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 a. Biological network analysis

b. Market trend prediction c. Topic modeling

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

**3**. Netflix’s movie recommendation system uses- a. Supervised learning

b. Unsupervised learning c. Reinforcement learning d. All of the above

**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

**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

**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 neighbor is same as k-means

d. None

**7.** Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?

1. Single-link

2. Complete-link

3. Average-link

Options: a. 1 and 2 b. 1 and 3 c. 2 and 3

d. 1, 2 and 3

**8**. Which of the following are true?

1. Clustering analysis is negatively affected by multicollinearity of features

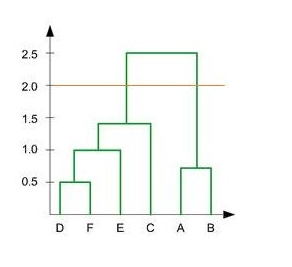
2. Clustering analysis is negatively affected by heteroscedasticity

Options: a. 1 only b. 2 only c. 1 and 2

d. None of them

**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



**10**. For which of the following tasks might clustering be a suitable approach?

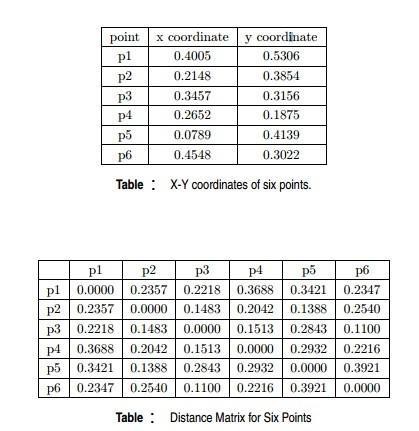
a. Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

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.

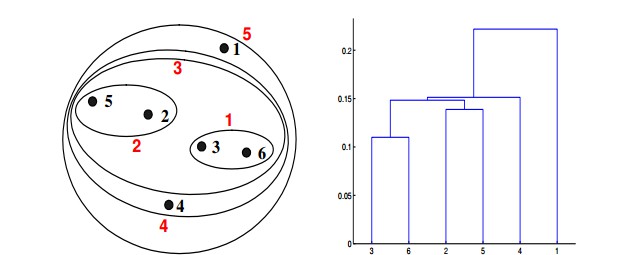
d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

**11**. Given, six points with the following attributes:

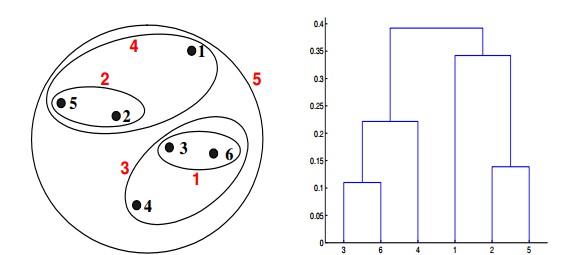


Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

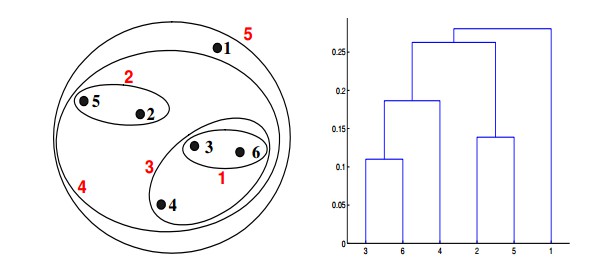
A.



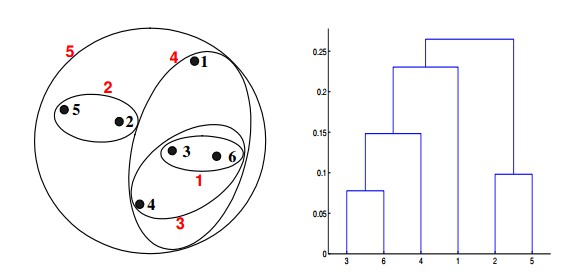
B



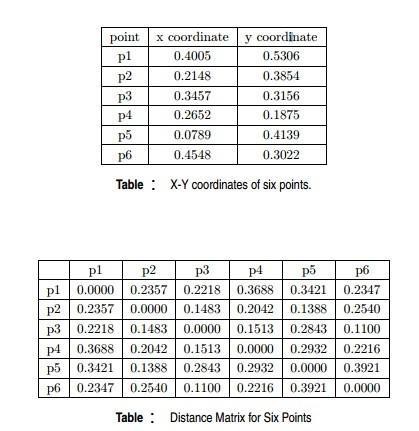
C.



D.



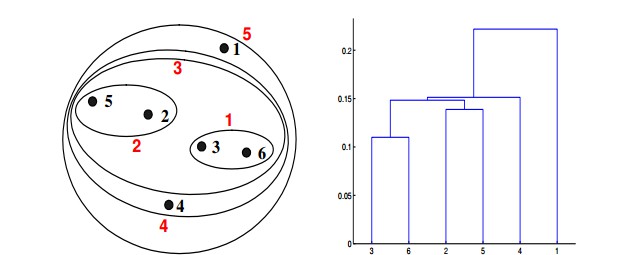
**12**. Given, six points with the following attributes:



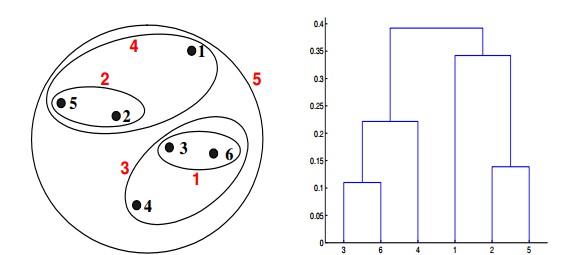
Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link

proximity function in hierarchical clustering:

A

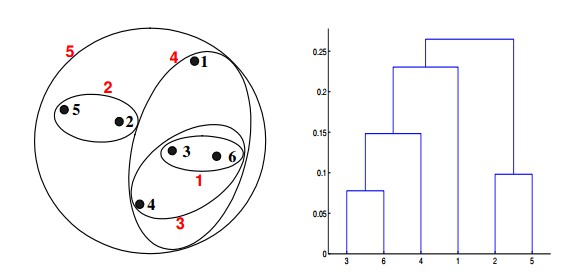
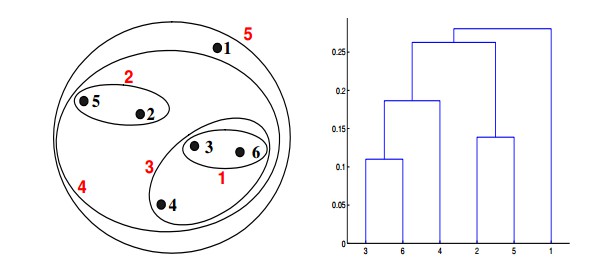


B.



C.

D.



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

**13**. What is the importance of clustering?

**Ans.** Clustering or unsupervised data analysis can be useful for several purposes.

The most frequent case is for explorative analysis, when NOBODY knows if the data you are analysing are characterised by a small number of representative patterns that can be used to summary the dataset in a more compact representation (groups, partitions, centroids, etc)

Discovering POSSIBLE partitions is usually based on some sort of similarity between the data variables. POSSIBLE underlined again. And everything depends on how you define the problem you want to study (variable engineering) and how you define “these two things are more similar than these other two things”.

Another case is to evaluate the presence of outliers. IF you are SURE that the data should show a certain set of patterns (similarity-based groups etc) you can check if some data samples are not following those patterns, and analyse them individually to understand why.

**14**. How do you cluster a profile?

**Ans**. Profiling involves generating descriptions of the clusters with reference to the input variables you used for the cluster analysis. Profiling acts as a class descriptor for the clusters and will help you to ‘tell a story’ so that you can understand this information and use it across your business. This approach is most useful where a multivariate cluster analysis has been conducted. To profile, interpret and understand your clusters, you can follow the steps below.

**Step 1:** Graphically represent your clusters according to your input variables. As you can see in the example below, we have clustered the raw data according to purchase frequency and monetary value. This has produced three distinct clusters. The selected clustering algorithm aims to maximise the similarity between the data points in the same cluster and minimise the similarity between data points in different clusters.

**Step 2:** Score your clusters in a table so that you can measure and compare them on each input variable with regards to numerical or descriptive values.

**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 or customer base. This will help buyers and marketers to use this information strategically with an in-depth understanding of the differences between each cluster and which variables define the groupings. This step sets cluster profiling apart from traditional segmentation.

**15**. How can I improve my clustering performance?

**Ans**. Improving clustering performance using independent component analysis and unsupervised feature learning. Principal Component Analysis (PCA) is an important approach to unsupervised dimensionality reduction technique. 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.