MACHINE LEARNING

1) Which of the following is an application of clustering?

Answer: d. All of the above

2) On which data type, we cannot perform cluster analysis?

Answer: d. None

3) Netflix's movie recommendation system uses-

Answer: c. Reinforcement learning

4) The final output of Hierarchical clustering is-

Answer: 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?

Answer: d. None

6) Which is the following is wrong?

Answer: 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?

Answer: d. 1, 2 and 3

8) Which of the following are true?

Answer: 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?

Answer: a. 2

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

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

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

Answer: Option A

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

Answer: Option B

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

13) What is the importance of clustering?

Answer: Clustering is used to find structure in unlabelled 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. It can be used for different market segments like: customer segmentation, data analysis, dimensionality reduction technique, anomaly detection, semi-supervised learning, search engines, segment an image etc.

14) How can I improve my clustering performance?

Answer: we can use K-Means++. In this 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.

'k-means++' : selects initial cluster centroids using sampling based on an empirical probability distribution of the points' contribution to the overall inertia. This technique speeds up convergence.

We can also use the MiniBatchKMeans class

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: