Penalized Logistic Regression in Gene Expression Analysis

Michael G. Schimek
Karl-Franzens-University Graz, Institute for Medical Informatics, Statistics and Documentation, A-8010 Graz, Austria

Abstract. A typical task in gene expression analysis is the classification of biological samples into two alternative categories. A statistical procedure is needed which, based on the expression profiles measured, allows to compute the probability that a new sample belongs to a certain class. The feature of high-dimensionality and small sample sizes makes this statistical task very challenging. Standard logistic regression fails in most instances because of condition problems. A state-of-the-art overview on penalized logistic regression approaches, including the choice of penalty functions and regularization parameters, is given.

Keywords. Akaike information criterion, classification, cross-validation, gene expression analysis, logistic regression, overfitting, penalization, ridge regression.

1 Introduction

With the advance of gene expression techniques, high expectation has been given to better sample classification (such as the classification of disease vs. normal) at molecular levels with microarray data. The feature of high-dimensionality (typically thousands of genes) and small sample sizes (typically less than one hundred cases) makes this statistical task very challenging. The complexity of the diseases, the poor understanding of the underlying biology and the imperfection of the measurements makes the problem even harder.

In gene expression analysis often we have biological samples belonging to either one of two alternative classes. A statistical procedure is needed which, based on the expression profiles measured, allows to compute the probability that a new sample belongs to a certain class. Such a procedure would be logistic regression. The problem is that different from conventional classification tasks there are far more variables than observations. Hence we have to cope with multicollinearity and overfitting (oversmoothing). How can we overcome these obstacles? In that we impose a penalty on large fluctuations of the estimated parameters and on the fitted curves.

Quadratic regularization is known as ridge regression. Other penalties lead to lasso (Tibshirani, 1995) or to bridge regression (Frank and Friedman, 1993). Here we discuss how logistic regression can be adequately penalized.
Further we address the problem of ridge (smoothing) parameter selection in this context.

2 Logistic regression

Logistic regression is a popular linear classification method (for an up-to-date overview see Hastie, Tibshirani and Friedman, 2001, chapter 4). Its predictor function consists of a transformed linear combination of explanatory variables. As stated in the introduction, logistic regression would be an appropriate tool to classify biological samples belonging to one of two alternative classes.

The typical situation is that a number of biological samples has been collected, preprocessed and hybridized to microarrays. It is assumed that \( n_1 \) microarrays belong to the one and \( n_2 \) microarrays belong to the other class (e.g. to patients having the disease vs. not having the disease) and \( n_1 + n_2 = n \). The statistical task is to compute the probability that a specific sample belongs to one of the two alternatives based on the expression profiles recorded. The final goal is the classification of new microarrays.

Why is the application of logistic regression not satisfying in this setting? A basic assumption, typical for standard classification techniques, is not fulfilled: \( m \ll n \), where \( m \) denotes the predictors (i.e. genes in the above example). On the contrary, here we always have \( m \gg n \). For estimation the consequence is that there are many more unknowns than equations and infinitely many solutions exist! Hence we end up with an ill-conditioned problem. A naive application of logistic regression, whatever the algorithm is, would result in unstable (unreliable) estimates.

Another problem in the context of classification is the low discrimination power due to the high variance of the data fit (unbalanced bias and variance).

How can we cope with these detrimental features? The answer is penalizing the logistic regression model. Before we can discuss ways of penalization we have to introduce standard logistic regression.

Let be \( y_i, i = 1, \ldots, m \), the response variable and \( x_{ij}, j = 1, \ldots, n \), the predictor variables. Let us further define the monotone logit transformation

\[
\eta_i = \frac{p_i}{1-p_i},
\]

where \( p_i \) is the probability of observing \( y_i = 1 \). The logit ensures that the probabilities of the two classes (0/1 response \( y_i \)) sum to one and remain in \([0,1]\). Then the logistic regression model is

\[
\eta_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij},
\]

where \( \beta_0 \) denotes the offset. \( \eta \) is the linear predictor. The connection between \( \eta \) and \( p \) for the binomial family of the generalized linear models is the
canonical link which is non-linear. The inverse relationship is
\[ p = \frac{1}{1 + e^{-\eta}}, \]
the logistic curve. The log-likelihood can be written
\[
\ell(\beta) = \sum_{i=1}^{m} \{y_i \log p(x_i; \beta) + (1 - y_i) \log(1 - p(x_i; \beta))\} = \sum_{i=1}^{m} \{y_i \beta^T x_i + \log(1 + e^{\beta^T x_i})\}.
\]
(1)
Here we have assumed \( \beta = \{\beta_0, \beta_j\} \) and that the vector of predictor values \( x_i \) includes the constant term 1 to accommodate the offset.
To maximize the log-likelihood, we set its derivatives to zero. The score equations are
\[
\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{m} x_i (y_i - p(x_i; \beta)) = 0.
\]
Their number is \( p + 1 \) and they are non-linear in \( \beta \). Since the first element of \( x_i \) is one, the first score equation specifies that
\[
\sum_{i=1}^{m} y_i = \sum_{i=1}^{m} p(x_i; \beta).
\]
This is to say that the mean of \( p \) has to be equal to the fraction of ones in \( y \).
The score equations can be solved via the Newton-Raphson algorithm. Typically the algorithm converges since the log-likelihood is concave (for details see Hastie, Tibshirani and Friedman, 2001, p.98f).

3 Shrinkage for ill-conditioned problems
For \( m \gg n \) an often used approach is subset selection. Only a subset of the predictors (would be the genes in our example) is retained and the rest discarded. The thus obtained model can be expected to have lower prediction error than the full model. However, for thousands of genes (the usual number in current gene expression analysis) compared to less than one hundred observations, this strategy cannot work. Even in more favorable situations there is the problem of high variance due to the discrete nature of the decisions during the process of selection.
A more continuous alternative are shrinkage methods. The key idea is that overfitting is avoided by imposing a penalty on large fluctuations of the estimated parameters respectively on the fitted function (the latter becomes more smooth). The oldest statistical tool is ridge regression. Hoerl and Kennard (1970) introduced it for ill-conditioned ordinary regression problems.
Ridge regression shrinks the regression coefficients by imposing a penalty on their size. A complexity parameter $\lambda$ ($\lambda \geq 0$) controls the amount of shrinkage. The (quadratic) regularization put forward by Hoerl and Kennard (1970) is the sum of squares of the regression coefficients. Since that it has been applied in numerous statistical techniques such as linear discriminant analysis (Friedman, 1989), neural networks (Girosi, Jones and Poggio, 1995) and logistic regression (leCessie and van Houwelingen, 1992). Regularization is closely related to certain spline concepts in nonparametric regression (Schimek, 1988, Wahba, 1990). In smoothing splines we are talking about penalty functions controlled by a smoothing parameter similar to the above complexity parameter.

4 Penalized logistic regression

As pointed out already, the application of standard logistic regression is not really promising. What we need is a reproducible (stable) classifier. It can be obtained by means of penalization (le Cessie and van Houwelingen, 1990).

The general motivation behind penalizing the likelihood – the approach logistic regression can be regularized in a straightforward manner – is the following: Avoid arbitrary coefficient estimates $\hat{\beta}$ and a classification that appears to be perfect in the training set but is poor in the validation set. Penalization aims at an improved predictive performance in a new data set by balancing the fit to the data and the stability of the estimates.

Let us now apply quadratic regularization to logistic regression. The log-likelihood $\ell(\beta)$ in equation (1) can be penalized in the following way:

$$\ell^*(\beta) = \ell(\beta) - \frac{\lambda}{2} J(\beta)$$  

(2)

where

$$J(\beta) = \|\beta\|^2 = \sum_{j=1}^{n} \beta_j^2$$

is a quadratic (ridge) penalty. As can be seen, only the regression coefficients $\beta_j$ are subject to penalization, not the offset $\beta_0$. The complexity (ridge) parameter $\lambda$ controls the size of the coefficients (increasing $\lambda$ values decrease their size).

There are also other penalties in the literature. In general a penalty function can be written

$$J(\beta) = \sum_k \gamma_k \psi(\beta_k),$$  

(3)

where $\gamma_k > 0$. The $L_1$ penalty $\psi(\beta_k) = |\beta_k|$ results in lasso, originally introduced by Donoho and Johnson (1994) and later extended by Tibshirani (1996) in the context of generalized least squares. The $L_q$ penalty, $0 \leq q \leq 1$, $\psi(\beta_k) = |\beta_k|^q$ is used in bridge regression (Frank and Friedman, 1993). For wavelets Antoniadis and Fan (2001) give some evidence on how to choose
a penalty function for (3). Usually, the penalty $\psi$ is chosen to be symmetric and increasing on $[0, +\infty)$. Furthermore, $\psi$ can be convex or non-convex, smooth or non-smooth. In a wavelet setting, Antoniadis and Fan (2001) provide some insight how to select an appropriate penalty function. Criteria for such a function are unbiasedness and stability of the results. More research, specifically for regression techniques, is required.

To maximize the penalized log-likelihood in (2), we set its derivatives to zero,

$$\frac{\partial \ell^*(\beta)}{\partial \beta_0} = 0,$$

$$\frac{\partial \ell^*(\beta)}{\partial \beta_j} = 0.$$

We obtain the penalized likelihood equations

$$u^T(y - p) = 0 \quad (4)$$

and

$$X^T(y - p) = \lambda \beta. \quad (5)$$

The $m$-dimensional vector $u$ consists of ones. As in standard logistic regression the equations are non-linear. A first-order Taylor expansion gives (we abbreviate $p_i = p(x_i; \beta)$)

$$p_i = \tilde{p}_i + \frac{\partial p_i}{\partial \beta_0} (\beta_0 - \tilde{\beta}_0) + \sum_{j=1}^{n} \frac{\partial p_i}{\partial \beta_j} (\beta_j - \tilde{\beta}_j).$$

Tilde denotes the approximation (e.g. $\tilde{p}_i$ for $p_i$). The partial derivatives are

$$\frac{\partial p_i}{\partial \beta_0} = p_i (1 - p_i)$$

and

$$\frac{\partial p_i}{\partial \beta_j} = p_i (1 - p_i) x_{ij}.$$ 

In writing $p_i (1 - p_i) = w_i$ we finally have

$$u^T \tilde{W} u \beta_0 + u^T \tilde{W} X \beta = u^T (y - \tilde{p} - \tilde{W} \tilde{\eta}) \quad (6)$$

and

$$X^T \tilde{W} u \beta_0 + (X^T \tilde{W} X + \lambda I) \beta = X^T (y - \tilde{p} - \tilde{W} \tilde{\eta}), \quad (7)$$

where $W$ is a diagonal matrix consisting of the elements $w_i$. As pointed out for standard logistic regression, the parameter $\beta_0$ is determined by the fraction of ones in $y$.

The above linearized system can be solved by means of iterative techniques. Appropriate starting values are $\tilde{\beta}_0 = \log(\bar{y}/(1 - \bar{y}))$, with $\bar{y} = \sum_{i=1}^{m} y_i/m$, and $\tilde{\beta} = 0$. 
Under the requirements of gene expression analysis the system of equations in (5) and (6) is huge. The statistical language R (Ihaka and Gentleman, 1996) offers singular value decomposition. It allows to economically evaluate penalized logistic regression. The data handling can be simplified when functions from Bioconductor, an R-based open source project for the analysis and comprehension of genomic data, are adopted. An algorithmic alternative to singular value decomposition based on (4) and (5), not yet implemented, was recently proposed by Eilers et al. (2002).

5 Choice of ridge parameter

In penalized logistic regression we have the same situation as in smoothing regression: The choice of the regularization (resp. smoothing) parameter is crucial (for an introduction see van der Linde, 2000). Extensive research has been done into selection techniques for nonparametric regression. There are two dominant data-driven techniques which also find application in penalized logistic regression. The one is based on Cross-Validation (CV) and the other on the Akaike Information Criterion (AIC). However, the experience from smoothing techniques should not be transferred one-to-one to ridge problems. There is certainly a lack of empirical evidence (simulation studies) for the latter.

A very popular and successful idea (Craven and Wahba, 1979) in determining $\lambda$ is that of CV, that is predicting an observed value from the remaining observations and choosing the $\lambda$ which yields the best predictions. If $\hat{y}_\lambda(x_i)$ denotes the prediction of $y(x_i)$ when $y(x_i)$ is removed from the data the simple leave-one-out CV criterion is

$$CV(\lambda) = m^{-1} \sum_{i=1}^{m} (y(x_i) - \hat{y}_\lambda(x_i))^2.$$ 

For $m$ data points (in genetic research a very large number) the computational demand is proportional to $m(m-1)$, hence high.

Criteria for the performance of CV are either the misclassification rate or the strength of prediction. Van Houwelingen and le Cessie (1990) differentiate between the Brier score $\sum_{i}(y_i - \hat{p}_{-i})^2$ and the log-likelihood prediction $\sum_{i}(y_i \log \hat{p}_{-i} + (1 - y_i) \log(1 - \hat{p}_{-i}))$.

The AIC is based on a log-likelihood loss function. According to Akaike (1974) the predictive likelihood of a model can be estimated by subtracting the number of its parameters from the maximum-likelihood. The criterion

$$AIC(\lambda) = -2\ell^*(\beta) + 2edf$$

gives the right balance between model complexity and fidelity to the data. $\ell^*(\beta)$ is the log-likelihood as introduced earlier on in (2). In contrast to standard (i.e. no penalization) regression models the effective degrees of freedom of the model $edf$ is not the number of elements in the parameter vector $\beta$. 
Hastie, Tibshirani and Friedman (2001, p.59ff) describe how to calculate the effective degrees of freedom for ridge regression when singular value decomposition is applied. Their formula can be generalized to penalized logistic regression. Similar ideas are also behind a MATLAB implementation (Eilers et al., 2002) and an R implementation (Antoniadis, 2003) of this model.

6 Concluding remarks

In the previous sections we have seen how to cope with the severe problem of far more variables (genes) than observations. Standard regression approaches (parametric or nonparametric) are prone to overfit (oversmooth) the data. Regularization techniques help us to avoid such inadequate results. Penalized logistic regression does even allow statistical classification in the way we are used to from standard procedures. However there is always a price to be paid.

From the frequentist point of view penalization adds some bias to the estimates $\beta$. The general hope is that this bias can be neglected. Obviously it depends on the value of the ridge parameter $\lambda$. The optimal choice via CV or AIC is crucial. Yet we have no mathematical theory how the standard errors for estimators with an unknown bias are influenced, hence test statistics and confidence intervals. Some insight into penalization can be obtained from a Bayesian position. A penalty can be interpreted as the logarithm of the prior density of the parameters.

From a Bayesian point of view the penalized likelihood estimator is the posterior mode for a prior consisting of the following components: a flat prior $\beta_0$ and the $\beta_s$ independently $N(0, \tau^2)$-distributed where $\tau^2 = \frac{1}{\lambda}$. As can be seen the ridge parameter is the inverse of the variance of the prior. For AIC this means estimation of the unknown variance. Looking at the penalization problem this way it becomes clear why Markov chain Monte Carlo techniques - computationally substantially more demanding compared to the approach outlined here - can be applied as well. The benefit is a better handling of model uncertainties.

The approximate prior of $\beta$ is finally

$$\beta \sim N(\hat{\beta}, (Z^T W Z + \lambda P)^{-1}),$$

where $Z = [u|X]$ and $W$ from equation (5) and (6). $P$ is a $(n+1) \times (n+1)$ identity matrix with the element $p_{11} = 0$ (i.e. no penalty on $\beta_0$). Such a Bayes concept is very close to penalized likelihood with prespecified $\tau^2$ respectively $\lambda$. Again this is reminiscent of smoothing splines that can be easily interpreted in a Bayesian manner (see van der Linde, 1993 and 2000).

In practice, taking the frequentist approach as described in this paper is certainly sufficient as long as the overall goal is a list of regression coefficients ordered according to size, indicating the specific contribution of each gene.
References