Second Assignment

Statistical Learning, 2009

Niels Richard Hansen
June 2, 2009

Formalities

This is the second and final compulsory assignment for the course Statistical Learning. This is a real data analysis problem, where you are asked to carry out the analysis using the tools and techniques from the course and hand in a report documenting the steps you have taken in the analysis.

The deadline for handing in the solution to the assignment is July 1 2009. You hand it in by sending an email to me with the solution as a single pdf-file attachment. The attached file must be named `yourfullname.pdf`.

The report must contain

- An introduction section describing the data, perhaps some summaries of the data and the objective of your analysis.
- A section where you describe the two strategies (see below) for the construction of a predictor.
- Two sections, one for each strategy, where you describe in details how you have build up the models/methods used, how you have done model selection, etc. Use graphics whenever appropriate.
- A concluding section where you describe a single model as your final choice and report the prediction of 35 test cases (see below).

Data

The data can be found from the course homepage. The data are given as binary R data in a file called `Assignment2.RData`. Download the file and load it into R using the command

```
load("Assignment2.RData")
```

Then you will have five dataframes in your R session called `EpiXTrain`, `EpiPhenoTrain`, `EpiYTrain`, `EpiXTest`, `EpiPhenoTest`.

The three dataframes ending on "Train" contain data that you should use for building a model for prediction. The variable that you need to predict is the categorical variable in `EpiYTrain`.

This dataset comes from a study where the purpose is to get a simple diagnostic tool for lung cancer based on the microarray technology where one can measure the expression level of thousands of genes simultaneously. The primary x-variable measured is in this case 22.215 dimensional and we find for each of 128 subjects this high-dimensional measurement in `EpiXTrain`. In addition, there are four variables we call phenotype variables (as opposed to the genotype variables measured by the microarray). These are an identification number (ID), the gender (GENDER), a discrete variable indicating if the subject quited smoking more or less than 10 years ago (SMOKING STATUS), and finally a quantitative variable (PACKYEARS) that quantifies how much the subject has smoked in total. In `EpiYTrain` the categorical variable that we are interested in predicting based on the other measured variables is the cancer status. Does the subject suffer from lung cancer or not?

The data in `EpiXTest` and `EpiPhenoTest` contain the microarray measurements and phenotype variables for additionally 35 subjects. This dataset is only needed for a final prediction of the cancer status for these 35 subjects.

N.B. There are 60 cases with lung cancer and 68 without lung cancer in the training data. You can assume that the distribution of the new cases will be roughly as for the training data, and that the objective is to predict lung cancer as well as not lung cancer. In reality this may not be the case and one might also believe that it is far worse to make a wrong prediction for a subject with lung cancer than for one without. Thus that the two types of errors should not be treated equally. In your solution and your report you are welcome to investigate this issue more but it is not required. For the final requirement of predictions on the test data, the misclassification rate will be computed using the 0-1-loss function, which does not discriminate between the two types of mistakes.

The Assignment

In principle the assignment is a simple, open assignment. You need to make a predictor of a categorical y-variable based on any given number of the x-variables. The

origin of the data considered is more closely described in the paper

Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer, Nature Medicine 13, 361 - 366 (2007) by **Avrum Spira et al.**

In that paper they present an analysis and a resulting predictor – a so-called biomaker – for easy lung cancer diagnostic. Your task can be easily formulated as a question of "beating their suggestion".

The problem is, however, a little more difficult than it first seems because there are several different strategies as we are in a large p small N scenario. There are at least three approaches:

One strategy: We keep all the genes and rely of some form of (strong) regularization, for instance, regularized LDA or some form of logistic regression or kernel method with a squared norm penalty term on the parameters.

Another strategy: We do a first, ad hoc filtering where we select a small set of genes, which we can then subsequently use with any method suitable for classification. The initial screening step can but does not need to be based on univariate t-tests.

A third strategy: We rely on automatic variable selection as done by methods with L_1-norm or the elastic net penalty or the variable selection done by the shrunken centroid method.

For your solution of this assignment you need to select at least two different strategies either from the above list or invent your own or do combinations. For instance, if you select the first 10 principal components of the x-variables and that 10-dimensional vector is used with any method we have done a serious dimension reduction just as in the second strategy, but all genes are still used just as in the first strategy. You also have to deal with the three phenotype variables. Should they enter in the analysis on an equal footing as the other variables or should we include any of them at all?

Your report need to document in details you choices and analyzes. This means that you should not write "and then I did cross-validation and got ... " because nobody can read from this precisely how you did cross-validation. You need to describe in detail how you did cross-validation so that I could redo you analysis and produce the same results – without reading your R-code.

For you final predictions on the test data you must report the predictions in you report in the form of a table with 35 rows and 2 columns. In column 1 there is the subject ID and in column 2 the cancer status indicated as either "Cancer" or "No Cancer".

You are encouraged to provide the R-code used as an appendix but the report must be able to stand alone.