# MACHINE LEARNING – WORKSHEET (CLUSTERING)

## 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
   1. Biological network analysis
   2. Market trend prediction
   3. Topic modeling
   4. All of the above

**ANSWER: c**

1. On which data type, we cannot perform cluster analysis?
   1. Time series data
   2. Text data
   3. Multimedia data
   4. None

**ANSWER: a**

1. Netflix’s movie recommendation system uses-
   1. Supervised learning
   2. Unsupervised learning
   3. Reinforcement learning
   4. All of the above

**ANSWER: a**

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

**ANSWER: b**

1. Which of the step is not required for K-means clustering?
   1. a distance metric
   2. initial number of clusters
   3. initial guess as to cluster centroids
   4. None

**ANSWER: d**

1. Which is the following is wrong?
   1. k-means clustering is a vector quantization method
   2. k-means clustering tries to group n observations into k clusters
   3. k-nearest neighbor is same as k-means
   4. None

**ANSWER: c**

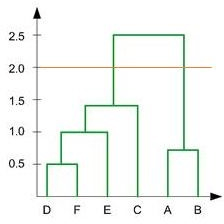
1. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?
2. Single-link
3. Complete-link
4. Average-link Options:
   1. 1 and 2
   2. 1 and 3
   3. 2 and 3
   4. 1, 2 and 3

**ANSWER: d**

1. Which of the following are true?
2. Clustering analysis is negatively affected by multicollinearity of features
3. Clustering analysis is negatively affected by heteroscedasticity Options:
   1. 1 only
   2. 2 only
   3. 1 and 2
   4. None of them

**ANSWER : a**

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



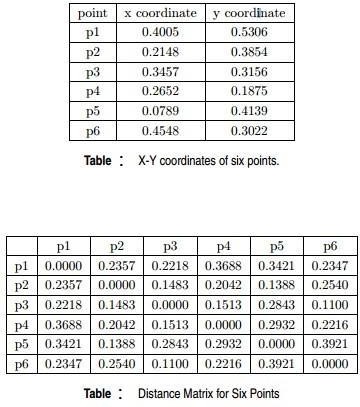
* 1. 2
  2. 4
  3. 3
  4. 5

**ANSWER: a**

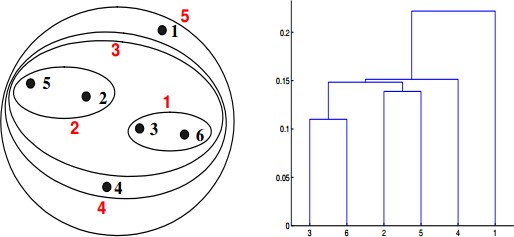
1. For which of the following tasks might clustering be a suitable approach?
2. Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.
3. Given a database of information about your users, automatically group them into different market segments.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

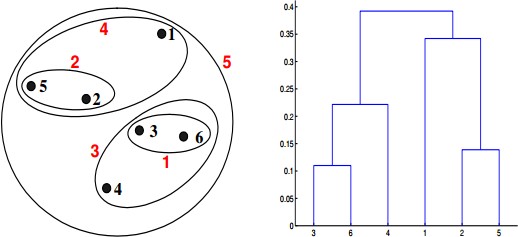
**ANSWER: b**

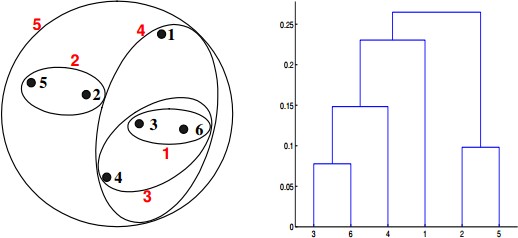
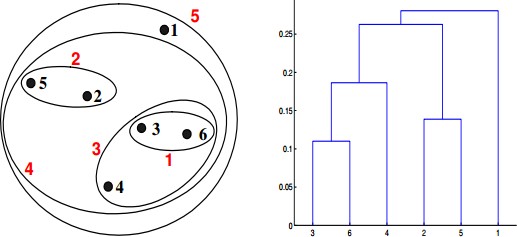
1. Given, six points with the following attributes:



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

A. 

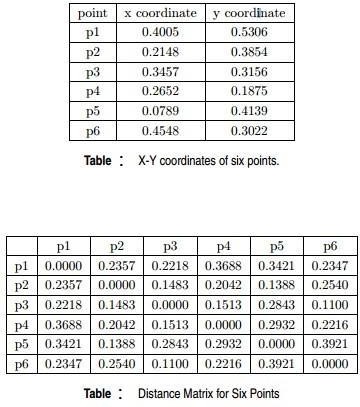
B

C. 

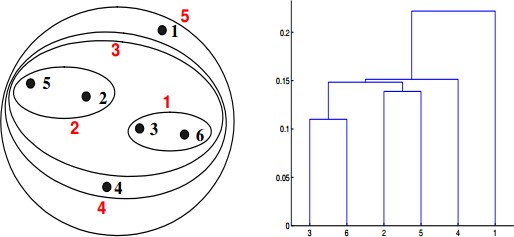
D.

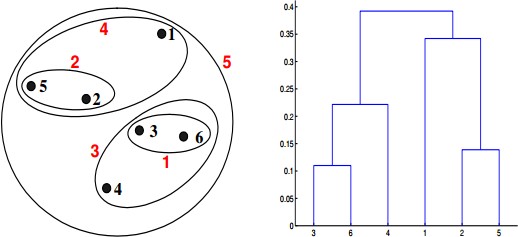
**ANSWER : a**

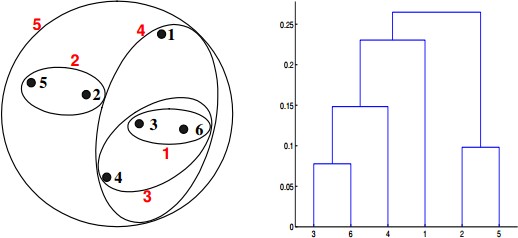
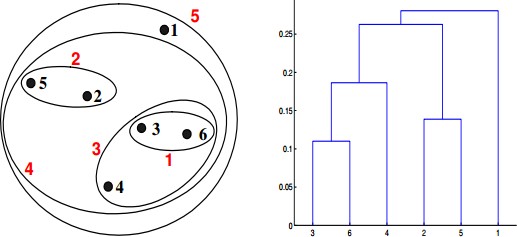
1. Given, six points with the following attributes:



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.

**ANSWER : b**

## Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly

1. What is the importance of clustering?

**ANSWER :** Clustering is valuable for investigating data. If there are many cases and no obvious groupings, Clustering calculations can be utilized to discover common groupings. Clustering can likewise fill in as a valuable data preprocessing step to recognize homogeneous gatherings on which to construct models.

1. How do you cluster a profile?

**ANSWER:**

Stage 1: Graphically address your groups as indicated by your input factors.

Stage 2: Score your clusters in a table so you can gauge and contrast them on each information variable and respects to mathematical or illustrative qualities.

Step 3: Now it’s time to profile your clusters. At this step, variables should be described in a type of ‘story’ about the category.

The output of this step is a clearly described set of clusters with a focus placed on the input variables.

1. How can I improve my clustering performance?

**ANSWER :** clustering performance can be significantly improved by using a better initialization technique, and by repeating (re-starting) the algorithm.