STAT3006 Assignment 4, 2020

High-Dimensional Analysis

Weighting: 25%

Due: Monday 16/11/2020

This assignment involves the analysis of a high-dimensional dataset. Here we focus on an early microarray dataset analysed by Alon *et al* (1999), which involves predicting whether a given tissue sample is cancerous or not, as well as trying to determine which genes are expressed differently between the two classes.

You should select one classifier for the task of classification, which you have not used in previous assignments. Probability-based classifiers discussed in this course include linear, quadratic, mixture and kernel density discriminant analysis. Non-probability-based classifiers discussed include k nearest neighbours, neural networks, support vector machines and classification trees. All of these are implemented via various packages available in R. If you wish to use a different method, please check with the coordinator. In addition, you will make use of lasso-penalised logistic regression.

It will quickly become apparent that the number of observations is less than the number of variables, and so some form of dimensionality reduction is needed for most forms of probability-based classifier and can be used if desired with the non-probability-based classifiers.

The Alon dataset contains measurements of the expression levels of 2000 genes from 62 colon tissue samples, 40 of which are labelled as being from tumours (cancers) and 22 normal (non-cancerous). Here we consider analysis of this data to (i) develop a model which is capable of accurately predicting the class (cancer or normal) of new observations, without the need for examination by clinicians (ii) determine which genes are expressed differently between the two groups. Discriminant analysis/supervised classification can be applied to solve (i), and in combination with feature (predictor) selection, can be used to provide a limited solution to (ii) also. You are encouraged to use R for the assignment. Python is also ok, but the course staff are less familiar with it.

Tasks:

1. (4 marks) Perform principal component analysis of the Alon dataset and report and comment on the results. Detailed results should be submitted via a separate file, including what each principal component direction is composed of in terms of the original explanatory variables, with some explanation in the main report about what is in the file. Give a plot which shows the individual and cumulative proportions of variance explained by each component. Also produce and include another plot about the principal

components which you think is of interest, along with some explanation and discussion. The R package FactoMineR is a good option for PCA.

2. (4 marks) Perform single variable analysis of the Alon dataset, looking for a relationship with the response variable (the class). Use the Benjamini-Hochberg (1995) approach to control the false discovery rate to be at most 0.01. Report which genes are then declared significant along with the resulting threshold in the original p-values. Also give a plot of gene order by p-value versus unadjusted p-value (or the log of these), along with a line indicating the FDR control.

Within the stats package is the function p.adjust, which offers this method. More advanced implementations include the fdrame package in Bioconductor.

3. (3 marks) Define binary logistic regression with a lasso penalty mathematically, including the function to be optimised.

4. (3 marks) Explain the potential benefits and drawbacks of using PCA to reduce the dimensionality of the data before attempting to fit a classifier. Explain why you have chosen to reduce the dimensionality or not to do so for this purpose.

5. Apply each classification method (your choice and lasso logistic regression) using R to the Alon dataset, report the results and interpret them.

For lasso logistic regression, I suggest you use the glmnet package, available in CRAN, and make use of the function cv.glmnet and the family="binomial" option. If you are interested, there is a recording of Trevor Hastie giving a tutorial on the lasso and glmnet at http://www.youtube.com/watch?v=BU2gjoLPfDc .

Results should include the following:

a) (1 mark) characterisation of each class: parameter estimates.

b) (2 marks) cross-validation (CV)-based estimates of the overall and class-specific error rates: obtained by training the classifier on a large fraction of the whole dataset and then applying it to the remaining data and checking error rates. You may use 5-fold, 10-fold or leave-one-out cross-validation to estimate performance.

c) (2 marks) For lasso logistic regression, you will need to use cross-validation to estimate of the optimal value of λ. Explain how you plan to search over possible values. Then produce and explain a graph of your cost function versus λ. You should also produce a list of the genes included as predictor variables in the optimal classifier, along with their estimated coefficients.

6. (5 marks) Compare the results from all approaches to analysis of the Alon dataset (PCA, single-variable analysis and the two classifiers). Explain what each approach seems to offer, including consideration of these results as an example.

Notes:

(i) R commands you might find useful:

objects() – gives the current list of objects in memory.

attributes(x) – gives the set of attributes of an object x.

(ii) How to open the Alon data file in R:

File → Load Workspace → Files of type: All files (*.*), select **alon.rda > Open**. This should load the Alon data into memory as x, xm and y. x is the predictor variable data for 62 subjects, y lists the labels (classes). You can ignore xm, but there are some clues on what these values are in the Alon *et al*. paper.

(iii) Please put all the R commands in a separate file or files and submit these separately via a single text file or a zip file. You should not give any R commands in your main report and should not include any raw output – i.e. just include figures from R (each with a title, axis labels and caption below) and put any relevant numerical output in a table or within the text.

(iv) As per http://www.uq.edu.au/myadvisor/academic-integrity-and-plagiarism, what you submit should be your own work. Even where working from sources, you should endeavour to write in your own words. Equations are either correct or not, but you should use consistent notation throughout your assignment and define all of it.

(v) Please include your name or student number in the filename of any files submitted. You should submit your assignment via Blackboard.

(vi) Some references:

R

Maindonald, J. and Braun, J. *Data Analysis and Graphics Using R - An Example-Based Approach*, 3rd edition, Cambridge University Press, 2010.

Venables, W.N. and Ripley, B.D., *Modern Applied Statistics with S*, Fourth Edition, Springer, 2002.

Wickham, H. and Grolemund, G. *R for Data Science*, O'Reilly, 2017.

High-dimensional Analysis

Bishop, C. *Pattern Recognition & Machine Learning*, Springer, 2006.

Buhlmann, P. and van de Geer, S. *Statistics for High-Dimensional Data*, Springer, 2011.

Efron, B. and Hastie, T. *Computer Age Statistical Inference*, Cambridge University Press, 2016.

Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition, Springer, 2009.

Hastie, T., Tibshirani, R. and Wainwright, M. *Statistical Learning with Sparsity*, CRC Press, 2015.

Other references

Alon, U. *et al*. Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12), 6745–6750, 1999.

Lazar, C. *et al*. A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9, 1106-1119, 2012.

Note: Lazar *et al*. is just an example overview of the range of techniques used in this field. It is also worth noting that microarray experiments have largely been superseded by more recent technology such as RNA-Seq. However, the methods of analysis are similar.