STA 414/2104, Spring 2007 — 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 apply several linear and quadratic discriminant methods to two synthetic datasets, and assess how well the different methods work on these problems.

For both datasets, each case has values for a binary class variable and for six real-valued input variables. The first dataset has 250 training cases and 3000 test cases. The second dataset has 2000 training cases and 5000 test cases. The training and test data can be downloaded from the course webpage, in eight files, with the first data set identified by 'a2a' and the second by 'a2b', with 't' for class variables (targets) and 'x' for input variables.

You should try out four classification methods on these datasets, a follows:

- A linear discriminant classifier, found from the class means and the pooled covariance matrix. You should set the threshold (determined by w_0) so as to minimize the error rate on training cases.
- A quadratic discriminant classifier, found from the class means and the covariance matrices for each class. Again, you should set the threshold (determined by w_0) so as to minimize the error rate on training cases.
- Maximum likelihood logistic regression with just the original inputs as predictors (plus an intercept).
- Maximum likelihood logistic regresson with the original inputs, the squares of inptus, and all the pairwise products of two different inputs as predictors (plus an intercept).

You should write R functions that implement the first two methods (not using any builtin R functions except general facilities such as matrix inverse). You should use R's builtin `glm` function (with the `family="binomial"` and `maxit=1000` options) to find the maximum likelihood estimates for logistic regression, but you should make predictions for test cases using this maximum likelihood estimate without using any builtin functions related to `glm`.

You should report the error rate on test cases from each of these four methods applied to each of the two datasets. You should also discuss why the results are as they are. As a basis for this discussion, you should look at scatterplots of pairs of input variables (with class indicated by colour or type of point drawn), so as to see whether the data has the characteristics that one would expect are necessary for each method to work well. You may also want to look at the estimated coefficients for the discriminants, comparing what was found by the two linear or the two quadratic methods.

You should hand in a listing of your program, with suitable but not excessive comments, the error rates you found, your discussion, and whatever (non-excessive) plots or other output is needed to support your discussion.