|  |
| --- |
| **Unit VII- Cluster Analysis:**  **Content:** Introduction, Types of clustering, Correlations and distances, clustering by partitioning methods, hierarchical clustering, overlapping clustering, K-Means Clustering-Profiling and Interpreting Clusters. |
| **What is clustering?**   * Clustering: the act of grouping “similar” object into sets. * A clustering problem can be viewed as unsupervised classification. * Clustering is appropriate when there is no a priori knowledge about the data. * Grouping is based on the distance (proximity)   you don’t know who or what belongs to which group. Not even the number of groups.  [In a supervised learning model, the algorithm learns on a labeled dataset, providing an answer key that the algorithm can use to evaluate its accuracy on training data. An unsupervised model, in contrast, provides unlabeled data that the algorithm tries to make sense of by extracting features and patterns on its own.]   * Given a collection of n objects, individuals, items, animals’ etc., each of which is described by a set of p variables, derive a useful division into a number of classes. * Both the number of classes and the properties of the classes are to be determined.      * The solution generally sought is a partition of the n items, that is a set of clusters where an object belongs to one cluster only, and the complete set of clusters contains all objects. * The grouping is done in such a way that items within a particular group are similar to each other but items between the groups are dissimilar to each other. This will be done based on suitable measure of similarity and dissimilarity.   **Why clustering/grouping?**  In the widest sense, a classification scheme may represent simply a convenient method for organising a large set of data so that the retrieval of information may be made more efficiently. |
| **Examples :**   1. A large number of cities are available that could be used as test markets, but due to economic factors testing must be restricted to only a small number of these. Clustering the cities into a small number of groups such that cities within a group are very similar to each other, and then choosing one city from each group, could be used as a means of selecting the test markets, cities can be classified on the basis of the variables such as city size, newspaper circulation, per capita income and so on. 2. Items produced in a particular shift can be investigated if there is any grouping with respect to process parameters or with respect to some output characteristics. 3. A bank wants to give credit card offers to its customers. Currently, they look at the details of each customer and based on this information, decide which offer should be given to which customer.   Now, the bank can potentially have millions of customers. Does it make sense to look at the details of each customer separately and then make a decision? Certainly not! It is a manual process and will take a huge amount of time.  So, what can the bank do? One option is to segment its customers into different groups. For instance, the bank can group the customers based on their income:  customer segmentation |
| **Applications of Clustering Applications**   * Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs * Land use: Identification of areas of similar land use in an earth observation database * Insurance: Identifying groups of motor insurance policy holders with a high average claim cost * City-planning: Identifying groups of houses according to their house type, value, and geographical location * The field of psychiatry: The characterization of patients on the basis of clusters of symptoms can be useful in the identification of an appropriate form of therapy. * Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults * Climate - Understanding the Earth’s climate requires finding patterns in the atmosphere and ocean. To that end, cluster analysis has been applied to find patterns in the atmospheric pressure of Polar Regions and areas of the ocean that have a significant impact on land climate. * Biology - used to find groups of genes that have similar functions. |
| **Common Role that Cluster analysis Plays:**   * **Data Reduction**:-A researcher may be faced with a large number of observations that can be meaningless unless classified into manageable groups. CA can perform this data reduction procedure objectively by reducing the info from an entire population of sample to info. About specific groups. * **Hypothesis Generation: -** Cluster analysis is also useful when a researcher wishes to develop hypotheses concerning the nature of the data or to examine previously stated hypotheses. |
| **There are two main types of classification:**   1. *k*-means clustering 2. Hierarchical clustering |
| **Hierarchical clustering**  **Hierarchical method** :  In a hierarchical clustering the data are not partitioned into a number of clusters at a single step. Instead the classification consists of a series of partitions, which may run from a single cluster containing all individuals to n clusters containing a single individual.  In Hierarchical clustering, objects that belong to a child cluster also belong to the parent cluster.    **Hierarchical clustering** are nested by this we mean that it also clusters to exist within bigger clusters  **Agglomerative method**:  In starts with the individual object, thus there are initially as many clusters as objects. The most similar objects are first grouped, and these initial groups are merged according to their similarities. Eventually, all subgroups are merged into a single cluster.  Hierarchical (Agglomerative**)** clustering will help to determine the optimal number of clusters.   1. It starts by putting every point in its own cluster, so each cluster is a singleton 2. It then merges the 2 points that are closest to each other based on the distances from the distance matrix. The consequence is that there is one less cluster 3. It then recalculates the distances between the new and old clusters and save them in a new distance matrix which will be used in the next step 4. Finally, steps 1 and 2 are repeated until all clusters are merged into one single cluster including all points.   There exists 5 main methods to measure the distance between clusters, referred as linkage methods:   1. **Single linkage**: computes the minimum distance between clusters before merging them. 2. **Complete linkage**: computes the maximum distance between clusters before merging them. 3. **Average linkage**: computes the average distance between clusters before merging them. 4. **Centroid linkage**: calculates centroids for both clusters, then computes the distance between the two before merging them. 5. Ward’s (minimum variance) criterion: minimizes the total within-cluster variance and find the pair of clusters that leads to minimum increase in total within-cluster variance after merging.   **Divisive method:**  Divisive method works in the opposite direction of agglomerative method. An initial single group of objects is divided into two sub-groups such that the objects in one subgroup are ‘far from’ the objects in the other group. These subgroups are then further divided into dissimilar subgroups, the process continues until there are as many sub-groups as objects; that is until each objects forms a group. |
| **How does Cluster Analysis Works?**  The primary objective of cluster analysis is to define the structure of the data by placing the most similar observations into groups. To accomplish this task, we must address three basic questions:  ◦ How do we measure similarity?  ◦ How do we form clusters?  ◦ How many groups do we form? |
| **Data :**  The basic data for cluster analysis is, in general a matrix **X**, giving the variable values for each of the n objects / items under investigation, that is      **Example :**  **Table 1: Data on chemical compositons of 6 heats.**   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | X1 | X2 | X3 | X4 | X5 | | Heat number | % C | % Mn | % S | P | Si | | 1  2  3 | 0.200 (x11)  0.180  0.190 | 1.480(x12)  1.580  1.490 | 0.026(x13)  0.029  0.026 | 0.031  0.029  0.029 | 0.059  0.071  0.063 | | 4  5  6 | 0.160  0.170  0.150 | 1.290  1.310  1.320 | 0.030  0.028  0.029 | 0.030  0.028  0.029 | 0.055  0.058  0.058 |     **Similarity Measures : (How do we measure similarity?)**  Similarity represents the degree of correspondence among objects across all of the characteristics used in the analysis. It is a set of rules that serve as criteria for grouping or separating items.  A variety of measures have been proposed for deriving dissimilarity matrix form a set of multivariate observations.  ◦Distance Measures: Most often used as a measure of similarity, with higher values representing greater dissimilarity (distance between cases), not similarity. Several distance measures are available, each with specific characteristics. Commonly used measure is Euclidean distance.  **Euclidean distance**: The most commonly recognized to as straight-line distance.      Euclidean distance between ith and jth item is given by    Obviously  = 0 for *j* = *i*  and  =  For the data matrix  We obtain dissimilarity matrix D    For the chemical composition data in table1, the distance matrix is      **Ex: 2** Suppose a marketing researcher wishes to determine market segments in a community based on patterns of loyalty to brands and stores a small sample of seven respondents is selected as a pilot test of how cluster analysis is applied. Two measures of loyalty- V1(store loyalty) and V2(brand loyalty)- were measured for each respondents on 0-10 scale. |
| **How do we form Clusters?**  Identify the two most similar (closest) observations not already in the same cluster and combine them. We apply this rule repeatedly to generate a number of cluster solutions, starting with each observation as its own “cluster” and then combining two clusters at a time until all observations are in a single cluster. This process is termed a **hierarchical procedure** because it moves in a stepwise fashion to form an entire range of cluster solutions. It is also an **agglomerative method** because clusters are formed by combining existing clusters.  Table details the steps of the hierarchical agglomerative process, first depicting the initial state with all seven observations in single-member clusters, joining them in an agglomerative process until only one cluster remains. The six-step clustering process is described here:  **Step 1:** Identify the two closest observations (E and F) and combine them into a cluster, moving from seven to six clusters.  **Step 2:** Find the next closest pairs of observations. In this case, three pairs have the same distance of 2.000 (E-G, C-D, and B-C). For our purposes, choose the observations E-G. G is a single-member cluster, but E was combined in the prior step with F. So, the cluster formed at this stage now has three members: G, E, and F.  **Step 3:** Combine the single-member clusters of C and D so that we now have four clusters.  **Step 4:** Combine B with the two-member cluster C-D that was formed in step 3. At this point, we now have three clusters: cluster 1 (A), cluster 2 (B, C, and D), and cluster 3 (E, F, and G).  **Step 5:** Combine the two three-member clusters into a single six-member cluster. The next smallest distance is 2.236 for three pairs of observations (E-B, B-G, and C-E). We use only one of these distances, however, as each observation pair contains a member from each of the two existing clusters (B, C, and D versus E, F, and G).  **Step 6:** Combine observation A with the remaining cluster (six observations) into a single cluster at a distance of 3.162. You will note that distances smaller or equal to 3.162 are not used because they are between members of the same cluster.    In steps 1,2,3 and 4, the OSM does not change substantially, which indicates that we are forming other clusters with essentially the same heterogeneity of the existing clusters.  When we get to step 5, we see a large increase. This indicates that joining clusters (B-C-D) and (E-F-G) resulted a single cluster that was markedly less homogenous. |
| **How many groups do we form?**  Therefore, the three – cluster solution of Step 4 seems the most appropriate for a final cluster solution, with two equally sized clusters, (B-C-D) and (E-F-G), and a single outlying observation (A). This approach is particularly useful in identifying outliers, such as Observation A. It also depicts the relative size of varying clusters, although it becomes unwieldy when the number of observations increases.  **Dendogram:** - A Dendrogram, or tree graph, is a graphical device for displaying clustering results. Vertical lines represent clusters that are joined together. The position of the line on the scale indicates the distances at which clusters were joined. The dendrogram is read from left to right. Graphical representation (tree graph) of the results of a hierarchical procedure. Starting with each object as a separate cluster, the dendogram shows graphically how the clusters are combined at each step of the procedure until all are contained in a single cluster |
| **Example:02 Simple Linkage Method**  The daily expenditures on food (X1) and clothing (X2) of five persons are shown in Table    The data of Table are plotted in Figure    Euclidean distance is the appropriate measure of proximity. We begin with each of the five observations forming its own cluster. The distance between each pair of observations is shown below:    Assuming the nearest neighbor method is used, the distance between the cluster (be) and another observation is the smaller of the distances between that observation, on the one hand, and b and e, on the other.  D(be;a) = min(D(b;a);D(e; a)) = min(6:325; 7:159) = 6:325:    Two pairs of clusters are closest to one another at distance 1.414; these are (ad) and (bce). We arbitrarily select (a;d) as the new cluster.  The distance between (be) and (ad) is  D(be; ad) = min(D(be;a);D(be; d)) = min(6:325;7:616) = 6:325  while that between c and (ad) is  D(c;ad) = min(D(c;a);D(c; d)) = min(7:071; 8:246) = 7:071    We merge (be) with c to form the cluster (bce) shown in Figure  The distance between the two remaining clusters is  D(ad; bce) = min(D(ad; be);D(ad; c)) = min(6:325;7:071) = 6:325:  The grouping of these two clusters, it will be noted, occurs at a distance of 6.325, a much greater distance than that at which the earlier groupings took place.    The groupings and the distance at which these took place are also shown in the tree diagram (dendrogram)    In this illustration, it is clear that the elements in each of the clusters (ad) and (bce) are close (they were merged at a small distance), but the clusters are distant (the distance at which they merge is large).  <https://online.stat.psu.edu/stat555/node/86/> |
| **Practice Problem Example 03: Simple Linkage Method**  Suppose that a biologist wants to determine the subspecies in a population of birds belonging the same specie A small sample of 8 birds is selected as a pilot test For each of the 8 birds, two characteristics of their beaks are measured: V1 - length and V2 - width. |
| **Example 04 Complete Linkage Method**  The table below is an example of a distance matrix.  Only the lower triangle is shown, because the upper triangle can be filled in by reflection.    Now let’s start clustering.  The smallest distance is between three and five and they get linked up or merged first into a the cluster '35'.   To obtain the new distance matrix, we need to remove the 3 and 5 entries, and replace it by an entry "35" .  Since we are using complete linkage clustering, the distance between "35" and every other item is the maximum of the distance between this item and 3 and this item and 5.  For example, d(1,3)= 3 and d(1,5)=11.  So, D(1,"35")=11.  This gives us the new distance matrix.  The items with the smallest distance get clustered next.  This will be 2 and 4.    The modified distance matrix becomes    The next merger occurs between the most similar groups 2 and 4 which gives the cluster (24).    And the distance matrix is    The next merger produces the cluster (124). At the final stage, the groups (35) and (124) are merged as a single cluster (12345) t level    The dendogram for the same is given below:    The single linkage dendrogram for the same distance matrix.  It starts with cluster "35" but the distance between "35" and each item is now the minimum of d(x,3) and d(x,5).  So c(1,"35")=3.    From both the figure we can conclude that single linkage and complete linkage differ in the allocation of object 1.  **Determining clusters**  One of the problems with hierarchical clustering is that there is no objective way to say how many clusters there are.  Note that cutting the dendrogram horizontally partitions the data points into clusters.  If we cut the single linkage tree at the point shown below, we would say that there are two clusters.    However, if we cut the tree lower we might say that there is one cluster and two singletons.    There is no commonly agreed-upon way to decide where to cut the tree. |
| **Ex:**      b) Complete Linkage: Distance between two clusters is the maximum between a pair of elements from the two clusters. |
| 1. **Average Linkage**:   *D (A, B)* =  Average of the *nA nB* distance bet. *nA* points in A and the *nB* points in B. |
| At step 1, the highest similarity value in the matrix is *S*(212, 214) = 0.600; hence the two objects fuse at level 0.600. As a consequence of this fusion, the similarity values of these two objects with each of the remaining  objects in the study must be averaged (values in the inner boxes in the Table, step 1); this results in a reduction of the size of the similarity matrix.    Considering the reduced matrix (step 2), the highest similarity value is *S* = 0.500; it indicates that objects 431 and 432 fuse at level 0.500. Again, this similarity value is obtained by averaging the boxed values; this produces a new reduced similarity matrix for the next step.    In step 3, the largest similarity is 0.250; it leads to the fusion of the already-formed group (431, 432) with object 233 at level 0.250.      In the example, this last fusion is the difficult point to understand. Before averaging the values, each one is multiplied by the number of objects in the corresponding group. There is one object in group (233) and two in group (431, 432), so that the fused similarity value is calculated as [(0.0355 ´ 1) +(0.06925 ´ 2)]/3 = 0.058. This is equivalent to averaging the six boxed similarities in the top matrix (larger box) with equal weights; the result would also be 0.058.      So, this method is “unweighted” in the sense that it gives equal weights to the original similarities. To achieve this at step 3, one has to use weights that are equal to the number of objects in the groups. At step 4, there is a single remaining similarity value; it is used to perform the last fusion at level 0.058.  In the dendrogram, fusions are drawn at the identified levels.    For the given example, the *chain of primary connections* corresponding to the dendrogram would be made of the following links: (212, 214) for the first fusion level, (431, 432) for the second level, (233, 431) for the third level, and (214, 432) for the last level. The topology obtained from Average linking clustering may differ from that of single linkage clustering. |
| **Practice Problem Example 05**  Using the data from the graph and the table below, perform **by hand** the 3 algorithms (single, complete and average linkage) and draw the dendrograms.  https://www.statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r_files/figure-html/unnamed-chunk-25-1.png  For all 3 algorithms, the distance matrix between the 5 points |
| * Overlapping clustering: (alternative clustering, multi-view clustering): objects may belong to more than one cluster; usually involving hard clusters. * **Overlapping**allows data objects to be grouped in 2 or more clusters. A real-world example would be the breakdown of personnel at a school. Overlapping clustering would allow a student to also be grouped as an employee while exclusive clustering would demand that the person must choose the one that is more important. In **Fuzzy clustering**every data object belongs to every cluster, I guess you can describe fuzzy clustering as an extreme version of overlapping, the major difference is that the data objects has a membership weight that is between 0 to 1 where 0 means it does not belong to a given cluster and 1 means it absolutely belongs to the cluster. Fuzzy clustering is also known as probabilistic clustering.     For example,   1. In fuzzy clustering, data points can potentially belong to multiple clusters. For example, an apple can be red or green (hard clustering), but an apple can also be red AND green (fuzzy clustering). Here, the apple can be red to a certain degree as well as green to a certain degree. 2. In biology, genes have more than one function by coding for proteins that participate in multiple metabolic pathways; therefore, when clustering microarray gene expression data, it is appropriate to assign genes to multiple, overlapping clusters |
| **Non-hierarchical cluster analysis:**  The nonhierarchical clustering methods are frequently referred to as k-means clustering. K-means clustering is one of the frequently used clustering algorithms. K-Means algorithm processes with the thought that new clusters should be formed according to distance between points and center of clusters.  K-Means algorithms are low cost in terms of calculation time (compared with HCA).   * Algorithm for K-means clustering   1. Partition items into K clusters  2. Assign items to cluster with nearest centroid mean  3. Recalculate centroids both for cluster receiving and losing item  4. Repeat steps 2 and 3 till no more reassignments  The biggest problem of K-Means algorithm is determination of starting points. |
| **Example 01**  We will apply k-means on the following 1 dimensional data set for K=2 Data set {2, 4, 10, 12, 3, 20, 30, 11, 25}  **Iteration 1**  M1, M2 are the two randomly selected centroids/means where M1= 4, M2=11  and the initial clusters are C1= {4}, C2= {11}  Calculate the Euclidean distance as D=[x,a]=√(x-a)²  D1 is the distance from M1 and D2 is the distance from M2  Image for post  Iteration 1  As we can see in the above table, 2 datapoints are added to cluster C1 and other datapoints added to cluster C2  Therefore C1= {2, 4, 3} C2= {10, 12, 20, 30, 11, 25}  **Iteration 2**  Calculate new mean of datapoints in C1 and C2. Therefore  M1= (2+3+4)/3= 3  M2= (10+12+20+30+11+25)/6= 18  Calculating distance and updating clusters based on table below  Image for post  Iteration 2  New Clusters  C1= {2, 3, 4, 10}  C2= {12, 20, 30, 11, 25}  **Iteration 3**  Calculate new mean of datapoints in C1 and C2.  Therefore  M1= (2+3+4+10)/4= 4.75  M2= (12+20+30+11+25)/5= 19.6  Calculating distance and updating clusters based on table below  Image for post  New Clusters  C1= {2, 3, 4, 10, 12, 11}  C2= {20, 30, 25}  **Iteration 4**  Calculate new mean of datapoints in C1 and C2.  Therefore  M1= (2+3+4+10+12+11)/6=7  M2= (20+30+25)/3= 25  Calculating distance and updating clusters based on table below  Image for post  New Clusters  C1= {2, 3, 4, 10, 12, 11}  C2= {20, 30, 25}  As we can see that the data points in the cluster C1 and C2 in iteration 3 are same as the data points of the cluster C1 and C2 of iteration 2.  It means that none of the data points has moved to other cluster. Also the means/centeroid of these clusters is constant. So this becomes the stopping condition for our algorithm. |
| **Example 02**  Consider 4 data points A,B,C,D as below  Image for post  Choose two centroids AB and CD, calculated as  AB = Average of A, B  CD = Average of C,D  Image for post  Calculate squared euclidean distance between all data points to the centroids AB, CD. For example distance between A(2,3) and AB (4,2) can be given by s = (2–4)² + (3–2)².  Image for post  4. If we observe in the fig, the highlighted distance between (A, CD) is 4 and is less compared to (AB, A) which is 5. Since point A is close to the CD we can move A to CD cluster.  5. There are two clusters formed so far, let recompute the centroids i.e, B, ACD similar to step 2.  ACD = Average of A, C, D  B = B  Image for post  6.As we know K-Means is iterative procedure now we have to calculate the distance of all points (A, B, C, D) to new centroids (B, ACD) similar to step 3.  Image for post  7. In the above picture, we can see respective cluster values are minimum that A is too far from cluster B and near to cluster ACD. All data points are assigned to clusters (B, ACD) based on their minimum distance. The iterative procedure ends here.  8. To conclude, we have started with two centroids and end up with two clusters, K=2. |
| **Example 03:**  Cluster the following eight points (with (x, y) representing locations) into three clusters:  A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)  Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-Ρ(a, b) = |x2 – x1| + |y2 – y1|  Use K-Means Algorithm to find the three cluster centers after the second iteration.  Solution-  We follow the above discussed K-Means Clustering Algorithm-  Iteration-01:  We calculate the distance of each point from each of the center of the three clusters.  The distance is calculated by using the given distance function.  The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-  Calculating Distance Between A1(2, 10) and C1(2, 10)-  Ρ(A1, C1) = |x2 – x1| + |y2 – y1|= |2 – 2| + |10 – 10|= 0  Calculating Distance Between A1(2, 10) and C2(5, 8)-  Ρ(A1, C2)= |x2 – x1| + |y2 – y1|= |5 – 2| + |8 – 10|= 3 + 2= 5  Calculating Distance Between A1(2, 10) and C3(1, 2)-  Ρ(A1, C3)= |x2 – x1| + |y2 – y1|= |1 – 2| + |2 – 10|= 1 + 8= 9  In the similar manner, we calculate the distance of other points from each of the center of the three clusters.  Next, We draw a table showing all the results.  Using the table, we decide which point belongs to which cluster. The given point belongs to that cluster whose center is nearest to it.    From here, New clusters are-  Cluster-01: First cluster contains points-A1(2, 10)  Cluster-02: Second cluster contains points-  A3(8, 4)  A4(5, 8)  A5(7, 5)  A6(6, 4)  A8(4, 9)  Cluster-03:  Third cluster contains points-  A2(2, 5)  A7(1, 2)  Now, We re-compute the new cluster clusters. The new cluster center is computed by taking mean of all the points contained in that cluster.  For Cluster-01:  We have only one point A1(2, 10) in Cluster-01.  So, cluster center remains the same.  For Cluster-02:  Center of Cluster-02= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)= (6, 6)  For Cluster-03:  Center of Cluster-03= ((2 + 1)/2, (5 + 2)/2)= (1.5, 3.5)  This is completion of Iteration-01.  Iteration-02:  We calculate the distance of each point from each of the center of the three clusters. The distance is calculated by using the given distance function.  The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-  Calculating Distance Between A1(2, 10) and C1(2, 10)-  Ρ(A1, C1)= |x2 – x1| + |y2 – y1|= |2 – 2| + |10 – 10|= 0  Calculating Distance Between A1(2, 10) and C2(6, 6)-  Ρ(A1, C2)= |x2 – x1| + |y2 – y1|= |6 – 2| + |6 – 10|= 4 + 4= 8  Calculating Distance Between A1(2, 10) and C3(1.5, 3.5)-  Ρ(A1, C3)= |x2 – x1| + |y2 – y1|= |1.5 – 2| + |3.5 – 10|= 0.5 + 6.5= 7  In the similar manner, we calculate the distance of other points from each of the center of the three clusters.  Next,  We draw a table showing all the results.Using the table, we decide which point belongs to which cluster.  The given point belongs to that cluster whose center is nearest to it.    From here, New clusters are-  **Cluster-01:** First cluster contains points-   * A1(2, 10) * A8(4, 9)   **Cluster-02:** Second cluster contains points-   * A3(8, 4) * A4(5, 8) * A5(7, 5) * A6(6, 4)   **Cluster-03:** Third cluster contains points-   * A2(2, 5) * A7(1, 2)    Now,   * We re-compute the new cluster clusters. * The new cluster center is computed by taking mean of all the points contained in that cluster.     **For Cluster-01:** Center of Cluster-01  = ((2 + 4)/2, (10 + 9)/2)= (3, 9.5)  **For Cluster-02:** Center of Cluster-02  = ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4)= (6.5, 5.25)  **For Cluster-03:**   Center of Cluster-03= ((2 + 1)/2, (5 + 2)/2)= (1.5, 3.5)   This is completion of Iteration-02.   After second iteration, the center of the three clusters are-   * C1(3, 9.5) * C2(6.5, 5.25)   C3(1.5, 3.5) |
| **Example 04**  Use K-Means Algorithm to create two clusters-  https://www.gatevidyalay.com/wp-content/uploads/2020/01/K-Means-Clustering-Algorithm-Problem-2.png  **Solution-We follow the above discussed K-Means Clustering Algorithm. Assume A(2, 2) and C(1, 1) are centers of the two clusters.**  **Iteration-01:**  **We calculate the distance of each point from each of the center of the two clusters. The distance is calculated by using the euclidean distance formula. The following illustration shows the calculation of distance between point A(2, 2) and each of the center of the two clusters-**  **Calculating Distance Between A(2, 2) and C1(2, 2)-**  **Ρ(A, C1)= sqrt [ (x2 – x1)2 + (y2 – y1)2 ]= sqrt [ (2 – 2)2 + (2 – 2)2 ]= sqrt [ 0 + 0 ]= 0**  **Calculating Distance Between A(2, 2) and C2(1, 1)-**  **Ρ(A, C2)= sqrt [ (x2 – x1)2 + (y2 – y1)2 ]= sqrt [ (1 – 2)2 + (1 – 2)2 ]= sqrt [ 1 + 1 ]= sqrt [ 2 ]= 1.41**  **In the similar manner, we calculate the distance of other points from each of the center of the two clusters. Next, We draw a table showing all the results.**  **Using the table, we decide which point belongs to which cluster. The given point belongs to that cluster whose center is nearest to it.**   |  |  |  |  | | --- | --- | --- | --- | | **Given Points** | **Distance from center (2, 2) of Cluster-01** | **Distance from center (1, 1) of Cluster-02** | **Point belongs to Cluster** | | A(2, 2) | 0 | 1.41 | C1 | | B(3, 2) | 1 | 2.24 | C1 | | C(1, 1) | 1.41 | 0 | C2 | | D(3, 1) | 1.41 | 2 | C1 | | E(1.5, 0.5) | 1.58 | 0.71 | C2 |     From here, New clusters are-  **Cluster-01:** First cluster contains points-   * A(2, 2) * B(3, 2) * E(1.5, 0.5) * D(3, 1)   **Cluster-02:**Second cluster contains points-   * C(1, 1) * E(1.5, 0.5)    Now,   * We re-compute the new cluster clusters. * The new cluster center is computed by taking mean of all the points contained in that cluster.   **For Cluster-01:**   Center of Cluster-01= ((2 + 3 + 3)/3, (2 + 2 + 1)/3)= (2.67, 1.67)  **For Cluster-02:**   Center of Cluster-02= ((1 + 1.5)/2, (1 + 0.5)/2)= (1.25, 0.75)   This is completion of Iteration-01. Next, we go to iteration-02, iteration-03 and so on until the centers do not change anymore. |
| **Example 05** |
| **Example 06**    Ans:   |  |  |  | | --- | --- | --- | | Subject | A | B | | 1 | 1.0 | 1.0 | | 2 | 1.5 | 2.0 | | 3 | 3.0 | 4.0 | | 4 | 5.0 | 7.0 | | 5 | 3.5 | 5.0 | | 6 | 4.5 | 5.0 | | 7 | 3.5 | 4.5 |     This data set is to be grouped into two clusters.  As a first step in finding a sensible initial partition, let the A & B values of the two individuals furthest apart (using the Euclidean distance measure), define the initial cluster means, giving:   |  |  |  | | --- | --- | --- | |  | Individual | Mean Vector (centroid) | | Group 1 | 1 | (1.0, 1.0) | | Group 2 | 4 | (5.0, 7.0) |     The remaining individuals are now examined in sequence and allocated to the cluster to which they are closest, in terms of Euclidean distance to the cluster mean. The mean vector is recalculated each time a new member is added. This leads to the following series of steps:   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Cluster 1 | | Cluster 2 | | | Step | Individual | Mean Vector (centroid) | Individual | Mean Vector (centroid) | | 1 | 1 | (1.0, 1.0) | 4 | (5.0, 7.0) | | 2 | 1, 2 | (1.2, 1.5) | 4 | (5.0, 7.0) | | 3 | 1, 2, 3 | (1.8, 2.3) | 4 | (5.0, 7.0) | | 4 | 1, 2, 3 | (1.8, 2.3) | 4, 5 | (4.2, 6.0) | | 5 | 1, 2, 3 | (1.8, 2.3) | 4, 5, 6 | (4.3, 5.7) | | 6 | 1, 2, 3 | (1.8, 2.3) | 4, 5, 6, 7 | (4.1, 5.4) |     Now the initial partition has changed, and the two clusters at this stage having the following characteristics:   |  |  |  | | --- | --- | --- | |  | Individual | Mean Vector (centroid) | | Cluster 1 | 1, 2, 3 | (1.8, 2.3) | | Cluster 2 | 4, 5, 6, 7 | (4.1, 5.4) |     But we cannot yet be sure that each individual has been assigned to the right cluster.  So, we compare each individual’s distance to its own cluster mean and to that of the opposite cluster. And we find:   |  |  |  | | --- | --- | --- | | Individual | Distance to mean (centroid) of Cluster 1 | Distance to mean (centroid) of Cluster 2 | | 1 | 1.5 | 5.4 | | 2 | 0.4 | 4.3 | | 3 | 2.1 | 1.8 | | 4 | 5.7 | 1.8 | | 5 | 3.2 | 0.7 | | 6 | 3.8 | 0.6 | | 7 | 2.8 | 1.1 |    Only individual 3 is nearer to the mean of the opposite cluster (Cluster 2) than its own (Cluster 1).  In other words, each individual's distance to its own cluster mean should be smaller that the distance to the other cluster's mean (which is not the case with individual 3).  Thus, individual 3 is relocated to Cluster 2 resulting in the new partition:   |  |  |  | | --- | --- | --- | |  | Individual | Mean Vector (centroid) | | Cluster 1 | 1, 2 | (1.3, 1.5) | | Cluster 2 | 3, 4, 5, 6, 7 | (3.9, 5.1) |   The iterative relocation would now continue from this new partition until no more relocation occurs.  However, in this example each individual is now nearer its own cluster mean than that of the other cluster and the iteration stops, choosing the latest partitioning as the final cluster solution. |

UNIT-VII