Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:09
Lecture:33

CLUSTERING TECHNIQUES Part 1

Data

	Respondents						
Clustering variable	A	B	C	D	E	F	G
V1	3	4	4	2	6	7	6
V2	2	5	7	7	6	7	4

Euclidean distance between respondents =

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2}$$

(k: variable; i, j: respondents)

Good morning and welcome back. So we are going to continue with the example we started, to understand cluster analysis in the last class. So we will explore this data further to understand what is the method that is followed in building clusters from a set of data. So in the simple example that we discussed, you can see that there are A, B, C, D, E, F, G. So there are 7 respondents, we call them 7 objects. It is actually, should have been in the rows but that is the way we have set this up or it is taken from a textbook which explains cluster analysis very clearly.

So there are 7 objects or 7 respondents in this case, 7 objects and V1 and V2 are 2 variables, 2 variables or 2 attributes as we call in database language. So 7 objects and 2 attributes and each object has a set of 2 attribute values. So A has V1 and V2 which is 3 and 2, B has V1 and V2 which is 4 and 5, C has 4 and 7, D has 2 and 7 and so on and G finally has 6 and 4. Those are the attribute values for each object.

In this specific example, we know that these values are the responses given by these respondents to a survey. The data could be obtained in any way, it could be from a database where a prior data exists in terms of customer behaviour or customer attributes etc. So we have this data. So our aim is to group these 7 objects into less than 7 number of groups. So there are 7 objects, so we can say there are 7 groups, it does not make any sense.

But therefore we have to group them together. And one way or you know an obvious way to think of how to group them together is, how have they responded, what are their attribute values. And we would imagine that those who have similar attributes, they have similar characteristics or similar responses, their beliefs are similar, their features are similar or their lifestyle is similar. There is some reason why they have responded similar, in this context. Clustering is context specific.

For a different set of questionnaire, they may respond differently. So that is a different project. As far as store loyalty and product loyalty or item loyalty, I think that is what we saw the data to be. This is store loyalty and brand loyalty V1 and V2. So they have, we will look at who have similar characteristics.

And those who have similar characteristics, we would like to them like them to be together or they should be in similar clusters. So how do we go? That is a simple question. How do we know that 2 or 3 persons have similar characteristics? Of course, for a small data set by looking at the data itself we can actually make sense. For example look at respondents E and F. One has responded 7, 7 other has responded 6, 6.

Looks like they are very similar in terms of their responses as compared to a 3, 2 or a 4, 5 etc. They are very different. So we would say E and F are similar in terms of their responses. We can also discover such similarities by looking at the responses of individuals. For example C and D are also not very dissimilar.

They look like similar in terms of their responses. So this is like looking at the data because it is a small data set, looking at the data and sort of, well they look alike. So like, you know this is sort of our intuition or common sense or our visual inference which

sometimes can be inaccurate as well and when the data, the size of the data increases it becomes very difficult. And therefore we need a proper algorithm to do this. We need to automate this.

That is why we discuss cluster algorithms, for how to work with large data sets. So first we should know how to work with small data sets. So we understand what are the principles and then we go to large data sets. So below the table you can see, a measure of distance is given, a formula for measure of distance called Euclidean distance. I believe you are familiar with this formula already.

What is the Euclidean distance between two vectors? So you can see A, vector A is you know consisting of two features V1 and V2. So 3 and 2 are the values, for vector B it is 4 and 5 and so on. It is to find out the distance between this feature vectors, in if I use another language. So A32, B45 what is the distance between them? And we calculate the distance using Euclidean distance here because they are continuous valued. These are, actually this is interval data because it is a scale used and so for interval data or continuous valued data, we can actually use Euclidean distance.

Euclidean distance is nothing but the square root of squared differences between each pair of variables, each pair of variables. For example, if you apply this formula to find the distance between A and B, what would be the distance between A and B? It would of course, be the square root of first pair is, we look at variable V1. For A, the value of variable V1 is 3 for B it is 4. So $(3- 4)^2$ plus for variable V2 it is 2 and 5, $(2 -5)^2$ the whole square. It is equal to whatever value it is.

We know, it is a measure of the distance between A and B or feature vector A and feature vector B, features here meaning the variables. A has two attributes V1 and V2, similarly B also has two attributes V1 and V2. What is the distance between that two? Can be calculated using the symbol, Euclidean distance as a formula. There are other measures, as in decision trees we saw there are different measures of purity. In a similar way there are different measures of distance as well, which I will show you a summary of very soon.

But for this problem, we need to have a measure of distance so that we can find the distances between each pairs. And once we calculate these distances and visualize it, then we can easily make out who is close to whom. Our ultimate aim is to bring objects which are similar together or in other words, our aim is to group these objects together. A, B, C, D, E, F, G are the objects. We want to put them together, not V1, V2, V3, V4 as in factor analysis.

Proximity matrix

Observation	A	B	C	D	E	F	G
A							
B	3.162						
C	5.099	2.000					
D	5.099	2.828	2.000				
E	5.000	2.236	2.236	4.123			
F	6.403	3.606	3.606	5.000	1.414		
G	3.606	2.236	2.236	5.000	2.000	3.162	-

We are not actually doing dimensionality reduction where we actually try to reduce the number of dimensions, that is your PCA. Here we are actually trying to group objects together, based on their attribute values. So intuitively, here you can have a diagonal matrix as you do in correlation matrix. Just as in correlation matrix, you have A,B,C,D,E,F,G or the respondents or the objects in the vertical as well as horizontal axis and you actually place the distances in the corresponding cells. For example B and A, the Euclidean distance is 3.162. What is the Euclidean distance between A and A? 0. There is no distance. They have the same value. If you copy A and A, then it is the same value. So therefore it is 0. So A and B is 3.162, A and C is 5.099 and C and B is 2.0 and so on. So we got a diagonal matrix of distances between objects, like the correlation matrix and this is known as proximity matrix.

Proximity matrix. These are actually values that indicate distances pair wise. Now looking at this table visually, do you change your opinion or which are the pair of objects who have the most similarity? No, no. Which is a pair which has the highest similarity? E and F. Similarity means lowest distance. So keep in mind when you use this distance, it is opposite of similarity.

It is inverse of similarity. So you always have to think in that direction. These are

inverses. So distance is a measure of dissimilarity if you have to articulate it in a direct sense but indirectly it is a measure of, proximity is a measure of dissimilarity. So 1.414 means there is a shortest distance or we say E and F are very similar and they can be grouped together.

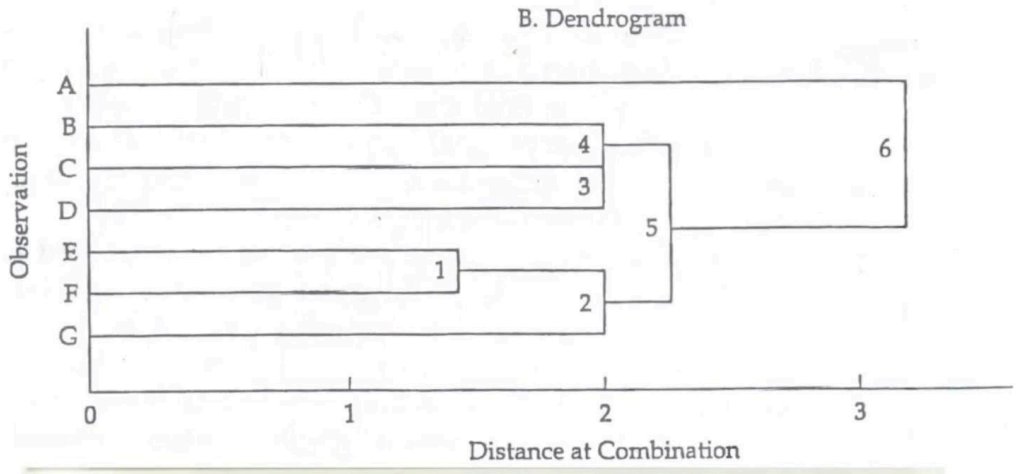
They are part of, they should be part of one cluster. And next what is the shortest distance? You see there is a value 2 here, 2 here and 2 here. So there are 3 pairs which have very similar values and obviously we would think that they are very similar in terms of how they feel, how they sense, how they responded. So as far as loyalty is concerned, these 3 objects are very close to each other but at the same time E and F are also in the same league or they are also very similar. And then of course, you can again extend this argument or this, you can induce this further between B and E, you can find a short distance, between C and E you can find a short distance and so on.

So this proximity matrix gives us a sense of the distances and also if we can try to put them together using some method, some algorithm, then we know how we can proceed in one way. Start with the shortest distance and then build upwards and then bring together who have similar values, you know into one group and then build upwards. Upwards in the sense, upwards based on the value of distance. So you know with this principle in mind, there is a clustering technique known as agglomerative clustering. Agglomerative clustering. It is a category of clustering which falls under hierarchical clustering.

In hierarchical clustering, there is a particular method where you follow a bottom up approach. You start with the shortest distance and then build a dendrogram. This is a dendrogram . You must have studied this already. You must be familiar with this already.

This kind of a structure is known as dendrogram. So in this particular exercise what is being done is, in the x axis you have the distance. X axis is the distance and y axis consists of the objects. A to G these are the objects and x axis is the, sort of the values, the values of the distances. And what are we doing here? The screenshot or the particular data in the form of table captured below the dendrogram shows you what is the procedure that is being followed in building the dendrogram .

Agglomerative clustering



Step	AGGLOMERATION PROCESS: Minimum Distance Between Unclustered Observations[a]	AGGLOMERATION PROCESS: Observation Pair	CLUSTER SOLUTION: Cluster Membership	CLUSTER SOLUTION: Number of Clusters	CLUSTER SOLUTION: Overall Similarity Measure (Average Within-Cluster Distance)
	Initial Solution		**(A) (B) (C) (D) (E) (F) (G)**	**7**	**0**
1	1.414	E-F	(A) (B) (C) (D) (E-F) (G)	6	1.414
2	2.000	E-G	(A) (B) (C) (D) (E-F-G)	5	2.192
3	2.000	C-D	(A) (B) (C-D) (E-F-G)	4	2.144
4	2.000	B-C	(A) (B-C-D) (E-F-G)	3	2.234
5	2.236	B-E	(A) (B-C-D-E-F-G)	2	2.896
6	3.162	A-B	(A-B-C-D-E-F-G)	1	3.420

[a]Euclidean distance between observations

So in agglomerative clustering, the starting point is no clusters. In agglomerative clustering which is hierarchical clustering the starting is, suppose you have n objects then you start with n clusters. We assume that each object is a cluster. So here there are, sort of we found there were 7 objects. So initially we assume, the initial solution is that there are 7 objects or number of clusters is equal to 7.

Number of clusters is equal to 7 means there is no distance. If you look at one cluster, there is only one object. So therefore, within that cluster, there is no distance. Distance between objects within a cluster is 0 because there is only one element. So that is the starting point, we are starting to assume that there are n clusters.

There are as many clusters as the number of objects, starting point. Now we come to the, this is step 0. And we come to step 1. When we come to step 1, we glance through the proximity matrix and spot the shortest distance. And we found that the shortest distance was 1.414 and that was between E and F. Between E and F, there is a distance which is 1.414 and here is 1.414. We connect the 2 objects, E and F and once they are connected, they are one.

It becomes one cluster. The E and F may have slightly different values in terms of their attribute values. Does not matter. Once you connect them you have to treat them as one object. One unit or one group. So E and F got connected and now what happens, what is the number of clusters now? A is a cluster, B is a cluster, C is a cluster, D is a cluster, E and F is one cluster.

And then there is G. So number of clusters become 6. Now there is an inter-cluster distance, within cluster distance not inter-cluster distance. I am sorry. There is a within cluster distance which is non-zero, which is higher than zero. What is that within cluster distance? It is the distance between E and F.

There is only one pair there. So therefore, it is just that distance between E and F which is 1.414. Next step, we spot the next shortest distance which, of course there are 3 pairs. So you look at one of them. There is a distance of 2 and the distance between E and G is 2. So you connect E and G. Connect E and G but there is nothing called E now. E and F are one. So you connect from EF cluster to G. That is what you do. So G also part become part of the league or part of the group.

It becomes EFG. So therefore in after step 2, you have 5 clusters. And now what is the average within cluster distance, which we are using as a measure to understand what is the total error. What is the total error that is formed? So within cluster distance, the average within cluster distance is EF, FG, EG.

There are 3 pairs. EF, FG, EG. Sum of the 3 distances, (EF + FG+ EG)/ 3. That is the value 2.12. The average within cluster distance. We are using as a measure to see how the cluster is building and how the error is actually increasing.

Of course when you put things together, they are not exactly matching. So there is some sort of grouping that we have done. Some error. Some measure of error. I will show you in the next slide how to actually have a measure of distance between clusters.

There are different methods for it. Here I am just following one method, consistently one method. There are multiple methods to arrive at a distance between clusters. Within cluster distance, it is easy pair of, you know these are pairs. But between cluster it is a challenge because you can go by the shortest distance, you can go by the farthest distance, you can go by centroid distances and so on. There are different measures. And based on that the algorithm will also change.

Alright. So coming back. So you can see how this algorithm is proceeding. Again we are spotting the next shortest distance there is another 2 and that connects C and D together and then the number of clusters comes down to 4 because C and D is one cluster

now.

EFG is another cluster. You see that. And therefore correspondingly the value, the total inter-cluster distance is actually changing. Then you have another 2 here. We connect them also, so it becomes B, C, D, E, F, G.

Number of clusters is 3. Total within cluster distance is 2.234. And then of course, going by the shortest distance, B and E are connected then you have B, C, D, E, F, G. All of them get connected. And in the final step, you can see we connect all. And what is the number of clusters here? All are connected.

Like we are saying all objects together, you all fall into one group. And obviously you see the error has increased, from 0 when you treated each object as a cluster, the error was 0. When you grouped everyone into one bucket, then it became 3.42. And obviously we know that both these boundaries are not good. Neither 0 is good or treating each object as a cluster is not what we want.

At the same time, putting all objects into one group is also not good. That actually makes the group very heterogeneous. Here, you have the highest homogeneity because each one is one group. I am the state. So, but at the end when you put everyone together, it is like you know the highest heterogeneity.

And we know that the optimum solution is somewhere in between. So manually if you have to find a solution, what is the best scenario here or the best grouping here. It lies between, you know step 1, between step 1 and step 5. We would say, it is between step 1 and 5 because there are there is more than one cluster and there is less than 7 cluster. So which solution would be good? Algorithmically, we can put conditions but manually when we look at it, how can you make sense? That is from 2.34 to 2.89, there is a sudden jump in the error. There is a step jump in the error, till that time it was 2.144, 2.234. The difference in error was small but you also got more number of clusters.

In the sense, you know it was 4, now in step 4 it became 3. So we are looking at two aspects. One is the number of clusters. We also do not want n number of clusters. We also do not want one number of cluster.

We want some sort of a reasonable number of clusters.Reasonable number of clusters. What is that reasonable? That reasonable is a number which is negotiable, in the sense if you are working on a problem, say if it is a segmentation problem, say in marketing. Do you want 50 segments? Do you want 100 segments? Or you want to see 5, 6, 7 segments. You cannot deal with too many number of segments.

You cannot even actually make sense out of it. A 50 is too large a number. A 2 is too small a number. So you would say, well I want if you are talk to expert or a subject matter expert, you say 4, 5, 6, 7. So you are okay to work with this reasonable numbers.

Reasonable number of groups than very unreasonable number of groups. That is one. So there is a practical consideration also, in deciding the number of clusters depending on the problem. That is one aspect. And the other is that when you form the clusters, the error should not increase too much. So what we try to minimize is the total within cluster distance which is a measure of overall homogeneity. This is a measure of your overall homogeneity that you have acquired. If the clusters within, have a lot of heterogeneity then this value is going to be high. You do not want that. You want clusters to be similar or objects in the clusters to be similar.

So this value should be low. But at the same time, you do not want too many clusters also. So two principles in cluster formation. Clusters should be very similar within. This measure, this is one of the measures for overall within cluster similarity.

Clusters should be similar within but they should be dissimilar across. The clusters should be, each cluster should be different from another cluster which is formed. So homogeneous within, heterogeneous across. That is the principle in clustering. And therefore we need to have measures for both.

Within cluster distance and across the cluster distance. So and now we move on to understand how clustering algorithms have been developed to address these principles, in line with these principles of cluster formation, forming practical number of clusters which are useful having minimum within cluster distance and also clusters should be heterogeneous across. That is how cluster algorithms work.