KEELE UNIVERSITY
SCHOOL OF COMPUTING AND MATHEMATICS
CSC-40054: DATA ANALYTICS AND DATABASES ASSESSMENT

Module leader: Amin Noroozi
Email: a.noroozi.fakhabi@keele.ac.uk

NOTES:

1- This is an open-book take-home exam.

2- Students should attempt to answer **two out of three questions**.

3- The time available to complete this assessment is 28 hours, but you would normally not be expected to invest more than 2 hours of active working time on this assessment

4- Your answer should include the solution and calculations, not the final results only.

5- Python or other programming language codes and outputs are not accepted as solutions. You are not permitted to use Excel functions to solve the questions. You can use Excel only to draw tables.

6- Students must not take any actions during the assessment that would be classed as academic misconduct. This includes plagiarising the work of others, distributing or sharing questions/answers or other relevant information relating to the assessment during the assessment window, working with their peers, and obtaining or attempting to obtain unpermitted assistance (including the use of AI and associated software)

7- You can submit your answers by the next day (19th of May) before 13:00 using the related Dropbox on KLE. Answers submitted even one minute late will be capped at 0.

1-

(a) Assume six data points with two binary attributes, X_1 and X_2, are given as listed in Table 1. These data points belong to three classes, $Y \in \{1,2,3\}$, and our purpose is to classify these data points using a decision tree classifier with only one split.

(i) Calculate the information gain values when the data points are split using X_1 and X_2. **[30%]**

(ii) Explain which split is better and why. Draw the decision tree using the best split, label the branches, and determine what the predicted class label is in each leaf. **[10%]**

Table 1

X_1	X_2	Y
1	1	1
1	1	1
1	1	2
1	0	3
0	0	2
0	0	3

(b) Assume we are given a shop's database containing two tables named Products and Orders. The Products table consists of four columns, namely ProductID ProductName, SupplierID, and Price. The second table consists of three columns, namely OrderID, ProductID, and Quantity. The column Quantity shows how many quantities of ProductID exist in the relevant OrderID. Examples of these two tables are shown below.

(i) Write a SQL query to find the product IDs with maximum and minimum prices. **[5%]**

(ii) Write a SQL query to find the product name with the highest number of quantities ordered **[10%]**

(iii) Write SQL queries to find the order ID that has the highest total price (To calculate the total price of each order, you need to multiply the price of each product in that order by its quantity, repeat this process for all products in that order, and sum the results). **[20%]**

Products

ProductID	ProductName	SupplierID	Price
1	Name1	1	4
2	Name2	1	5
3	Name3	2	10
4	Name4	2	15
5	Name5	3	25

Orders

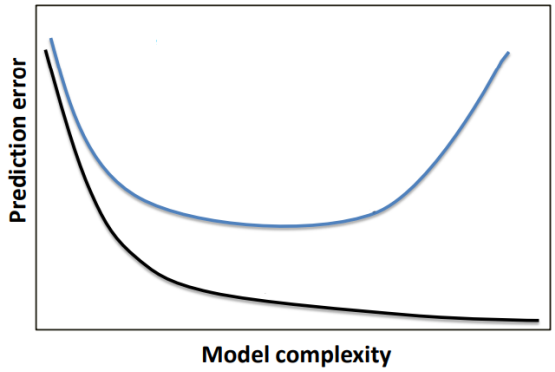
OrderID	ProductID	Quantity
1	1	1
1	2	1
2	3	6
3	2	2
3	5	4
3	4	2
4	5	12
4	2	3

(c) The following frequency table shows different age groups watching different genres of movies. Use the appropriate test to determine if there is a significant relationship between the age group and the genre of the movie. **[25%]**

Genre / Age group	Action	Drama	Comedy	Thriller
Age group 1	59	145	102	56
Age group 2	52	172	85	68
Age group 3	64	110	150	120
Age group 4	42	132	68	140

2-

(a) The following figure illustrates training and validation curves of a neural network with increasing model complexity.



(i) Which of the curves is more likely to be the training error and which is more likely to be the validation error? Briefly explain your answer. **[10%]**

(ii) In which regions of the graph are bias and variance low and high? Briefly explain your answer. **[10%]**

(iii) In which regions does the model overfit or underfit? Briefly explain your answer. **[10%]**

(b) Consider the following sorted sequence of integers:

$$1, 2, 3, X_t, X_{t+1}, 30, 38, 46$$

(i) Find X_t and X_{t+1} such that $X_{t+1} - 2X_t = 1$, and the mean value of the sequence will be equal to 20. **[5%]**

(ii) Are there any values of X_t and X_{t+1} for which the mean and median of the sequence are equal? If yes, give two examples. **[5%]**

(iii) If the sequence was not sorted and we had $X_{t+1} = 4{X_t}^2$, is it possible for the mean value of the sequence to be equal to 12? Please describe your answer. **[10%]**

(c) Use the K-means clustering to find two different clusters in the following sequence of two-dimensional points:

X=[(7,11),(16,20),(28,26),(18,10),(13,19),(22,29),(32,21),(12,15),(38,2),(3,4)]

Choose two random centres for your clusters to start the algorithm. Include the centres of clusters and calculations for each iteration in your answer. You can use the Euclidean distance to calculate the distance between points. **[35%]**

(d) Use the equations of a two-layer neural network and explain what problem we may encounter if not using the activation function in neural networks. Back up your explanation with mathematical proof. **[15%]**

3-

(a) Table 2 shows the number of a website's viewers in different weeks for the last four months of 2022.

(i) Which statistical test can be used to determine if there is a significant relationship between the month and the number of the website's viewers? Briefly describe your answer. **[5%]**

(ii) Use the appropriate test selected in (i) to calculate the p-value. Interpret the results. **[40%]**

Table 2. The restaurant's average sales in December for five different foods on different weekdays

September	October	November	December
300	240	410	200
740	480	270	560
370	520	320	310
850	760	390	180

(b) The stationary time series x contains $N=8$ samples $x_N, x_{N-1}, \ldots, x_1$ equal to 1, 2, 6, 7, 10, 12, 17, 19, respectively. Calculate the coefficients φ_1 and φ_2 such that x could be modelled using an auto-regressive model as follows

$$x_{i+1} = \varphi_1 x_i + \varphi_2 x_{i-1} \ \ (i = 1, 2, \ldots, N) \qquad \textbf{[35\%]}$$

(c)

(i) Fig. 3 shows the error term after decomposing a time series into its components. According to the figure, have the components been calculated correctly? Briefly explain your answer. **[10%]**

(ii) Fig. 4 shows the ACF plot of a time series. Briefly describe whether there is any seasonality in the time series. **[10%]**

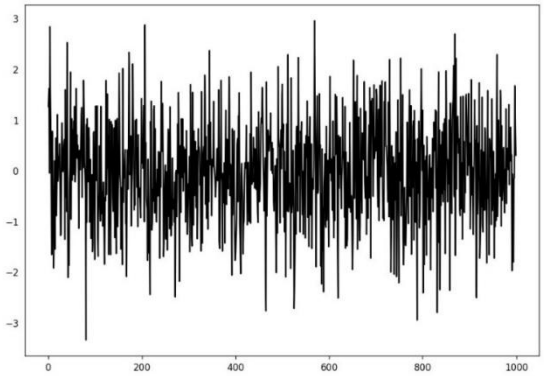


Fig. 3. The error term after decomposing a time series into its components.

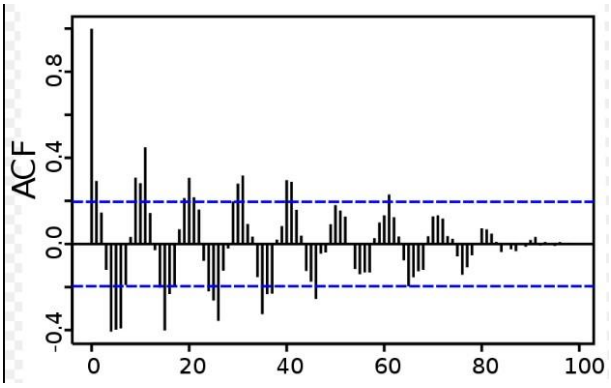


Fig. 3. ACF plot of a time series.