THE UNIVERSITY OF MINNESOTA

Statistics 5401 November 28, 2005

Assignment 9

Reading

Week of November 28 - December 2 J&W, Chapter 11
Week of December 5 - 9 J&W, Sec. 12.1-12.3
Week of December 12 - 14 J&W, remainder of Chapter 12

Final Examination: A **take home** final examination will be handed out in the next to final class, Monday, December 12. It will be due by 12 noon on Tuesday December 20.

Written Assignment (due in class Monday, December 12)

1. J&W Exercise 11.29, p. 658. Assume prior probabilities proportional to to sample size. Thus on (b), you can use the ordinary between group matrix **B** (that is, hypothesis matrix **H**) instead of $\hat{\mathbf{B}}$.

Here is a partial listing of matrix `T11_06` in file `JWData5.txt`. Note that the group number is in column 1 and there are no case numbers.

```
T11_06    85    3 format
) Data from Table 11.6 p. 659 in
) Applied Mulivariate Statistical Analysis, 5th Edition
) by Richard A. Johnson and Dean W. Wichern, Prentice Hall, 2002
) These data were edited from file T11-6.DAT on disk from book
) Group numbers were moved to column 1
) Admission data from Graduate School of Business
) Col. 1: Group number (1 = admit, 2 = do not admit, 3 = borderline)
) Col. 2: x1 = GPA
) Col. 3: x2 = GMAT
)"%lf %lf %lf"
(f2.0,f5.2,f4.0)
 1 2.96 596
 1 3.14 473
 1 3.22 482
 1 3.29 527
 1 3.69 505
. . . . . .
. . . . . .
. . . . . .
 3 3.05 399
 3 2.85 483
 3 3.01 453
 3 3.03 414
 3 3.04 446
```

2. J&W Exercise 11.30, p. 659-660. Use macro `jackknife()` to perform Lachenbruch's holdout procedure. In addition use macros `discrimquad()` and `probsquad()` to compute quadratic discriminant functions and classify the cases in the training sample. For part (c), use square root transform for X2, X3, X4 and X5, verifying that this improves normality.

Here is a partial listing of data set `T11_07` in file `JWData5.txt`. Notice the zone number is in column 1. You will need to separate it from the remaining 5 columns.

```
T11_07    56    6 format
) Data from Table 11.7 p. 661 in
) Applied Mulivariate Statistical Analysis, 5th Edition
) by Richard A. Johnson and Dean W. Wichern, Prentice Hall, 2002
) These data were edited from file T11-7.DAT on disk from book
) Group identification was moved from last column to first and
) made numeric
) Crude oil data
) Col. 1: Zone (1 = Wilhelm, 2 = sub-Mulinia, 3 = Upper
) Col. 2: X1 = vanadium (percent ash)
) Col. 3: X2 = iron (percent ash)
) Col. 4: X3 = beryllium (percent ash)
) Col. 5: X4 = saturated hydrocarbons (percent area)
) Col. 6: X5 = aromatic hydrocarbons (percent area)
)"%lf %lf %lf %lf %lf %lf"
(f2.0,2f5.1,2f5.2,f6.2)
 1  3.9 51.0 0.20 7.06 12.19
 1  2.7 49.0 0.07 7.14 12.23
 1  2.8 36.0 0.30 7.00 11.30
 1  3.1 45.0 0.08 7.20 13.01
 1  3.5 46.0 0.10 7.81 12.63
 . . . . . . . . . . . . . . .
 . . . . . . . . . . . . . . .
 . . . . . . . . . . . . . . .
 3  7.3 22.0 0.00 4.13  2.70
 3  4.1 29.0 0.70 5.78  7.76
 3  5.4 29.0 0.20 4.64  2.65
 3  5.0 34.0 0.70 4.21  6.50
 3  6.2 27.0 0.30 3.97  2.97
```

3. Table 11.4, p. 655 in J&W contains annual financial data of 46 firms, 21 that declared bankruptcy and 25 that were financially sound." There are 4 financial variables that might be used to discriminate between sound and failing firms.

(a) Using all the data, estimate the probability of misclassification using linear discrimination. Assume equal prior probabilities and equal costs of misclassification.

(b) Determine the best subset of variable for classification. Use forward and backward stepwise selection with $\alpha = .05$. Also use `dascreen()` (in the updated version of Mulvar.mac) to find the best subset.

Here is a partial listing of the data. Notice that the population is coded as 0 or 1. You will need to add 1 to recoded it as 1 or 2 as is required by MacAnova macros.

```
T11_04    46    5 format
) Data from Table 11.4 p. 655 in
) Applied Mulivariate Statistical Analysis, 5th Edition
) by Richard A. Johnson and Dean W. Wichern, Prentice Hall, 2002
) These data were edited from file T11-4.DAT on disk from book
) Bankruptcy data
) Population identifier (0 or 1) moved to column 1
) Col. 1: Population (0 - population 1, 1 = population 2)
) Col. 2: x1 = CF/TD
) Col. 3: x2 = NI/TA
) Col. 4: x3 = CA/CL
) Col. 5: x4 = CA/NS
)"%lf %lf %lf %lf %lf"
(f2.0,4f6.2)
 0 -0.45 -0.41  1.09  0.45
 0 -0.56 -0.31  1.51  0.16
 0  0.06  0.02  1.01  0.40
 0 -0.07 -0.09  1.45  0.26
 . . . . . . . . . . . . .
 . . . . . . . . . . . . .
 . . . . . . . . . . . . .
 1  0.20  0.08  1.99  0.30
 1  0.47  0.14  2.92  0.45
 1  0.17  0.04  2.45  0.14
 1  0.58  0.04  5.06  0.13
```