GET – Quiz1 (2 hours) **Max : 100**

Total: 110

I. K-Means

1. Consider the two points A(7,50) and B(23,34). **Which point is closer** (or more similar) to point C(12,12)? **(1)**

2. **Find the centroid** of the five observations given in the following table: **(3)**

X	Y	Z
12	23	45
31	31	31
17	15	25
19	27	45
13	11	27

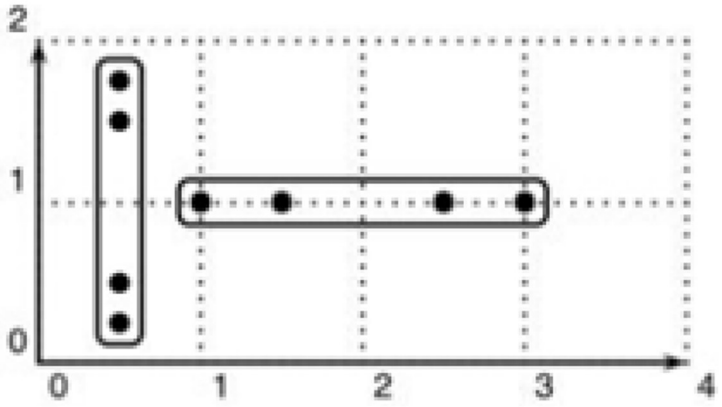
3. Consider three cluster centres A(2,3), B(4,5) and C(6,2). A point (1,2) is to be assigned to one of these clusters. According to k-means clustering concepts and using **Euclidean distance** as the measure of closeness, which cluster should it be assigned to? **(2)**

4. **Chebyshev distance (2)**

   Calculate the Chebyshev distance between the two points given in the table below.

	Feature 1	Feature 2	Feature 3	Feature 4
Point 1	2	3	1	-1
Point 2	4	-1	0	-2

5. Consider the points and the cluster membership depicted in the following image. Does this arrangement of points indicate **a convergence if k-means** is used to cluster these points in two clusters? **(5)**



6. Suppose a k-means algorithm is made to run on a data set. You are provided with two possible clustering's. Which of these clustering's is **more likely than the other? (5)**

Note: The **points represented here are equally spaced out** in the Euclidean space

7. Explain **Termination Condition** of K Means? **(1)**

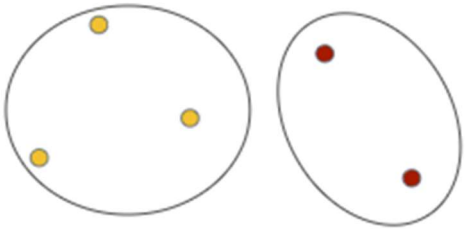
8. **True or False and Why : (2)**

"If you are worried about K-means getting stuck in bad local optima, then a way to solve this problem is using multiple random initialisations."

9. **Comment on Inter-Cluster vs Intra-Cluster distance (3)**

1. 

2. 

3. 

10. Comment on Silhouette Coefficient (3)

Silhouette Coefficient	Conclusion
Close to 1	
Close to 0	
Close to -1	

11. Consider the following two points that have been taken from a data set. (3)

ID	Height	Weight
1	150	50
2	180	60

You can find the statistics for the data given below.

	Mean	Standard Deviation
Weight	70	10
Height	160	20

Find the values of the data after standardization has been completed.

12. K-means algorithm (20)

In this exercise, you will perform k-means clustering **manually on a small data set**. Consider the following data set with two features and six observations.

Observation number	X1	X2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

Use **k = 2** for the entire exercise.

1. Choose the first two observations as your initial cluster centroids.
2. Assign each observation to the centroid to which it is closest (using Euclidean distance). Report the new cluster labels for each observation.
3. Recompute the cluster centroid on the basis of the points assigned to the cluster
4. Next, repeat the above steps until the clusters stop changing.

The observations (identified by their row numbers) belonging to the final two clusters, respectively, are_____.

II. Coding (50)

1 . K-Means Coding:

Download the attached data set about the batting figures of batsmen in ODI matches given below. Analyse the data and answer the following.

1.Calculate **Hopkins score** to know whether the data is good for clustering or not?

2.Choose the **number of clusters as 4** and find out Who falls in the same cluster as **Virat Kohli**?

3.Based on the clustering, find out **IVA Richards and SR Tendulkar** both belong to which group given that the clusters formed are (high SR, high Ave) - 1, (low SR, low Ave) - 2, (High SR, Low Ave) - 3, (Low SR, High Ave) - 4?

4.Validate using **Silhouette score** analysis the number of clusters you created is optimum number or not?

Note:

Choose **strike rate** and **average** as the **two factors** on which you will cluster the data.

Choose **random_state=100** for running K-Means in Python with SKLearn.

III. Achiever's Section (Bonus Section- Optional) (4+4+2)

1. If **N** is the number of data points, **K** is the number of clusters, and **T** is the number of iterations, then **K Means time complexity** is_____
2. Explain with picture **K Means Sensitivity to initial seeds**.
3. Explain **k-means++ initial seed selection** methodology