UECM3993 Practical Quiz Marking Guide

Name: Student ID: Marks: /12

Course Code & Course Title: UECM3993 Predictive Modelling
Faculty: LKC FES, UTAR   Course: AM, AS, FM
Session: Jan 2024   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 of the R command: `paste(LETTERS[6:13], collapse="-")` .

   (0.5 mark)

   Ans.   `F-G-H-I-J-K-L-M` ....[0.5 mark]

2. Write down the return value of the R command: `sum((c(34, 32, 29, 23, 45, 44)-mean(c(34, 32, 29, 23, 45, 44)))**2)/(6-1)` . Then write down an R statistical command which gives the same result. (1.5 mark)

   Ans. 73.9 ....[1.2 mark]

   var(c(34, 32, 29, 23, 45, 44)) ....[0.3 mark]

3. Write down the return value of the R command: `table(c('C3', 'D1', 'A1', 'D1', 'A1', 'D1', 'A1', 'C3', 'D1', 'C3', 'C3'), c('D1', 'C3', 'A1', 'C3', 'C3', 'D1', 'A1', 'A1', 'C3', 'C3', 'D1'))` .

   Note that R sort labels in dictionary order. (1 mark)

   Ans.

   ```
        A1 C3 D1
    A1   2  1  0
    C3   1  1  2
    D1   0  3  1
   ```

   ....[1 mark]

4. Write down the return value of the R command: `matrix(seq(74,41,-3),2,6)`   (1 mark)

   Ans.

   ```
        [,1] [,2] [,3] [,4] [,5] [,6]
   [1,]   74   68   62   56   50   44
   [2,]   71   65   59   53   47   41
   ```

   ....[1 mark]

5. 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.

```
  Age Grade Gender Height
1  11    B      1    196
2   5    B      0    135
3   7    C      0    124
4  18    A      0    194
5  16    B      1    141
6   6    C      1    133
```

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

```
     Age          Grade  Gender      Height
 Min.   : 5.00   A:1    0:3    Min.   :124.0
 1st Qu.: 6.25   B:3    1:3    1st Qu.:133.5
 Median : 9.00   C:2           Median :138.0
 Mean   :10.50                 Mean   :153.8
 3rd Qu.:14.75                 3rd Qu.:180.8
 Max.   :18.00                 Max.   :196.0
```

Then apply min-max scaling to the numeric features and one-hot encoding for the categorical features to obtain the following table.

```
  (Intercept)        Age GradeB GradeC Gender1    Height
1           1 0.46153846      1      0       1 1.0000000
2           1 0.00000000      1      0       0 0.1527778
3           1 0.15384615      0      1       0 0.0000000
4           1 1.00000000      0      0       0 0.9722222
5           1 0.84615385      1      0       1 0.2361111
6           1 0.07692308      0      1       1 0.1250000
```

(2 marks)

Ans. A sample R script is listed below.

```
d.f = data.frame(
Age = c(11,5,7,18,16,6),
Grade = c("B", "B", "C", "A", "B", "C"),
Gender = c(1,0,0,0,1,1),
Height = c(196,135,124,194,141,133) )              [0.5 mark]
d.f$Grade = factor(d.f$Grade)
d.f$Gender = factor(d.f$Gender)                    [0.5 mark]
summary(d.f)                                       [0.5 mark]
d.f$Age = (d.f$Age-min(d.f$Age))/diff(range(d.f$Age))
d.f$Height = (d.f$Height-min(d.f$Height))/diff(range(d.f$Height))
print(as.data.frame(model.matrix(~., d.f)))        [0.5 mark]
```

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

X_1	X_2	Y
-1.07	0.14	0
-0.37	0.73	0
-0.24	0.78	0
0.71	-0.8	1
0.13	-0.38	1
1.59	-0.63	1
1.72	0.36	1
2.02	0.12	1

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

Given $X_1 = 1.29$ and $X_2 = -0.21$. Use the Euclidean distance and the supervised learning model kNN to find the probabilities of $Y = 0$ and $Y = 1$ and predict Y for $k = 2$ and $k = 5$, respectively. Note that if there is a tie in the probabilities, R will choose the outcome based dictionary ordering. ....(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
-1.07	0.14	2.385812	8	0
-0.37	0.73	1.907669	7	0
-0.24	0.78	1.822361	6	0
0.71	-0.8	0.827345	4	1
0.13	-0.38	1.172391	5	1
1.59	-0.63	0.51614	1	1
1.72	0.36	0.714003	2	1
2.02	0.12	0.801124	3	1

....[2.2 marks]

$$P_{k=2}(Y = 0) = 0; \qquad P_{k=2}(Y = 1) = 1$$
$$P_{k=5}(Y = 0) = 0; \qquad P_{k=5}(Y = 1) = 1$$

....[0.2+0.2=0.4 mark]

The prediction with kNN (k=2) is 1 ....[0.2 mark]

The prediction with kNN (k=5) is 1 ....[0.2 mark]

7. Given the Irish education transitions data with the following features

- **Sex**: gender of the Irish schoolchildren aged 11 in 1967 (`female`, `male`);
- **DVRT**: Drumcondra Verbal Reasoning Test Score (integers);
- **PS**: Prestige score (integers);
- **TS**: Type of school (`primary_terminal_leaver`, `secondary`, `vocational`);

and the labelled output is the `Leaving.Certificate` with two values — `not.taken` ($Y = 0$) or `taken` ($Y = 1$). Suppose the trained supervised learning logistic regression model has the following analysis result.

```
Call:
glm(formula = Leaving.Certificate ~ Sex + DVRT + PS + TS, family = binomial,
    data = data.train)

Coefficients:
              Estimate Std. Error  z value Pr(>|z|)
(Intercept) -21.672763 730.716488  -0.030 0.976339
Sexmale       0.004639   0.275457   0.017 0.986563
DVRT          0.038905   0.010020   3.883 0.000103 ***
PS            0.025408   0.009013   2.819 0.004818 **
TSsecondary  17.238768 730.715962   0.024 0.981178
TSvocational 14.373213 730.716053   0.020 0.984307
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 522.87  on 378  degrees of freedom
Residual deviance: 346.59  on 373  degrees of freedom
  (21 observations deleted due to missingness)
AIC: 358.59

Number of Fisher Scoring iterations: 16
```

(a) Determine the odds ratio for the `Leaving.Certificate` of the prestige score when it is increment by 1 from some unknown value x and then determine the ordering of the probabilities of $Y = 1$ for a prestige score x and for a prestige score $x + 1$. ....(1 mark)

(b) For a male student with a Drumcondra Verbal Reasoning Test Score of 123 and a pretige score of 43 with a type school from secondary, calculate the conditional probability and the odds of the Leaving Certificate is taken given the conditions. ....(1 mark)

(c) For a female student with a Drumcondra Verbal Reasoning Test Score of 114 and a pretige score of 43 with a type school from secondary, calculate the conditional probability and predict whether the Leaving Certificate is taken or not given the conditions. ....(1 mark)

Ans.

$$\frac{Odds(PS = x + 1)}{Odds(PS = x)} = \exp(0.02540801) = 1.025734 \qquad \text{[0.5 mark]}$$

$$P(Y = 1|PS = x + 1) > P(Y = 1|PS = x) \qquad \text{[0.5 mark]}$$

Part	Sexmale	DVRT	PS	TSsecondary	TSvocational	prob	prediction	mark
(b)	1	123	43	1	0	0.809766	4.256685	[1 mark]
(c)	0	114	43	1	0	0.749079	taken	[1 mark]
-21.6728	0.0046	0.0389	0.0254	17.2388	14.3732	NA	NA	-