NAME: Dipesh Ramesh Limaje

INTERNSHIP BATCH : 33

TOPIC: MACHINE LEARNING

SME : Mr. Shwetank Mishra

WORKSHEET NO : 3

Q1. d. All of the above

Q2. d. None

Q3. c. Reinforcement learning and Unsupervised learning

Q4. b. The tree representing how close the data points are to each other

Q5. d. None

Q6. c. k-nearest neighbour is same as k-means

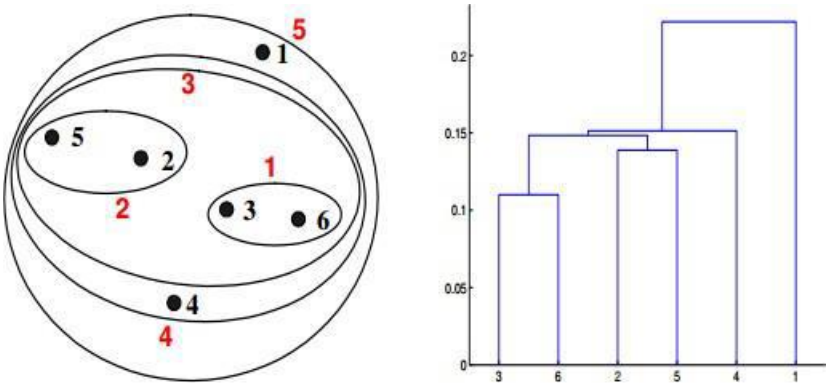
Q7. d. 1, 2 and 3

Q8. a. 1 only

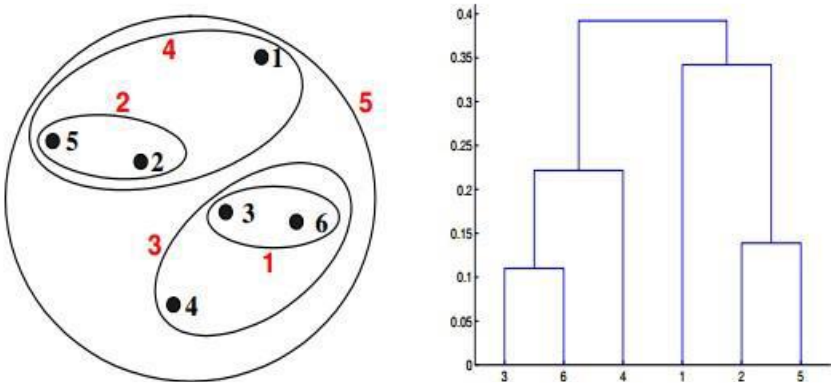
Q9. a. 2

Q10. b. Given a database of information about your users, automatically group them into different market segments.

Q11. a.



Q12. b.



Q13. What is the importance of clustering?

ANS: It is basically a type of unsupervised learning. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

Importance: Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters.

Q14. How can I improve my clustering performance?

1. We need to specify the number of clusters beforehand.
2. It is required to run the algorithm multiple times to avoid a sub-optimal solution
3. K-Means does not behave very well when the clusters have varying sizes, different densities, or non-spherical shapes.
4. The clusters sometimes vary based on the initial choice of the centroids. An important improvement to the K-Means algorithm, called K-Means, was proposed in a 2006 paper by David Arthur and Sergei Vassilvitskii. They introduced a smarter initialization step that tends to select centroids that are distant from one another, and this makes the K-Means algorithm much less likely to converge to a suboptimal solution.
5. Another important improvement to the K-Means algorithm was proposed in a 2003 paper by Charles Elkan.It considerably accelerates the algorithm by avoiding many unnecessary distance calculations: this is achieved by exploiting the triangle inequality (i.e., the straight line is always the shortest; in a triangle with sides a,b and c=> a+b>c) and by keeping track of lower and upper bounds for distances between instances and centroids.
6. Yet another important variant of the K-Means algorithm was proposed in a 2010 paper by David Sculley. Instead of using the full dataset at each iteration, the algorithm is capable of using mini-batches, moving the centroids just slightly at each iteration. This speeds up the algorithm typically by a factor of 3 or 4 and makes it possible to cluster huge datasets that do not fit in memory. Scikit-Learn implements this algorithm in the MiniBatchKMeans class. You can just use this class like the KMeans class