Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Faculty of Engineering, Mathematics and Science

School of Computer Science & Statistics

MSc Computer Science **Semester 2, 2023**
2022-2023

Data Analytics – CS7DS1

Venue: RDS, Simmonscourt

3rd May, 2023 **09:30 – 11:30**

Dr. Bahman Honari

Instructions to Candidates:

This exam is 30% of your total grade.
Attempt **All** questions.
Answer all questions in this exam booklet and return it.
You can use non-programmable calculators if required.

QUESTION 1

In a data set, it seems because of a technical issue, the data of variable X1 for every fifth case is missing. This indicates a missingness mechanism that is called:

- a. At Random
- b. Not At Random
- c. Missingness depends on an unobserved factor
- d. Completely At Random

3 marks

QUESTION 2

In a dataset, it seems that, data missingness rate in variable X1 depends on the levels in the categorical variable X2. This indicates the missingness mechanism of

- a. Not At Random
- b. Depends on unobserved variable
- c. At Random
- d. Completely At Random

3 marks

QUESTION 3

In which one of the following methods, the concepts of "distance" and "similarity" of the cases is used in

1. Rejection Sampling
2. Clustering Analysis
3. Missing Data Imputation for a categorical variable
4. Regression Tree
5. Classification Tree

- a. 1
- b. 2
- c. 1 and 2
- d. 1, 2 and 3
- e. 1 and 3
- f. 2 and 3
- g. 2, 3 and 5
- h. 2, 4 and 5

5 marks

QUESTION 4

In "Rejection Sampling" method to take samples from Random Variable X with probability density function f(x), the samples are chosen from another random variable with pdf g(x). Which of the followings are correct? (You may need to choose more than one choice)

- a. Samples x_is from g(x) are decided to be rejected based on comparing $f(x_i)$ to $g(x_i)$.
- b. h(x) is defined as mg(x) where m=max [g(x)/f(x)]
- c. h(x) is defined as mg(x) where m=max [f(x)/g(x)]
- d. The y_i for each sample is chosen from h(x) distribution.
- e. Samples x_is from g(x) are decided to be rejected based on comparing the random sample $y_i \sim U(0, h(x_i))$ to $f(x_i)$.

4 marks

QUESTION 5

Which one is correct:

1. In bagging, individual trees are built independently of each other.

2. Bagging is the method for improving the performance by aggregating the results of weak learners.

- a. 1
- b. 2
- c. Both
- d. None

2 marks

QUESTION 6

Which of the following is true about each individual tree in Random Forest?

1. Individual tree is built on a subset of the features
2. Individual tree is built on all the features
3. Individual tree is built on a subset of observations
4. Individual tree is built on full set of observations

- a. 1 & 3
- b. 2 & 3
- c. 2 & 4
- d. 1 & 4

2 marks

QUESTION 7

Which of the following is true about “max_depth” hyperparameter in rpart ?
1. Lower is better parameter in case of same validation accuracy.
2. Higher is better parameter in case of same validation accuracy.
3. Increase the value of max_depth may overfit the data.
4. Increase the value of max_depth may underfit the data.

- a. 1 & 4
- b. 2 & 4
- c. 2 & 3
- d. 1 & 3

3 marks

QUESTION 8

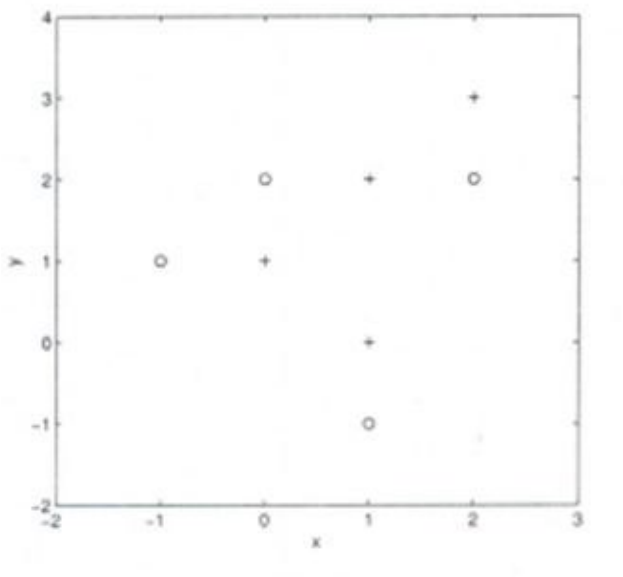
Which of the following algorithm are NOT an example of ensemble learning algorithm?

- a. LASSO
- b. Gradient Boosting
- c. Random Forest
- d. Adaboost

2 marks

QUESTION 9

Based on the scatter plot below, if you want to predict the class of new data point at (1,1) using Euclidian distance in 4-NN, which class this data point belong to?



- a. o class
- b. + class
- c. Cannot say

4 marks

QUESTION 10

You have given the following 2 statements, find which of these options is/are true in case of k-NN?

1. In case of very large value of k, we may include points from other classes into the neighbourhood.

2. In case of too small value of k the algorithm is very sensitive to noise

- a. 1
- b. 2
- c. Both
- d. None

2 marks

QUESTION 11

minsplit parameter in rpart indicates

- a. The minimum number of splits in a decision tree.
- b. The minimum number of observations required in a node to attempt the next split.
- c. The minimum number of observations in a node after splitting its parent node.
- d. None.

3 marks

QUESTION 12

LASSO regression is an example of ..... and aims to overcome the problem of ..... .

- a. regularization, underfitting
- b. cross validation, underfitting
- c. regularization, overfitting
- d. cross validation, overfitting

2 marks

QUESTION 13

Increasing which parameter may cause overfitting:

1. minsplit
2. minbucket
3. maxdepth

- a. 1
- b. 1 & 3
- c. 2 & 3
- d. 1 & 2
- e. 1 & 2 & 3
- f. 3
- g. 2

4 marks

QUESTION 14

The result of a Chi-square test below indicates there is no association between X1 and Y.

X1/Y	1	2	3	4	*RowTotl*
1	23	5	19	4	*51*
2	12	2	15	13	*42*
Coltotl	*35*	*7*	*34*	*17*	*93*

Chi^2 = 9.193281 d.f.= 3 (p=0.02682849)

- a. False
- b. True

2 marks

QUESTION 15

In a classification tree, the variable that appears as the first split, reduces the GINI index ..... than other variables.

- a. less
- b. more

2 marks

QUESTION 16

Which one of the following assumptions are required in a Regression Tree:

1. Normality of the error term
2. Equality of Variances of response for different levels of the predictor
3. Linearity
4. Low correlation between predictors

- a. 1
- b. 2
- c. 3
- d. 4
- e. None

2 marks

QUESTION 17

In a regression tree with the following output, how many numbers of splits is suggested?

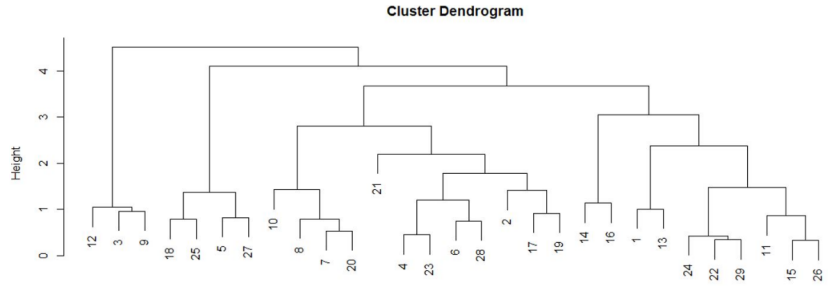
	CP	nsplit	rel error	xerror	xstd
1	0.57654632	0	1.0000000	1.0051577	0.06114786
2	0.10765980	1	0.4234537	0.4442595	0.03876053
3	0.04236603	2	0.3157939	0.3758393	0.03526972
4	0.01710748	3	0.2734279	0.3237557	0.03341855
5	0.01128780	4	0.2563204	0.3121954	0.03305469
6	0.01000000	5	0.2450326	0.3030461	0.03218785

- a. 3
- b. 5
- c. 2
- d. 4

2 marks

QUESTION 18

In the following Dendrogram, for splitting the dataset into 6 clusters, how many cases would be in the cluster with the lowest number of cases?

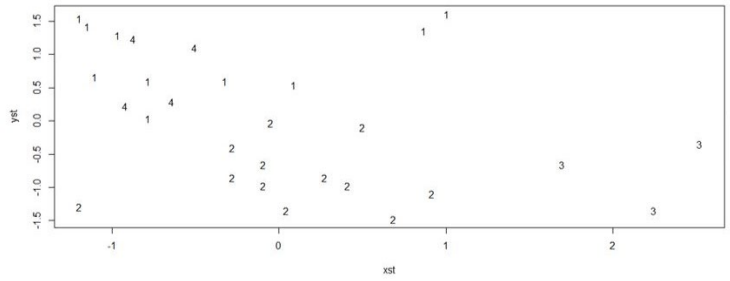


- a. 1
- b. 2
- c. 5
- d. 3

2 marks

QUESTION 19

In the following graph, the output of a clustering analysis, Xst and Yst are standardized X and Y features.



Which of the following rules purely determines one of clusters?

- a. Xst>0, Yst<0
- b. Xst<0, Yst >0
- c. Xst<0, Yst>1
- d. Xst>1.50, Yst<1

2 marks

QUESTION 20

If probability of missing (i.e. R=0) in a dataset depends on a parameter shown by ψ, what is the missingness mechanism shown by the equation Pr (R=0|Yobs, Ymis, ψ)=Pr(R=0|Yobs, ψ)

- a. Missing completely at random
- b. Missing at random
- c. Missing depending on the values of the variable with missings
- d. Missing not at random

3 marks

QUESTION 21

For which one of the distance measures below, the triangle inequality is violated?

- a. Chi-square
- b. Bray-Curtis
- c. Euclidean
- d. L_1

3 marks

QUESTION 22

1. The Entropy for a 6-sided die is ……. that of a 4-sided die.

- less than
- greater than
- equal to

3 marks

QUESTION 23

For the following pair of observations in a binary presence/absence variable, the Jaccard index for dissimilarity equals:

0	1	0	1	1	0	0	0	0	1
0	1	1	0	1	1	0	1	1	0

- 0.55
- 0.65
- 0.50
- 0.75

3 marks

QUESTION 24

1. The inversion method of sampling is based on the theorem that if you apply the transformation Y=F(x) to the samples generated from random variable X with the pdf f(x), the distribution of the y samples is: (F(x) is the cumulative density function of random variable X.)

- a. $f^{-1}(x)$
- b. $F^{-1}(x)$
- c. Normal(0,1)
- d. Uniform (0,1)

2 marks

QUESTION 25

1. The greater probability of an event, the more information is conveyed by the statement that says it happened.

- True
- False

2 marks

QUESTION 26

The following lines of R codes show a snapshot of the result of performing PCA for the *Active Variables* of the dataset *decathelon2*.

```
PCs <- PCA(decathlon2.active, graph = FALSE)
eig.val <- get_eigenvalue(PCs)
eig.val
       eigenvalue
Dim.1   4.1242
Dim.2   1.8385
Dim.3   1.2391
Dim.4   0.8194
Dim.5   0.7015
Dim.6   0.4228
```

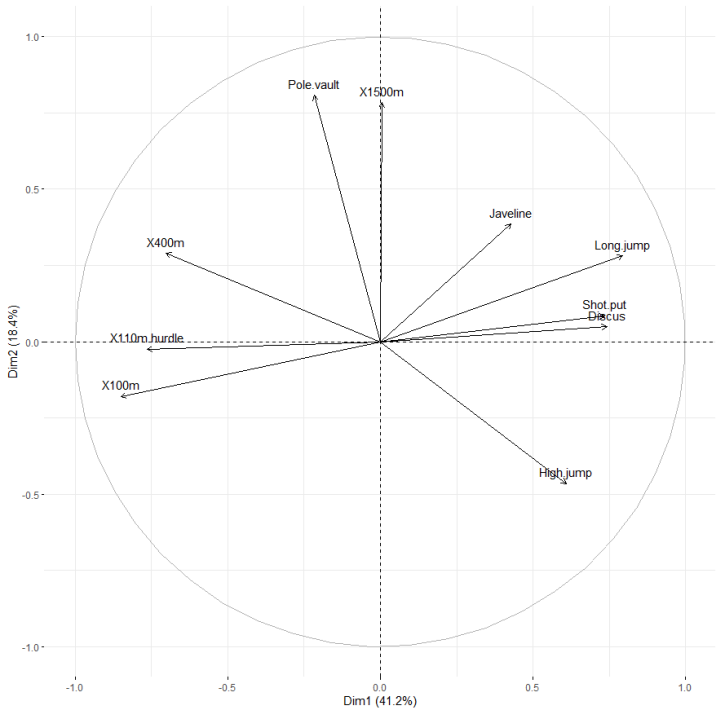
An eigenvalue greater than 1 indicates that the corresponding PC accounts for more variance than that accounted by one of the original variables in standardized dataset.

- a. True
- b. False
- c. Not known; it depends on the number of *Active Variables* in *decathlon2.active*.

3 marks

QUESTION 27

Using the variable correlation plot below, which original variable has the smallest correlation with PC1?



- a. X100m
- b. X110m.hurdle
- c. Pole.vault
- d. X1500m

3 marks

QUESTION 28

In above variable correlation plot, if a variable is perfectly represented by only the first two principal components (Dim.1 & Dim.2), it will be positioned on the:

a. centre of the circle

b. horizontal diameter of the circle

c. vertical diameter of the circle

d. circumference of the circle

3 marks

QUESTION 29

In a Correspondence Analysis of a 5X6 contingency table, a dimension with a contribution larger than ……. should be considered as important.

a. 16.7%

b. 20.0%

c. 22.5%

d. 25.0%

5 marks

QUESTION 30

For the contingency table below

Political party identification by gender, with estimated expected frequencies for independence in parentheses.

	Political Party Identification			
Gender	Democrat	Republican	Independent	Total
Female	495 (456.9)	272 (297.4)	590 (602.6)	1357
Male	330 (368.1)	265 (239.6)	498 (485.4)	1093
Total	825	1088	2450	

assume G^2 and χ^2 show Likelihood-Ratio Test Statistic and Pearson Chi-squared Test Statistic, respectively. Furthermore, G_1^2 is the Likelihood-Ratio Test Statistic that compares the first two columns (i.e. Democrats and Republicans), and G_2^2 is the Likelihood-Ratio Test Statistic for the second 2 × 2 table that combines Democrats and Republicans and compares them to the Independent column. Similarly, χ_1^2 and χ_2^2 show the Pearson Chi-squared Test Statistic for the above comparisons. We therefore have:

a. $G^2 = G_1^2 + G_2^2$

b. $\chi^2 = \chi_1^2 + \chi_2^2$

c. Both of above.

d. None of above.

3 marks

QUESTION 31

Each subject in a sample of 100 men and 100 women is asked to indicate which of the following factors (one or more) are responsible for increases in teenage crime:
A – the increasing gap in income between the rich and poor,
B – the increase in the percentage of single-parent families,
C – insufficient time that parents spend with their children.
A cross-classification of the responses by gender is

Gender	A	B	C
Men	60	81	75
Women	75	87	86

In a Chi-square Test for independence, what is the expected frequency for the cell related to factor A and Men?

- a. 145.8
- b. 62.8
- c. 67.5
- d. 70.0

5 marks

QUESTION 32

A 95% confidence interval for the odds ratio between the treatment (placebo, aspirin) and the outcome for Myocardial Infarction (yes, no) is (1.44, 2.33). If we form the table with aspirin in the first row (instead of placebo), the confidence interval is:

- a. (1.44, 2.33), the odds ratio CI does not depend on the table's row orientation.
- b. (0.43, 0.69)
- c. (0.097, 0.237)
- d. None

4 marks

QUESTION 33

When x_1 or x_2 is the sole predictor for binary Y, the likelihood-ratio test of the effect has p-value < 0.0001. When both x_1 and x_2 are in the model, it is possible that the likelihood-ratio tests for H_0: $\beta_1 = 0$ and for H_0: $\beta_2 = 0$ could both have p-values larger than 0.05.

- a. True.
- b. False

3 marks

QUESTION 34

According to the Pew Research Centre, when adults in the US were asked in 2010 whether there is solid evidence that the average temperature on Earth has been getting warmer over the past few decades, the estimated odds of a "Yes" response for a Democrat was 2.96 times higher than for an Independent, and it was 2.08 times higher for an Independent than for a Republican. The estimated odds ratio between opinion on global warming and whether one is a Democrat or a Republican equals:

- a. 6.2
- b. Cannot say.

4 marks