Departamento de Matemática

Multivariate Analysis

Mater in Eng. and Data Science & Master in Mathematics and Applications

2^{nd} Test | 1^{st} Semester – 2019/2020

Duration: 1.5 hours | 30/01/2020 – 11:30

Please justify conveniently your answers

Group I **8.0 points**

1. How can we use cluster evaluation measures to determine the correct number of natural clusters? Do these methods always indicate the correct number of natural clusters? (2.0)

2. Suppose that $\boldsymbol{x}_1 = (2,5,2,5,3)^t$, $\boldsymbol{x}_2 = (3,5,2,4,3)^t$, $\boldsymbol{x}_3 = (9,1,1,1,1)^t$, and $d_{rs} = 1-2C_{rs}/(A_r + B_s)$, where $A_r = \sum_j x_{rj}$, $B_s = \sum_j x_{sj}$, $C_{rs} = \sum_j \min(x_{rj}, x_{sj})$, and $x_{ij} \geq 0$.

   (a) Verify that $d_{13} > d_{12} + d_{23}$. (1.0)

   (b) Show that d_{rs} is a dissimilarity but not a metric. (3.0)

3. An observation x comes from one of the two populations with probability density functions:
$$f_{X|Y=i}(x) = \frac{1}{\lambda_i}\exp\left(-\frac{x}{\lambda_i}\right), \quad x \geq 0,$$
with $\lambda_1 > \lambda_0 > 0$, $i = 0, 1$, known as Exponential distribution with parameter λ_i.

   Let us admit that the group each observation belongs to, Y, was not observed and $P(Y=1) = p$ is unknown. Then X can be seen as a mixture of two Exponential distributions. Consider that $\boldsymbol{x} = (x_1, \ldots, x_n)^t$ is a sample of size n from this population.

   It can be shown that the complete log-likelihood is:
$$l(\boldsymbol{\lambda}|\boldsymbol{x},\boldsymbol{y}) = \sum_{j=1}^{n} \ln\left\{f_{X|Y=y_j}(x_j|\boldsymbol{\lambda})p^{y_j}(1-p)^{1-y_j}\right\},$$
where $\boldsymbol{\lambda} = (p, \lambda_0, \lambda_1)^t$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^t$ represents the not observed classes of $\boldsymbol{x}$.

   (a) Estimate the unknown parameters, p, λ_0, and λ_1, using the EM algorithm. Define the E- and M-step. Start by showing that: (2.0)
$$\begin{aligned} E\left(l(\boldsymbol{\lambda}|\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{\lambda}^{(g)}\right) &= \sum_{j=1}^{n}\ln\left\{f_{X|Y=1}(x_j)p\right\}P(Y=1|X=x_j,\boldsymbol{\lambda}^{(g)}) + \\ & \quad \sum_{j=1}^{n}\ln\left\{f_{X|Y=0}(x_j)(1-p)\right\}P(Y=0|X=x_j,\boldsymbol{\lambda}^{(g)}). \end{aligned}$$

Group II **12.0 points**

Conn's Syndrome is a form of hypertension that has two possible causes: an adenoma (Type A patient), which has to be removed by surgery, and bilateral hyperplasia (Type B patient), which is a more diffuse condition and is treated with drugs. It can be hard to tell whether a patient is Type A or Type B. Researchers investigated a group of 31 sufferers of Conn's Syndrome, recording their age (in years) and the concentrations of the following three chemicals in blood plasma (in meq/l): sodium, potassium, carbon dioxide. All these patients then underwent surgery, which revealed that 20 of them were Type A and the other 11 Type B. The results displayed below is from an analysis of the data for all 31 patients in the study.

1. What does the matrix plot suggest about the potential usefulness of each of these variables for classifying patients as Type A or Type B? (1.5)

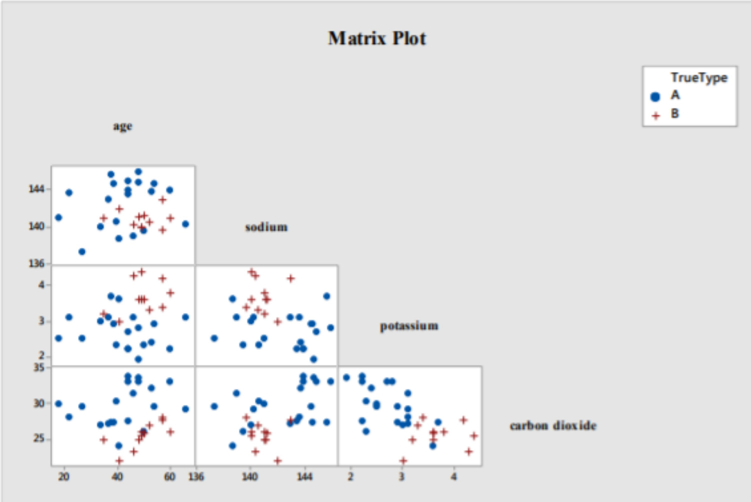

Figure 1: Matrix Plot.

2. A linear discriminant analysis of the data was carried out, with the results shown in the Table 1(a) and Table 1(b). Explain what is meant by leaving-one-out cross-validation. Give one reason for carrying out this procedure as part of discriminant analysis. What do the results in Table 1 suggest about the usefulness of the linear discriminant? (3.0)

Table 1: Results of the linear discriminant analysis.

(a) Without Cross-Validation

	True Group	
Put into Group:	A	B
A	17	0
B	3	11
Total	20	11
No. correct	17	11
Proportion	0.850	1.000

Overall Proportion Correct = 0.903

(b) With Cross-Validation

	True Group	
Put into Group:	A	B
A	16	1
B	4	10
Total	20	11
No. correct	16	10
Proportion	0.800	0.909

Overall Proportion Correct = 0.839

3. Confirm your comments in (2) calculating the overall F_1 measure, based on confusion matrix of Table 1(b). (2.0)

4. The linear discriminant function is (1.5)

$$y = -20.0 - 0.1 \times age + 0.2 \times sodium - 3.0 \times potassium + 0.6 \times carbon\_dioxide,$$

where positive values of y indicate Type A patients. Use it to classify another Conn's Syndrome sufferer, who is 40 years old and has the following test results: sodium 144.1, potassium 3.4, carbon dioxide 25.2 meq/l.

5. List the main statistical assumptions required to justify the use of linear discrimination. In what way may these assumptions be relaxed if quadratic discrimination is used instead of linear discrimination? (2.0)

6. Table 2(a) and Table 2(b) give the results of quadratic discrimination, in a form comparable to (2.0)

Table 1(a) and Tables 1(b), respectively. Compare the results from the two discriminants. Which would you recommend using in practice, and why?

Table 2: Results of the quadratic discriminant analysis.

(a) Without Cross-Validation

Put into Group:	True Group A	True Group B
A	19	0
B	1	11
Total	20	11
No. correct	19	11
Proportion	0.950	1.000

Overall Proportion Correct = 0.968

(b) With Cross-Validation

Put into Group:	True Group A	True Group B
A	17	3
B	3	8
Total	20	11
No. correct	17	8
Proportion	0.850	0.727

Overall Proportion Correct = 0.806