2 marks Questions or 4 Marks Questions

1. What is the difference between “supervised” and unsupervised” learning scheme?
2. What is clustering?
3. What are the requirements of clustering?
4. State the categories of clustering methods?
5. Difference between K-Means and K-Medoids Algorithms.
6. What do you mean by Hierarchical Clustering?
7. What do you mean by Agglomerative Clustering?
8. What do you mean by Outlier Detection?
9. What do you mean by divisive Clustering?
10. Both k-means and k-medoids algorithms can perform effective clustering.
    (a) Illustrate the strength and weakness of k-means in comparison with the k-medoids algorithm.
    (b) Illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme (such as AGNES).
11. Show that BCubed metrics satisfy the four essential requirements for extrinsic clustering evaluation methods.
12. Give an example of how specific clustering methods may be integrated, for example, where one clustering algorithm is used as a preprocessing step for another.
13. Give an application example where global outliers, contextual outliers and collective outliers are all interesting. What are the attributes, and what are the contextual and behavioral attributes? How is the relationship among objects modeled in collective outlier detection?

13. Give an application example of where the border between “normal objects” and outliers is often unclear, so that the degree to which an object is an outlier has to be well estimated.
14. In outlier detection by semi-supervised learning, what is the advantage of using objects without labels in the training data set?

12 marks Questions

1. Discuss about k-nearest neighbor classification algorithm with an example?
2. What do you mean by Clustering? Explain the requirements used in Clustering?
3. Briefly describe and give examples of each of the following approaches to clustering:

partitioning methods, hierarchical methods, density-based methods and grid-based methods.

4. Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are $A_1(20, 100)$, $A_2(4, 10)$, $B_1(6, 9)$, $B_2(8, 10)$, $B_3(6, 4)$, $C_1(1, 2)$, $C_2(4, 9)$, $C_3(3,5)$. The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the k-means algorithm to show only
    a. the three cluster centers after the first round of execution.
    b. the final three clusters.
5. Use an example to show why the k-means algorithm may not find the global optimum, that is, optimizing the within-cluster variation.
6. Explain in detail about Hierarchical Clustering.
7. Explain in detail about partitional Clustering method.
8. Clustering has been popularly recognized as an important data mining task with broad applications.
9. Give one application example for each of the following cases:

   (a) An application that takes clustering as a major data mining function

   (b) An application that takes clustering as a preprocessing tool for data preparation for other data mining tasks
10. Prove that in DBSCAN, for a fixed MinPts value and two neighborhood thresholds $\epsilon 1 < \epsilon 2$, a cluster C with respect to $\epsilon 1$ and MinPts must be a subset of a cluster C′ with respect to $\epsilon 2$ and MinPts.
11. Discuss about Outlier Detection.
12. Why is outlier mining important? Briefly describe the different approaches behind statistical-based outlier detection, distance-based outlier detection, and deviation-based outlier detection.
13. Explain in detail about Clustering methods with an example.
14. Clustering is recognized as an important data mining task with broad applications. Give one application example for each of the following cases:

    (a) An application that takes clustering as a major data mining function

    (b) An application that takes clustering as a preprocessing tool for data preparation for other data mining tasks.
15. Explain briefly how data is analyzed by data mining in Finance Sector
16. Data cubes and multidimensional databases contain categorical, ordinal, and

numerical data in hierarchical or aggregate forms. Based on what you have learned about the clustering methods, design a clustering method that finds clusters in large data cubes effectively and efficiently.

17. Describe each of the following clustering algorithms in terms of the following criteria: (1) shapes of clusters that can be determined; (2) input parameters that must be specified; and (3) limitations.
    a) k-means
    b) k-medoids
    c) DBSCAN
18. Research and describe an application of data mining that was not presented in this chapter. Discuss how different forms of data mining can be used in the application.
19. What is a recommender system? In what ways does it differ from a customer or product-based clustering system? How does it differ from a typical classification or predictive modeling system? Outline one method of collaborative filtering. Discuss why it works and what its limitations are in practice.