BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
K. K. BIRLA GOA CAMPUS

First Semester 2019-20
Advanced Data Mining (CS G520)
Mid Semester Exam

Duration: 90 minutes Max Marks: 40

Note: Q1 to Q5 carry two marks each (Max 3 lines for each answer)
Q6 to Q8 carry four marks each (Max 6 lines for each answer)
Q9 to Q11 carry six marks each (Max 10 lines for each answer)
Show all the numerical calculations in fair and box/highlight the final answer

1. Document term matrix is an example of asymmetric discrete or asymmetric continuous feature? Justify
2. Would it make more sense to move all representative points in CURE clustering algorithm to a fixed distance towards the centroid? Why or why not?
3. Discuss the impact of replacing "weak" learners with "strong" learners in boosting.
4. What are the properties of rules extracted from decision trees?
5. Given a dataset D with n number of attributes and |D| training tuples. Derive and justify the worst-case complexity of the decision tree algorithm.

6. Given Alice is a sporty person and sportiness is common among CSE students (75% are sporty) than ECE students (15% are sporty). Using Bayesian classification, predict whether Alice will go for CSE or ECE engineering. Also, 1 out of every 11 students is CSE student and remaining is from ECE.
7. In BIRCH, each entry in the CF tree is characterised by 3-tuple (N, LS, SS). Derive the equation for Centroid and Radius for that entry in terms of the above tuple.
8. On a given dataset, discuss "RF vs bagged DT" classifier based on efficiency and robustness.

9. Consider the training examples in below table for a binary classification problem.

Instance	A1	A2	A3	Target Class
1	T	T	1	+
2	T	T	6	+
3	T	F	5	-
4	F	F	4	+
5	F	T	7	-
6	F	T	3	-
7	F	F	8	-
8	T	F	7	+
9	F	T	5	-

   1. What are the information gains of A1 and A2 relative to these training examples
   2. What is the best split (b/w A1 and A2) wrt classification error
   3. What is the best split (b/w A1 and A2) wrt Gini index

10. Perform hierarchical clustering with the single and complete link on below similarity matrix

	P1	P2	P3	P4	P5	P6
P1	1.00	0.70	0.65	0.40	0.20	0.05
P2	0.70	1.00	0.95	0.70	0.50	0.35
P3	0.65	0.95	1.00	0.75	0.55	0.40
P4	0.40	0.70	0.75	1.00	0.80	0.65
P5	0.20	0.50	0.55	0.80	1.00	0.85
P6	0.05	0.35	0.40	0.65	0.85	1.00

11. Using BIRCH, draw the CF-Tree with following points:
    x1 = 0.5, x2 = 0.25, x3 = 0, x4 = 0.65, x5 = 1, x6 = 1.4
    Assume Threshold = 0.15
    Number of entries in a leaf node = 2
    Branching factor = 2