Hall Ticket No | | | | | | | | | | | Question Paper Code: BCS001

INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

M.Tech I Semester End Examinations (Regular) - February, 2017

Regulation: IARE–R16

FOUNDATIONS OF DATA SCIENCES
(Computer Science and Engineering)

Time: 3 Hours **Max Marks: 70**

Answer ONE Question from each Unit
All Questions Carry Equal Marks
All parts of the question must be answered in one place only

UNIT – I

1. (a) Discuss the roles of data scientist and data architects. **[7M]**

   (b) W.r.t. bar charts write a code snippet in R for **[7M]**
      i. side by side bar chart and faceted bar chart
      ii. specifying different styles of bar chart

2. (a) Assume that you have been given customer dataset file consisting of 50000 records of customers along with their income data. However, during analysis, you may observe that some values are missing systematically. How do you handle this scenario? **[7M]**

   (b) Explain for and while loop in R with syntax and example for each. **[7M]**

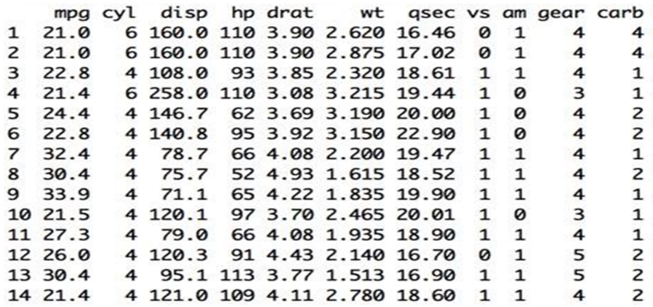
UNIT – II

3. (a) Assume an employee.xml file containing records of employee including id, name, salary, startdate and department are stored for n employees. Write a R script to get the number of nodes present in this XML file. Further, get the details of the first node and get different elements of a node. **[9M]**

   (b) What is heteroscedasticity? Explain **[5M]**

4. (a) Write a code snippet in R to create a new empty .xlsx file with one empty sheet named ‘Input’, add day and month to the empty sheet input. **[5M]**

   (b) Assume that you have been given the R built in data set mtcars. **[9M]**
   We observe that the field ”am” represents the type of transmission (auto or manual). It is a categorical variable with values 0 and 1. The miles per gallon value(mpg) of a car can also depend on it besides the value of horse power(”hp”). Suggest the best way to study the effect of the value of ”am” on the regression between ”mpg” and ”hp”.



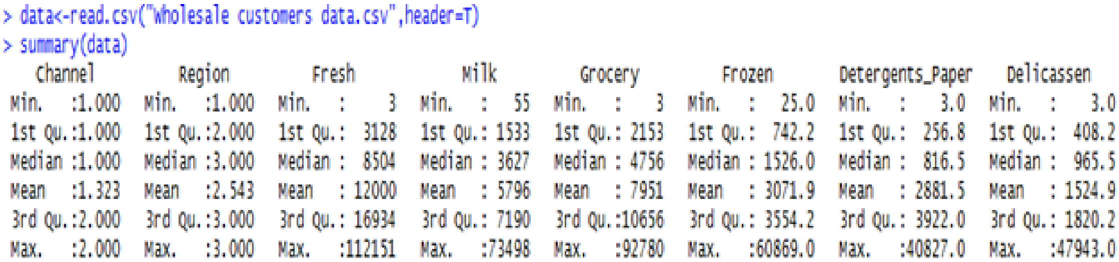
```
    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
1  21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
2  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
3  22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
4  21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
5  24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
6  22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
7  32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
8  30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
9  33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
10 21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
11 27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
12 26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
13 30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
14 21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

Figure 1

UNIT – III

5. (a) When do you use Naive Bayes, Support Vector Machines and Decision trees for classification. **[5M]**

   (b) Assume that you have been given a wholesale customer database. **[9M]**



```
> data<-read.csv("Wholesale customers data.csv",header=T)
> summary(data)
    Channel          Region          Fresh             Milk           Grocery           Frozen          Detergents_Paper    Delicassen
 Min.   :1.000   Min.   :1.000   Min.   :     3   Min.   :   55   Min.   :    3   Min.   :   25.0   Min.   :    3.0   Min.   :    3.0
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:  3128   1st Qu.: 1533   1st Qu.: 2153   1st Qu.:  742.2   1st Qu.:  256.8   1st Qu.:  408.2
 Median :1.000   Median :3.000   Median :  8504   Median : 3627   Median : 4756   Median : 1526.0   Median :  816.5   Median :  965.5
 Mean   :1.323   Mean   :2.543   Mean   : 12000   Mean   : 5796   Mean   : 7951   Mean   : 3071.9   Mean   : 2881.5   Mean   : 1524.9
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.: 16934   3rd Qu.: 7190   3rd Qu.:10656   3rd Qu.: 3554.2   3rd Qu.: 3922.0   3rd Qu.: 1820.2
 Max.   :2.000   Max.   :3.000   Max.   :112151   Max.   :73498   Max.   :92780   Max.   :60869.0   Max.   :40827.0   Max.   :47943.0
```

Figure 2

There's obviously a big difference for the top customers in each category (e.g. Fresh goes from a min of 3 to a max of 112,151). Remove the top 5 customers from each category and using custom functions create a new data set called data.rm.top and using this new data set perform cluster analysis using k-means.

6. (a) What is model overfitting? **[4M]**

   (b) What is association rules? Explain association rules mining with a suitable example. **[10M]**

UNIT – IV

7. (a) Do you think that generally, the more pieces of information (i.e., input dimensions) we add, the greater is the chance that your perceptron can do its task perfectly? Explain your answer, including a geometrical interpretation (one- vs. two- vs. three dimensional input spaces, and so on). Under what conditions does additional information help with the classification, and under what conditions does it not? **[8M]**

   (b) We have a function which takes a two-dimensional input x = (x1, x2) and has two parameters w = (w1, w2) given by f(x, w) = $\sigma\left(\sigma\left(x1w1\right)w2 + x2\right)$ where $\sigma\left(x\right) = 11 + e - x$. We use backpropagation to estimate the right parameter values. We start by setting both the parameters to 0. Assume that we are given a training point x1 = 1, x2 = 0, y = 5. Given this information answer the next two questions. **[6M]**

      i. What is the value of $\partial f/\partial w^2$?

      ii. If the learning rate is 0.5, what will be the value of w2 after one update using back propagation algorithm?

8. (a) The chart below shows a set of two-dimensional input samples from two classes: **[10M]**

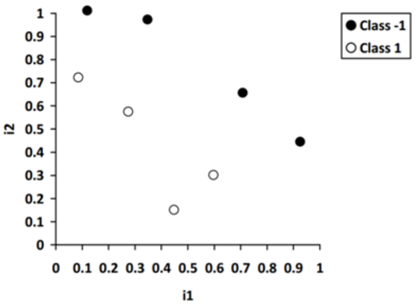


Figure 3

It looks like there exists a perfect classification function for this problem that is linearly separable, and therefore a single perceptron should be able to learn this classification task perfectly. Let us study the learning process, starting with a random perceptron with weights w0 = 0.2, w1 = 1, and w2 = -1, where of course w0 is the weight for the constant offset i0 = 1. For the inputs, just estimate their coordinates from the chart. Now add the perceptron's initial line of division to the chart. How many samples are misclassified? Then pick an arbitrary misclassified sample and describe the computation of the weight update (you can choose = 1 or any other value; if you like you can experiment a bit to find a value that leads to efficient learning). Illustrate the perceptron's new line of division and give the number of misclassified samples. Repeat this process four more times so that you have a total of six lines (or fewer if your perceptron achieves perfect classification earlier).

(b) W.r.t to question 8(a), let us assume that less information were available about the samples that are to be classified. Let us say that we only know the value for i1 for each sample, which means that our perceptron has only two weights to classify the input as best as possible, i.e., it has weights w0 and w1, where w0 is once again the weight for the constant offset i0 = 1. Draw a diagram that visualizes this one-dimensional classification task, and determine weights for a perceptron that does the task as best as possible (minimum error, i.e., minimum proportion of misclassified samples). Where does it separate the input space, and what is its error? **[4M]**

UNIT – V

9. (a) Discuss briefly the graphical facilities available in R. **[5M]**

(b) What is knitr? What is its Purpose? Write a code snippet in R to illustrate a simple LaTeX document with knitr chunks. **[9M]**

10. (a) What is the significance of presenting work of data scientists to his/her peer? Discuss the peer presentation structure. **[8M]**

(b) Explain briefly the two functions in R for representing multivariate data. **[6M]**