STAT3006/7305 Assignment 4, 2023

High-Dimensional Analysis

Weighting: 20%

Due: Monday 13/11/2023

A person can be in the process of developing breast cancer, but not show clear signs of this, even after mammography (production of x-ray images of the breast). Sharma *et al*. (2005) and Aaroe *et al*. (2010) wished to determine whether gene expression profiles from peripheral blood cells (a blood sample) could be used to predict whether or not a person has breast cancer. They and other researchers were also interested in the types of changes in gene expression that occur during the development of breast cancer.

In both studies, blood was drawn from a set of women who had a suspect initial mammogram, but not yet had a diagnosis of whether the abnormality observed was benign (currently harmless) or malignant (cancerous). Aaroe et al. followed on from the work by Sharma et al. with a larger number of patients and a much larger set of genes. Both datasets were made public, with the Aaroe dataset now available from https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE16443 and also from Blackboard.

Each patient's mammography results were assessed further by clinicians and a diagnosis made. The patient condition labels are stored in Aaroelabels.csv : normal or cancer. The batch-normalised, logged and otherwise processed gene expression data is stored in Aaroe.csv. This processed dataset contains gene expression data derived from blood samples from 121 women, processed with microarrays to record values for 11217 probes, most of which represent individual genes.

Your tasks with the dataset are focused on classification of a sample as coming from a patient with breast cancer or without, and identification of genes of potential interest.

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 lecturer. In addition, you will make use of lasso-penalised logistic regression. Note that you cannot choose another form of logistic regression as your other classifier.

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.

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 based on a blood sample, without the need for a mammogram or its examination by clinicians

(ii) determine which genes are expressed differently between the two groups, individually, or as part of a combination.

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. Other methods such as single-variable analysis can also be applied to attempt to answer (ii). You should use R (recommended) or Python for the assignment.

Tasks:

1. (5 marks) Following this, perform principal component analysis of the gene expression 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 (transformed) original explanatory variables, with some explanation in the main report about what is in the file. Give a plot or plots 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 would be of interest to clinicians and scientists such as Aaroe *et. al*, 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 dataset, looking for a relationship with the response variable (the class). Use the Benjamini-Hochberg (1995) or Benjamini-Yekutieli (2001) approach to control the false discovery rate to be at most 0.1. Explain the assumptions of this approach and whether or not these are likely to be met by this dataset, along with possible consequences of any violations. Also explain how the method works mathematically, but leave out why (i.e. give something equivalent to pseudocode). 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 and briefly introduce a method than can be used to optimise it. Note that this might require a little research.

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) to the dataset using R or Python, report the results and interpret them.

For lasso logistic regression in R, 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 . There are other options in Python including in scikit-learn.

Results should include the following:

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

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 K-fold cv with K ≥ 5 or leave-one-out cross-validation to estimate performance. Additionally report the overall apparent error rates (when trained on all the data and applied back to it).

c) (3 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 ordered by importance of the genes included as predictor variables in the optimal classifier, along with their estimated coefficients.

For your other classifier, also determine an ordered list of the most important genes, stopping at 50, or earlier if justified. For each classifier, comment on any differences between the apparent and CV-derived overall error rates.

6. (4 marks) Compare the results from all approaches to analysis of the dataset (PCA, single-variable analysis and the two classifiers). Explain what each approach seems to offer, including consideration of these results as an example. In particular, if you had to suggest 10 genes for the biologists to study further for possible links to this form of cancer, which ones would you prioritise, and what makes you think they are worth studying further?

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

(iii) 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.

(iv) Please name your files something like student_number_STAT3006_A4.pdf to assist with marking. You should submit your assignment via two links on Blackboard: one for your pdf report and the other for your .zip file containing code (readable via text editor) and any ancillary files.

(v) 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.

McLachlan, G.J., Do, K.-A. and Ambroise, C. *Analyzing Microarray Gene Expression Data*, Wiley, 2004.

Other references

Aaroe, J. *et al*. Gene expression profiling of peripheral blood cells for early detection of breast cancer, *Breast Cancer Research*, 12:R7, 1-11, 2010.

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.

Sharma, P. *et al*., Early detection of breast cancer based on gene-expression patterns in peripheral blood cells, *Breast Cancer Research*, 7(5), R634-644, 2005.

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.