STAT 802

Assignment 3

1. Observations on 2 sexes of two types of wolf (Rocky Mountain and Arctic) are in the file

   `~lockhart/Teaching/Courses/802/Data/Asst3Q1`

   The first 6 are Rocky Mountain males, then 3 females then 10 Arctic males and 6 Arctic females. The 9 variables are 9 different lengths measured on the skulls of the animals (in millimetres).

   (a) Use pairwise T^2 tests to decide if any of the pairs are indistinguishable on the basis of the data here.

   (b) Assess whether there are any variables on which all 4 groups have means so similar that the variable in question will not really help discriminate. A hypothesis test with a large α is probably useful.

   (c) For the standard linear discriminant rule classify all of the data points and determine the raw and cross-validated error rates.

2. Fisher's iris data is available in S. Pick a pair of the 4 available variables and plot on a graph the boundaries of the linear and quadratic discriminant rules for discriminating the three varieties of iris along with the data. Estimate misclassification rates parametrically for both discriminant rules but, in estimating the error rates, do not assume that all the variance covariance matrices are the same.

3. From the text 11.28.

4. Consider the situation where you have two normal populations and $p = 1$; the populations are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. You have samples of size n_1 and n_2 and observe means $\bar{X}_1$ and $\bar{X}_2$ where we assume that $\bar{X}_2 > \bar{X}_1$. We plan to use the following rule:

   R: Classify a new X as coming from population 2 if $X > (\bar{X}_1 + \bar{X}_2)/2$

We want to estimate the error rates for this classifier. The usual linear rule for known parameters would classify X in group 2 if $X > (\mu_1 + \mu_2)/2$ and have an error rate of

$$P(N(\mu_1, \sigma_1^2) > (\mu_1 + \mu_2)/2) = 1 - \Phi(-\delta/2)$$

where

$$\delta = |\mu_2 - \mu_1|/\sigma_1$$

NOTE: if $\sigma_1 \neq \sigma_2$ then this linear rule is **not** the best rule but we are analyzing its performance when it is used in less than ideal circumstances.

(a) Derive an expression for the conditional error rates given $\bar{X}_1$ and $\bar{X}_2$ of rule R as a function of $\bar{X}_1$, $\bar{X}_2$, μ_1 and σ_1.

(b) Find a Taylor expansion for these quantities as a function of $\bar{X}_1$ and $\bar{X}_2$ about the points μ_1 and μ_2. Keep terms out to quadratic in $(\bar{X}_i - \mu_i)^2$.

(c) Use this expression to compute the expected error rate and show that it is the error rate $1 - \Phi(-\delta/2)$ for the linear classifier given above plus a term of the form $c_1/n_1 + c_2/n_2$. Your answer should include a formula for the c_i.

(d) The cross validation estimate of this error rate is

$$\frac{\sum_{i=1}^{n_1} 1(\text{item } i \text{ misclassified when left out})}{n_1}$$

Use the result in the first 3 parts of this question but with n_1 1 smaller get a formula for the expected value of this estimate plus terms like $c_1^*/(n_1 - 1) + c_2^*/n_2$.

(e) Put these results together to argue that the expected value of the cross validation estimate of the error rates is the same, to terms of order $1/n_i$ as the expected error rate.

5. From the text questions 8.18 and 8.19. The data are in

`lockhart/Teaching/Courses/802/Data/Asst3Q5PartA`

6. From the text questions 9.28 and 9.29.

7. From the text 10.10.

8. From the text 12.16. CANCELLED.