	2.7	51.3	74.5	6.3
New	12.7	96.6	9.5	39.7	83.2	13.8	58.0	69.2	4.1	49.5	76.8	23.3
	Ionosphere (351/33)			Pima (768/8)			Spambase (500/57)			WDBC (569/30)		
	$-l(\widetilde{\boldsymbol{\beta}})$	$\mathrm{CR}^s(\%)$	Dim	$-l(\widetilde{\boldsymbol{\beta}})$	$\mathrm{CR}^s(\%)$	Dim	$-l(\widetilde{\boldsymbol{\beta}})$	$\mathrm{CR}^s(\%)$	Dim	$-l(\widetilde{\boldsymbol{\beta}})$	$\mathrm{CR}^s(\%)$	Dim
ML	39.8	88.2	34.0	48.7	77.4	9.0	32.6	88.7	58.0	15.2	95.7	31.0
AIC	36.9	87.6	15.6	49.0	76.8	6.7	33.6	86.9	19.4	15.5	96.0	10.7
BIC	40.1	85.9	7.3	49.6	76.4	4.5	37.2	86.6	8.6	16.6	95.6	4.6
CIC	39.3	86.1	8.0	48.7	77.4	8.8	37.9	86.6	7.7	16.8	95.5	4.4
RIC	41.5	85.6	6.0	49.2	76.5	5.3	38.2	86.7	7.4	16.7	95.5	4.4
New	38.8	87.8	24.7	48.7	77.3	8.5	34.5	86.9	19.3	16.7	95.6	4.6

Table 1. Results for practical datasets in Experiment 2.

high-dimensional parameter spaces. The ideas resemble the empirical Bayes methodology, although we do not adopt a Bayesian perspective nor even assume the existence of a *random* prior distribution.

We conducted a case study on the use of this general approach to fit logistic regression models. Although all our theoretical conclusions are asymptotic, simulation results on finite datasets (both artificial and practical) are promising: the new method is nearly always

amongst the best, as well as reduces the model dimensions appropriately.

Because of its generality and favourable theoretical properties, the new method shows great promise. However, before it is truly useful in a wide range of application areas, more work is needed on the best choice for the transformation matrix $\mathbf{Q}$, improved approximations for finite samples, inclusion of multi-labeled enumerated variables, the handling of missing values, and application to other model structures such as decision trees.

Acknowledgments

The first author would like to thank Alastair Scott and Alan Lee for timely technical guidance in finalizing his PhD research and for their constant encouragement. We have benefited greatly in this work from the stimulating research environment provided by the WEKA project at the University of Waikato, New Zealand.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. John Wiley & Sons.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, *21*, 243–247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (pp. 267–281). Akadémiai Kiadó, Budapest.

Blake, C., Keogh, E. & Merz, C. J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/∼mlearn/MLRepository.html]. Department of Information and Computer Science, University of California, Irvine.

Blum, J. R. & Susarla, V. (1977). Estimation of a mixing distribution function. *Ann. Probab.*, *5*, 200–209.

Böhning, D., Schlattmann, P. & Lindsay, B. (1992). C.A.MAN (computer assisted analysis of mixtures): Statistical algorithms. *Biometrics*, *48*, 283–303.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, *37(4)*, 373–384.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont CA.

Choi, K. & Bulgren, W. B. (1968). An estimation procedure for mixtures of distributions. *J. R. Statist. Soc. B*, *30*, 444–460.

Deely, J. J. & Kruse, R. L. (1968). Construction of sequences estimating the mixing distribution. *Ann. Math. Statist.*, *39*, 286–288.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, *81*, 425–455.

Foster, D. & George, E. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, *22*, 1947–1975.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

Kullback, S. (1968). *Information Theory and Statistics* (2nd Ed.). New York: Dover. Reprinted in 1978, Gloucester, MA: Peter Smith.

Kullback, S. & Leibler, R. (1951). On information and sufficiency. *Ann. Math. Statist.*, *22*, 79–86.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.*, *73*, 805–811.

Lesperance, M. L. & Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *J. Amer. Statist. Assoc.*, *87*, 120–126.

Macdonald, P. D. M. (1971). Comment on a paper by Choi and Bulgren. *J. R. Statist. Soc. B*, *33*, 326–329.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, *15*, 661–675.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Mateo, Calif.: Morgan Kaufmann Publishers.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

Robbins, H. (1955). An empirical Bayes approach to statistics. In *Proc. Third Berkeley Symposium on Math. Statist. and Prob.*, 1 (pp. 157–164). University of California Press.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.*, *35*, 1–20.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6(2)*, 461–464.

Stuart, A., Ord, K. & Arnold, S. (1999). *Classical Inference and the Linear Model*, Volume 2A of *Kendall's Advanced Theory of Statistics*. Oxford University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, *58*, 267–288.

Tibshirani, R. & Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, Series B*, *61*, 529–546.

Wang, Y. (2000). *A new approach to fitting linear models in high dimensional spaces*. PhD thesis, Department of Computer Science, University of Waikato, New Zealand.