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

First Semester 2019-20
Comprehensive Examination
Advanced Data Mining

Course Code: CS G520 Duration: 180 minutes
Date: 02/12/2019 Max Marks: 35

Note: Q1 to Q5 carry one mark each (Max 4 lines for each answer)
Q6 to Q7 carry three marks each
Q8 to Q11 carry six marks each
Write to the point answer (negative marking for generic statements)
Show all the calculations/derivation in fair and **box/highlight the final answer**
Answer the complete question at one place and fill the index on the first page

1. In DBSCAN, density reachability is a symmetric property? T/F Justify your answer
2. One advantage of boosting is that it does not overfit the data. T/F Justify your answer
3. Is accuracy a good measure to determine the quality of a rule. T/F Justify your answer
4. Compare and discuss, the number of access required to cluster a particular data point in k-means and BIRCH clustering algorithm.
5. Is feature scaling relevant for kNN classification algorithm?

6. Let c1, c2, and c3 be the confidence values for the rules $\{p\} \rightarrow \{q\}$, $\{p\} \rightarrow \{q, r\}$, and $\{p, r\} \rightarrow \{q\}$, respectively
   a) If we assume that c1, c2, and c3 have different values, what are the possible relationships that may exist among c1, c2, and c3? Which rule has the lowest confidence?
   b) Repeat the above analysis assuming that now all rules have the same support. Which rule has the highest confidence now?
7. Find all frequent itemsets in the below dataset using FP-growth tree:
   Assume min sup = 60% and min conf = 80%

TID	**items bought**
T100	M, O, N, K, E, Y
T200	D, O, N, K, E, Y
T300	M, A, K, E
T400	M, U, C, K, Y
T500	C, O, O, K, I, E

8. Use the following methods to normalise the given data (solve till two decimal places):
   100, 300, 500, 700, 1000
   a) min-max normalisation (min = 0 and max = 1)
   b) z-score normalisation
   c) z-score normalisation (using the mean absolute deviation)
   d) normalisation by decimal scaling
9. Cluster the following data points (0, 4, 5, 20, 25, 39, 43, and 44) using hierarchical agglomerative clustering with:
   a) Single linkage
   b) Complete linkage
   c) Average linkage
   Show the proximity matrix, calculations to merge clusters at each step and dendrogram.

10. Consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules,

    R1: A → + (covers 4 positive and 1 negative example),
    R2: B → + (covers 30 positive and 10 negative examples),
    R3: C → + (covers 100 positive and 90 negative examples),

    Determine the best and worst candidate rule according to:

    a) Rule Accuracy
    b) FOIL's information gain

11. Calculate the page ranks in the below web graph consisting of eight pages (till 3 iterations)