First Assignment

Statistical Learning, 2009

Niels Richard Hansen
May 1, 2009

Formalities

This is the first compulsory assignment for the course Statistical Learning. This is a practical assignment, where you are asked to hand in a small, commented computer program written in R. The final solution of the assignment is evaluated based on whether your program computes the correct solutions.

The deadline for handing in the solution to the assignment is May 22, 2009. You hand it in by sending the solution as a single file attachment to an email. The attached file must be named `yourfullname.r`.

In the program you need to make sure that the following things are included.

- Initially all libraries that you need must be loaded.
- The dataset must be loaded.
- All variables that you rely on are defined and given appropriate values.
- If you implement some computations within a function remember to include the implementation in this file also. This should be included close to where the function is needed the first time.

Try to make it as easy as possible for me to read the R code and provide a reasonable amount of comments. This means that you *do not need to write:*

```
> tmp <- 2  #Here I set the variable tmp equal to 2
```

But if you implement a function, you may want to briefly describe what the function does and what the parameters mean. Finally, in the questions given below you will be asked to compute some final values or produce a plot. When you are asked to compute a value you should store the result in a variable called `questionx` where `x` is replaced by the number of the question and then this should be followed by a `print(questionx)`. If you are asked to produce a plot then simple write the code for plotting on the screen. When you plot *do not used system specific commands* like `windows()`. If you want explicitly to open a new plotting window use `dev.new()`.

Data

The data can be found from the course homepage. The data are given as binary R data in a file called `Assignment1.RData`. Download the file and load it into R using the command

```
load("Assignment1.RData")
```

Then you will have two dataframes in your R session called `Assignment1Train` and `Assignment1Test`. Each contain 16 columns of data from different individuals, with the first 15 being the *genetic fingerprint* – the count of the number of repeats for certain so-called tandem repeats in the genome – and the last being the population variable. The purpose is to predict the population from the genetic fingerprint. We refer below to the repeat counts as the count data (the x variables) and the population as the group (the y variable).

Problems

In the following **all** estimation should be done using only the training data. Whenever you are asked to compute the training error you compute the average error of predictions – using the relevant loss function – on the training data. Whenever you are asked to compute the test error you compute the average error on the test data.

> **Question 1.** *Within the setup of LDA compute the estimates of the three group means and the estimate of the common covariance matrix.*
>
> **Question 2.** *Make a plot of the projection of the count date onto the first two principal components. Add to this plot the projection of the three group means.*
>
> **Question 3.** *Within the setup of LDA make a plot of the projection of the count data onto the two canonical coordinates. Add to this plot the projection of the three group means.*

For the remaining questions you are asked to consider only two of the three groups; the `Caucasians` and the `African Americans`.

> **Question 4.** *Estimate the LDA classifier and compute the training error.*
>
> **Question 5.** *Estimate the logistic regression model and compute the training error.*

The next question rely on that you make some qualified choices of what "a comparison" means. Comment briefly on these choices.

Question 6. *Set up a comparison of the two classifiers in terms of estimated parameters and predictions on the training data.*

Question 7. *Construct a naive Bayes classifier and compute the training error on the dataset. You are here free to choose how you estimate the marginal distributions of the counts within each group. You can use the fact that the distributions are discrete, but you can also approximate the distributions by continuous distributions.*

Question 8. *Compute for the three methods – LDA, logistic regression and naive Bayes – the test error. Compare and comment on the results.*