STA 414/2104, Spring 2006 — Assignment #2

Due at ***start*** *of class on March 9. Note that this assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own.*

In this assignment you will try out logistic regression and linear discriminant analysis on an artificial dataset that I have created. This dataset resembles data obtained from spectrometry methods (eg, mass spectrometry or NMR spectrometry), which are commonly used to analyse complex mixtures of components, such as occur in blood. The aim is to assign spectra to one of two classes, which might in a real application correspond, for example, to blood from patients who do or do not have a certain disease.

Each spectrum consists of 200 non-negative data points, which are ordered (eg, by mass for mass spectrometry). Nearby points in the spectrum are often correlated, since a single component of the mixture analysed may show up as a peek that is spread over a range of masses. We hope that which peeks are present in the spectrum, or their magnitude, is related to the class, but there may also be variation unrelated to the class. Each point is also subject to measurement noise.

From the course web page, you can obtain a training set of 60 cases, with the inputs for each case consisting of the 200 points of the measured spectrum. You also are given the class (0 or 1) for these 60 cases. You are to test the methods you try on a test set of 1000 cases, also available from the web page, for which you initially assume that you have only the 200 inputs, from which you try to predict the class. The classes for these test cases are also available, however, so that you can see how well each method did (in terms of error rate). But the methods you test should not look at these, of course!

Rather than try to apply logistic regression and LDA directly to the 200 inputs, you should apply them after reducing dimensionality to 10 by one of two methods:

- Simply average the inputs in blocks of 20. That is, the first average is of inputs 1–20, the next of inputs 21–40, etc.
- Find the first ten principle components.

For each of these new data sets (with only 10 inputs), you should try out both logistic regression, with parameters found by maximum likelihood, and LDA. (You will therefore try a total of four methods.)

You should write your own R or Matlab functions for finding principle components and for LDA — you should **not** use or consult the source code for any predefined functions in R or Matlab that do PCA or LDA. However, you **may** use a predefined function for finding maximum likelihood estimates for logistic regression, and information on doing this in R and Matlab will be put on the course web page soon. However, you should **not** use a predefined function for using the maximum likelihood estimates to make predictions. You should instead write your own code to make predictions for test cases using the estimated regression coefficients.

Once you have the results of the four methods, you should try to make sense of them — for example, by plotting the original data and the principle components, examining the details of the discriminant functions, etc. You should hand in a listing of the functions and scripts you wrote, the results you obtained (summaries only, not results for every test case), and your discussion of the results, including whatever plots or other information you think helps explain these results. Finally, you should suggest methods that might be better than any of the four you tried (but you're not expected to actually try them).