Machine learning worksheet 3

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is an application of clustering?
   a. Biological network analysis
   b. Market trend prediction
   c. Topic modeling
   d. All of the above
   - d) All of the above

2. On which data type, we cannot perform cluster analysis?
   a. Time series data
   b. Text data
   c. Multimedia data
   d. None
   - d) None

3. Netflix's movie recommendation system uses-
   a. Supervised learning
   b. Unsupervised learning
   c. Reinforcement learning and Unsupervised learning
   d. All of the above
   - c) Reinforcement learning and Unsupervised learning.

4. The final output of Hierarchical clustering is-
   a. The number of cluster centroids
   b. The tree representing how close the data points are to each other
   c. A map defining the similar data points into individual groups
   d. All of the above
   - b) The tree representing how close the data points are to each other.

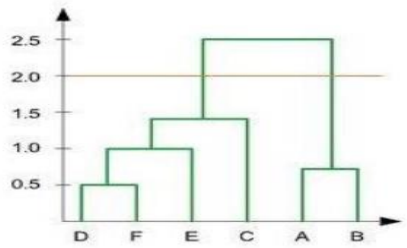
5. Which of the step is not required for K-means clustering?
   a. A distance metric
   b. Initial number of clusters
   c. Initial guess as to cluster centroids
   d. None
   - d) None

6. Which is the following is wrong?
   a. k-means clustering is a vector quantization method
   b. k-means clustering tries to group n observations into k clusters
   c. k-nearest neighbour is same as k-means
   d. None
   - c) k-nearest neighbour is same as k-means

7. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?
   i. Single-link
   ii. Complete-link
   iii. Average-link
   Options:
   a.1 and 2
   b. 1 and 3
   c. 2 and 3
   d. 1, 2 and 3
   - d) 1,2 and 3

8. Which of the following are true?
   i. Clustering analysis is negatively affected by multicollinearity of features
   ii. Clustering analysis is negatively affected by heteroscedasticity
   Options:
   a. 1 only
   b. 2 only
   c. 1 and 2
   d. None of them
   - a) 1 only

9. In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?



   a. 2
   b. 4
   c. 3
   d. 5
   - b) 2

10. For which of the following tasks might clustering be a suitable approach?
    a. Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.
    b. Given a database of information about your users, automatically group them into different market segments.
    c. Predicting whether stock price of a company will increase tomorrow.
    d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.
    - a) Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

11. Given, six points with the following attributes:

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

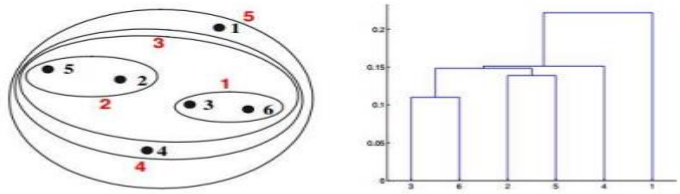
Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

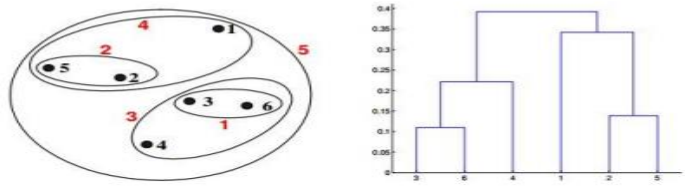
Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

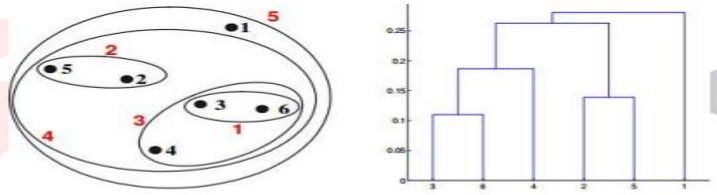
a.



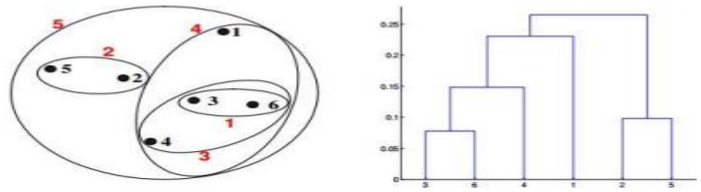
b.



c.



d.



➢ c)

12. Given, six points with the following attributes:

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

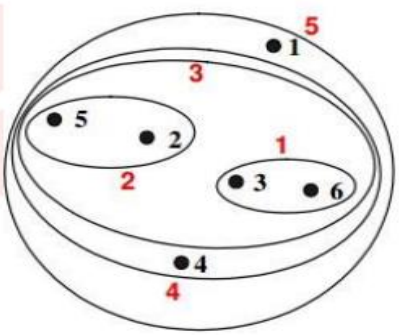
Table : X-Y coordinates of six points.

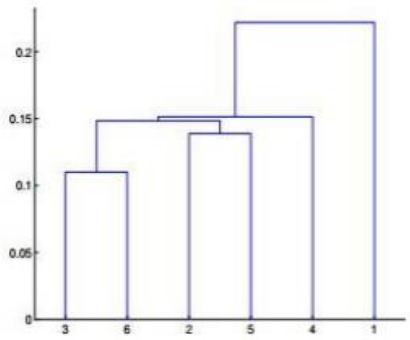
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

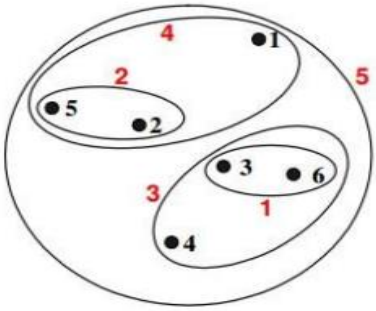
Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.

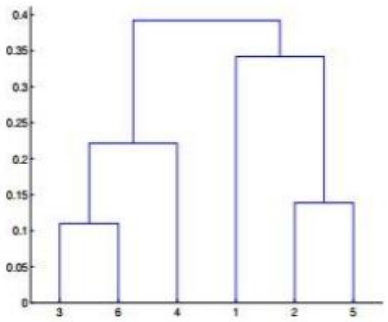
a.



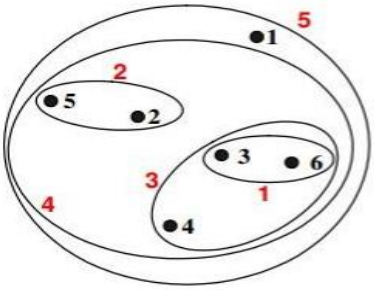


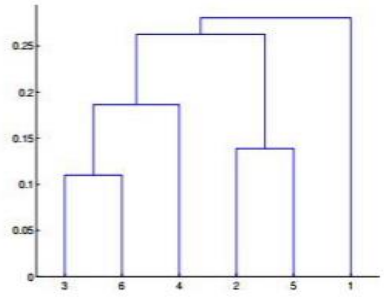
b.



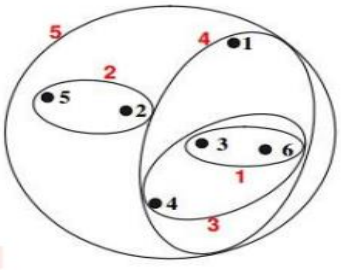


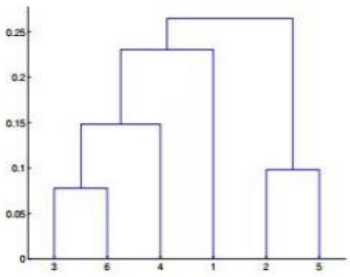
c.





d.





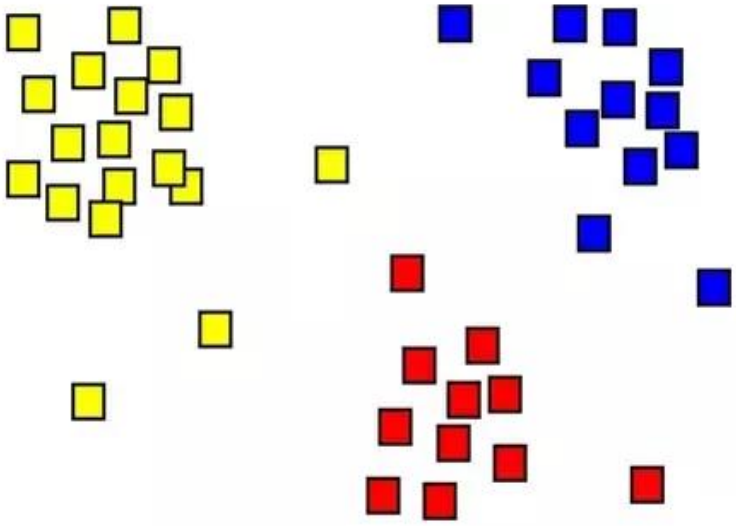
- a)

Q13 to Q14 are subjective answers type questions, Answers them in their own words briefly

13. What is the importance of clustering?

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Cluster analysis is an unsupervised form of learning , which means, that it doesn't use class labels. This is different from methods like discriminant analysis which use class labels and come under the category of supervised learning. K-means is the most simple and popular algorithm in clustering and was published in 1955, 50 years ago.

Clustering is used to find structure in unlabeled data. It's the most common form of unsupervised learning. Given a dataset you don't know anything about, a clustering algorithm can discover groups of objects where the average distances between the members of each cluster are closer than to members in other clusters, such as this:



This illustrates a simple 2-dimensional example. Clusters will usually have a higher dimensionality.

Clustering has many practical applications. For instance, it's used in marketing to assess the demographics of consumers. By knowing more about different market segments you can target consumers more accurately with commercials.

14. How can I improve my clustering performance?

- The k-means algorithm

  A solution for this problem is the k-means algorithm, which uses a different initialization. The idea is pretty simple:

  Instead of random initialization, we only choose the first center randomly. All following centers are then still sampled, but with a probability that is proportional to their squared distance from all current centers. Points further

away from current centers get a higher probability to become a center in the next iteration of initialization.

This attempts to fill the space of the observations more evenly, while still retaining some randomness. Even with k-means, the outcome can differ between multiple runs on the same data. While it does require some more computation at the beginning of the algorithm, it leads to much faster convergence, making it highly competitive to vanilla k-means in regards to runtime. Therefore, many common libraries use k-means initialization as their default, for example sk-learn or the MatLab implementation.

- Principal Component Analysis

  As a pre processing stage of data mining and machine learning, dimension reduction not only decreases computational complexity, but also significantly improves the accuracy of the learned models from large data sets. PCA [11] is a classical multivariate data analysis method that is useful in linear feature extraction. Without class labels it can compress the most information in the original data space into a few new features, i.e., principal components. Handling high dimensional data using clustering techniques obviously a difficult task in terms of higher number of variables involved. In order to improve the efficiency, the noisy and outlier data may be removed and minimize the execution time and we have to reduce the no. of variables in the original data set

  The central idea of PCA is to reduce the dimensionality of the data set consisting of a large number of variables. It is a statistical technique for determining key variables in a high dimensional data set that explain the differences in the observations and can be used to simplify the analysis and visualization of high dimensional data set.