Name: Roll No.:

Introduction to Data Science

Quiz 2 (22nd Nov 2018)

Instructions: Write your answer at the space provided in the question paper. Select the *most appropriate answer* from the given options. Correct answer carries 2 marks and wrong answer carries -1 mark.

1. The statement "A point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster" holds true (most appropriately) in which of the following clusters:
   (a)Well-Separated (b) Center-based (c) Contiguous (d) Density-based

2. K-means clustering is a _____ based clustering algorithm.
   (a)Density (b) Partition (c) Agglomerative hierarchy (d) Divisive hierarchy

3. Which one out of the following clustering methods is more susceptible to noise and outliers?
   (a)K-means (b) K-medoids (c) MAX based Agglomerative (d) DBSCAN

4. A border point in DBSCAN algorithm is
   (a) a core point (b) a noise point (c) part of a cluster (d) a point between two or more clusters

5. In DBSCAN algorithm, *Eps* remaining same, if *MinPts* is increased, then number of core points will
   (a) increase (b) decrease (c) remain same

6. Which of the following algorithm is susectible to order of data being processed?
   (a) K-means (b) DBSCAN (c) K-medoids (d) Agglomerative hierarchical

7. If a dataset has non-elliptical clusters then which inter-cluster distance in agglomerative clustering is suitable?
   (a) Single link (b) Complete link (c) Group average (d) Centroid based

8. Let a cluster *C* has following points: (2,2), (4,4), (6,4), (4,2). What is the Manhattan distance between the centroid of cluster *C* and point (5,1).
   (a) 1 (b) 3 (c) 5 (d) 8

9. In classification, if the correlation between two features is 1 then the features are
   (a) relevant (b) irrelevant (c) redundant (d) Can't say

10. Suppose our dataset has 10 records out of which two records have exactly the same value for all the attributes. Then is the statement "removing one of these two records from the dataset will not change the decision tree we learn from this dataset." True?
    (a) Yes (b) No

11. Suppose a dataset has 10 records and one of the attribute (say *X*) is a continuous attribute with 6 distinct values. We are interested to build a decision stump with *X* as the splitting attribute, such that both the branches have at least one record. In this case, how many splitting conditions we need to examine?
    (a) 10 (b) 9 (c) 6 (d) 5

12. We have a dataset with four records as follows: (0,0,Yes), (2,2,Yes), (0,3,No) and (*p*,1,No). The first two attributes are numeric and the third one is class attribute (having values 'Yes' or 'No'). What is the largest value out of the given options that *p* can take so that both the classes are linearly separable?
    (a) 0 (b) 0.99 (c) 1 (d) 1.01

Name: Roll No.:

Table 1: Confusion matrix

	Predicted Class		
Actual class	C	¬C	Total
C	70	30	100
¬C	50	150	200
Total	120	180	300

13. Based on the confusion matrix in Table 1, what is the value of Precision?
    (a) 70/100 (b) 70/120 (c) 70/80 (d) 70/300

14. Based on the confusion matrix in Table 1, what is the value of accuracy?
    (a) 70/300 (b) 150/300 (c) 220/300 (d) 80/300

Table 2: Dataset

Gender	House_Type	Car_Type	Class
M	Duplex	Sports	P
M	Mansion	Luxury	N
F	2BHK	Family	N
M	2BHK	Sports	P
F	Mansion	Luxury	P
F	Duplex	Family	N
F	Duplex	Family	N
F	2BHK	Family	P
M	Duplex	Sports	N
M	Mansion	Luxury	P

15. Assuming the dataset given in Table 2, what is the gini index of the root node?
    (a) 0 (b) 0.5 (c) 1 (d) None of the above
16. Assuming the dataset given in Table 2 as training data, predict the class label of record (M,2BHK, Family) using Naïve Bayes algorithm.
    (a) P (b) N (c) Either P or N (d) Can't be computed

17. Assuming the dataset given in Table 2 as training data, predict the class label of record (F, Duplex, Family) using Naïve Bayes algorithm.
    (a) P (b) N (c) Either P or N (d) Can't be computed

18. Clustering algorithms that we have covered in class are part of _____ learning technique.
    (a) supervised (b) unsupervised (c) Semi-supervised (d) reinforcement

19. Leave-one out is a special case of _____ method.
    (a) Cross-validation (b) Stratified cross-validation (c) Bootstrap (d) Bagging

20. Cluster separation is captured by _____.
    (a) SSE (b) BSS (c) Jointly by SSE and BSS (d) None of the above