United International University (UIU)
Department of Computer Science and Engineering
CSE 4891: Data Mining, Final Spring 2024

Total Marks: **30** Duration: 1 hour 45 minutes

[Any examinee found adopting unfair means including copy from another examinee will be expelled from the trimester/program as per UIU disciplinary rules.]

Answer <u>All</u> Questions

1. Explain the effect of voting based ensemble classifier based on Figure 1, where M1, M2, M3 is base models and x1, x2, x3 is class labels. [6]

	x_1	x_2	x_3
M_1	✓	✓	✗
M_2	✗	✓	✓
M_3	✓	✗	✓

Figure 1(a)

	x_1	x_2	x_3
M_1	✓	✓	✗
M_2	✓	✓	✗
M_3	✓	✓	✗

Figure 1(b)

	x_1	x_2	x_3
M_1	✓	✗	✗
M_2	✗	✓	✗
M_3	✗	✗	✓

Figure 1(c)

2. (a) What are basic differences between noise and outliers? [2]

(b) How do boxplots help in identifying outliers, and what are the key components of a boxplot that are used for this purpose? Could you explain the process with proper notation and examples? [4]

3. The k-means clustering algorithm is known for being sensitive to outliers. How can the objective function be modified to reduce this sensitivity to outliers? Can you explain with a proper equation and example based on k-means and improved clustering algorithm you are proposing? [6]

4. (a) If you have a dataset and it is not linearly separable, how can you cluster that given dataset using kernel function? Mention the kernel function names and show its suitability using graphical representation. [4]

(b) From a mixture of Gaussian distributions of a given dataset, how can distributions are utilized as criteria of clustering. [2]

5. Create a graph using the following correlation table, and then identify key nodes and prominent edges within the context of graph data mining. Ensure that the edges are depicted in accordance with their level of significance. [6]

	Node1	Node2	Node3	Node4	Node5
Node1	1.00	0.80	0.20	-0.30	0.70
Node2	0.80	1.00	0.25	-0.40	0.60
Node3	0.20	0.25	1.00	-0.10	0.10
Node4	-0.30	-0.70	-0.10	1.00	-0.20
Node5	0.70	0.60	0.10	-0.20	1.00