Roll No									

NATIONAL INSTITUTE OF TECHNOLOGY GOA

Farmagudi, Ponda, Goa, 403401

Programme Name: B.Tech.

Mid Semester Examinations, October-2020

Course Code and Name: CS505: Data Warehousing and Mining — Semester: V

Date: 09.10.2020 — Time: 3.30-5.00 PM

Duration: 1 Hour 30 Minutes — Max. Marks: 50

---

i) Answer all the questions **serially to the point**.
ii) Assume suitable data in case of missing.
iii) Irrelevant answers score negative marks.

---

1. **Clustering (Unsupervised Learning) can be used to improve the accuracy of Linear Regression model (Supervised Learning). How many of the below statements are correct and why? [6]**
   a) **Creating different models for different cluster groups.**
   b) **Creating an input feature for cluster ids as an ordinal variable.**
   c) **Creating an input feature for cluster centroids as a continuous variable.**
   d) **Creating an input feature for cluster size as a continuous variable.**

2. **In which of the following cases will K-Means clustering fail to give good results and why? [6]**
   a) **Data points with outliers**
   b) **Data points with different densities**
   c) **Data points with round shapes**
   d) **Data points with non-convex shapes**

3. **Which of the below is true about K-Means Clustering and why? [6]**
   a) **K-means is extremely sensitive to cluster center initializations**
   b) **Bad initialization can lead to Poor convergence speed**
   c) **Bad initialization can lead to bad overall clustering**

4. **If Epsilon is 2 and minpoint is 2, what are the clusters that DBSCAN would discover with the following 8 examples: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). (The distance matrix is provided below for your reference). Draw the 10 by 10 space and illustrate the discovered clusters. What if Epsilon is increased to √10 ? [8]**

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

5. **Suppose while performing DBSCAN we randomly choose a point which has less than MinPts number of points in its neighborhood. Which among the following is true for such a point and justify your answer? [6]**
   a) **It is treated as noise, and not considered further in the algorithm**
   b) **It becomes part of its own cluster**
   c) **Depending upon other points, it may later turn out to be a core point**
   d) **Depending upon other points, it may be density connected to other points**

6. **Use the Nearest Neighbor clustering algorithm and Euclidean distance to cluster the examples from the question no. 4: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Suppose that the threshold t is 4. [6]**

7. **Define intra-inter ratio validity index and discuss its drawbacks. [6]**

8. **Compare agglomerative clustering with divisive clustering using an example. What are the challenges in hierarchical clustering [6]**