Birkbeck

(University of London)

MSc EXAMINATION

Department of Computer Science and Information Systems

BIG DATA ANALYTICS USING R

MODULE CODE: BUCI042H7

CREDIT VALUE: 15 credits

Date of examination: TUESDAY 02 JUNE 2015
Duration of paper: 10:00AM – 1:00PM

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

This question contains two multiple choice problems (a) and (b) and two true/false problems (c) and (d). Only one answer should be chosen in each multiple choice problem. In the true/false problems, if your answer is correct, you will be awarded 1 point. If your answer is incorrect, you will be awarded -1 point. If no answer has been given, you will be awarded 0 point.

(a) A sample consists of four observations: 1, 3, 5, 7. What is the standard deviation? (**1** marks)

(A) 1.12
(B) 1.49
(C) 2.24
(D) 2.58
(E) None of the above

(b) Assume you have a dataset of points in 2-dimensional space and you do a Principal Component Analysis (PCA). In Figure 1 you can see the data points as well as two of the eigenvectors `a` and `b` identified by PCA. For each of the following statements, indicate whether it is correct or not: (**2** marks)

i. Vector `a` is the dominant eigenvector (the first principal component).
ii. There can only be two eigenvectors for the given dataset.

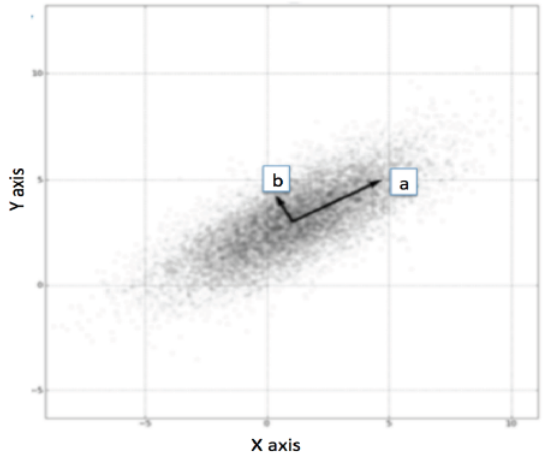

Figure 1: PCA in Question 1(c)

(c) "Score received on an exam (measured in percentage points)" (Y) is regressed on "percentage attendance" (X) for 22 students in a Big Data Analytics course. If the Y intercept $\beta_0 = 39.39$ and the slope $\beta_1 = 0.341$, which of the following statements is correct? (**1** marks)

(A) If attendance increases by 1%, the estimated average score received will increase by 39.39 percentage points.

(B) If attendance increases by 1%, the estimated average score received will increase by 0.341 percentage points.

(C) If the score received increases by 39.39%, the estimated average attendance will go up by 1%.

(D) If attendance increases by 1%, the estimated average score received will increase by 34.1 percentage points.

(d) You are working with a dataset that contains descriptions of toxic and non-toxic substances. The data, which consists of 1000 toxic examples and 1000 non-toxic examples, is described in terms of a class label and 20 numeric attributes. The values for all attributes range from 0 to 10. The data is divided into a training set and a test set of equal size. The training set and the test set thus each contain descriptions of 500 toxic and 500 non-toxic examples. Assume a decision tree has been built from the training set and that the error rates of the tree on both training and test data have been calculated. For each of the following statements, indicate whether it is correct or not: (**6** marks)

i. The error rate on the test data cannot be higher than the error rate on the training data.

ii. The error rate on the test data is likely to be higher than the error rate on the training data.

iii. The error rate on the training data is a good estimate of the error rate on the test data.

iv. When pruning the tree, the training error rate will never decrease compared to when not using pruning.

v. When pruning the tree, the test error rate will never increase compared to when not using pruning.

vi. The difference between the training and test error rates is independent of the size of the decision tree.

2. . . . . (**13** marks)

To study factors that affect the recurrence of heart attacks (HA), an investigator collected data from 20 HA victims. The investigator fit a logistic regression model with an indicator of a second HA within one year (1 = HA; 0 = no HA) as the binary outcome. There are two predictors:

- $X_1 = 1$ if the patient completed an anger management program; 0 otherwise
- X_2 = anxiety score (0 = low anxiety, 100 = high anxiety)

Computer output is given below:

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-6.36347	3.21362	-1.980	0.0477
X_1	-1.02411	1.17101	-0.875	0.3818
X_2	0.11904	0.05497	2.165	0.0304

(a) In terms of X_1 and X_2, what are the odds of a patient having a second heart attack? (**3** marks)

(b) What is the probability of a second heart attack for a patient who has completed an anger management program and scored a 100 on the anxiety test? (**3** marks)

(c) For patients who have completed the anger management program, is high anxiety associated with an increased probability of a second heart attack? Explain. (**2** marks)

(d) Is there statistical evidence that an anger management program is associated with a reduction in the probability of a second heart attack? Explain. (**2** marks)

(e) Explain why linear regression is inappropriate for modeling the probability of a second heart attack. (**3** marks)

3. . . . . (**17** marks)

Please give brief answers to the following questions.

(a) What are the four main characteristics (also known as 4Vs) in big data? (**2** marks)

(b) Cross Validation (CV) can be used to tune parameters in statistical models. Give two examples how CV is used to tune parameters. (**5** marks)

(c) It is known that k-means is sensitive to initialisation. Explain why this is so and illustrate your answer using an example. (**5** marks)

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

	A	B	C	D	E
A	0				
B	9	0			
C	3	7	0		
D	6	5	9	0	
E	11	10	2	8	0

On the basis of the similarity matrix, sketch the dendrogram that results from hierarchically clustering these 5 observations using single linkage. (**5** marks)

4. . . . . (**13** marks)

Please give brief answers to the following questions.

(a) Explain the meaning of the following terms: (**4** marks)

i. Bias

ii. Variance

(b) Bias-variance tradeoff is one of the most recurrent topics of this course. Compare the bias and variance for each of the following pairs. Briefly explain why the bias/variance of the first term is higher/lower than the bias/variance of the second term. (**6** marks)

i. LOOCV and k-fold CV

ii. Tree and its post-pruned tree

iii. Tree and tree after bagging

(c) What is overfitting? What is the relationship between overfitting and bias/variance? (**3** marks)

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

You are given the dataset presented in Figure 2. Assume that we are training an SVM with a quadratic kernel - that is, our kernel function is a polynomial kernel of degree 2 (which in this case means that the separating hyperplane is a degree 2 polynomial curve). The cost (penalty) C will determine the location of the separating hyperplane.

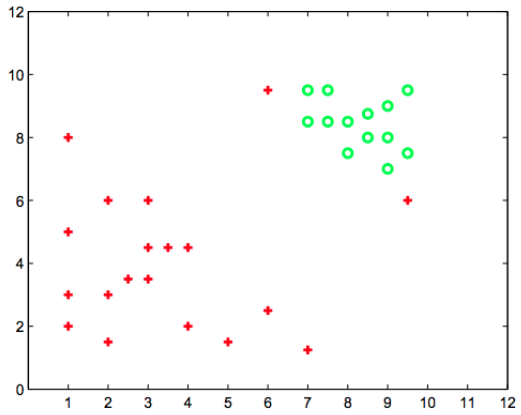

Figure 2: Dataset in Question 4

Please answer the following questions qualitatively. Give a one-sentence answer/justification for each and draw your solution in the appropriate part of the Figure.

(a) Where would the decision boundary be for very large values of C (i.e., $C \to \infty$)? (Remember that we are using an SVM with a quadratic kernel.) Draw on Figure 2. Justify your answer. (**3** marks)

(b) For $C \approx 0$, indicate in Figure 2, where you would expect the decision boundary to be. Justify your answer. (**3** marks)

(c) Which of the two cases above would you expect to work better in the classification task? Why? (**3** marks)

(d) Draw a data point which will not change the decision boundary learned for very large values of C. Justify your answer. (**2** marks)

(e) Draw a data point which will significantly change the decision boundary learned for very large values of C. Justify your answer. (**2** marks)

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

(a) How is randomisation used in two ways in constructing a random forest, given a set of attributes and a set of training examples? (**6** marks)

(b) Comparing to bagging, how does random forest improve the result? (**5** marks)

7. . . . . (**10** marks)

We usually use MSE or error rate to measure the goodness of fit (or model accuracy). Please give answers to following questions.

(a) What does MSE stand for? Write down the equation for MSE. (**2** marks)

(b) Define error rate. (**2** marks)

(c) Given a dataset `DS` with 100 observations, response variable `Y`, and predictor variable `X`, **write down your R code** to build a classification tree model and compute its testing error rate. (**6** marks)

8. . . . . (**13** marks)

Given four points $(1, 0.8)$, $(4, 4.2)$, $(5, 4.7)$ and $(7, 7.8)$, **write down your R code** to

(a) Build the linear regression model. (**4** marks)

(b) Predict the results on the new data with a sequence of 51 numbers equally spaced values starting from 0 to 8. (**4** marks)

(c) Generate the plot in Figure 3, where the curved lines are the upper (`upr`) and lower (`lwr`) bounds of the confidence and prediction intervals.

(Hint: The resulting object of the `predict` command is a matrix of predictions and bounds with column names `fit`, `upr` and `lwr`.) (**5** marks)

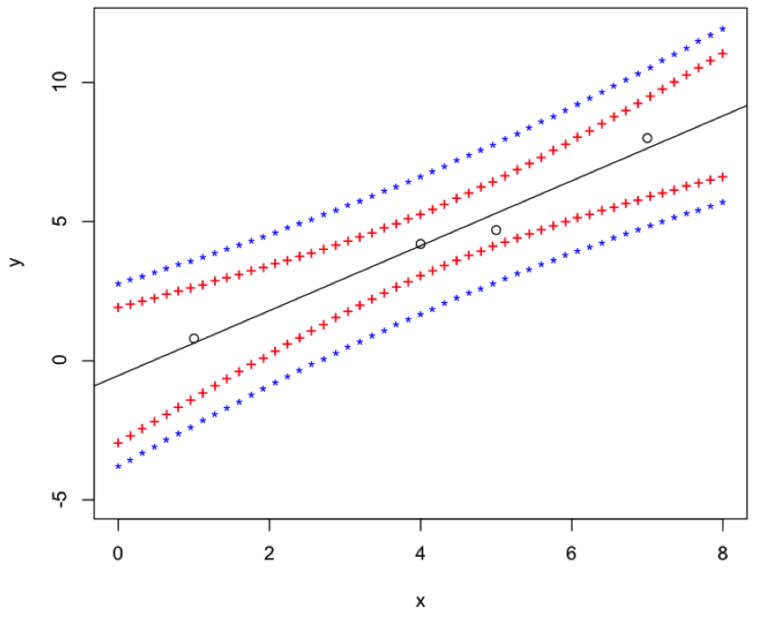

Figure 3: Plot in Question 8

