Intelligent Data Analysis
Mid Term Exam – Practice Problems

Consider the following dataset:

X1	X2	X3	X4	Class
Red	Y	2	36	0
Green	Y	3	49	1
Blue	N	4	50	1
Pink	Y	2	26	0
Orange	N	4	78	1
Magenta	Y	3	23	0
Gold	Y	4	92	1
Black	N	3	74	1
Brown	Y	2	36	0

1. (12) Find the entropy of (i) the original database; and (ii) gain ratio for the splits formed by the attributes X2, and X3.

2. (8) How would you use the attribute X4 to find a test condition for constructing a decision tree with this dataset? Show the process of finding the test condition for X4 by your suggested method and the entropy reduction that will be obtained. Suggest another alternative method of using X4 for decision tree construction and give reasons to show that your chosen method is the better choice.

3. (8) What is the advantage of using gain ratio over plain information gain?

4. (8) A dataset was split into test and training parts and a decision tree built using the training-data part and tested using the test-data part. The error rate for both training and test datasets keeps declining continuously as the size of the decision tree (number of nodes in the tree) grows. What can be inferred about the nature of data in the dataset? Give reasons to justify your answer.

5. (8) For another dataset the training error rate keeps declining but the test dataset has a constant error rate of 50% - no matter how much the tree grows. What can be said about the nature of this dataset? Give reasons to justify your answer.

Consider the following dataset:

Y1	Y2	Y3	Y4	Y5
1	0	0	1	1
1	1	0	1	1
0	1	1	1	0
1	1	0	1	0
0	0	1	1	1
0	1	1	1	0
1	1	0	1	0
1	1	1	1	1
0	0	0	1	1

6. (6) Find the support for the following item sets (Y1, Y2, Y4), (Y1, Y2), (Y1, Y4, Y5), and (Y1, Y5). Also, which 4-item set and which 5-item set have the largest support in the above dataset?

7. (12) Find all the rules from any one of the above listed four item-sets that have confidence above 10%.

8. (10) What is the effect of Y4 in generating item-sets and in generating the rules from them? Is Y4 redundant, very significant, or can influence rules only under certain specific circumstances? Justify your answer.

9. Find all the closed and maximal item sets embedded in the dataset given above.

10. (10) Give examples of one monotone property and one anti-monotone property for generating the association rules.

11. (8) What is the advantage of using Lift as a measure for correlation or usefulness of a rule? Intuitively what aspect does it represent and quantitatively how is it measured?

12. (10) Take the first six rows of the database shown above and show the FP-tree representation for the strings. What are the relative advantages and disadvantages of using the FP tree method over other methods for determining the frequent itemsets?
13. Given a classifier and its results with a test dataset. How would you compute its accuracy, precision and recall metrics. Give one potential disadvantage of each of these metrics.
14. What is meant by normalization of dataset? Why is normalization necessary?
15. Given a dataset of points for perceptron training, show all the steps till a convergence is achieved. An example showing this process is posted on Blackboard in the course documents folder.
16. Explain why three layers are sufficient in a perceptron network to classify any set of data points from two classes.