WORKSHEET 3

MACHINE LEARNING

1.Application of clustering - d) All of the above (Biological network analysis, Market trend prediction, Topic modelling)

2. cannot perform cluster analysis - d) None

3. Netflix's movie recommendation system uses – c) Reinforcement learning and Unsupervised learning

4. The final output of Hierarchical clustering is – 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 – d) None

6. Which is the following is wrong – 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 – d) 1, 2 and 3(Single-link , Complete-link, Average-link)

8. Which of the following are true? – a) 1 only(Clustering analysis is negatively affected by multicollinearity of features)

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

10. For which of the following tasks might clustering be a suitable approach? b) Given a database of information about your users, automatically group them into different market segments.

11. Given, six points with the following attributes: a)

12.Given, six points with the following attributes: - b)

13. Importance of clustering

Clustering is an easy way to perform many surface-level analyses that can give you quick wins in a variety of fields. Marketers can perform a cluster analysis to quickly segment customer demographics, for instance. Insurers can quickly drill down on risk factors and locations and generate an initial risk profile for applicants. Clustering is useful for exploring data. If there are many cases and no obvious groupings, clustering algorithms can be used to find natural groupings. Clustering can also serve as a useful data pre-processing step to identify homogeneous groups on which to build supervised models. 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.

Much like with other useful algorithms and data science models, you'll get the most out of clustering when you deploy it not as a standalone, but as part of a broader data discovery strategy. Customer cluster analysis can help you segment your audience, classify your data better, and generally structure your datasets, but it won't do much more if you don't give your input data a broader context.

14.Improve my clustering performance

Clustering analysis is one of the main analytical methods in data mining. K-means is the most popular and partition-based clustering algorithm. But it is computationally expensive and the quality of resulting clusters heavily depends on the selection of initial centroid and the dimension of the data. Several methods have been proposed in the literature for improving performance of the k-means clustering algorithm. Principal Component Analysis (PCA) is an important approach to unsupervised dimensionality reduction technique Improving clustering performance using independent component analysis and unsupervised feature learning. 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. The main objective of applying PCA on original data before clustering is to obtain accurate results. But the clustering results depend on the initialization of centroid.