

Statistics 579  
Applied Multivariate Methods

Exam 2

1. The file Oil.txt contains data on crude-oil samples from sandstone in the Elk Hills, California, petroleum reserve. On the basis of their chemistry, these crude oils can be assigned to one of the three stratigraphic units (i.e., populations): Wilhelm sandstone, Sub-Mulinia sandstone, and Upper sandstone. This data set contains measurements on 5 (chemistry related) variables: Vanadium (in percent ash), Iron (in percent ash), Beryllium (in percent ash), Saturated Hydrocarbons (in percent area), and Aromatic Hydrocarbons (in percent area).
  - a) Using an appropriate variable selection procedure, determine which variables are useful for discriminating between the 3 groups. Use  $\alpha=0.02$ .
  - b) Using the variables selected, and assuming multivariate normal data with a common covariance matrix, with priors proportional to sample size, classify the samples according to stratigraphic units. Estimate the classification error rate.
  - c) Characterize each of these 3 groups based on their average scores on the original 5 measurements.

2. Salmon fishing is a valuable resource for both the United States and Canada. Because it is a limited resource, and because more than one country is involved, Alaskan fisherman cannot catch too many Canadian salmon and vice versa. To help regulate catches, samples of fish taken during harvest must be identified as coming from Alaskan or Canadian waters. The fish also carry information about their birthplace in the growth rings on their scales. The file Salmon.txt contains data on 45 salmon captured in Alaskan waters, and on an equal number captured in Canadian waters. There are 4 variables associated with each fish:

Waters = 1 if Alaskan, 2 if Canadian

Gender = 1 if Female, 2 if Male

Freshwater = diameter of rings for the first-year Freshwater growth  
(in hundredths of an inch,  $\times 100$ )

Marine = diameter of rings for the first-year Marine growth  
(in hundredths of an inch,  $\times 100$ )

For the purpose of this analysis, you should ignore Gender.

- a) For Alaskan salmon, and separately for Canadian salmon, determine if the assumption of bivariate normality is reasonable (using the variables Freshwater and Marine)
- b) Obtain the sample linear discriminant function, and estimate the classification error rate. Assume equal priors.
- c) Using the above linear discriminant function, classify the 10 salmon whose origin is unknown that are contained in the file Salmontest.dat. (This file contains data on the 3 variables Gender, Freshwater and Marine for each of 10 salmon whose origin is unknown.)
- d) Plot all 100 values in a 2-dimensional plot of Marine vs. Freshwater, using different plotting symbols to identify whether the individual salmon is known to come from Alaskan or Canadian waters, or is Unknown.
- e) Develop a logistic regression model that predicts origin based upon Freshwater and Marine. Estimate the cross-classification error rate, and compare it with the error rate obtained from discriminant analysis.
- f) Using the above logistic regression model, classify the 10 salmon whose origin is unknown that are contained in the file Salmontest.txt. Compare these results with the results obtained from discriminant analysis.