

Practical 7

Statistical Learning, 2011

Niels Richard Hansen

October 13, 2011

Naive Bayes

The data for this exercise can be downloaded from the course homepage. The data set is a binary R data file called `prac7.RData`. Download the file and load it into R using the command

```
load("prac7.RData")
```

Then you will then have two data frames in your R session called `prac7Train` and `prac7Test`. Each contains 16 columns of data from different individuals, with the first 15 being the *genetic fingerprint* – the count of the number of repeats of certain so-called tandem repeats in the genome – and the last being the population variable. The purpose here is to predict the population from the genetic fingerprint. We refer below to the repeat counts as the X variables and the population as the group or Y variable.

Use the training data for estimation and the test data for assessment and comparison.

1. Plot all the X -variables against each other and color code the points according to population (use the `pairs` function).
2. Construct a naive Bayes classifier and compute the test and training error on the dataset. You are free to choose how you estimate the marginal distributions of the counts within each group.

You can assume that the counts are continuous variables and use `density` in R to compute a fitted density at a grid of values.

You can also use the fact that the distributions are discrete and tabulate. The tabulations can be “smoothed” using `convolve`.

If the problem seems difficult break it down into pieces. How would you proceed if X was one-dimensional and not 15-dimensional?

3. Estimate the group means of the X -variables within each population and estimate a common correlation matrix.
4. Compare with an LDA classifier and a logistic regression classifier.