Advanced Data Analysis (FINAL)

Kevin Eng

Spring 2014

document created on: May 8, 2014
created from file: `FINAL.tex`
cover page automatically created with `CoverPage.sty`
(available at your favourite CTAN mirror)

Introduction

This report investigates the relationship between 14 types of cancer and 16063 gene expression measurements. There are a total of 198 observations. The observations are broken down into two sets. One set contains 144 observations and is used as the training set. The left over 54 observations are used as the testing set. The models used in this report classify the 14 types as follows: 1 - breast, 2 - prostate, 3 - lung, 4 - collerectal, 5 - lymphoma, 6 - bladder, 7 - melanoma, 8 - uterus, 9 - leukemia, 10 - renal, 11 - pancreas, 12 - ovary, 13 - meso, and 14 - cns. The proportion of each cancer type in the full data set differ significantly. The least common cancer type recorded was melanoma which accounted for 5.06% of observations. On the other hand the most common type of cancer recorded was leukemia which accounted for 15.2% of all observations.

Achieving a high classification rate is desirable. Current cancer treatment is specialized. Different procedures are used to treat different types of cancer. Correctly predicting cancer type will give doctors crucial time to devise effective treatment. Indeed, cancer is like a virus; it can spread through out the body. Thus, early treatment is of paramount importance.

Variable Screening

A linear regression of indicators model was examined first. However, in order to use this model the 16063 genes had to be filtered down to 144 genes or less. The genes were selected based on a metric called *between-class gene variation*. Let K be the total number of classes and n be the number of observations. The between-class gene variation of gene j, V_j, is defined to be

$$V_j = \sum_{k=1}^{K} n_k \cdot (\overline{x}_{kj} - \overline{x}_j)^2.$$

Here, n_k is the number of patients in class k; $\overline{x}_{kj}$ is the average in class k along gene j; and $\overline{x}_j$ is the overall average along gene j. More specifically, $\overline{x}_{kj}$ and $\overline{x}_j$ are defined as

$$\overline{x}_{kj} = \frac{1}{n_k} \sum_{i:y_i=k} x_{ij} \quad , \overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \quad .$$

Figure (1) on the next page depicts a plot of the 50 highest between-class gene variation scores.

There were four variables which were perfectly correlated. These are the points in figure (1) which are at exactly the same height. Removing one of each of the perfectly correlated variables generated a further reduced set of 46 variables. Of the 46 left over genes, the 4390th and 4044th genes had the highest absolute correlation standing at .9756. These two genes correspond to the 20th and 14th highest scores respectively.

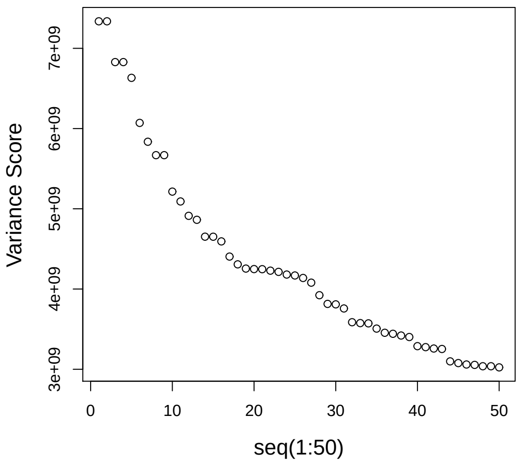

Figure 1: 50 highest between-class gene variation scores

Linear Regression of Indicators

A linear regression of indicators model was constructed using the screened pool of 46 genes determined in the previous section. This model is constructed as follows. First define the class labels $\tilde{y}_1^{(k)}, \ldots, \tilde{y}_n^{(k)}$ as

$$\tilde{y}_i^{(k)} = \begin{cases} 1 & \text{if } y_i = k \\ 0 & \text{otherwise} \end{cases}$$

For each k an ordinary least squares linear regression model is fit on the screened set of 46 predictors. Donate the coefficient of the least square equation associated with k as $\hat{\beta}^{(k)}$. Each β contains 47 components: 46 predictor coefficients and 1 intercept. Since there are 14 classes, there are 14 β's. Given a single patient with gene measurements x_0, the predicted classification, $\hat{y}_0$, using linear regression of indicators is

$$\hat{y}_0 = \underset{k=1,\ldots,K}{\operatorname{argmax}} \{x_0^T \hat{\beta}^{(k)}\}$$

This model generated 19 misclassifications on the entire training set. This corresponds to about a 13.2% misclassification rate. Not all of the 16,063 genes could be used because otherwise we would have more predictors than observations. This would lead to an under determined system in which the β coefficients are not well defined (i.e there are multiple coefficients satisfying the least squares criterion).

Using four fold cross-validation on the training set, the average number of misclassification was 16. This is slightly less than the 19 misclassification for the entire training set. However the average cross-validation misclassification rate was about 44.4%. A significantly higher rate than the 13.2% generated on the entire training set.

Lasso Multinomial Regression

In addition to using a linear regression of indicators model, a lasso based multinomial logistic regression model was also fitted. The multinomial model was generated using the `cv.glmnet` function in R. Note `cv.glmnet` uses the symmetric multinomial model which looks at

$$\log\left(\frac{P(Y=j|X)}{P(Y\neq j|X)}\right)$$

for $j \in \{1, 2 \ldots, K\}$. This differs slightly from other multinomial models which look at

$$\log\left(\frac{P(Y=j|X)}{P(Y=K|X)}\right)$$

In particular the symmetric multinomial model generates K many β coefficients unlike the $K-1$ many β coefficients generated using the later multinomial model.

Figure (2) contains two graphs of the CV error curve using two different metrics. The left hand graph uses multinomial deviation. The right hand graph uses the misclassification rate. The points on both graphs are evaluated at the same λ points.

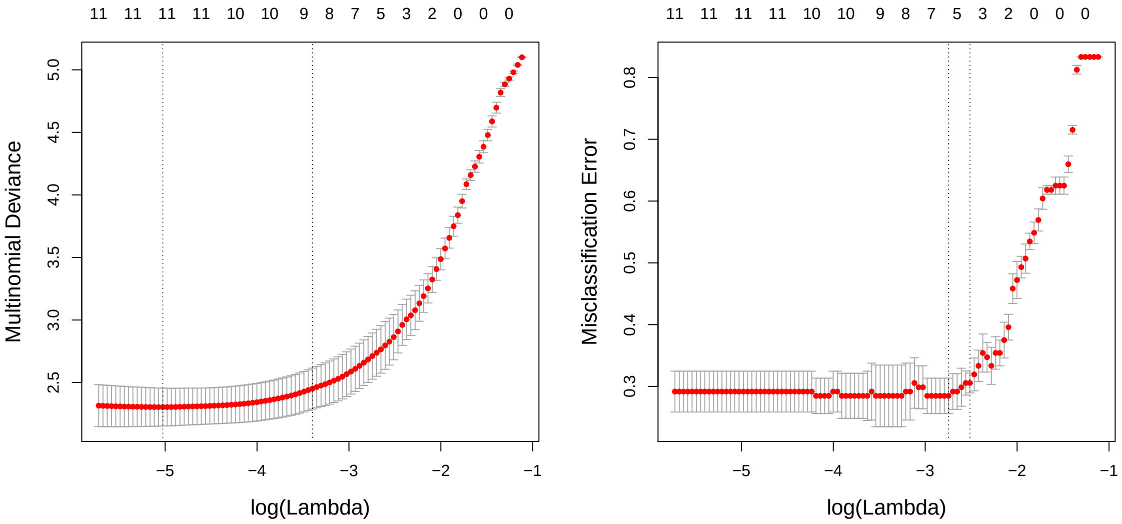

Figure 2: Two CV error curves

The λ_{min} value selected by the usual rule was 6.28×10^{-3}. The λ_{1se} value selected using the 1 standard error rule was 2.65×10^{-2}. The misclassification rate at λ_{min} and λ_{1se} was .271 and .264 respectively. These rates are significantly lower than the rates seen using the indicator model.

Using λ_{min} , the multinomial model had 155 different gene variables among its β coefficients. For λ_{1se} there were 137 different gene variables among its β coefficients. Both of these are significant reductions on the total number of variables. Indeed both of these represent less than 1% of the original 16063 genes. This is quite nice for a cancer researcher since this means it may be possible to predict cancer type using only a handful of gene indicators. Particularly, it lends well to test reliability. The less genes that require measurement, the less room there is for measurement error. Also the less a genes a model uses the easier and less expensive it is to conduct a test. Although this may soon become a moot point at the rate the cost of gene sequencing is decreasing.

The multinomial model and the indicator model do not share many genes in common. The set of 137 different gene selected using the one standard error rule shares three genes in common with the 46 screened genes used in the indicator model. One may be inclined to say the variable screening was ineffective because the lasso model yielded significantly better misclassification rates using different genes. However this is not necessarily the case.

The lasso method is computationally expensive; the model is calculated at 100 different λ values. Additionally, a four fold cross-validation is run at each λ value. A super computer may be needed if this study is extended to include hundreds of thousands of observations. On the other hand the variable screening procedure is much less computationally demanding. Also the indicator model is much more simple. The lasso model contains almost three times as many variables as the indicator model. Moreover, the misclassification rate seen using the indicator model is much lower than randomly guessing. Overall, this suggest variables screening is an effective tool in generating reasonable models using relatively small amounts of computing power.

Test Set Evaluations

Out of the 54 observations in the test set, the indicator model misclassified 30 and the lasso model misclassified 20. The corresponding misclassification rates, were 55.6% and 37.0% respectively. This suggests lasso method has better predictive capabilities. The misclassification rates are significantly different from the cross-validation misclassification rates (55.6% vs. 44.4% and 37.0% vs. 26.4%). However they are also not terribly different.

The difference between the cross-validation error rate and test error rate can be attributed to two factors. Firstly, the cross-validation methodology was less than desirable. Using four folds isn't a lot and leaves room for more variability compared to, for example, using 10 folds. Unfortunately, increasing the number of folds comes with a cost; adding more folds requires more computational resources since the model must be trained on more sets. Lastly, the number of observations used was small which, by nature, introduces more noise into the models. Also, it goes without saying, there is always the very real possibility both models are just poor representations of what was observed.

Discussion

Overall both models provide hope for improved cancer prediction. The most common type of cancer out of the 198 observations is leukemia which accounts for 30 of the cases. Assuming the data set is a good representation of the true population, we can expect about 15.2% of all cancer patients of the 14 types to have leukemia. So even if the test misclassification rate for the indicator model (55.6%) were true, this is still significantly better than just guessing the most common type of cancer.

Use of bootstrapping or a future follow up study are two possible continuations of this report. Since the data set used in this report is small, the use of bootstrapping could ameliorate this problem. However nothing is as good as getting real data. With the decreasing cost of gene sequencing, and the proliferation of electronic records, more data will surely become available.

Besides the problem of having a small sample, the fact that the lasso and indicator models did not share many genes in common is concerning. This indicates perhaps neither model is appropriate for cancer prediction. Assuming one can predict cancer type from genes, a “correct” model should be able to discern the important genes from the unimportant. That a “correct” model is accurate is unimportant. What matters is that a good model knows what genes can be used to predict cancer type. Since there is a disagreement between which genes to use between the two models, it suggests one or both models are having a hard time discerning the important genes.