IDNO		NAME	

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

Second Semester 2021-22

COMPREHENSIVE EXAMINATION

PART A (CLOSED BOOK)

CS F415 Data Mining

Date : 12-05-2022 (FN) **Max.Duration : 90 minutes** **Max.Marks : 40 (20%)**

Write your answers only in the space provided.

1. Underfitting occurs when [1]
   (A) Training Error is High and Generalization Error is High.
   (B) Training Error is High and Generalization Error is Low.
   (C) Training Error is Low and Generalization Error is High.
   (D) Training Error is Low and Generalization Error is Low.

2. Overfitting occurs when [1]
   (A) Training Error is High and Generalization Error is High.
   (B) Training Error is High and Generalization Error is Low.
   (C) Training Error is Low and Generalization Error is High.
   (D) Training Error is Low and Generalization Error is Low.

3. Which of the following is a fundamental difference between bagging and boosting? [1]
   (A) Bagging is used for supervised learning. Boosting is used with unsupervised clustering.
   (B) Bagging gives varying weights to training instances. Boosting gives equal weight to all training instances.
   (C) Bagging does not take the performance of previously built models into account when building a new model. With boosting each new model is built based upon the results of previous models.
   (D) Boosting is used for supervised learning. Bagging is used with unsupervised clustering.

4. This approach is best when we are interested in finding all possible interactions among a set of attributes. [1]
   (A) decision tree (B) association rules
   (C) K-Means algorithm (D) Rule based learning

5. Training Error of a model can be reduced by …….(increasing/decreasing) the model complexity. [1]

6. Assume that we have a dataset containing information about 200 individuals. One hundred of these individuals have purchased life insurance. A supervised data mining session has discovered the following rule:
   IF age < 30 & credit card insurance = yes
   THEN life insurance = yes
   Rule Accuracy: 70%
   Rule Coverage: 63%
   How many individuals in the class *life insurance= no* have credit card insurance and are less than 30 years old? [1]
   (A) 63
   (B) 70
   (C) 30
   (D) 27

Use these tables to answer questions 7 and 8.

Single Item Sets	Number of Items
Magazine Promo = Yes	7
Watch Promo = No	6
Life Ins Promo = Yes	5
Life Ins Promo = No	5
Card Insurance = No	8
Sex = Male	6

Two Item Sets	Number of Items
Magazine Promo = Yes & Watch Promo = No	4
Magazine Promo = Yes & Life Ins Promo = Yes	5
Magazine Promo = Yes & Card Insurance = No	5
Watch Promo = No & Card Insurance = No	5

7. One two-item set rule that can be generated from the tables above is:
   If Magazine Promo = Yes Then Life Ins promo = Yes
   The confidence for this rule is: [1]
   (A) 5 / 7
   (B) 5 / 12
   (C) 7 / 12
   (D) 1

8. Based on the two-item set table, which of the following is *not* a possible two-item set rule? [1]
   (A) IF Life Ins Promo = Yes THEN Magazine Promo = Yes
   (B) IF Watch Promo = No THEN Magazine Promo = Yes

(C) IF Card Insurance = No THEN Magazine Promo = Yes
(D) IF Life Ins Promo = No THEN Card Insurance = No

9. If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share > 99:9% of the same genes.) [2]

10. Consider a set of points that are uniformly distributed on the interval [0,1]. Is the statistical notion of an outlier as an infrequently observed value meaningful for this data? [2]

11. Define Outlier with respect to [2*3=6]

A. Density	
B. Probability	
C. Proximity	

12. Expand the following Terms: [1*4=4]

A. DBSCAN	
B. CURE	
C. ROCK	
D. BIRCH	

13. Consider the (relative distance) K-means scheme for outlier detection described in Section 10.5 and the accompanying figure, Figure 10.10. [2+2+2=6]
(a) The points at the bottom of the compact cluster shown in Figure 10.10 have a somewhat higher outlier score than those points at the top of the compact cluster. Why?

(b) Suppose that we choose the number of clusters to be much larger, e.g., 10. Would the proposed technique still be effective in finding the most extreme outlier at the top of the figure? Why or why not?

(c) The use of relative distance adjusts for differences in density. Give an example of where such an approach might lead to the wrong conclusion.

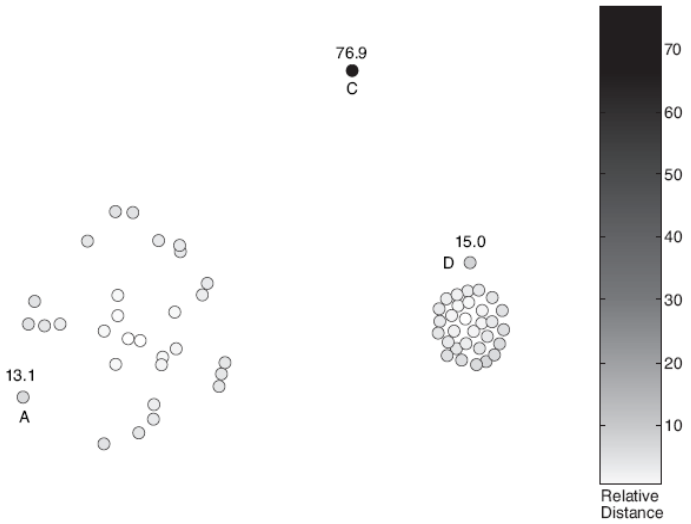

Figure 10.10. Relative distance of points from closest centroid.

14. Consider the following four faces shown in Figure 1. Again, darkness or number of dots represents density. Lines are used only to distinguish regions and do not represent points. [2+2+2=6]

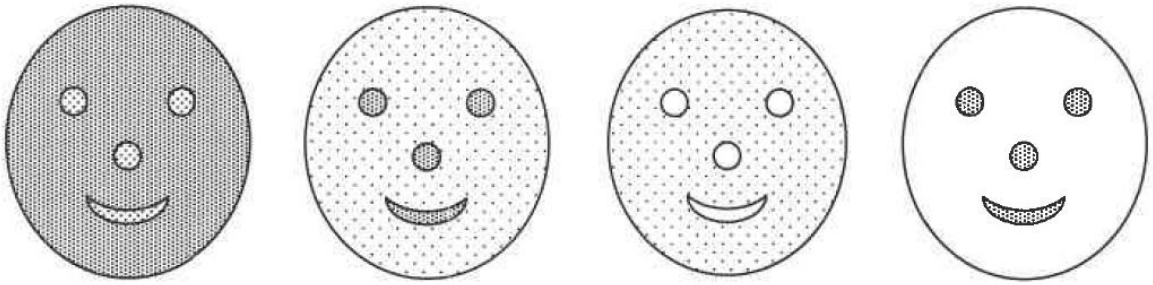

Figure 1

(a) For each figure, could you use single link to find the patterns represented by the nose, eyes, and mouth? Explain.

(b) For each figure, could you use K-means to find the patterns represented by the nose, eyes) and mouth? Explain.

(c) What limitation does clustering have in detecting all the patterns formed by the points in Figure 1 (c)?

15. We generally will be more interested in association rules with high confidence. However, often we will not be interested in association rules that have a confidence of 100%. Why? Then specifically explain why association rules with 99% confidence may be interesting (i.e., what might they indicate)? [2]

16. What are two other names by which following terms are identified? [1+1=2]

A. Training Error	
B. Outlier Detection	

17. Association rule mining often generates a large number of rules. Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules. [2]

****