STAT5104 17/18 Second term Final Examination

Answer **ALL** Questions (Time: 2 hour). Show all the detail of your calculation. Hand in this question paper together with your answer book.

Dataset for all Questions

The following dataset is selected from a large dataset from bank marketing survey with the following information:

Column	Name	Description
1	age	continuous: age of the customer
2	balance	continuous: average yearly balance, in euros
3	duration	continuous: last contact duration, in seconds
4	campaign	no of contacts performed during this campaign (numeric)
5	pdays	no of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
6	previous	no of contacts performed before this campaign and for this client (numeric)
7	poutcome	outcome of the previous marketing campaign (binary: 1=”success", 0=”unknown" or "other" or "failure")
8	deposit	has the client subscribed a term deposit? (binary: 1=yes, 0=no)

The last column (**deposit**) is considered as the target variable.

Question 1 [25%]

The following is the output from `rpart()`. Note that some numbers are missing (denoted by ?).

```
node), split, n, loss, yval, (yprob)
      * denotes terminal node
1) root     ?    ?   ?  (?  ?)
   2) duration< 633.5 41034 3529 0 (0.91399815 0.08600185)
     4) poutcome< 0.5 39699 2677 0 (0.93256757 0.06743243) *
     5) poutcome>=0.5    ?    ?   ?  (?  ?)
      10) duration< 162.5 349  108 0 (0.69054441 0.30945559) *
      11) duration>=162.5 986  242 1 (0.24543611 0.75456389) *
   3) duration>=633.5 2606  985 1 (0.37797391 0.62202609)
     6) duration< 892.5    ?    ?  ? (?  ?)
      12) poutcome< 0.5 1716  755 0 (0.56002331 0.43997669) *
      13) poutcome>=0.5 76    9 1 (0.11842105 0.88157895) *
     7) duration>=892.5 814   15 1 (0.01842752 0.98157248) *
```

(a) Fill in ALL the missing information in **root node, node 5 and node 6**. Show all the details.

(b) Write down the rule with highest support, highest confidence and highest capture respectively. Compute the support, confidence and capture of these rules.

(c) Construct the classification table for this classification tree and compute the error rate.

(d) Suppose a record is selected at random, what is the probability that **deposit=1** in this record? Furthermore, if we know that the **duration=650** in this record, then what is the probability that **deposit=1** in this record?

(e) Suppose that a new record with **duration=261** and **poutcome=0**, what is the predict probability that **deposit=1** in this new record?

Question 2 [20%]

A single hidden layer ANN with 2 neurons, logistic transfer function and **linear output** is used to predict **deposit** using variables **balance, duration, campaign and poutcome** as input (**i1=balance, i2=duration, i3=campaign, i4=poutcome**). The following is the output:

```
a 4-2-1 network with 13 weights
options were - linear output units
 b->h1 i1->h1 i2->h1 i3->h1 i4->h1
 -2.66   0.00   0.00  -0.08   1.39
 b->h2 i1->h2 i2->h2 i3->h2 i4->h2
-48.48   0.00   0.05  -0.14  41.36
  b->o  h1->o  h2->o
 -0.04   0.62   0.44
```

(a) Write down the system of equations of this 4-2-1 ANN.

(b) Suppose we have the following new observation:

balance	duration	campaign	poutcome
2143	261	1	0

What is the predicted value of **deposit** in this record?

(c) If we increase the size of the hidden layer increase from 2 to 4, what will be the number of weights in this ANN? What will be the change in the final objective function compare to the 4-2-1 ANN? Be specific.

(d) Suppose we have the following record (with poutcome missing):

balance	duration	campaign	poutcome	deposit
29	151	1	?	0

Explain in details how you could guess poutcome=0 or poutcome=1 based on this ANN.

Question 3 [25%]

The following is a logistic regression output of **deposit** on **duration** and **poutcome**:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.082      3.593e-02 -113.60   <2e-16 ***
duration     0.00565    7.792e-05   72.44   <2e-16 ***
poutcome     3.251      6.365e-02   51.07   <2e-16 ***
```

(a) Based on the above output, write down two separate logistic regression models for **poutcome=0** and **poutcome=1** respectively.

(b) Suppose we have the following new observation with **duration=261** and **poutcome=0**, what is the predict probability that **deposit=1** based on this logistic regression?

(c) Suppose we an observation with **duration=261** and **deposit=0**. Explain in details how you could guess **poutcome=0** or **poutcome=1** based on this logistic regression?

(d) Suppose we define a new variable **w=poutcome+1**, i.e. **w=1** if **poutcome=0** and **w=2** if **poutcome=1**. If we perform the logistic regression of deposit on **duration** and **w**, then what will be the maximum likelihood estimate of the coefficients? Explain your answer and show all the details of your calculation.

(e) Besides the basic assumptions of the logistic regression, what is the **additional** assumption in the logistic regression model in part (a)? How can we test whether this additional assumption is valid or not? Be specific but **do not** write any R code.

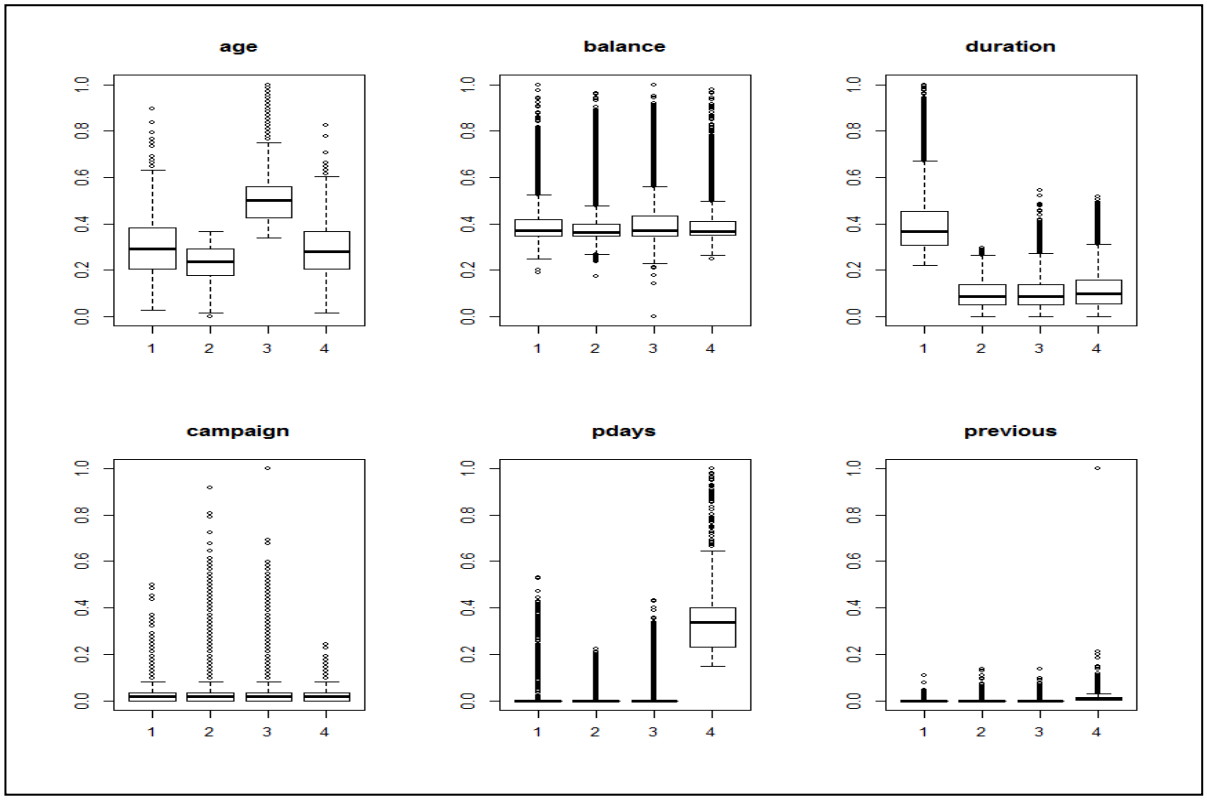
Question 4 [20%]

The first 7 columns (excluding **deposit** in the last column) in the dataset is transformed to [0,1] using range transformation. Then the K-means clustering with k=4 is performed on the transformed data and the result in saved in km4. The following are the output:

```
> km4$totss
[1] 2622.373
> km4$withinss
[1] 206.4747 196.5595 401.6597 400.4321
> km4$betweenss
[1] 1417.247
> km4$size
[1]  4934  4126 20692 13888
```

(a) Compute the R-statistic of this K-means clustering based on the above output?

(b) The following is the boxplot of the first 6 variables with the cluster label and the frequency table of **poutcome** with the cluster label:




```
> table(d$poutcome,km4$cluster)
      1     2     3     4
0  3994 20307 13447  4441
1   132   385   441   493
```

Describe the characteristic of **each** cluster based on the above outputs.

(c) Explain why we should not include **poutcome** in the boxplots? Be specific.

(d) Explain why we use the transformed data instead of the original data. Be specific.

Question 5 [10%]

The following is a frequency table output of **poutcome** by **deposit**:

	deposit=0	deposit=1
poutcome=0	37998	4191
poutcome=1	492	959

(a) Suppose we use **poutcome** to predict **deposit**, i.e.,

Rule 1: If poutcome=0 then deposit=0

Rule 2: If poutcome=1 then deposit=1

Find the support, confidence and lift value of these two rules.

(b) If we consider poutcome as a proxy for deposit, i.e., deposit=1 and poutcome=1 as true positive, deposit=0 and poutcome=0 as true negative. Compute the Recall, Precision and F1 score.

- END OF QUESTIONS -

Please return this question paper with your answer book