Roll Number:_____

Thapar Institute of Engineering and Technology, Patiala

Computer Science and Engineering Department

BE(3rd Year) Feb 23, 2022 *Auxiliary Examination* UML501: Machine Learning

Time: 3 Hours Marks:100

Note: All questions are compulsory. All parts of a question must be answered in order. Show all intermediate steps, where applicable.

Q(1) a) Suppose we have the following dataset of 6 two-dimensional points: [(1, 1), (1, 2), (2, 1), (6, 7), (7, 6), (7, 7)] use k-means clustering to group these points into two clusters. Assuming initial cluster centers (1, 1) and (7, 7) Give the results of the algorithm till **2 iterations** of K means clustering.
(Note- Use **Manhattan distance** for finding closeness of data points and show all intermediate steps)

b) How do we choose the value of k in k-means clustering? Justify your answer (10+5=15 marks)

Q(2) Given the following dataset with 10 instances with 3 attributes < Raining, temperature, humidity> and 1 target variable (Play) Apply ID3 algorithm and build the final decision tree. Show all intermediate steps.

Raining	temperature	humidity	Play
True	Hot	High	No
True	Hot	High	No
False	Hot	High	Yes
False	Cool	Normal	Yes
False	Cool	Normal	Yes
True	Cool	High	No
True	Hot	High	No
True	Hot	Normal	Yes
False	Cool	Normal	Yes
False	Cool	High	Yes

(15 marks)

Q(3) a) Suppose we have a dataset of customer purchases in an online store, and we want to find association rules between items. Here is a sample dataset:

Transaction	Items
1	book, DVD, pen
2	book, DVD, phone
3	DVD, phone
4	book, pen
5	DVD, pen

a) Give the formulas of support, confidence and lift.
b) Find the following
- support(book, pen, DVD)
- confidence(book, pen -> DVD)
- lift(book, pen -> DVD)

c) What does the output of b) part depicts. Explain in your wordings

b) What is backpropagation, and how is it used to train artificial neural networks? Explain in brief how gradient descent can be used in optimize the weights in backpropagation neural network

(8+7=15 marks)

Q(4) a) Given the following dataset showing the salary of employees based on year of experience.

year of experience(X)	salary of employee(Y)
1	2
2	4
3	5
4	4
5	5

Assuming the best fit line using linear regression is y=mx+c with m=0.6 and c=2.2

.Find RMSE value and R2 score.

P.T.O

b) Given the marks of 27 students in sorted order

Marks: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35,36, 40, 45, 46, 52, 70.
Find

- Mean,Median,Mode
- 1st Quartile,2nd Quartile,3rd Quartile and Inter Quartile range(IQR)

(8+7=15 marks)

Q(5) a) Apply Bayes Theorem to solve the following problem: Consider a set of patients coming for treatment in a certain clinic. Let A denote the event that a "Patient has heart disease" and B the event that a "Patient is having high blood pressure(BP)."It is known from experience that 10% of the patients entering the clinic have heart disease and 5% of the patients are having high BP. Also, among those patients diagnosed with heart disease, 7% are having high BP. Given that a patient with high BP, what is the probability that he will have heart disease?

b) Suppose the training data(Table 1) contains two binary features <f1, f2>and a binary class<Yactual >. Using this training data, learn Naïve Bayes classifier and apply the learned model to predict the value of y (ie. Ypred) for test samples in Table 2. .(NOTE- In case of any ties, label predicted class, Ypred as 0).Also find precision and recall value of test data (given in table 2)

(5+10=15 marks)

Table 1: Training Data		
fl	f2	Yactual
1	0	1
1	0	1
1	1	0
1	1	0
1	0	0
0	1	1
0	0	0
0	0	1

Table 2:Test Data			
fl	f2	Yactual	Ypred
0	1	0	?
1	1	0	?
0	0	1	?
1	0	1	?

Q(6) a) Suppose we have a dataset of eight 3-bit binary vectors and their corresponding labels. Classify a new data point, 010, using a k-nearest neighbor classifier with k=3 and Hamming distance.

Data	000	001	010	011	100	101	110	111	111	111
point	A	A	B	B	C	C	D	D	D	D

b) Why is the KNN Algorithm known as Lazy Learner? How to find the best value for K in the KNN algorithm? Justify your answer

(5+10=15 marks)

Q(7) Categorize the following real world problems as classification, regression, clustering or reinforcement learning. Justify your answer in one line

1. Predicting the electricity consumption of a household based on the number of occupants, time of day, and weather conditions.
2. Identifying whether a network traffic is normal or malicious
3. to teach robots to perform tasks such as grasping, walking, and navigation.
4. Predicting the duration of a flight based on its origin, destination, and other flight-related factors.
5. to group geographic locations based on similar attributes such as population density or land use

(10 marks)