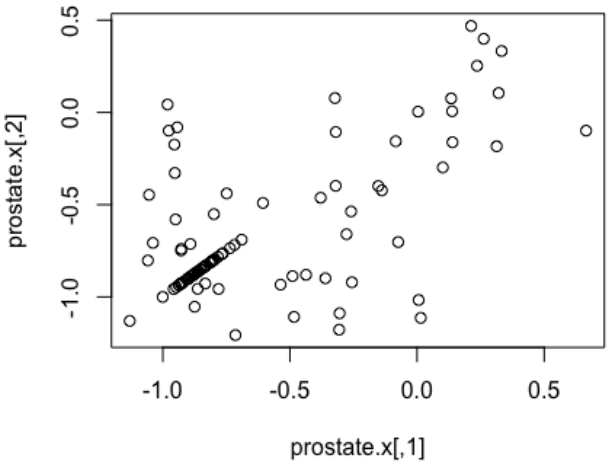
Program: AARMS 2018 Summer School in Data Analytics
Course: Statistical Learning for High Dimensional Data
Instructor: Dr. Wenqing He
Author: Hamza Qureshi
Project Title: High Dimensional Statistics with Prostate Gene Expression Data

Course Project Report

Microarray experiments are expected to contribute significantly to progress in cancer treatment by enabling a precise and early diagnosis. They create a need for class prediction tools that can deal with a large number of highly correlated input variables, perform feature selection, and provide class probability estimates that serve as a quantification of the predictive uncertainty. This project explores some of the techniques related to high dimensional statistics namely SIS, regression, K means and hierarchical clustering and SVM in order to achieve just that goal.

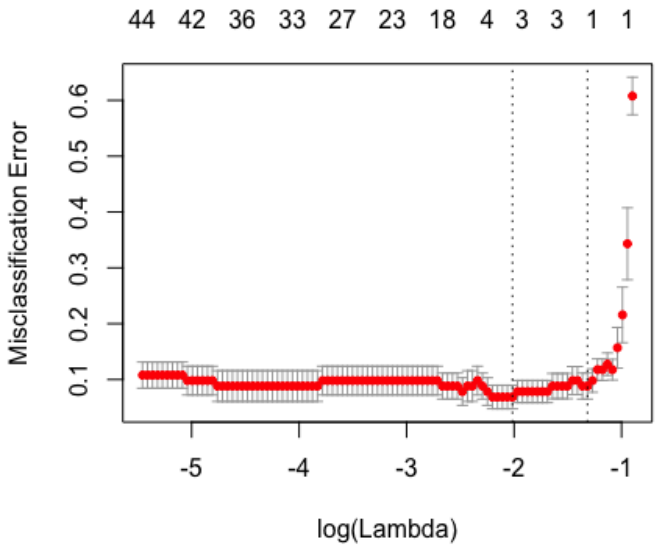
The project utilizes prostate gene expression data for the purposes of performing different forms of analyses. The data was made available to public by Marcel Dettling and can be accessed here. The paper associated with the data is listed as [1]. The mentioned high dimensional data comprises the expression of 52 prostate tumors and 50 non-tumor prostate samples, obtained using the Affymetrix technology. The data used is normalized and thresholded as described in [2]. Genes whose expression varied less than fivefold relatively or less than 500 units absolutely, between the samples, were excluded, leaving 6,033 of them only. Finally, each experiment was applied a base 10 logarithmic transformation, and standardized to zero mean and unit variance.

The data is binary and is shown in the figure below.

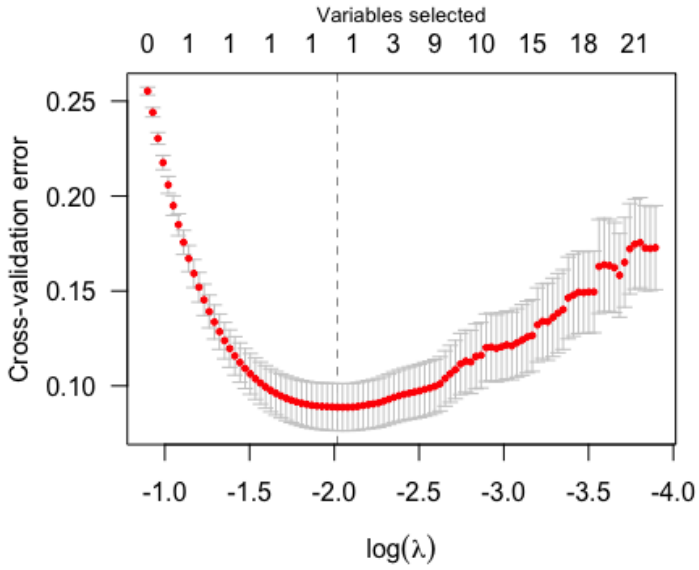


While running sure independence screening, the algorithm narrowed the variables down to 2619 genes in one instance and 5016 in the other.

For glmnet with lasso and adaptive lasso, misclassification error was calculated, and it was determined that the minimum value of lambda be 0.1333006 while lambda LSE value be 0.2678319. Here is the relevant graph.



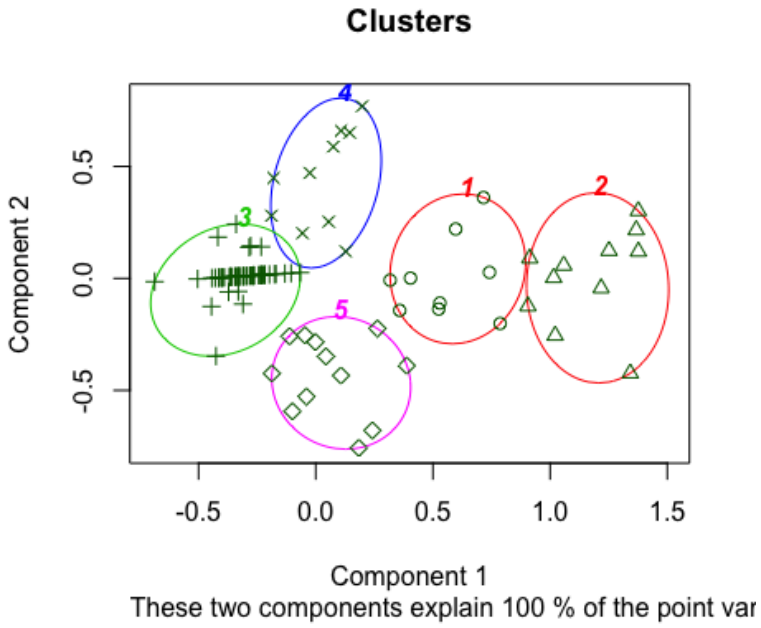
As far as ncvreg package goes, cross validation was calculated, and it turned out the minimum lambda is 0.1328731. Here are the results.



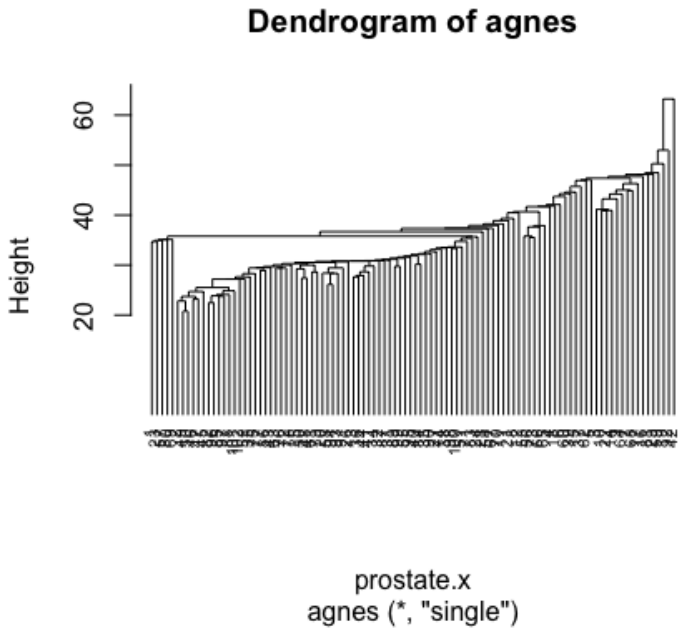
When including the SCAD penalty in logistic regression, while using the package ncvreg, it was observed that the coefficients would go to zero faster for the same values of lambda, when

alpha was increased from 0.1 to 1. The opposite is true when the value of a is increased to 10. The coefficients also converged to zero faster when MCP (with gamma=3) was used as opposed to SCAD.

The next classification method used was K means clustering. Elbow method, silhouette method and gap statistic were implemented to determine the optimal number of clusters. All methods gave different answers and therefore, gap statistic was chosen, with k=5 as the answer, as it is more reliable than others. The elbow and silhouette methods measure a global clustering characteristic only. Here are the clusters.



Thereafter, hierarchical clustering was also tried. The 'single' method with agnes package returned the highest agglomerative coefficient, and therefore, was selected. Here is the relevant dendrogram. It is difficult to see in this diagram how many clusters would be optimal, and so the two clustering methods, unfortunately, cannot be compared.



Last but not the least, SVM was implemented. It was seen that the square root mean error between the actual data and the prediction is 0.1425545.

SIS does a good job identifying the 'suitable' variables. The same can said of regression. Both glmnet and ncvreg oriented analyses were done and it was found out that the error is in fact lower with ncvreg package. That is when SCAD penalty is used. When it comes to clustering, hierarchical one beats the k means in principle as the bottom-up or top-bottom technique in hclust gives a result that is better off because here, one doesn't have to select k by trial and error. This holds regardless of the fact that the two clustering techniques could not be compared. SVM was implemented successfully as well. As the data is high dimensional, the graph cannot be shown. On the bright side, the error with SVM is also pretty low.

Since SVM error is the lowest, it looks like this one is the best technique among the ones implemented, with an error of only 0.1425545.

References

1. Dettling M, *Bioinformatics* (2004), Vol. 20, No. 18, p. 3583-3593.

2. Singh D, Febbo P, Ross K, Jackson D, Manola J, Ladd C, Tamayo P, Renshaw A, D'Amico A, Richie J, *et al*.: **Gene expression correlates of clinical prostate cancer behavior**. *Cancer Cell* 2002, **I**:203-209.