WORKSHEET –MACHINE LEARNING

(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:

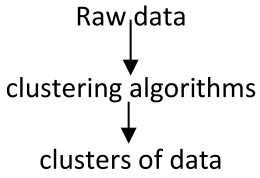
A B. C. D.

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?
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). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics
Data mining involves the anomaly detection, association rule learning, classification, regression, summarization and clustering. In this paper, clustering analysis is done. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups. A good clustering algorithm is able to identity clusters irrespective of their shapes. The stages involved in clustering algorithm are as follows

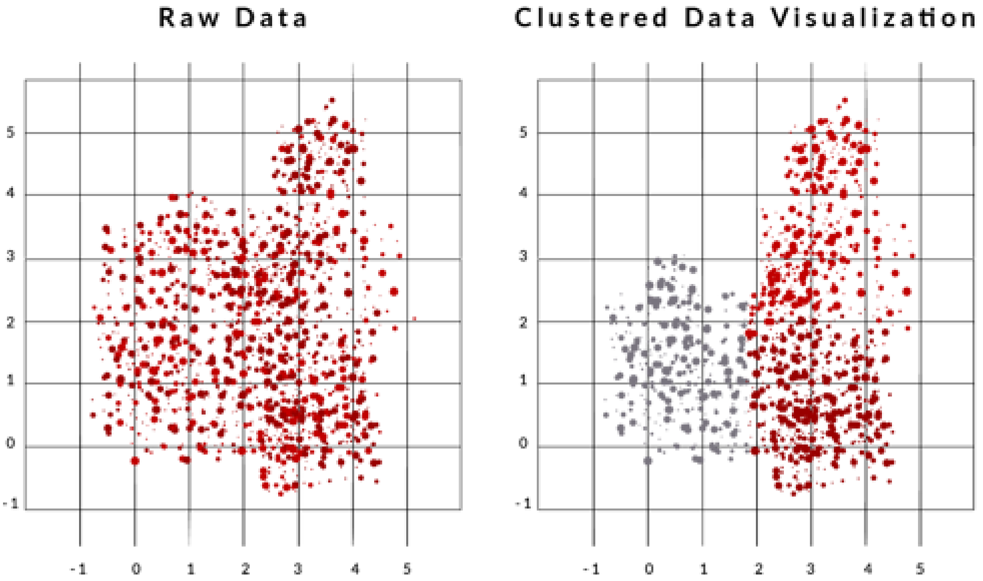


Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goal. Clustering helps in

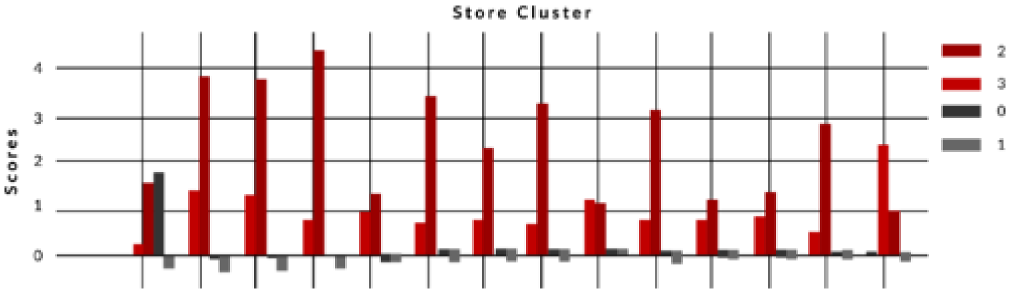
- Organizing data into clusters shows internal structure of the data
- Sometimes the partitioning is the goal
- Prepare for other AI techniques
- Techniques for clustering is useful in knowledge discovery in data

14. How do we cluster a profile?

To profile, we can graphically represent our clusters according to our input variables. As we 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.



Score our clusters in a table so that we can measure and compare them on each input variable with regards to numerical or descriptive values.



Now it’s time to profile 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.

The output is a clearly described set of clusters with a focus placed on the input variables. If we have access to a wider set of data, we can use other loyalty data to supplement the cluster profile even if it was not used in the original cluster analysis.

15. How can I improve my clustering performance?
Clustering is an unsupervised machine learning methodology that aims to partition data into distinct groups, or clusters. There are a few different forms including hierarchical, density, and similarity based. Each have a few different algorithms associated with it as well. One of the hardest parts of any machine learning algorithm is feature engineering, which can especially be difficult with clustering as there is no easy way to figure out what best segments your data into separate but similar groups.

The guiding principle of similarity based clustering is that similar objects are within the same cluster and dissimilar objects are in different clusters. This is not different than the goal of most conventional clustering algorithms. With similarity based clustering, a measure must be given to determine how similar two objects are. This similarity measure is based off distance, and different distance metrics can be employed, but the similarity measure usually results in a value in [0,1] with 0 having no similarity and 1 being identical. To measure feature weight importance, we will have to use a weighted euclidean distance function. The similarity measure is defined in the following:

$$\rho_{ij}^{(w)} = \frac{1}{1 + \beta * d_{ij}^{(w)}}$$

β here is a value that we will actually have to solve for, (w) represents the distance weight matrix, and d represents the pairwise distances between all objects. To solve for β, we have to use the assumption that if using the standard weights(all 1's), our similarity matrix would uniformly distributed between [0,1] resulting in a mean of .5. So to find β, we solve the equation:

$$\frac{2}{n(n-1)} \sum_{j<i} \frac{1}{1 + \beta * d_{ij}} = 0.5$$

If using a weighted euclidean distance, it is possible to use this similarity matrix to identify what features introduce more noise and which ones are important to clustering. The ultimate goal is to minimize the “fuzziness” of the similarity matrix, trying to move everything in the middle (ie .5) to either 1 or 0. For this purpose we use the loss metric:

$$E(w) = \frac{2}{N(N-1)} \sum_{q<p} \frac{1}{2} \left(\rho_{pq}^{(w)} \left(1 - \rho_{pq}^{(1)}\right) + \rho_{pq}^{(1)} \left(1 - \rho_{pq}^{(w)}\right) \right)$$

Here (1) represents the base weights (all 1's), and ρ represents the resulting fuzzy partition matrix that is a product of the weights used in the euclidean distance function between points p and q.

We can then attempt to use Gradient Descent on this loss function to try and minimize it with respect to the similarity matrix. Gradient Descent is one of the most common optimization algorithms in machine learning that is used to find best parameters of a given function by using the function gradient, a combination of the partial derivatives. By taking steps proportional to the negative of the gradient, we can try to find the local minimum of the function. We will continually update the weights until either our maximum number of iterations has been met, or the function converges. So the gradient descent will be of our loss function with a partial derivative in respect to the weights. We will update the weights every iteration with respect to the gradient and learning rate.

$$\triangle w_j = -\eta \frac{\partial E(w)}{\partial w_j},$$

Where n is the learning rate defined. n is a very important parameter, as something too small will require too much computation, while too big and the function may never converge.

We can think of it in terms of a 3D graph, it would be like stretching or shrinking each axis, in a way that would put our points into tighter groups, that are further away from each other. We are not actually changing the locations of the data, we are solely transforming how we measure the distances that drive our similarity metrics.