Machine Learning and Data Mining
In-class Examination

Total Duration 3 hours
(2.5 hrs + 0.5 hrs for compilation of answers and submission as an electronic version via BB)

Non-programmable, scientific calculators are allowed.
Closed Book examination.
Question paper contains 5 pages and Part A – 8 questions, Part B – 2 questions.

Part A

1. Briefly discuss the terms "Classification", "Regression" and "Clustering" in terms of similarities and differences among them. *(3 Marks)*

   Indicative Answer:
   Classification and Regression fall under supervised learning whereas Clustering is an unsupervised learning technique. Classification is the task of inferencing a discrete value like High or low whereas Regression is associated with inferencing a continues value like the temperature outside 34.8C etc. In Clustering we group datapoints based on their similarities so that intra-cluster similarity is high and inter-cluster similarity is as low as possible. We don't even know how many clusters we will end up with, at the beginning.

2. "Interpretability" or "Explainability" is one of the main challenges while building machine learning models. Explain the term "interpretability" in the context of machine learning models *(1 Mark)* and give examples for interpretable models and non-interpretable models. *(2 Marks).* Also give a real-life scenario where interpretability is more important than just achieving a higher accuracy. *(1 Mark)*

   Indicative Answer:
   Interpretability is the ability explain how a particular model make its decision. In other terms, given a particular decision we should be able explain why and how the model reached its conclusion in the terms that human can understand. For example, if we build a model to decide whether a person is accepted for a credit card or not, we should be able to explain our customer (applicant here) why his/her requested was rejected. Most of mathematical models like Decision Tree, Random Forest and Naïve Bayes etc. are explainable where are Neural Networks and different variants of them are un-explainable. The above credit card request is a good example where interpretability is critical because if rejected customers are asking for a reason, we should be able to defense.

3. Explain why data normalization is important before feeding the data into the machine learning model *(2 Marks).* Discuss two normalization techniques briefly *(2 Marks).* Someone argues normalization is not needed while using complex models like neural networks. Give one fact that justifies this argument and one fact against it. *(2 Marks)*

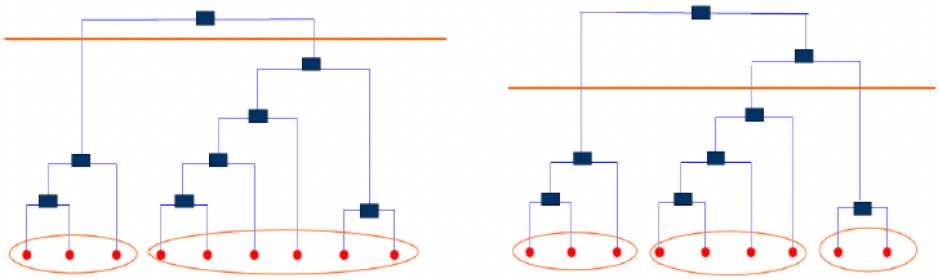
   Indicative Answer:
   If we do not do normalization, the features have implicit importance based on the numeric value of the feature. For example, features in house value prediction, number of rooms have less importance than age of the house because of the range of them are different. Normalization brings all the features in the same range so that importance of the features are actually identified after model training. We can use min-max or z-score normalization. (Appropriate discussion with equations is needed on each). For this fact: Neural networks understand the optimal importance of features during training process whether we normalize or not. Against it: It might need to use more resources and run more epochs to find the best solution if we do not normalize.

4. What is Clustering in Machine Learning? Briefly discuss what is end goal of clustering in terms of intra-cluster and inter-cluster similarities *(3 Marks).* Explain how clustering can be used as pre-processing tool for semi-supervised learning. *(2 Marks)*

   Indicative Answer:
   Clustering is an unsupervised learning technique and builds groups or clusters of data points that are similar. Example would be identifying groups of social media users with same interests. Clustering tries to achieve as high as possible intra-cluster similarity and as low as possible inter-cluster similarity. In semi-supervised learning, we have a fraction of output with label and other without. So, we will use clustering to generate fake labels. Using real and fake labeled data instances we can do a supervised learning activity and build a model now.

5. What is a Dendrogram? *(2 Marks)* Explain how the dendrogram can be used to identify the best *N* (>=2) number of clusters as a combination of all possible clusters identified in hierarchical clustering. *(2 Marks)*

   Indicative Answer:
   Dendrogram is a binary tree like structure that represents how the clusters are merged/split hieratically. Each leaf node is a singleton cluster and dendrogram shows the best way to merge them (if we consider from bottom to top) so that best clustering strategy is identified.



Using the above dendrogram, if N=2 we cut the dendrogram as in the first diagram and for the case of N=3 we will do the cut as in the second diagram here. Based on the desired N, we must decide from which level the cut has to be made and after that we can identify the data instances in each cluster.

6. Following is the Confusion Matrix for the test result of an Animal Detector.

		Predicted Class		
		Cat	Dog	Monkey
Actual Class	Cat	35	2	3
	Dog	1	38	1
	Monkey	2	1	37

i) Find the overall accuracy of the above given Confusion Matrix (CM) *(2 Marks).*

Overall classification accuracy = (35+38+37)/120

ii) Decompose the above given CM into individual CM for each class (2 x 3 = 6 Marks) and find the specificity and sensitivity of each class *(4 Marks).*

Confusion Matrix (General Definition)		Predicted class	
		YES	NO
Actual class	YES	TP	FN
	NO	FP	TN

Sensitivity = TP / (TP + FN)
Specificity = TN / (TN + FP)

Then we will calculate CM for each of the individual class and calculate respective Sensitivity and Specificity values:

CM for Cat		Predicted class	
		Cat	Not
Actual class	Cat	35	5
	Not	3	77

Sensitivity for Cat = 35/ (35+5)
Specificity for Cat = 77/ (77+3)

CM for Dog		Predicted class	
		Dog	Not
Actual class	Dog	38	2
	Not	3	77

Sensitivity for Dog= 38/ (38+2)
Specificity for Dog= 77 / (77+3)

CM for Monkey		Predicted class	
		Monkey	Not
Actual class	Monkey	37	3

	Not	4	76

Sensitivity for Monkey= 37/ (37+3)
Specificity for Monkey= 76/(76+4)

7. A, B, C and D are four cities in a country. The distance matrix between each other is given in the following matrix.

	A	B	C	D
A	0	4	3	7
B		0	8	2
C			0	5
D				0

i) Perform Agglomerative Hierarchical Clustering using the Complete Linkage method and draw a dendrogram to represent the clusters. *(10 Marks)*

To perform agglomerative hierarchical clustering using the complete linkage method, we start with each data point in its own cluster and iteratively merge the two closest clusters until we have a single cluster containing all the data points.

The steps are as follows:

1. Start with each point in its own cluster: {A}, {B}, {C}, {D}
2. Find the two closest points in the distance matrix. In this case, the two closest points are B and D, with a distance of 2. Merge B and D into a cluster. We can represent this cluster as (B, D).

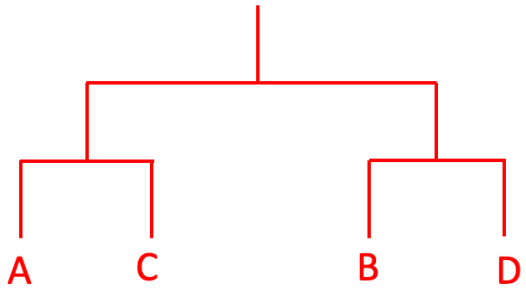
3. Update the distance matrix by calculating the distance between the newly formed cluster (B, D) and the remaining points A and C. Therefore, the distance between (B, D) and A is max(4, 7) = 7, and the distance between (B, D) and C is max(8,5) = 8.

	{B,D}	A	C
{B,D}	0	7	8
A		0	3
C			0

4. Find the two closest clusters in the updated distance matrix. In this case, the two closest clusters are A and C, with a distance of 3.
5. Update the distance matrix by calculating the between the newly formed cluster {B, D} and { A , C}. Therefore, the distance between (B, D) and {A,C} =max(4,5,7,8}=8

	{B,D}	{A,C}
{B,D}	0	8
{A,C}		0

6. The final dendrogram for agglomerative hierarchical clustering using complete linkage is:



8. Explain the steps in the process of building a decision tree from training data. You may need to use the terms "Entropy", "Information Gain" etc. in your discussion *(2 Marks).* Decision Trees are more explainable models compared to algorithms like Neural Network. Explain why this happens in that way. *(1 Marks)*

Indicative Answer:
The main idea in DT is to identify the most informative attributes and in which order we need to apply the check for each attribute so that we can effectively make decisions. We start with the class label as the parent attribute, do the entropy for that. Then we will take one input feature randomly and do the entropy calculation. Based on above entropies, then we calculate the information gain foe that split. Do the same thing of other input features as well and identify the feature with highest IG as the best split. This Hunt's algorithm is recursively applied for each branch and will stop when all the attributes are covered, or all the instances belong to the same class.

DT is a Mathematical model it is representing its decision-making strategy in a tree like structure which human also follow in decision making. Entropy in IG concepts that guides the splitting process in the tree. NN is a computational model that uses back/forward propagation and error calculations for the purpose of improving model parameters. But given particular decision, how it was made by the model cannot be explained in NN.

Part B

Question 1

Market Basket Analysis (MBA) is one of the key techniques used by large retailers to increase sales by better understanding customer purchasing patterns. Association Rules are widely used to analyse retails basket or customer purchasing data. You have been given the following transaction database that consists of items bought in a store by customers.

Consider the transactions shown in the below table.

Transaction ID	**Products**
1	Chips, Coke, Hot Dogs, Ketchup, Pizza
2	Chips, Coke, Hot Dogs, Ketchup, Lime
3	Chips, Coke, Donuts, Ketchup, Pizza
4	Chips, Coke, Hot Dogs, Submarine
5	Donuts, Hot Dogs, Ketchup, Pizza
6	Cake, Chips, Coke, Ketchup, Pizza
7	Burgers, Chips, Coke
8	Chips, Soda
9	Buns, Cake, Chips, Coke, Hot Dogs, Ketchup
10	Buns, Chips, Coke, Donuts, Ketchup

Answer the following questions using the information above. Show all your steps clearly.

i. Give the definitions for Support of item A and, Confidence of buying item B when the item A is already bought. (1 Mark)

$$\text{Support(A)} = \frac{\text{freq(A)}}{\text{Total number of transactions}}$$

$$\text{Confidence (A => B)} = \frac{\text{freq(A, B)}}{\text{freq(A)}}$$

ii. Find the support values for buying Chips, Coke, Lime and Pizza individually. (2 Marks)

Sup (Chips) = 0.9
Sup (Coke) = 0.8
Sup (Lime) = 0.1

Sup (Pizza) = 0.4

iii. Find the confidence values for buying Chips, given that Pizza is already bought. (1 Mark)

Conf (Pizza -> Chips) = 3/4 =0.75

iv. Find the support and confidence for the associations given below. (3 Marks)
a. {Coke} → {Chips, Pizza}
b. {Chips, Cake} → {Hot Dogs}
c. {Chips, Coke, Pizza} → {Hot Dogs, Ketchup}

Rule	Support	Confidence
{Coke} → {Chips, Pizza}	0.3	3/8 = 0.375
{Chips, Cake} → {Hot Dogs}	0.1	½ = 0.5
{Chips, Coke, Pizza} → {Hot Dogs, Ketchup}	0.1	1/3 = 0.33

v. Find all the frequent itemsets (sets of 1 item, 2 item, 3 item, etc.) Assume that minimum support count threshold is 4. (10 Marks)

Find itemsets of size-01

Itemset	Support Count
Buns	2
Burgers	1
Cake	2
Chips	9
Coke	8
Donuts	3
Hot Dogs	5
Ketchup	7
Lime	1
Pizza	4
Soda	1
Submarine	1

Here the support count of Buns, Burgers Cake, Donuts, Lime, Soda and Submarine are less than the minimum support count. Therefore, all the other elements are candidates for itemsets of size-02.

Find itemsets of size-02

Itemset	Support Count
Chips, Coke	8

Chips, Hot Dogs	4
Chips, Ketchup	6
Chips, Pizza	3
Coke, Hot Dogs	4
Coke, Ketchup	6
Coke, Pizza	3
Hot Dogs, Ketchup	4
Hot Dogs, Pizza	2
Ketchup, Pizza	4

7, 2-itemsets have a support count greater than the minimum support count. Therefore, they will be candidates for the itemsets of size 3.

Remaining elements: Chips, Coke, Hot Dogs, Ketchup, Pizza

Find itemsets of size-03

Itemset	Support Count	Comments
Chips, Coke, Hot Dogs	4	
Chips, Coke, Ketchup	6	
Chips, Coke, Pizza	3	Rejected, due to lower sup. count
Chips, Hot Dogs, Ketchup	3	Rejected, due to lower sup. count
Chips, Hot Dogs, Pizza		Rejected, subsets ({Chips, Pizza} and {Hot Dogs, Pizza} are not frequent
Chips, Ketchup, Pizza		Rejected, subset ({Chips, Pizza} is not frequent
Coke, Hot Dogs, Ketchup	3	
Coke, Hot Dogs, Pizza		Rejected, subsets (Coke, Pizza} and {Hot Dogs, Pizza} are not frequent
Coke, Ketchup, Pizza		Rejected, subset ({Coke, Pizza} is not frequent
Hot Dogs, Ketchup, Pizza		Rejected, subset ({Hot Dogs, Pizza} is not frequent

2, 3-itemsets have a support count greater than the minimum support count. Therefore, they will be candidates for the itemsets of size 4.

Remaining elements: Chips, Coke, Hot Dogs, Ketchup

Itemset	Support Count	Comments
Chips, Coke, Hot Dogs, Ketchup		Rejected, some subsets ({Chips, Hot Dogs, Ketchup}, {Coke, Hot Dogs, Ketchup}, etc.) are not frequent

None of the size 4 itemsets are frequent.

So, finally we have 5 1-itemsets, 7 2-itemsets, and 2 3-itemsets which are frequent.

Therefore, frequent itemsets: {Chips}, {Coke}, {Hot Dogs}, {Ketchup}, {Pizza}, {Chips, Coke}, {Chips, Hot Dogs}, {Chips, Ketchup}, {Coke, Hot Dogs}, {Coke, Ketchup}, {Hot Dogs, Ketchup}, {Ketchup, Pizza}, {Chips, Coke, Hot Dogs}, {Chips, Coke, Ketchup}

In summary, following are the frequent itemsets.

Chips	9	Chips, Coke	8	Chips, Coke, Hot Dogs	4
Coke	8	Chips, Hot Dogs	4	Chips, Coke, Ketchup	6
Hot Dogs	5	Chips, Ketchup	6		
Ketchup	7	Coke, Hot Dogs	4		
Pizza	4	Coke, Ketchup	6		
		Hot Dogs, Ketchup	4		
		Ketchup, Pizza	4		

vi. Using the frequent itemsets, find all the closed frequent itemsets which are not maximal. Show all the steps/results of your work clearly and justify any decision you have taken in your analysis. (8 Marks)

Closed frequent itemsets:

If we get {Coke} for an example, it's immediate superset {Chips, Coke} has the same support as itself: 8. Therefore, {Coke} is not a closed frequent itemset. Frequent itemset {Coke, Ketchup} also has one immediate superset - {Chips, Coke, Ketchup} which has the same support as itself. Therefore, {Coke, Ketchup} is also not a closed frequent itemset. Likewise, we can omit the frequent itemsets which has at least one immediate superset which has the same support as itself when finding the closed frequent itemsets.

In summary, following are the closed frequent itemsets we can get from the given dataset.

Chips	9	Chips, Coke	8	Chips, Coke, Hot Dogs	4
Hot Dogs	5	Hot Dogs, Ketchup	4	Chips, Coke, Ketchup	6
Ketchup	7	Ketchup, Pizza	4		

But we are interested in those closed itemsets which are not maximal ones. Let's try to find those which are maximal ones and remove them for the closed list. Then, the remaining ones will be the required answer.

Maximal frequent itemsets:

Hot Dogs, Ketchup	4
Ketchup, Pizza	4
Chips, Coke, Hot Dogs	4
Chips, Coke, Ketchup	6

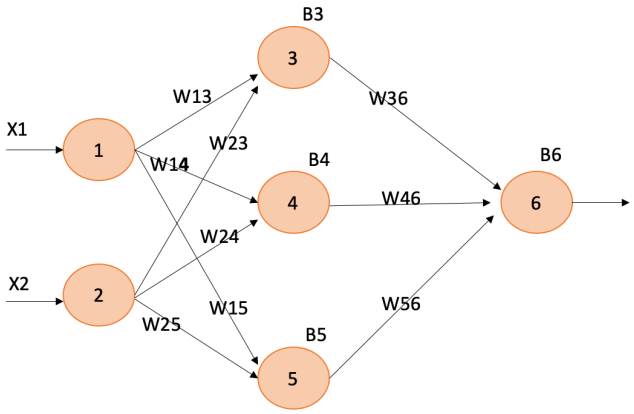
As the above given closed frequent itemsets have no immediate supersets which are frequent, they are maximal frequent itemsets.

Therefore, closed frequent itemsets which are not maximal frequent itemsets are as follows:

Chips	9
Hot Dogs	5
Ketchup	7
Chips, Coke	8

Question 2

Consider the below neural network architecture containing 6 neurons.



X1 and X2 are two input features. Wights between corresponding neurons are denoted using W and bias of neurons are mentioned by B values.

The below table shows a part of training data instances.

X1	X2	Y
1	0	0
0	1	0
1	1	1

The following table summarizes the initial weights and biases of neural network.

W13	W23	W14	W24	W15	W25	W36	W46	W56	B3	B4	B5	B6
0.2	- 0.3	0.4	0.1	0.5	- 0.6	- 0.5	0.6	0.3	- 0.3	0.1	0.4	- 0.7

Answer all the questions by showing all the steps of your work.

i) Assuming the Sigmoid is the activation function, execute the forward propagation and calculate the inferred value by the neural network using the first data instance in the training data set. Hence calculate the error associated with the inference task. (12 Marks)

Sigmoid activation function for Z is:

$$f(z) = \frac{1}{1+e^{-z}}$$

Error at the output layer is given by the equation:

$Err_j = O_j(1 - O_j)(T_j - O_j)$; For the calculation purpose you can safely assume $T_j = 1$

At the N (Neuron) 3:

$\sum XW$ = X1*W13 + X2*W23 = 1*0.2 + 0 = 0.2
Output in N3 = 1/(1+e^(-0.2)) = 0.5498 => 0.550 (approximately)

At N4:
$\sum XW$ = X1*W14 + X2*W24 = 1*0.4 + 0 = 0.4
Output in N4 = 1/(1+e^(-0.4)) = 0.5987 => 0.599

At N5:
$\sum XW$ = X1*W15 + X2*W25 = 1*0.5 + 0 = 0.5
Output in N5 = 1/(1+e^(-0.5)) = 0.6225 => 0.623

Similarly:

At N6:
$\sum XW$ = 0.55*W36 + 0.599*W46 + 0.623*W56
=0.55*(-0.5) + 0.599*0.6 + 0.623*0.3 = -0.275 + 0.3564 + 0.1896 = 0.271

Output in N6 = 1/(1+e^(-0.271)) = 0.5673

Error = 0.567 * (1- 0.567) * (1- 0.567) = 0.1823

ii) Using the error value calculated above, execute back propagation and calculate error values in neurons denoted by 3, 4 and 5 as well. (6 Marks)

Error of a hidden layer neuron is given by:

$$Err_j = O_j(1 - O_j)\sum_k w_{j,k} \cdot Err_k$$

i, j, k denotes identifiers of neurons and W, B, Err and O are weights, biases, errors and outputs respectively.

At N3

Error = 0.55 * (1-0.55) * (-0.5 * 0.1823) = -0.0226

At N4

Error = 0.599 * (1-0.599) * (0.6 * 0.1823) = 0.0263

At N5

Error = 0.623 * (1-0.623) * (0.3 * 0.1823) = 0.0128

iii) Now considering the Learning Rate (l) is 0.7, calculate the updated weights and bias values resulted by the back propagation. (10 Marks)

Updated weight values are given by:

$$w_{i,j} = w_{i,j} + \Delta w_{i,j} \; where \; \Delta w_{i,j} = l \cdot Err_j \cdot O_i$$

Updated bias values are given by:

$$B_j = B_j + \Delta B_j \; where \;\; \Delta B_j = l \cdot Err_j$$

W36 = -0.5 + (0.7 * 0.1823 * 0.55) = -0.4298
W46 = 0.6 + (0.7 * 0.1823 * 0.599) = 0.6764
W56 = 0.3 +(0.7 * 0.1823 * 0.623) =0.3795

W13 = 0.2 + (0.7 * 0.55 * 1) = 0.585
W14 = 0.4 + (0.7 * 0.599 * 1) = 0.8193
W15 = 0.5 +(0.7 * 0.623 * 1) = 0.9361

W23 = -0.3 + 0 = -0.3
W24 = 0.1 +0 = 0.1
W25 = -0.6 + 0 = -0.6

B6 = -0.7 + 0.7 * 0.1823 = -0.5724
B3 = -0.3 + 0.7 * (- 0.0226) = -0.3158
B4 = 0.1 + 0.7 * 0.0263 = 0.7590
B5 = 0.4 + 0.7 * 0.0128 = 0.4090