UECM3993 Quiz Marking Guide

Name: Student ID: Marks: /12

Course Code & Course Title: UECM3993 Predictive Modelling
Faculty: LKC FES, UTAR Course: AM, AS, FM
Session: June 2022 Lecturer: Liew How Hui

Instruction: Answer all questions in the space provided. **If you do not write your answer in the space provided, you will get ZERO mark**. An answer without necessary working steps may also receive ZERO mark.

CO4: Demonstrate supervised and unsupervised learning with statistical software

1. Write down the return value (or output) of the following (unsupervised learning) R commands.

   (a) sum(c(15,12,16,13,14)) (0.5 mark)

   Ans. 70 ..... [0.5 mark]

   (b) mean(c(20,18,13,19,15)) (0.5 mark)

   Ans. 17 ..... [0.5 mark]

   (c) sd(c(20,18,13,19,15)) (1 mark)

   Ans. 2.915476 ..... [1 mark]

   (d) table(c('B', 'B', 'A', 'A', 'B', 'B', 'A'), c('B', 'A', 'A', 'A', 'B', 'A', 'A')) (1 mark)

   Ans.

```
   A B
A 3 0
B 2 2
```

   ..... [1 mark]

   (e) matrix(seq(54,40,-2),2,4) (1 mark)

   Ans.

```
[1,]   54   50   46   42
[2,]   52   48   44   40
```

   ..... [1 mark]

2. Write a simple R script to generate the following table (with the correct data type for each column) without importing any library or reading data from any file.

```
  Grade Height Age Gender
1     B    160  16      0
2     B    159   8      0
3     B    177  11      1
4     C    138  14      1
5     B    134   2      0
6     C    158   3      0
```

Write down the R command(s) to obtain the following statistics of the table.

```
 Grade      Height            Age         Gender
 B:4    Min.   :134.0   Min.   : 2.00   0:4
 C:2    1st Qu.:143.0   1st Qu.: 4.25   1:2
        Median :158.5   Median : 9.50
        Mean   :154.3   Mean   : 9.00
        3rd Qu.:159.8   3rd Qu.:13.25
        Max.   :177.0   Max.   :16.00
```

(2 marks)

Ans. A sample R script is listed below.

```
d.f = data.frame(
Grade = c("B", "B", "B", "C", "B", "C"),
Height = c(160,159,177,138,134,158),
Age = c(16,8,11,14,2,3),
Gender = c(0,0,1,1,0,0) )                                    [1 mark]
d.f$Grade = factor(d.f$Grade)
d.f$Gender = factor(d.f$Gender)                              [0.5 mark]
summary(d.f)                                                  [0.5 mark]
```

3. Given the training data with features X_1, X_2 and the label Y in Table 3.1.

X_1	X_2	Y
6.8	25.05	1
6.5	26.1	1
5.25	26.75	1
7.5	25.65	1
10.6	18.9	2
11.65	17.45	2
4.45	19.9	2

Table 3.1: Training data with features X_1, X_2 and a label Y.

Given $X_1 = 6.15$ and $X_2 = 26.9$. Use the Euclidean distance and the supervised learning model kNN to predict Y for $k = 3$ and $k = 5$, respectively. ..... (3 marks)

Ans. After calculating the distances of the point we want to predict to each point in the training data, we obtain the following table.

X_1	X_2	dist	rank	Y
6.8	25.05	1.960867	4	1
6.5	26.1	0.873212	1	1
5.25	26.75	0.912414	2	1
7.5	25.65	1.839837	3	1
10.6	18.9	9.154371	6	2
11.65	17.45	10.934007	7	2
4.45	19.9	7.203471	5	2

..... [2 marks]

The prediction with kNN (k=3) is 1 ..... [0.5 mark]

The prediction with kNN (k=5) is 1 ..... [0.5 mark]

4. Given the insurance data with the following features

   - `age`: age of the policyholder
   - `sex`: gender of the policyholder (female=0, male=1)
   - `bmi`: body mass index, body weight to square of the height ratio, ideally 18.5 to 25
   - `children`: number of children/dependents of the policyholder
   - `smoker`: smoking state of the policyholder (non-smoker=0, smoker=1)
   - `region`: the residential area of policyholder in the USA (northeast=0, northwest=1, southeast=2, southwest=3)
   - `charges`: individual medical costs billed by the health insurance

   and the labelled output is `insuranceclaim`, 1 for valid insurance claim and 0 for invalid insurance claim. Suppose the trained supervised learning logistic regression model has the following analysis result.

```
Call:
glm(formula = insuranceclaim ~ ., family = binomial, data = data.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0137  -0.5764   0.1023   0.5095   3.2432

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.188e+00  7.310e-01  -9.833  < 2e-16 ***
age          2.620e-02  9.395e-03   2.788   0.0053 **
sex1         3.960e-02  1.946e-01   0.203   0.8387
bmi          2.702e-01  2.326e-02  11.612  < 2e-16 ***
children1   -2.101e+00  2.516e-01  -8.352  < 2e-16 ***
children2   -3.722e+00  3.179e-01 -11.708  < 2e-16 ***
children3   -4.759e+00  4.389e-01 -10.842  < 2e-16 ***
children4   -5.623e+00  8.721e-01  -6.448 1.13e-10 ***
children5   -3.692e+00  9.185e-01  -4.020 5.82e-05 ***
smoker1      4.098e+00  5.067e-01   8.089 6.03e-16 ***
region1     -4.436e-01  2.759e-01  -1.608   0.1079
region2     -6.801e-01  2.870e-01  -2.369   0.0178 *
region3     -3.765e-01  2.727e-01  -1.381   0.1673
charges      1.178e-05  1.933e-05   0.609   0.5423
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1270.09  on 935  degrees of freedom
Residual deviance:  669.93  on 922  degrees of freedom
AIC: 697.93

Number of Fisher Scoring iterations: 6
```

By calculating the conditional probabilty $P(Y = 1|X = x)$, predict whether the insurance claim is valid or invalid when a female policyholder of age 61, of bmi 39.1, having two children, is a non-smoker, living in the southwest of USA and has a medical costs bill of 14235.07. In addition, compare the odds of insurance claim for a policyholder of bmi 25 against a policyholder of bmi 18.5 by computing the odds ratio and then compare the probabilities getting valid policy claim in the two cases. (3 marks)

Ans.

age	0.0261958855	61.00
sex1	0.0395991299	0.00
bmi	0.2701582775	39.10
children1	-2.1011023960	0.00
children2	-3.7218081674	1.00
children3	-4.7586887311	0.00
children4	-5.6234310023	0.00
children5	-3.6924041239	0.00
smoker1	4.0983992038	0.00
region1	-0.4436383930	0.00
region2	-0.6801038451	0.00
region3	-0.3765294689	1.00
charges	0.0000117807	14235.07

$\beta^T x = 1.042339$

The conditional probability is 0.739301 ..... [1.5 marks]

The insurance claim is predicted to be valid. ..... [0.5 mark]

The odds ratio is

$$\frac{odds(bmi = 25)}{odds(bmi = 18.5)} = \exp(0.2701583 \times (25 - 18.5)) = 5.789401 \quad \text{[0.5 mark]}$$

A policyholder of bmi 25 has higher probability of getting a valid insurance claim than a policyholder of bmi 18.5. ..... [0.5 mark]