Final Assignment

Statistical Learning, 2011

Niels Richard Hansen
October 17, 2011

Formalities

This is final compulsory assignment for the course *Statistical Learning*, 2011. 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 ultimate goal is to build a predictive model.

The deadline for handing in the solution to the assignment is Thursday, November 10, 2011. 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`.

This is an **individual assignment**.

To help and guide you in the analysis and the writing of the report the assignment is divided into four parts, described in details below, but here is an overview of the different parts of the assignment. Each part has equal weight in the final assessment of the assignment. Each part contains several subparts.

1. An introduction and initial analysis. Describe the objective of the analysis and describe the data. Perform simple, initial analyses to get an idea about what to expect.

2. Two required models. You are asked to use two specific models on the training data and to tune the parameters.

3. An open ended question. You are asked to investigate possible improvements of or alternatives to the required models.

4. A concluding section with a comparison of models, your final choice of preferred model and a prediction of 35 test cases (see below).

Data

The data can be downloaded from the course homepage. The data are provided as a binary R data file called `Assignment.RData`. Download the file and load it into R using the command

```
load("Assignment.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` taking the values `Cancer` or `No Cancer`.

This data set 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 genetic variables measured by the microarray). The four phenotype variables consist of an identification number (`ID`), the gender (`GENDER`), a discrete variable indicating if the subject quitted 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. The categorical variable `EpiYTrain`, which 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 frames `EpiXTest` and `EpiPhenoTest` contain the microarray measurements and phenotype variables for additionally 35 subjects. This data set 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, in which case 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 overall goal of the assignment is a simple. 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 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, difficult because we are in a large p small N scenario. To guide you towards a solution we break the problem down into some reasonable components and leave the final model selection open ended. Remember, that even though there are bullet point questions given below, there are in most cases still decisions to be made. It is not always clear that there is an "obviously" correct decision, so don't try to guess what I think is the "correct" answer. Try instead to be aware of the possibility that there may be different solutions, that you have chosen one, and be clear on the choices you make.

Getting started

You should always start your report with an introduction that describes what the report is about. That is, the purpose of the analysis, the type of questions to be considered, the type of data to be considered (in words, no statistics). Then proceed to an initial, or preliminary, analysis, in a separate section.

In this assignment the initial analysis is required to contain the following elements.

- Construct a simple reference model of the presence of cancer given the phenotype variables only – excluding the `ID` variable of course. Investigate if there is any reason to include `PACKYEARS` as a non-linear function in this model.
- Compute for each gene (column in `EpiXTrain`) the t-test statistic of whether the means in the two groups (`Cancer` or `No Cancer`) differ. Report the 10 column names corresponding to the 10 genes with the largest absolute value of the t-test statistic and investigate the empirical correlations between these 10 genes.
- Construct a plot of the first two principal components of the gene expression measurements and color code the points plotted according to group (`Cancer` or `No Cancer`). Comment on the result.

You can extend the initial analysis by, for instance, transformations of the x-variables, consideration of possible interactions of the phenotype variables, alternative marginal tests and by including the top 10 genes found in the simple model.

Lasso and LDA

This section deals with fitting two models on the training data set using the entire set of gene expressions.

The estimation and tuning of a lasso model and a linear discriminant model are required. More precisely, this analysis is required to contain the following elements.

- Use lasso for the fitting of a logistic regression model to the entire data set of gene expression variables. Use cross-validation to select the penalization parameter λ.
- Fit a nearest shrunken centroids model (remember, this is a version of LDA) to the entire data set of gene expression variables. Use cross-validation to select the shrinkage parameter Δ. You are welcome to use the `pamr` package for R for this.
- Discuss the possibility of including the phenotype variables in the analyses above.

You can extend this analysis by considering the more general elastic net penalty, other linear discriminant models such as RDA or a general naive Bayes model.

Extensions

For this part it is required to choose at least one model among the following models and fit the model to the data. It is optional whether you consider the gene expression variables only or include the phenotype variables too.

- A classification tree.
- A random forest.
- A choice of boosted model.
- A generalized additive model.
- An LDA or QDA model.

Not all the models deal equally well with $p \gg N$, and if you run into problems, you will have to preselect a smaller subset of genes. This can, for instance, be done using the marginal t-tests from the initial analysis, or it can be done by using one of the

approaches from the previous section for the *selection* of a subset of genes only, and then use this subset for the training of the model. Optimize, for any of the selected models, the relevant tuning parameter(s). Recall, that preselection of variables is part of the estimation.

Conclusion

For the conclusion you should put together a comparison of the different models considered and present a discussion of pros and cons of the different models. Then select a final model that you would recommend. For this final model compute the predictions on the test data and report the results in the form of a table with 35 rows and 2 columns. In column 1 you put the subject ID and in column 2 the cancer status indicated as either "Cancer" or "No Cancer".

It should be pointed out that the assessment of the report is affected very little by how many correct predictions you get, though this is the "real world" assessment of your selected model. What is important for the assessment is whether the model building, the model comparisons and the final performance of the model is coherent. If, for instance, the report leaves the impression that the model will perform well with close to 0 misclassifications and the result is 10, this will have a negative effect on the overall assessment. If the model selection is superficially, haphazardly or perhaps even wrongly carried out and if the selected model performs badly in terms of the prediction, this will also have a negative effect on the overall assessment.

General comments

The most important rule when writing the report is to document what you do in a clear language. Include plots and tables whenever needed to support any kind of conclusion you draw or observation that you do. Remember that the reader does not see your computer screen and is not able to read your mind. Therefore, describe the results you obtain and how you interpret them, include all the figures needed to support an argument, and make sure that the included figures in the report actually convey the information intended. On the other hand, don't include excessive amounts of figures or lists of parameter estimates or the like that are not needed or commented.

Describe the methods used and how they are used in detail. That is, it is not enough to say "... and the by cross-validation we found the estimate of ..." Describe how cross-validation was done exactly.

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