Birkbeck

(University of London)

BSc/FD EXAMINATION

Department of Computer Science and Information Systems

INTRODUCTION TO DATA ANALYTICS USING R (BUCI045H6)

CREDIT VALUE: 15 credits

Date of examination: WEDNESDAY 8 JUNE 2016
Duration of paper: 10.00 – 13.00

RUBRIC

1. This paper contains 8 questions for a total of 100 marks.
2. Students should attempt to answer **all** of them.
3. The use of non-programmable electronic calculators is permitted.
4. This paper is not prior-disclosed.
5. Time allowed: 3 hours.

1. ..... (**15** marks)

   (a) What is unsupervised learning? Give two examples of unsupervised learning techniques. (**7** marks)

   (b) What is bias and what is variance? Give two statistical learning models where bias and variance are both lower in one model than the other. (**8** marks)

2. A sample consists of four observations: $\{2, 3, 6, 10\}$. ..... (**10** marks)

   (a) What is the unbiased sample variance? (**3** marks)

   (b) Come up with another set of 4 observations that has the same mean as the given one, but a larger variance. (**3** marks)

   (c) What is the covariance of the given set of observations and the set of observations you created? (**4** marks)

3. ..... (**11** marks)

   (a) What does PCA stand for? (**2** marks)

   (b) What can PCA be used for? (**4** marks)

   (c) How does PCA work? (**5** marks)

4. ..... (**10** marks)

   You are given a data set with 400 observations and you want to train a linear SVM, but do not know the best value for the cost parameter C.

   (a) Explain how to set the value of C using cross-validation. (**5** marks)

   (b) If you want to test $C = 0.1, 1, 10$. How many different SVMs do you need to train before you can make predictions if you use 10-fold cross-validation? Explain your answer. (**5** marks)

5. ..... (**15** marks)

   In 1965, data on the connection between radioactive waste exposure and cancer mortality were published. The data were collected from 9 counties that were located near an Atomic Energy Commission facility in Hanford, Washington.

   The data give the index of exposure and the cancer mortality rate during 1959-1964 for the nine counties affected. Higher index of exposure values represent higher levels of contamination.

	County	Name of county
Variable Description:	Exposure	Index of exposure
	Morality	Cancer mortality per 100,000 man-years

The data is as follows:

	County	Exposure	Mortality
1	Umatilla	2.49	147.1
2	Morrow	2.57	130.1
3	Gilliam	3.41	129.9
4	Sherman	1.25	113.5
5	Wasco	1.62	137.5
6	HoodRiver	3.83	162.3
7	Portland	11.64	207.5
8	Columbia	6.41	177.9
9	Clatsop	8.34	210.3

Output from fitting the simple linear regression for predicting Mortality from Exposure is shown below:

```
> lm.out=lm(Mortality ~ Exposure)
> summary(lm.out)
Call:
lm(formula = Mortality ~ Exposure)

Residuals:
     Min   1Q   Median   3Q    Max
-16.295 -12.755 4.011 9.398 18.594

Coefficients:
          Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  114.716   8.046   14.258   1.98e-06 ***
Exposure     9.231     1.419   6.507     0.000332 ***
---
Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 .  0.1 1

Residual standard error:  14.01 on 7 degrees of freedom
Multiple R-Squared:  0.8581,Adjusted R-squared:  0.8378
F-statistic:  42.34 on 1 and 7 DF, p-value:  0.0003321
```


(a) Draw the scatterplot. (**3** marks)

(b) What is the expected mortality rate for a county with an exposure index of 3? (**3** marks)

(c) Calculate two points that fall on the fitted line (and would fall in the window of the scatterplot shown), draw the two points on the scatterplot, and connect them to show the fitted line. Show your work for calculating the points. (**3** marks)

(d) Interpret the estimated slope of the fitted model. (**2** marks)

(e) Is there a significant linear relationship between Mortality and Exposure? Provide a null hypothesis, a test statistic, p-value, and conclusion. (**4** marks)

6. . . . . (**10** marks)

Suppose that we have 5 observations, for which we compute a distance matrix as follows:

	A	B	C	D	E
A	0				
B	14	0			
C	8	6	0		
D	7	2	9	0	
E	11	10	4	8	0

On the basis of the distance matrix, sketch the dendrogram that results from hierarchically clustering these 5 observations using average linkage.

7. . . . . (**12** marks)

(a) What is overfitting? (**3** marks)

(b) You are working on a particular learning task and cross-validation experiments indicate that your SVM is overfitting. Name the actions that can help decrease overfitting in an SVM. (**9** marks)

8. . . . . (**17** marks)

Given a dataset DS with 100 observations, response variable Y, and 10 predictor variables, write down your R code to

(a) build a regression tree model; (**5** marks)

(b) compute its testing MSE; (**5** marks)

(c) prune your tree to the best number of leaves; (**3** marks)

(d) make a prediction on a new test dataset TD based on the best pruned tree. (**4** marks)

Some related R-Documentation is attached for reference.

`cv.tree {tree}`

Description

Runs a K-fold cross-validation experiment to find the deviance or number of misclassifications as a function of the cost-complexity parameter k.

Usage

Some related R-Documentation is attached for reference.

```
cv.tree(object, rand, FUN = prune.tree, K = 10, ...)
```

Arguments

- object
  An object of class ”tree”.
- rand
  Optionally an integer vector of the length the number of cases used to create object, assigning the cases to different groups for cross-validation.
- FUN
  The function to do the pruning.
- K
  The number of folds of the cross-validation.

Value

A copy of FUN applied to object, with component dev replaced by the cross-validated results from the sum of the dev components of each fit.

