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: 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. **(4 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. **(6 marks)**

(c) What does PCA stand for? Why is scaling needed in PCA in some occasions? **(5 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)**

An insurance company has examined a random sample of 190 automobile accident claims for fraud. A logistic regression model is fitted to this data with the dependent variable being coded as one for a case that was fraudulent, and as zero otherwise. The five independent (predictor) variables included in the model are:

i. $CityCode := 1$ if the claimant lived in a large city, $:= 0$ otherwise (o.w., for short);

ii. $SexCode := 1$ for males, $:= 0$ for females;

iii. Age in years;

iv. $FaultCode := 1$ if the fault in the accident was that of the policy holder, $:= 0$ o.w.;

v. $DeductibleAmount$ (in pound sterling).

The model estimate for the logarithm of the odds of fraud is:

$$53.119 - 0.081 \times CityCode + 0.367 \times SexCode + 0.060 \times Age \\ - 1.738 \times FaultCode - 0.142 \times DeductibleAmount$$

(a) Describe in words the base case claimant whose odds for fraud are $e^{53.119}$. **(3 marks)**

(b) Do the odds for fraud increase or decrease with age? **(3 marks)**

(c) What is the probability of fraud in a claim by a male policyholder aged 30 years, who lives in a major city, has a deductible of £400 and who was not at fault in the accident? **(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 $C = 0.1, 1, 10, 100$. 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.

Variable Description:	County	Name of county
	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) What causes overfitting in a decision tree? Does overfitting increase with the number of training examples? Explain your answer. (**5** marks)

(c) 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. (**4** marks)

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

Given a dataset DS with 100 observations, response variable Y, and 15 predictor variables, write down your R code to

(a) build a bagged model; (**4** marks)

(b) compute its testing MSE; (**4** marks)

(c) find the best value of `ntree`, the number of generated trees; (**6** marks)

(d) make a prediction on a new test dataset TD based on the bagged model with the best `ntree`. (**3** marks)

Some related R-Documentation is attached for reference.

`randomForest {randomForest}`

Description

`randomForest` implements Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code) for classification and regression. It can also be used in unsupervised mode for assessing proximities among data points.

Usage

```
randomForest(x, y=NULL, xtest=NULL, ytest=NULL, ntree=500,
        mtry=if (!is.null(y) && !is.factor(y))
        max(floor(ncol(x)/3), 1) else floor(sqrt(ncol(x))),
        importance=FALSE, proximity, ...)
```

Arguments

- `data`
  an optional data frame containing the variables in the model. By default the variables are taken from the environment which randomForest is called from.

- `subset`

  an index vector indicating which rows should be used. (NOTE: If given, this argument must be named.)

- `na.action`

  A function to specify the action to be taken if NAs are found. (NOTE: If given, this argument must be named.)

- `x, formula`

  a data frame or a matrix of predictors, or a formula describing the model to be fitted (for the print method, an randomForest object).

- `y`

  A response vector. If a factor, classification is assumed, otherwise regression is assumed. If omitted, randomForest will run in unsupervised mode.

- `xtest`

  a data frame or matrix (like x) containing predictors for the test set.

- `ytest`

  response for the test set.

- `ntree`

  Number of trees to grow. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times.

- `mtry`

  Number of variables randomly sampled as candidates at each split. Note that the default values are different for classification (sqrt(p) where p is number of variables in x) and regression (p/3)

- `importance`

  Should importance of predictors be assessed?

- `proximity`

  Should proximity measure among the rows be calculated?

