WORKSHEET SET 3

MACHINE LEARNING WORKSHEET 3

OBJECTIVE TYPE QUESTIONS

Q1. D) ALL OF THE ABOVE

Q2. D) NONE

Q3. C) REINFORCEMENT AND UNSUPERVISED LEARNING

Q4. B) THE TREE REPRESENTING HOW CLOSE THE DATA POINTS ARE TO EACH OTHER

Q5. D) NONE

Q6. C) K-NEAREST NEIGHBOUR IS SAME AS K-MEANS

Q7. D) 1- SINGLE LINK, 2- COMPLETE LINK AND 3- AVERAGE LINK

Q8. A) 1 ONLY – CLUSTERING ANALYSIS IS NEGATIVELY AFFECTED BY MULTICOLLINEARITY OF FEATURES

Q9. A) 2

Q10. B) GIVEN A DATABASE OF INFORMATION ABOUT YOUR USERS, AUTOMATICALLY GROUP THEM INTODIFFERENT MARKET SEGMENTS AND C) PREDICTING WHETHER STOCK PRICE OF A COMPANY WILL INCREASE TOMORROW

Q11. A)

Q12. B)

SUBJECTIVE TYPE QUESTIONS

Q13. **IMPORTANCE OF CLUSTERING**

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).

IMPORTANCE OF CLUSTERING

1. Cluster sampling requires fewer resources.
A cluster sampling effort will only choose specific groups from within an entire population or demographic. That means this method requires fewer resources to complete the research work. That’s why it is one of the cheapest investigatory options that’s available right now, even when compared to simple randomization or stratified sampling. Even when the costs of obtaining data are similar, cluster sampling typically requires fewer administrative and travel expenses.

2. It is a feasible way to collect statistical information.
The division of a demographic or an entire population into homogenous groups increases the feasibility of the process for researchers. Because every cluster is a direct representation of the people being studied, it is easy to include more subjects in the project as needed to obtain the correct level of information.
The design of cluster samples makes it a simple process to manage massive data input. It takes large population groups into account with its design to ensure that the extrapolated information gets collected into usable formats.

3. The cluster sampling approach reduces variabilities.
Every research effort creates estimates as the discovered statistics get extrapolated to the rest of the population. When investigators use cluster samples to generate this information, then the estimation has more accuracy to it when compared to the other methods of collection. Researchers must make their best effort to ensure that each cluster is a direct representation of the population or demographic to achieve this benefit. Then the data obtained from this method offers reduced variability with its results since the findings are closer to a direct reflection of the entire group.

4. Researchers can conduct cluster sampling almost anywhere.
When resources are tight and research is required, cluster sampling is a popular method to use because of its structures. You can take a representative sample from anywhere in the world to generate the results that you want. Although geographic variability will increase the error rate in the sample by a small margin, it also opens the door to localized efforts that can still be useful to the overall demographic.

5. Researchers receive the benefits of stratified and random sampling with this method.
Cluster sampling is a popular research method because it includes all of the benefits of stratified and random approaches without as many disadvantages. This benefit works to reduce the potential for bias in the collected data because it simplifies the information assembly work required of the investigators. Because there are fewer risks of adverse influences creating random variations, the results of the work can generate exclusive conclusions when applied to the overall population.

6. It gives researchers a large data sample from which to work.
When you work with a larger population group, then you're creating more usable data that can eventually lead to unique findings. After researchers design and place the cluster sampling method on their preferred demographic, then similar information gets collected from each group. Investigators can then compare data points between the clusters to look for specific conclusions within a particular population group.
This advantage generates tracking data that looks at how individual clusters evolve in the future when compared to the rest of the population group. Then researchers can use that variability to understand more of the differences that can lead to a higher error rate.

Q14. HOW CAN I IMPROVE MY CLUSTERING PERFORMANCE?

K-means clustering algorithm can be significantly improved by using a better initialization technique, and by repeating (re-starting) the algorithm. When the data has overlapping clusters, k-means can improve the results of the initialization technique. When the data has well separated clusters, the performance of k-means depends completely on the goodness of the initialization. Initialization using simple furthest point heuristic (Max-Min) reduces the clustering error of k-means from 15% to 6%, on average.

Graph-based clustering performance can easily be improved by applying ICA blind source separation during the graph Laplacian embedding step.

There are two important elements in improving the quality of clustering:-

- improving the weights of the features in a document vector and
- creating a more appropriate distance measure.