Birkbeck

(University of London)

MSc EXAMINATION

Department of Computer Science and Information Systems

BIG DATA ANALYTICS USING R (BUCI042H7)

CREDIT VALUE: 15 credits

Date of examination: THURSDAY 8 JUNE 2017
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. . . . . (**14** marks)

(a) What are the four scales of measurement? Which scale of measurement does the military ranks data belong to? (**4** marks)

(b) State two reasons why we may not want to use a more flexible statistical learning method even if it is more realistic. (**6** marks)

(c) What does PCA stand for? How does PCA work? (**4** marks)

2. . . . . (**8** marks)

(a) Which R command can create the following matrix? There may be more than one way to create such a matrix. Write down one possible way. (**4** marks)

```
> A
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
[4,]   13   14   15   16
```

(b) Based on the matrix A as above, use positive indices to derive the following matrix: (**2** marks)

```
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    9   10   11   12
```

(c) Based on the matrix A as above, use negative indices to derive the following matrix: (**2** marks)

```
     [,1] [,2] [,3]
[1,]    1    2    4
[2,]    9   10   12
```

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

A sample X consists of five observations: $\{2, 3, 6, 10, 4\}$. Another sample Y consists of five observations: $\{6, 4, 2, -3, 1\}$.

(a) What is the unbiased sample covariance coefficient between X and Y? (**2** marks)

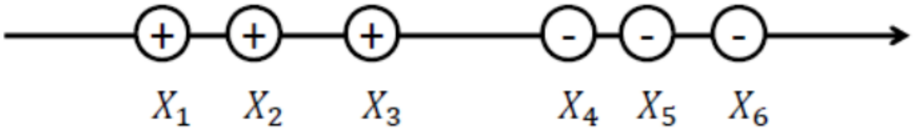
(b) What is the correlation coefficient of X and Y? (**3** marks)

(c) Comment on the correlation between X and Y. For instance, is it perfect positive correlation, or low negative correlation, or no correlation, etc? (**2** marks)

(d) Write down the commands in R that define X and Y, and calculate the unbiased sample covariance and correlation between X and Y. (**4** marks)

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

(a) Consider applying a support vector classifier (SVC) to the 1-dimensional data shown below. What will be the support vectors for the parameter cost $C = 0$ and $C = \infty$, respectively? (**4** marks)



(b) What impact will the following operation have on overfitting, increase, decrease or no impact?

(i) Increase C for support vector machines. (**2** marks)

(ii) Increase the amount of training data for logistic regression. (**2** marks)

(iii) Remove non-support vector instances in the training set for SVM. (**2** marks)

5. . . . . (**13** marks)

(a) What is a good clustering? What is a cluster centroid? (**4** marks)

(b) How does random forest improve the results from decision trees? (**4** marks)

(c) Why is logarithm transformation sometimes applied on the data prior to the analysis? (**5** marks)

6. . . . . (**11** marks)

The Swiss military carried out a study in order to analyse which soldiers are fit enough to join the special force team AAD10. In this regard, the dependent binary variable (y) reflects state of fitness of a soldier. $y = 1$ means that the soldier is fit enough for the special force team AAD10, whereas $y = 0$ indicates that the soldier is not fit enough. The following predictor variables were used for the analysis:

- x_1: The soldiers age (in years older than 18)
- x_2: The body mass index
- x_3: The average amount of sport/exercise per week (in hours)

(a) Look at the following R-Output. Write down the logistic regression model for this case. (**3** marks)

```
Coefficients:
              Estimate Std. Error z value  Pr(>|z|)
(Intercept)  -15.5543  7.2946    -2.132   0.0330
x1           -0.5859   0.3569      ???      ???
x2           0.5643    0.3317      ???      ???
x3           1.9639    0.8800      ???      ???
```

(b) Does the odds for fitness increase or decrease with soldier age? (**2** marks)

(c) We have $x_1 = 5$ and $x_2 = 25$. Which value do we have to choose for x_3 in order to get a probability of 50% for $y = 1$? (Hint: $e^0 = 1$.) (**3** marks)

(d) Suppose the dataset is called `SoldierFitness`. Write down the R command for this logistic regression model. (**3** marks)

7. . . . . (**13** marks)

Suppose that there is a book recommending system, where a score of dissimilarity is given for every pair of books in the system. If two books are the same, then the score of dissimilarity between the two books will be 0. If two books are very different, the score will be high. The higher the score is, the more different two books are.

You may choose a book in this system to start with, and enter a maximum dissimilarity score d you allow. The system will then recommend a cluster of books, each of which is at most d-dissimilar to the one you choose. In other words, the score of dissimilarity between any recommended book and the book you choose will be at most d.

The scores of dissimilarity are given as follows:

	Book A	Book B	Book C	Book D	Book E	Book F
Book A	0					
Book B	0.24	0				
Book C	0.22	0.15	0			
Book D	0.37	0.20	0.15	0		
Book E	0.34	0.14	0.28	0.29	0	
Book F	0.23	0.25	0.11	0.22	0.39	0

Suppose you choose Book D initially, and allow the maximum dissimilarity score to be 0.3 ($d = 0.3$). List all the books that will be recommended by the system.

You may use hierarchical clustering to achieve this.

(a) Choose the right linkage. (**3** marks)

(b) Sketch the dendrogram. (**7** marks)

(c) List all the books that will be recommended by the system. (**3** marks)

Justify your answers.

8. . . . . (**20** marks)

One end A of an elastic string was attached to a horizontal bar and a mass m grams, was attached to the other end B. The mass was suspended freely and allowed to settle vertically below A. The length AB, $l$mm, was recorded, for various masses as follows:

m	100	200	300	400	500	600
l	228	236	256	278	285	301

Part of the output from fitting the simple linear regression for predicting the length `l` from mass `m` is shown below:

```
Coefficients:

             Estimate   Std. Error   t value   Pr(>|t|)

(Intercept)  210.60000    4.06706     51.78     8.32e-07 ***

m              0.15257    0.01044     14.61     0.000128 ***

---

Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 .  0.1   1


Residual standard error:  4.369 on 4 degrees of freedom

Multiple R-squared:  0.9816, Adjusted R-squared:  0.977

F-statistic:  213.4 on 1 and 4 DF, p-value:  0.0001277
```

(a) Write down the least squares line of regression of l on m. (**2** marks)

(b) What is the expected length AB for a mass of 290 grams? (**2** marks)

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

(d) Is there a significant linear relationship between length l and mass m? Provide a null hypothesis, a test statistic, p-value, and conclusion. (**4** marks)

(e) How would you comment on whether this model fits the data based on the R results given above? (**2** marks)

(f) Suppose now you are given more observations and your dataset `ElasticString` has 200 observations in total. Write down your R code to

   i. build a linear regression model, and (**3** marks)

   ii. estimate the testing mean squared error. (**5** marks)