XST30111 (Code?)

Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Faculty of Engineering, Mathematics and Science

School of Computer Science & Statistics

BA (Mod) JS MSISS, JS-SS MATHS & TSM **?? Term 2017**

Multivariate Linear Analysis (ST3011)

DD MMM YYYY **Venue** **00.00 – 00.00**

Examiners

Prof. Brett Houlding

Instructions to Candidates:

Attempt **two** questions. All questions carry equal marks. Each question is scored out of a total of 25 marks.

You may not start this examination until you are instructed to do so by the invigilator.

Materials Permitted for this examination:

Non-programmable calculators are permitted for this examination – please indicate the make and model of your calculator on each answer book used.

1. Data were recorded on the subjective assessment (on an integer scale of 0 to 20) of 54 classical painters.

Variable	**Description**
comp	A Composition score.
draw	A Drawing score.
col	A Colour score.
expr	An expression score.

The output from two principal components analyses applied on the numeric variables of the Painters data is provided on the next page.

a) Provide a brief description on the benefits of Dimension Reduction as a statistical technique.

[3 marks]

b) Assuming the raw data is stored in a matrix named **cp** (short for classical painters), state R code that would provide the required information for the output given in the two principal components analyses overleaf.

[4 marks]

c) Making reference to either of the principal components outputs, explain the interpretation of the standard deviation row and how the values are calculated.

[5 marks]

d) Provide a description of the remaining output, and hence suitable conclusions that may be drawn from the principal components analysis you consider most appropriate. Justify your choice of analysis.

[7 marks]

e) Describe in detail an alternative dimension reduction technique that could be applied for this data and explain its difference to principal components analysis.

[6 marks]

Covariance Matrix:

	comp	draw	col	expr
comp	16.70	5.87	-1.86	12.89
draw	5.87	11.95	-8.31	9.52
col	-1.85	-8.31	21.64	-4.45
expr	12.89	9.52	-4.45	23.02

Principle Components applied on the Covariance Matrix:

Rotation:

	PC1	PC2	PC3	PC4
comp	0.48	-0.38	-0.78	-0.10
draw	0.42	0.19	0.28	-0.85
col	-0.38	-0.85	0.21	-0.31
expr	0.66	-0.33	0.51	0.43

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	6.40	4.57	2.58	2.17
Proportion of Variance	0.56	0.29	0.09	0.06
Cumulative Proportion	0.56	0.85	0.94	1.00

Principle Components applied on the Correlation Matrix:

Rotation:

	PC1	PC2	PC3	PC4
comp	0.50	-0.49	0.60	0.40
draw	0.56	0.27	-0.59	0.52
col	-0.35	-0.77	-0.49	0.23
expr	0.56	-0.32	-0.25	-0.72

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	1.51	1.02	0.63	0.54
Proportion of Variance	0.57	0.26	0.10	0.07
Cumulative Proportion	0.57	0.83	0.93	1.00

2. a) Explain the difference between classification and clustering.

[2 marks]

b) State the modelling assumptions of linear and quadratic discriminant analysis and indicate one type of data such techniques are not appropriate for.

[3 marks]

c) Assuming we have a data matrix named **data** in which the first column contains a class recording and the second and third columns contain associated covariate information, state R code for training an LDA classifier from the information available in the matrix **data**.

[4 marks]

d) Bivariate data has mean $\boldsymbol{\mu}^T = (2,1)$ and covariance matrix:

$$\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}, \text{ hence } \Sigma^{-1} = \begin{pmatrix} 2/5 & -1/5 \\ -1/5 & 3/5 \end{pmatrix}$$

Determine i) the Mahalanobis distance and ii) the Euclidean distance from the point $\boldsymbol{x}^T = (2,3)$ to the mean.

[6 marks]

The United States' CIA publishes demographics of different countries through its World Factbook. In particular, the following information is made available for different countries:

Variable	**Description**
Birth	Annual number of births per 1,000 people.
Death	Annual number of deaths per 1,000 people.
Life	Life expectancy at birth (years).
Infla	Inflation rate.
GDP	GDP per capita.

Of the countries listed in the CIA data, those listed as European or African were subject to a *k*-Nearest Neighbour classification. The objective was to determine if the data variables listed would be sufficient to predict whether a country was European or African.

Output from a *k*-Nearest Neighbours using a (50%, 25%, 25%) training-validation-test split of the standardized data is provided at the end of this question. There were 49 African countries and 39 European countries.

Question continued on the next page

e) Explain the meaning and role of training, validation, and test data for calibrating *k*-Nearest Neighbours, and explain what the output below tells us about the performance of *k*-Nearest Neighbours in classifying a country as European or African based on the information available in the scaled CIA data.

[6 marks]

f) What would happen to the misclassification rate in this instance if the value of *k* continued to increase?

[3 marks]

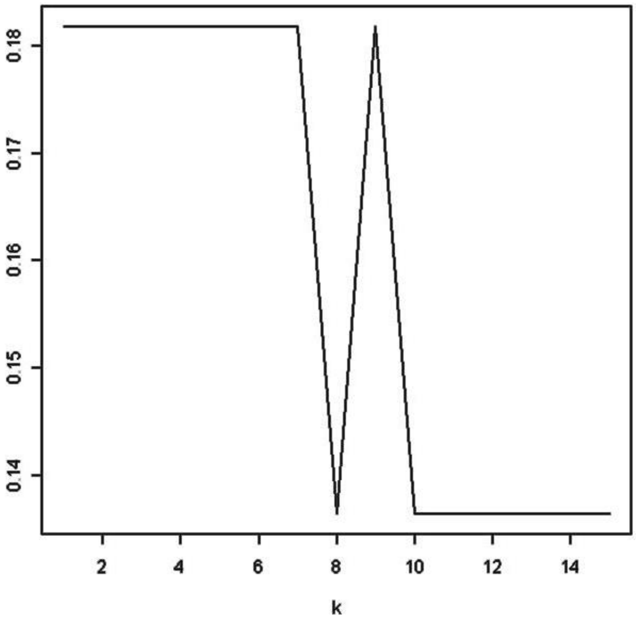
g) Other than LDA and QDA, name another classification technique that would be suitable for this analysis.

[1 mark]

Output for Question 2 e)

Training Size = 44, Validation Size = 22, Test Size = 22.

Misclassification rate at Validation Step:



Misclassification at Test Step:

k	1	2	3	4	5	6	7	8	9	10	11
Proportion Wrong	0.18	0.18	0.14	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18

3.

a) Explain the differences between hierarchical and iterative clustering algorithms.

[2 marks]

b) Detail the pseudo-code for PAM clustering. What does the acronym PAM stand for and why may a person be interested in this form of clustering? What is the downside of PAM clustering in comparison to an alternative such as *k*-Means clustering?

[6 marks]

c) A (scaled) subset of the CIA Factbook data described in question 2 e) is as follows:

Country	**Birth**	**Death**	**Life**	**Infla**	**GDP**
China	-0.41	-0.32	-1.48	1.46	-1.41
Ireland	1.49	-1.15	0.65	-0.80	0.66
UK	-0.66	1.23	0.57	-0.40	-0.02
USA	-0.41	0.25	0.25	-0.27	0.76

Using the Absolute Distance (Manhattan) dissimilarity generate a dissimilarity matrix for the subset.

[6 marks]

d) Using your answer to part c) and single linkage produce a sketch of the resulting dendrogram (you should show your work in calculating the heights at which merges occur). Hence provide a reasonable suggestion, with explanation, for the clustering solution.

[7 marks]

e) Describe a function(s) in R that could have performed the task requested in parts c) and d) above.

[4 marks]