Final99.ans

STAT 501 **Final Exam Solutions** **Spring 1999**

1. (12 points) Complete linkage cluster analysis proceeds as follows:

$\{1\}\ \{2\}\ \{3\}\ \{4\}\ \{5\}\ \{6\}\ \{7\} \rightarrow \{1\}\ \{3\}\ \{5\}\ \{6\}\ \{7\}\ \{2,4\}$

-	37	31	36	39	$33 = \max\{24, 33\}$
	-	45	**30**	42	$41 = \max\{42, 41\}$
		-	46	48	$41 = \max\{39, 41\}$
			-	34	$42 = \max\{32, 42\}$
				-	$40 = \max\{35, 40\}$

$\rightarrow \{1\}\ \{5\}\ \{7\}\ \{2,4\}\ \{3,6\}$

-	**31**	39	33	37
	-	48	41	46
		-	40	42
			-	42

$\rightarrow \{7\}\ \{2,4\}\ \{3,6\}\ \{1,5\}$

-	**40**	42	48
	-	42	41
		-	46

$\rightarrow \{2,4,7\}\ \{3,6\}\ \{1,5\}$

2. a. (6 points) The scores for the first principal component represent an overall impression of the importance of the five taste, appearance and nutritional properties of the food.

The scores for the second principal component reflect a contrast between the importance of the appearance and taste of the food with the nutritional propertities of the food. Individuals with high scores on this component give more importance to taste and appearance, while individuals with low scores give more importance to nutritional attributes of the food.

b. (4 points) (3.3+1.2)/5 x 100% = 90%

c. (4 points) There are positive correlations between each pair of variables, but the correlation between the taste and appearance attributes (attributes 1 and 2) and the correlations among the nutrition attributes (variables 3,4 and 5) are stronger than correlations among attributes

from these two sub-groups. You could check this by computing the approximation to the correlation matrix provided by the first two principal components

$$\hat{\lambda}_1 \underset{\sim}{e}_1 \underset{\sim}{e}_1^T + \hat{\lambda}_2 \underset{\sim}{e}_2 \underset{\sim}{e}_2^T = \begin{bmatrix} 1 & .94 & .38 & .38 & .32 \\ .94 & 1 & .46 & .45 & .40 \\ .38 & .46 & 1 & .84 & .89 \\ .38 & .45 & .84 & 1 & .85 \\ .32 & .40 & .89 & .85 & 1 \end{bmatrix}$$

d. (4 points) $\hat{\lambda}_2 = 1.2$

e. (4 points) correlation $= (0.61)\sqrt{1.2} = 0.67$

f. (4 points) $(3.3)(0.47)^2 + (1.2)(-0.38)^2 = 0.90225$

g. (4 points) The main objective of a varimax rotation is to create a new set of orthogonal factors where each of the measured attributes has a high loading on only one factor. Make a plot of the original factor loadings to see that the taste and appearance attributes will have high loadings on one factor and the "nutrition" factors have high loadings on the other factor.

h. (4 points) (step 1) Compute initial estimates of communalities for the correlation matrix. One option is to use the R^2 value for the regression of the i-th attribute on the values for the other attributes as the initial estimate of the communality for the i-th attribute.

(step 2) Substitute the current estimates of the communalities for the ones on the diagonal of the estimated correlation matrix, and compute the eigenvalues and eigenvectors of the matrix of the resulting matrix.

(step 3) Compute new communalities from the results of step 2.

Repeat steps 2 and 3 until convergence.

i. (6 points) For this model the observations on the measured traits and the corresponding values of the unobserved factors have a joint normal distribution, i.e.,

$$\begin{bmatrix} \mathbf{X_j} \\ \mathbf{F_j} \end{bmatrix} = \mathbf{N}\left(\begin{bmatrix} \mu_j \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{L}\mathbf{L}^T + \psi & \mathbf{L} \\ \mathbf{L} & \mathbf{I} \end{bmatrix} \right).$$ Then, factor scores are obtained by computing

the conditional mean of $\mathbf{F_j}$ given $\mathbf{X_j}$, replacing μ_j with the vector of sample means $\bar{\mathbf{X}}$, replacing $\mathbf{L}$ with the estimated factor loadings $\hat{\mathbf{L}}$, and replacing ψ with a diagonal matrix of estimated specific variances $\hat{\psi}$. The result is

$$\hat{\mathbf{F}}_j = \hat{\mathbf{L}}^T\left(\hat{\mathbf{L}}\hat{\mathbf{L}}^T + \hat{\psi}\right)^{-1}\left(\mathbf{X_j} - \bar{\mathbf{X}}\right)$$

3. (a) (8 points) The null hypothesis can be expressed in several equivalent ways. Let μ_{ij} denote the mean response for subjects of the j-th sex at the i-th inspection time. Then, the null hypothesis is

$$H_0: C\beta M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \mu_{11} & \mu_{21} & \mu_{31} & \mu_{41} \\ \mu_{12} & \mu_{22} & \mu_{32} & \mu_{42} \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}$$

Alternatively, the mean response for the j-sex at the i-th inspection time could be expressed as $\mu_{ij} = \mu_i + \alpha_{ij}$. The GLM procedure in SAS would impose the constraints $\alpha_{i2} = 0$ for $i = 1,2,3,4$. Then, the null hypothesis is written as

$$H_0: C\beta M = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} \mu_1 & \mu_2 & \mu_3 & \mu_4 \\ \alpha_{11} & \alpha_{21} & \alpha_{31} & \alpha_{41} \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}$$

In either case, we have n=48, k=2, r=2, p=4, and u=3 which yield

$$a = (n-r) - \frac{u - k + 1}{2} = 45$$

$$b = \sqrt{\frac{u^2 k^2 - 4}{u^2 + k^2 - 5}} = 2 \qquad \text{and} \qquad c = \frac{uk - 2}{2} = 2$$

Then,

$$F = \frac{1 - \sqrt{0.12}}{\sqrt{0.12}}\frac{ab - c}{uk} = 27.67 \quad \text{on} \quad (uk, ab - c) = (6,88) \text{ d.f.}$$

(b) (8 points) This null hypothesis can also be expressed in several equivalent ways. Let μ_{ij} denote the mean response for subjects of the j-th sex at the i-th inspection time. Then, the null hypothesis is

$$H_0:\ C\beta M = \begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}\mu_{11} & \mu_{21} & \mu_{31} & \mu_{41}\\ \mu_{12} & \mu_{22} & \mu_{32} & \mu_{42}\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ -1 & 1 & 0\\ 0 & -1 & 1\\ 0 & 0 & -1\end{bmatrix}$$

Alternatively, the mean response for the j-sex at the i-th inspection time could be expressed as $\mu_{ij} = \mu_i + \alpha_{ij}$. The GLM procedure in SAS would impost the constraints $\alpha_{i2} = 0$ for $i = 1,2,3,4$. Then, the null hypothesis is written as

$$H_0:\ C\beta M = \begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}\mu_1 & \mu_2 & \mu_3 & \mu_4\\ \alpha_{11} & \alpha_{21} & \alpha_{31} & \alpha_{41}\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ -1 & 1 & 0\\ 0 & -1 & 1\\ 0 & 0 & -1\end{bmatrix}$$

In either case, we have n=48, k=1, r=2, p=4, and u=3 which yield

$$a = (n-r) - \frac{u-k+1}{2} = 44.5$$

$$b = \sqrt{\frac{u^2k^2-4}{u^2+k^2-5}} = 1 \qquad \text{and} \qquad c = \frac{uk-2}{2} = 0.5$$

Then,

$$F = \frac{1-\sqrt{0.38}}{\sqrt{0.38}}\,\frac{ab-c}{uk} = 23.73 \quad \text{on } (uk,\ ab-c) = (3,44) \text{ d.f.}$$

4. (12 points) The expected cost of misclassification is minimized by classifying a cow into the diseased group if

$$\frac{f_1(x)}{f_2(x)} \geq \frac{C(1|2)\,p_2}{C(2|1)\,p_1} = \frac{(0.9)(1)}{(0.1)(20)} = 0.45$$

Here, x denotes the possible outcomes for the test A and test B, respectively, and 1 corresponds to the diseased population and 2 corresponds to the healthy population. The four possible results are:

Test A is positive and test B is positive: $\frac{f_1(+,+)}{f_2(+,+)} = \frac{0.90}{0.04} > 0.45$

Test A is positive and test B is negative: $\frac{f_1(+,-)}{f_2(+,-)} = \frac{0.01}{0.06} < 0.45$

Test A is negative and test B is positive: $\frac{f_1(-,+)}{f_2(-,+)} = \frac{0.05}{0.08} > 0.45$

Test A is negative and test B is negative: $\frac{f_1(-,-)}{f_2(-,-)} = \frac{0.04}{0.82} < 0.45$

Hence, the classification rule that minimizes the expected cost of misclassification ignores the result from test A and classifies the cow as diseased or healthy depending on whether or not test B is positive or negative.

5. (20 points) Your answer should address the following issues:

(1) Talk to bankers and credit card marketing experts to obtain information about prior probabilities and misclassification costs. Here you might set the prior for a positive response at (8257)/(400000). You should also discuss the possible creation of new variables such as total debt divided by annual income.

(2) Use exploratory methods, such as box plots, Xgobi and projection pursuit, principal components analysis, to examine the data. You might also use cluster analysis based on the centroid method to identify extreme cases.

(3) Recode categorical variables X_{11} X_{12} and X_{16} as sets of binary variables. Consider transforming continuous variables to obtain more nearly symmetric distributions of values. (Note that monotone transformations have not effect on classification trees.) Check for missing data and make some decisions about whether you will delete cases, delete variables or try to impute values.

(4) Search for good classification rules. You could use the STEPDISC procedure in SAS to examine linear classification rules using stepwise or backward elimination searches. Since there are only two populations in this case, you could use the all possible regressions option in the REG procedure in SAS to examine a larger set of possible classification rules. You could also use the stepwise or backward elimination searches built into to the LOGISTIC procedure in SAS. Similar options are available in S-PLUS and other software packages. Since the training samples are large, it would be more efficient to use a set aside method (setting aside 30% of each training sample) than a cross-validation method to assess misclassification rates.

(5) Report results and determine if the information provided by these 17 variables enables you to effectively screen data files of potential customers.

Final exam scores are given in the following stem-leaf display:

```
10|0 0
 9|5 9
 9|4
 8|6 7 7 8 8
 8|3 4 5
 7|
 7|0 0 1 2 2 3
 6|8 8
 6|3
 5|6
```