# UECM3993 Quiz Jan 2021 Marking Guide

Marks Allocation: First 10 questions, each 0.5; Next 10 questions, each 0.7. Total marks: 12 marks

1. Which ONE of the following course learning outcome of UECM3993 Predictive Modelling is most relevant when R (or Python) programming environment is used to carry out statistical computation and predictive modelling?
   (a) Describe the key concept of statistical learning ⋉
   (b) Compare statistical models for prediction and estimation through supervised learning ⋉
   (c) Identify relationship and structures from unlabelled data through unsupervised learning ⋉
   (d) Demonstrate supervised and unsupervised learning with statistical software ✓
   (e) Interpret results from supervised and unsupervised learning ⋉

2. Which ONE of the following command is used for the (unweigthed) kNN supervised learning for classification problems in the statistical software R?
   (a) summary ⋉
   (b) pairs ⋉
   (c) class::knn ✓
   (d) FNN::knn.reg ⋉
   (e) glm ⋉

3. Which ONE of the following supervised learning method is NOT used for regression problems?
   (a) logistic regression ✓
   (b) linear regression ⋉
   (c) kNN ⋉
   (d) regression tree ⋉
   (e) elasticnet ⋉

4. Which ONE of the following data can be used directly for supervised learning in R?
   (a) Microsoft Excel documents ⋉
   (b) data frame ✓
   (c) Microsoft Word documents ⋉
   (d) pictures and images ⋉

(e) PDF documents ⋉

5. Using the following code, answer the questions below:

```
df <- data.frame(a = c(1, 2, 3, 4, 5),
                 b = c(6, 7, 8, 9, 10))
```

Provide the SINGLE line of code that will create a column named 'c' in this data frame that is the sum of coulumn 'a' and column 'b':

(a) c = a + b ⋉

(b) c = df$a + df$b ⋉

(c) cbind(df$a, df$b, df$c) ⋉

(d) df$c = df$a + df$b ✓

(e) none of the above ⋉

6. Which ONE of the following command can be used to count number of missing values in the categorical data x in R?

(a) length(x == NA) ⋉

(b) count(x == NA) ⋉

(c) sum(x == NA) ⋉

(d) length(is.na(x)) ⋉

(e) sum(is.na(x)) ✓

7. Which ONE of the following command can be used to summarise the categorical data x in which class(x) is integer in R?

(a) summary(x) ⋉

(b) summary(sort(x)) ⋉

(c) summary(factor(x)) ✓

(d) summary(category(x)) ⋉

(e) sum(x) ⋉

8. The Auto data frame has a column for horsepower. Which R statement correctly orders the data frame by horsepower in descending order?

(a) Auto[ sort(-Auto$horsepower),] ⋉

(b) Auto[ sort( Auto$horsepower,decreasing=TRUE),] ⋉

(c) Auto[ sort( Auto$horsepower,descending=TRUE),] ⋉

(d) Auto[order(-Auto$horsepower),] ✓

(e) Auto[order( Auto$horsepower),] ⋉

9. Given a matrix $A$ and a matrix $B$. The matrix multiplication in the R statistical software can be performed in a single command using

(a) matrix.multiply(A, B) ⋉

(b) A.dot(B) ⋉

(c) A * B ⋉

(d) A @ B ⋉

(e) A %*% B ✓

10. The supervised learning for the Auto data with R statistical software is demonstated below:

```
Call:
lm(formula = mpg ~ horsepower + weight, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-11.0762  -2.7340  -0.3312   2.1752  16.2601

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.6402108  0.7931958  57.540  < 2e-16 ***
horsepower  -0.0473029  0.0110851  -4.267 2.49e-05 ***
weight      -0.0057942  0.0005023 -11.535  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.24 on 389 degrees of freedom
Multiple R-squared:  0.7064,    Adjusted R-squared:  0.7049
F-statistic: 467.9 on 2 and 389 DF,  p-value: < 2.2e-16
```

The mpg for a car with a horsepower of 90 and a weight 2500kg is estimated to be

(a) 18.7428 ⋉

(b) 26.8975 ✓

(c) 31.1548 ⋉

(d) 41.3830 ⋉

(e) 45.6402 ⋉

11. Suppose the prediction of the supervised learning linear SVM model is given by the following table:

| Predicted | Actual |
|---|---|
| virginica | virginica |
| versicolor | versicolor |
| virginica | virginica |
| setosa | setosa |
| setosa | setosa |
| versicolor | versicolor |
| versicolor | versicolor |
| setosa | setosa |
| virginica | virginica |
| versicolor | versicolor |
| setosa | setosa |
| virginica | versicolor |
| versicolor | versicolor |
| setosa | setosa |
| virginica | virginica |
| virginica | virginica |
| setosa | setosa |
| virginica | virginica |
| virginica | virginica |

The accuracy of the supervised learning model is

(a) 20 ⋉

(b) 19 ⋉

(c) 1 ⋉

(d) 0.05 ⋉

(e) 0.95 ✓

12. Given a linear regression model with the following parameters.

```
Call:
lm(formula = Rings ~ Length + Diameter + Height + WholeWeight,
    data = d.f.train)

Residuals:
     Min       1Q   Median       3Q      Max
-18.6361  -1.6027  -0.5874   0.9520  15.3239

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1270     0.3582   8.730  < 2e-16 ***
Length      -16.0830     2.5451  -6.319 3.03e-10 ***
Diameter     30.3546     3.1127   9.752  < 2e-16 ***
Height       17.5085     1.9446   9.003  < 2e-16 ***
WholeWeight   0.4485     0.2598   1.726   0.0844 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.553 on 2918 degrees of freedom
Multiple R-squared:  0.3819,    Adjusted R-squared:  0.3811
F-statistic: 450.8 on 4 and 2918 DF,  p-value: < 2.2e-16
```

Calculate $\mathbb{E}[Y|X = x]$ given the inputs $x$ are the Length is 0.35, the Diameter is 0.265, the Height is 0.09 and the WholeWeight is 0.2255.

(a) 0.9993 ⋉

(b) 3.1270 ⋉

(c) 4.0918 ⋉

(d) 5.1918 ⋉

(e) 7.2188 ✓

13. Given a k-nearest neighbour model with $k = 1$ and the following data.

| Length | Diameter | Height | WholeWeight | old |
|---|---|---|---|---|
| 0.615 | 0.495 | 0.200 | 1.2190 | TRUE |
| 0.425 | 0.340 | 0.105 | 0.3890 | FALSE |
| 0.535 | 0.430 | 0.140 | 0.7165 | FALSE |
| 0.430 | 0.350 | 0.110 | 0.4060 | TRUE |
| 0.650 | 0.510 | 0.175 | 1.1550 | TRUE |
| 0.575 | 0.475 | 0.160 | 0.8950 | FALSE |
| 0.420 | 0.325 | 0.100 | 0.3680 | TRUE |
| 0.405 | 0.305 | 0.100 | 0.2710 | FALSE |
| 0.465 | 0.405 | 0.135 | 0.7775 | TRUE |
| 0.530 | 0.415 | 0.110 | 0.5745 | FALSE |

Assume that the usual Euclidean distance is used, Length, Diameter, Height and WholeWeight are the input variables and the 'old' is the target variable. Find the performance accuracy if the testing data are given below.

| Length | Diameter | Height | WholeWeight | old |
|---|---|---|---|---|
| 0.530 | 0.420 | 0.135 | 0.6770 | FALSE |
| 0.440 | 0.365 | 0.125 | 0.5160 | TRUE |
| 0.330 | 0.255 | 0.080 | 0.2050 | FALSE |
| 0.530 | 0.415 | 0.150 | 0.7775 | TRUE |
| 0.535 | 0.405 | 0.145 | 0.6845 | TRUE |
| 0.470 | 0.355 | 0.100 | 0.4755 | TRUE |
| 0.565 | 0.440 | 0.155 | 0.9395 | TRUE |
| 0.560 | 0.440 | 0.140 | 0.9285 | TRUE |
| 0.665 | 0.525 | 0.165 | 1.3380 | TRUE |
| 0.355 | 0.290 | 0.090 | 0.3275 | FALSE |

(a) 0.5 ⋉

(b) 0.6 ✓

(c) 0.7 ⋉

(d) 0.8 ⋉

(e) 0.9 ⋉

14. Given a logistic regression model with the following parameters.

```
all:
glm(formula = old ~ Sex + Length + Diameter + Height + WholeWeight,
    family = binomial, data = d.f.train)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.9331  -0.7351  -0.1455   0.7859   2.3781

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.45598    0.53551  -8.321  < 2e-16 ***
SexI        -0.79309    0.13048  -6.078 1.21e-09 ***
SexM        -0.02155    0.10772  -0.200 0.841424
Length      -9.16262    2.60904  -3.512 0.000445 ***
Diameter    15.53750    3.12630   4.970 6.70e-07 ***
Height      17.95518    3.07036   5.848 4.98e-09 ***
WholeWeight  0.70865    0.34656   2.045 0.040872 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4052.1  on 2922  degrees of freedom
Residual deviance: 2849.7  on 2916  degrees of freedom
AIC: 2863.7
```

---

Calculate $\mathbb{P}(Y = 1|X = x)$ given the inputs $x$ are the Sex is M, the Length is 0.35, the Diameter is 0.265, the Height is 0.09 and the WholeWeight is 0.2255.

(a) 0.011476 ⋉

(b) 0.145581 ⋉

(c) 0.14292 ✓

(d) 0.85708 ⋉

(e) 0.854419 ⋉

15. Given a logistic regression model with the following parameters.

---

```
all:
glm(formula = old ~ Sex + Length + Diameter + Height + WholeWeight,
    family = binomial, data = d.f.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.9331  -0.7351  -0.1455   0.7859   2.3781

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.45598    0.53551  -8.321  < 2e-16 ***
SexI        -0.79309    0.13048  -6.078 1.21e-09 ***
SexM        -0.02155    0.10772  -0.200 0.841424
Length      -9.16262    2.60904  -3.512 0.000445 ***
Diameter    15.53750    3.12630   4.970 6.70e-07 ***
Height      17.95518    3.07036   5.848 4.98e-09 ***
WholeWeight  0.70865    0.34656   2.045 0.040872 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4052.1  on 2922  degrees of freedom
Residual deviance: 2849.7  on 2916  degrees of freedom
AIC: 2863.7
```

The input and output should be clear from the listing above and the previous question. Given the following test set:

| | Sex | Length | Diameter | Height | WholeWeight | old |
|---|---|---|---|---|---|---|
| 1 | F | 0.530 | 0.420 | 0.135 | 0.6770 | FALSE |
| 2 | M | 0.440 | 0.365 | 0.125 | 0.5160 | TRUE |
| 3 | I | 0.330 | 0.255 | 0.080 | 0.2050 | FALSE |
| 4 | F | 0.530 | 0.415 | 0.150 | 0.7775 | TRUE |
| 5 | F | 0.535 | 0.405 | 0.145 | 0.6845 | TRUE |
| 6 | F | 0.470 | 0.355 | 0.100 | 0.4755 | TRUE |
| 7 | F | 0.565 | 0.440 | 0.155 | 0.9395 | TRUE |
| 8 | F | 0.560 | 0.440 | 0.140 | 0.9285 | TRUE |
| 9 | M | 0.665 | 0.525 | 0.165 | 1.3380 | TRUE |
| 10 | M | 0.355 | 0.290 | 0.090 | 0.3275 | FALSE |

The accuracy of this problem is

(a) 0.5 ⋉

(b) 0.6 ⋉

(c) 0.7 ✓

(d) 0.8 ⋉

(e) 1 ⋉

16. A distance metric is an essential measure in both supervised and unsupervised learning methods. Which of the following function is not a distance metric for the vectors x and y?

(a) d(x,y) = 1 for all x and y ✓

(b) d(x,y) = 0 if x = y. Otherwise, d(x,y) = 1 ⋉

(c) Minkowski ⋉

(d) Jaccard ⋉

(e) Hamming ⋉

17. Which of the following will be the Manhattan distance between the two data points [-9, 8, 1, 4] and [1, 4, -6, 9]?

(a) 4 ⋉

(b) 26 ✓

(c) 13.784049 ⋉

(d) 7.745967 ⋉

(e) 11.527972 ⋉

18. Which of the following is the Minkowski distance with power $r = 3$ between two data points [4,5,6,7,8,9] and [18,17,16,15,14,13]?

(a) 1842.1070844348967 ⋉

(b) 54 ⋉

(c) 23.579652 ⋉

(d) 18.433901 ✓

(e) 14 ⋉

19. Which of the following is IMPOSSIBLE to obtain from the Minkowski distance measure with appropriate power $r$ between two data points [3,,2,,4,,20] and [1,,6,,5,,7]?

(a) 0.447104 ✓

(b) 20 ⋉

(c) 13.784049 ⋉

(d) 13.030954 ⋉

(e) 13 ⋉

20. Which of the following is usually NOT regarded as part of unsupervised learning?

(a) association rules ⋉

(b) clustering ⋉

(c) dimensionality reduction ⋉

(d) visualisation ⋉

(e) reinforcement learning ✓