Indian Institute of Information Technology, Allahabad
C-3 Examination (May-2022)
Paper: Data Mining

B.Tech (IT), VI Semester **Max. marks: 80**
Duration: 2 Hrs.

Course Instructor: Prf. OP Vyas, Dr. Manish Kumar & Dr. Muneendra Ojha

Q1.A database has 4 transactions, shown below.

TID	Date	items_bought
T100	10/15/04	{K, A, D, B}
T200	10/15/04	{D, A, C, E, B}
T300	10/19/04	{C, A, B, E}
T400	10/22/04	{B, A, D}

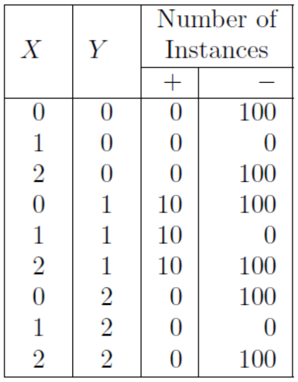
Assuming a minimum level of support min_sup = 60% and a minimum level of confidence min_conf = 80%:

(a) Find all frequent itemsets (not just the ones with the maximum width/length) using the Apriori algorithm. Show your work—just showing the final answer is not acceptable. For each iteration show the candidate and acceptable frequent itemsets
(b) List all of the strong association rules, along with their support and confidence values, which match the following metarule, where X is a variable representing customers and item i denotes variables representing items (e.g., “A”, “B”, etc.).
$\forall x \in$ transaction, buys(X, item1)^buys(X, item2) $\Rightarrow$ buys(X, item3). **[10 marks]**

Q2. In the dataset given below, the class **Play Tennis** is predicted as **‘yes’** for the sample **X= (Outlook=Rain, Temperature=Mild, Humidity=High, Wind=Strong)** using the Bayesian classifier. Considering the incremental learning property of the Bayesian classifier, what will be the predicted class for the sample Y**= (Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)? [10 marks]**

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

Q3. Following is a data set that contains two attributes, X and Y , and two class labels, “+” and “ −“. Each attribute can take three different values: 0, 1, or 2. **[10 marks]**



X	Y	Number of Instances	
		+	−
0	0	0	100
1	0	0	0
2	0	0	100
0	1	10	100
1	1	10	0
2	1	10	100
0	2	0	100
1	2	0	0
2	2	0	100

The concept for the “+” class is $Y = 1$ and the concept for the “−” class is $X = 0 \vee X = 2$.

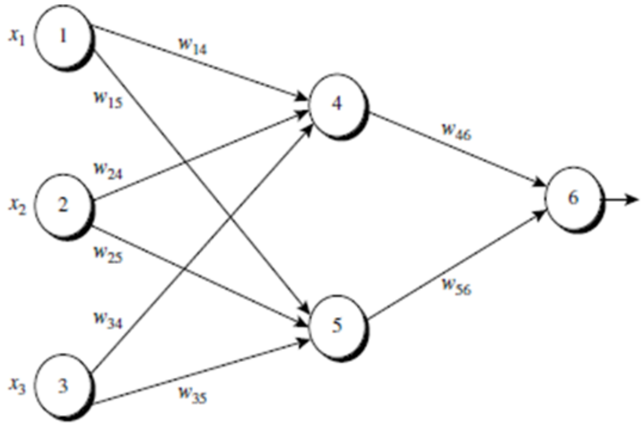
(a) Build a decision tree on the data set. Does the tree capture the “+” and “−” concepts?

(b) What are the accuracy, precision, recall, and F1-measure of the decision tree? (Note that precision, recall, and F1-measure are defined with respect to the "+" class.)

Q4. Apply the Backpropagation algorithm to find out the updated weights and bias after the first iteration. Initial input, weight, and bias values are given in the table. If the net input is I_j to unit j, then the output of unit j is $o_{j.}$ **[10 marks]**

Activation Function = o_j = 1/1 + exp^Ij

x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	−0.3	0.4	0.1	−0.5	0.2	−0.3	−0.2	−0.4	0.2	0.1



Q5. If epsilon is 2 and minpoint is 2, what are the clusters that DBSCAN would discover for the distance matrix given in table -1 with the following 8 examples: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Draw the 10 by 10 space and illustrate the discovered clusters. Also, find out the clusters if epsilon is increased to√10 ? **[10 marks]**

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	25	36	13	50	52	65	5
A2		0	37	18	25	17	10	20
A3			0	25	2	2	53	41
A4				0	13	17	52	2
A5					0	2	45	25
A6						0	29	29
A7							0	58
A8								0

Table 1: DBSCAN training dataset

Q6. Briefly compare the following concepts. You may use an example to explain your point(s). **[10 marks]**

a. Hard margin & Soft margin SVM.

b. Nominal, Ordinal, Ratio scale variables

Q7. Implement Complete linkage and Average linkage hierarchical clustering on the following data. **[10 marks]**

Dis	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Q8. Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters: **A1 (2, 10), A2 (2, 5), A3 (8, 4), B1 (5, 8), B2 (7, 5), B3 (6, 4), C1 (1, 2), C2 (4, 9).** The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only. **[10 marks]**

(a) The three cluster centers after the first round execution.

(b) The final three clusters are.