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:32

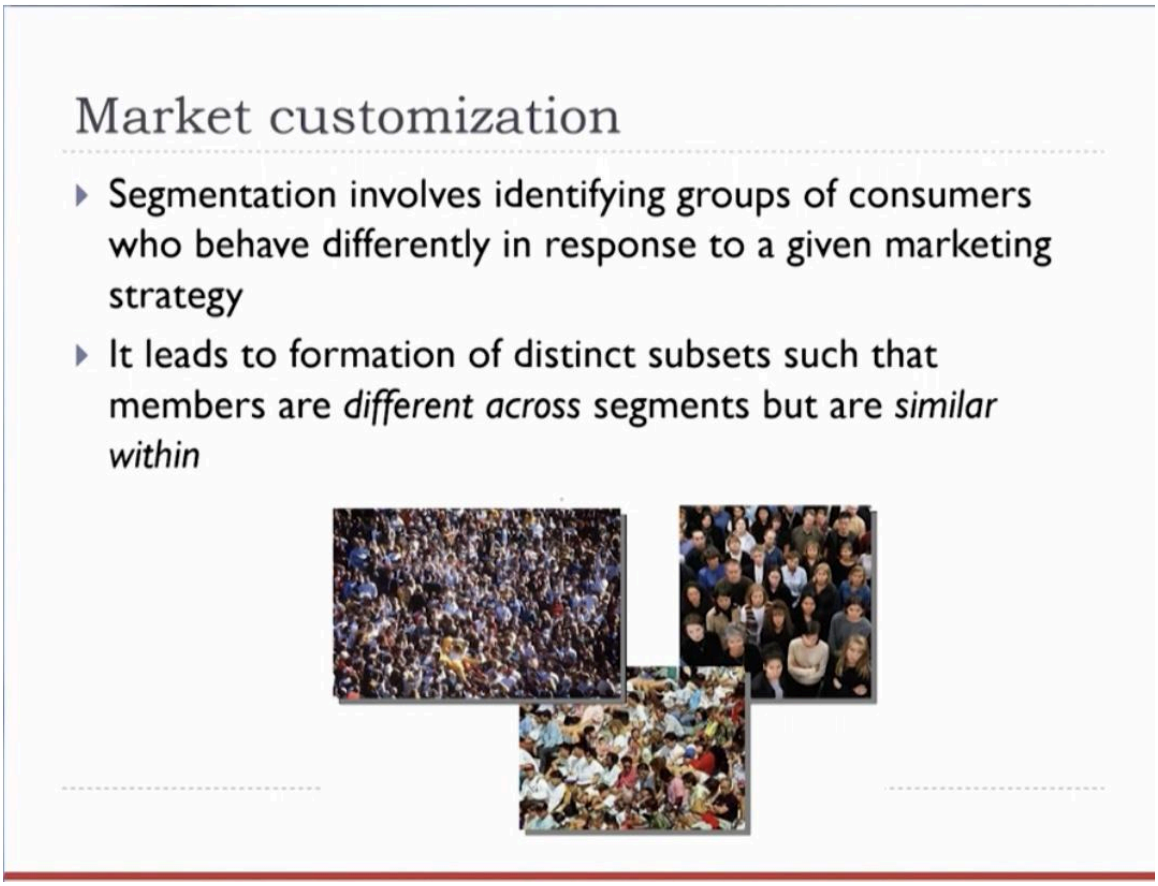
CLUSTER ANALYSIS

Hello and welcome back. So, we will continue the next, with the next topic which is Cluster Analysis. So, cluster analysis is one of the widely used applications of analytics in business. In business, cluster analysis has several applications. We learn decision trees it has certain strengths, but when it comes to clustering, you will see that it has a different approach and different capabilities. So, we are going to discuss what is cluster analysis and how this can be used to solve business problem today and in the next session.

So, I will try cover concepts of cluster analysis and subsequently you must be able to apply this. What is the difference between cluster analysis and decision trees? What do you expect the difference to be? Both are used for grouping objects. In order to group this class, I can actually use a clustering technique, I can also use decision trees. But what would be a difference if I have to form groups, a project groups from this class based on the concepts of decision trees and clustering.

There will be one difference, major difference. The difference would be that if I use decision trees I would say what is my objective, what is my target. Form groups such that the CGPA will be the highest. So, I have some output that I am looking at, some target variable I am looking at, based on that you have to form the groups. So, you work on it based on what is the target given.

But in cluster, I will not tell you what I want. I would just specify certain variables or certain attributes. You form into groups based on your interest, your region, your undergraduate degree and what kind of specialization you want, you know what kind of job you want, four or five attributes I gave. But I do not give you any target variable. I do not give any y variable.



You have only attributes. So, cluster analysis is a technique that is unsupervised or there is, nothing is known about a target. But what you want is to find out how this various objects or why, how this various participants of a class get formed into groups, when I say these are my attributes. I am curious about it. It is a discovery process or it is induction.

I need to index certain structure from the data. Instead of deducing a structure based on prior principles, I am actually inducing a structure into the data and that is what clustering is. A good application as far as business is concerned for clustering is segmentation. As soon as you hear segmentation, think of clustering. If you have to segment market, well you have a product and you think that whole of India is your market.

Is that ever correct? India is your market. Would a marketing manager say that? No. India is the most complex country in terms of demography. You have all categories here. So, where is your product going to be relevant? First thing is, they call it market customization in marketing.

You have to actually look at what are the segments within this geography. And now I understand demographically I can actually segment a population or a geography. Then it results to various segments and looking at the segments I know whether my product or service match with a particular segment or not. So, that is the first exercise. Segmentation in marketing, segmentation, targeting and third activity is, business students, positioning.

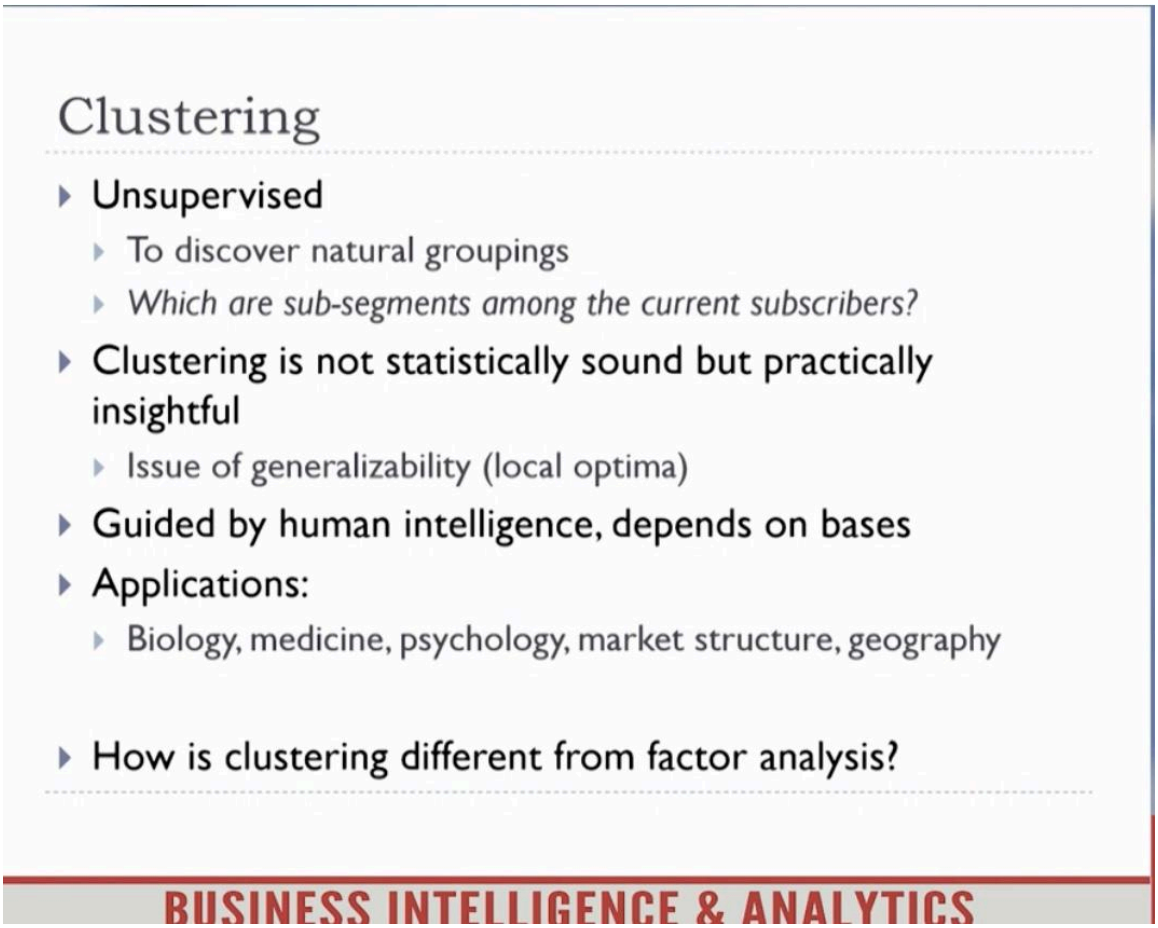
Segmentation targeting positioning, three major activities of marketing. So, segmentation, you cluster targeting is more of a supervised learning technique that we applied because we are trying to look at or target who are buyers or non buyers . The target is not. In this case, there is no such thing as a target. We just let the data group or, sorry, form into clusters based on mutual affinity.

If I simply ask you to form into groups, what are you going to do in the class? I just say I do not give any attributes also. You decide which attributes. Just form into some groups, five groups. How will you form the groups? This will be one group because they have a lot of things to talk. You can listen to them.

So they are close. There are some common topics. It is called affinity grouping. So, clustering is an affinity grouping. So, what is the principle there? So, there is something related to institute, job, internship, politics, something they are talking.

But that is of interest to the group, that brings them together. So, something can be food. Somebody else is interested in movies or you are coming from Punjabis. All Punjabis get into one group or Keralites get into another group. These are our natural affinities.

So cultural distance, you know as it is called in literature, that will be one attribute to form groups. There is no objective there. There is no target there. You just formed into groups based on affinity. So, that is a broad principle that is driving clustering, based on how close the attributes are.



The objects come together. So as I said, clustering is unsupervised, not target variable. So, it aims to find out natural groupings. Clustering is not statistically sound but practically insightful. Yeah well, that is a very broad statement because we do not do, generally many statistical tests when you do cluster analysis.

That is generally the case with data science applications where you work with large databases, millions of data. And then you form clusters from that large data. You know when the data size is too large then of course there is sampling involved. So, data mining algorithms can actually do sampling of that data and then build models based on the samples. But the thing is, you do not actually try to do a significance test to see if the model would generalize to the population etc.

That kinds of test are, they do exist but they are not generally done. You work with large databases and generally deploy those models after testing for prediction purpose. And therefore, you know statistically not sound, that is the basis for that statement. So, you know issue of generalizability or local optima as it is called in optimization techniques and it is guided by human intelligence and it, the your solution will be as good as the

variables and the data that you use. If you do not have the right variables for, say segmentation the segments that you get would not be useful for your product.

What do you do with it? When you look at it, you do not get anything out of it. So, that is an important aspect and it is widely used in different domains including business. Are you familiar with factor analysis? Have you done factor analysis in data analysis? Anyone familiar with factor analysis? What do you do? There is something called dimensionality reduction. Principal component analysis, dimensionality reduction. What is the purpose there? Clustering is in principle similar.

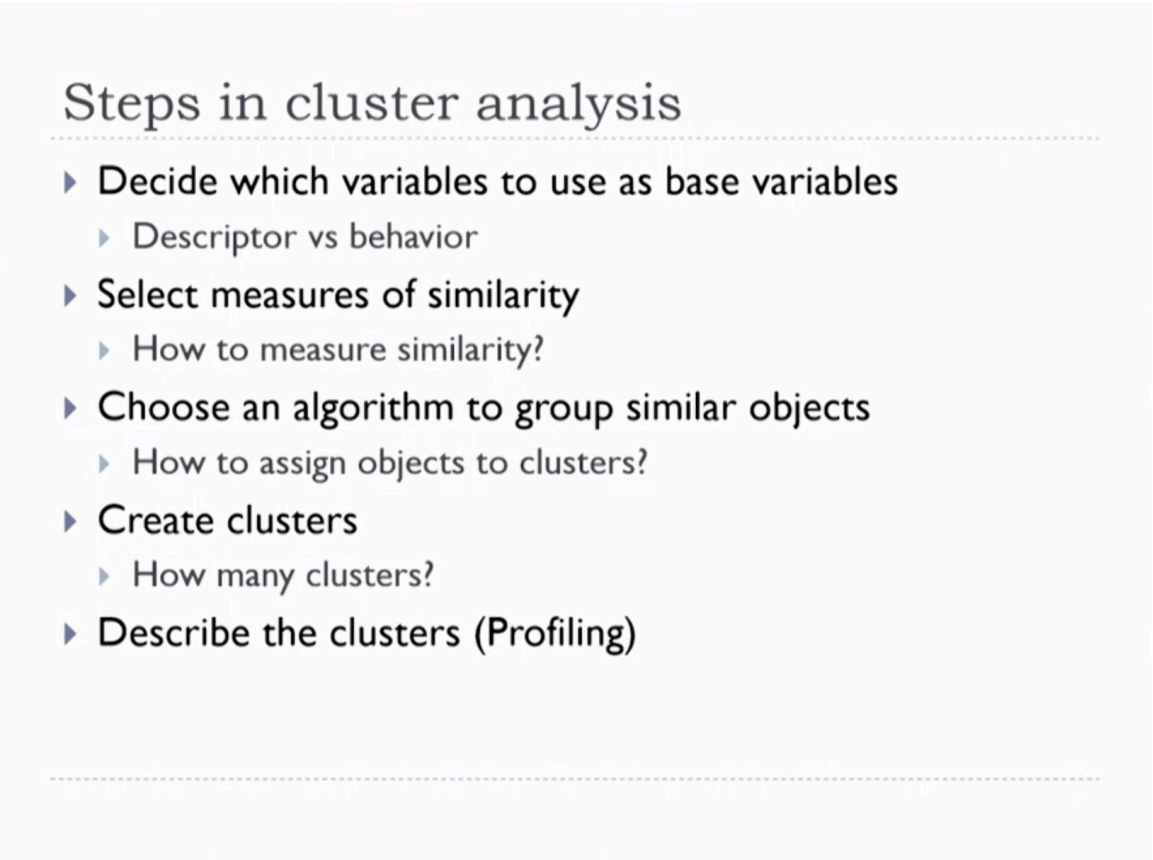
It is not the same thing, but it is very similar to factor analysis. Suppose I get into an exercise, like you know I want to do everything from the scratch. So, I believe that, you know that is the way things should be done. So, I develop an instrument to get student feedback about my course and satisfaction, student satisfaction about my course. I develop, custom develop some 50 items, 50 questions or 50 items and I give that questionnaire to you.

50 items or 50 variables I would say, each of you score me in 1 to 5 or 1 to 10 scale . When I look at the 50 items or 50 variables, too many dimensions, too many things there. I am not able to sort of make much sense from it because many of them may be very correlated. For example, was the faculty friendly? Did the faculty smile? So, what does it mean? So, many things will be very correlated. So that is called in, factor analysis what you do is you try to reduce this number of items to a few dimensions.

You say faculty behavior, faculty knowledge, examination. You can reduce the whole 50 into 5 or 6 dimensions. This is called dimensionality reduction. Too many variables, difficult to comprehend, difficult to interpret, related variables or related items you bring together into dimensions or map into dimensions. Basically, that is the principle in factor analysis sort of techniques.

In clustering we do similar thing. We actually bring objects together. In factor analysis, it is the variables that you reduce to a few dimensions. In cluster analysis you have a large database or objects. You actually group them into few classes or few clusters. You reduce the number of objects, not the number of variables. That is the difference.

So steps in cluster analysis. First one is of course decide which variables to use. Cluster analysis is based on variables, base variables. Variables determine how the clusters will be formed. I can form different clusters of this class.



Let me again use the example of the classroom. I can use, say CGPA and performance in under graduation as two variables. How do you expect the groups to be formed? All the nerds will be in one class, who will be wearing the specs and reading all the time and all the fun guys will be in another class, bike bunches for them and all. So you can expect based on academic performance, you are trying to create clusters. But instead of that if I objective is to like I also want guys who are smart, who are having social skills.

I look at both your social skill as well as your academic performance. Then I add one more variable related to your social skill, provided it is measurable and data is available. Then I will find better groups formed. It also depends on my objective. If I want to recruit for something which is purely data science or does not require any social skill, I may go by the CGPA alone.

So choice of variables is based on why are you doing the clustering exercise. What is your purpose? Your purpose drives your choice of variables. Now one important aspect of cluster analysis is a measure known as similarity. Similarity is similar to purity which we discussed in decision trees. In decision trees, the subgroups would be or should be pure, as the child node should be purer than the parent nodes.

So in order to do that, the algorithm needs something known as measure of purity. So

we looked at Gini score, we looked at entropy and so on. For clustering, again the purpose is homogeneity. You want to form groups which are homogeneous with respect to certain variables.

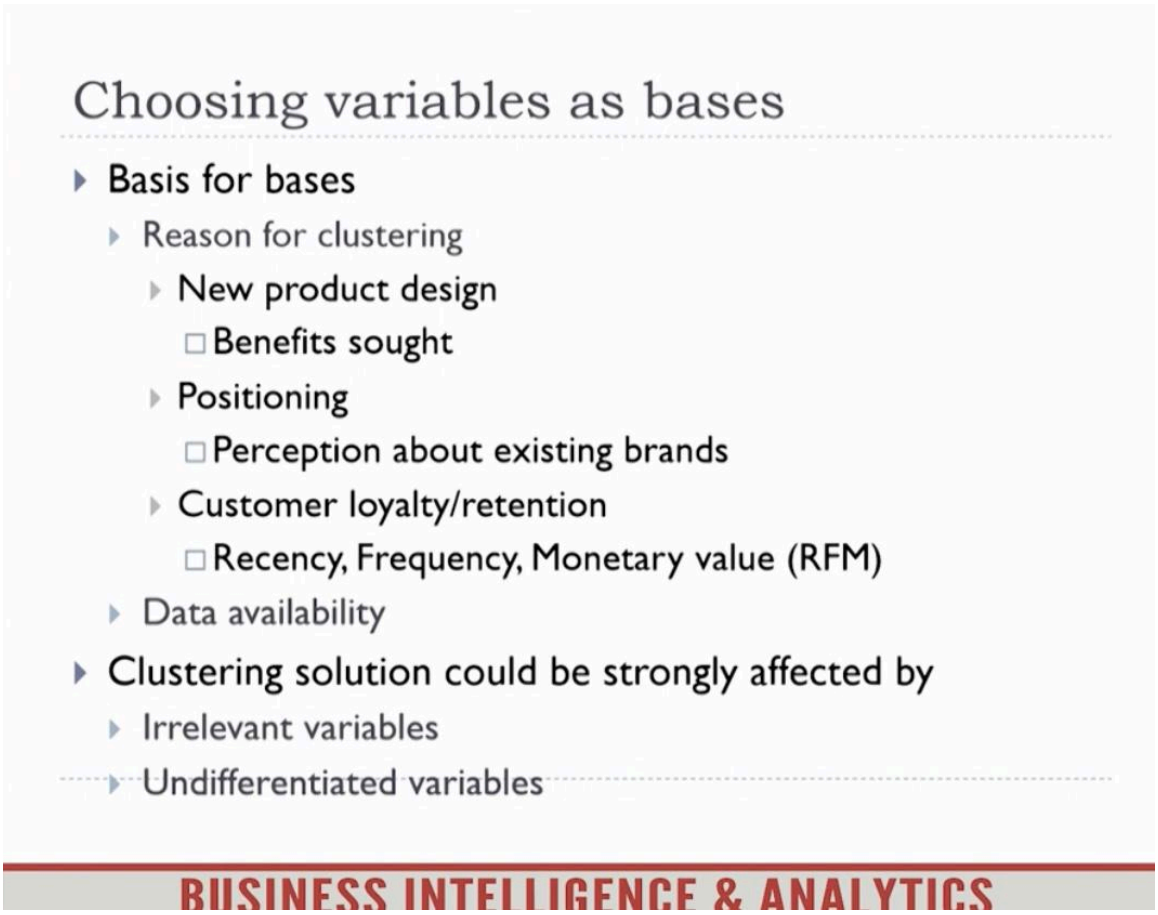
Objective is similar here also. When I form five groups in the class with respect to certain variables, they should be homogeneous with respect to those variables. And also one group here, another group there. You are homogeneous, they are homogeneous but within the group. But these two groups should also be, they should discriminate.

They should be different. The two groups should be different otherwise you can add more the groups. There is no point in having two groups if they are similar. But you should be, say high scorers or medium scorers or, you know but highly socially skilled etc. So each group should have different characteristic.

So that is called the profile of the cluster. You need to profile the cluster. Looking at the profile of the cluster, it should be useful and interesting. And then choice of algorithms, we will see what are the different types of algorithms for clustering and how many clusters. Just like we have the parameters, control parameters to control and prune the trees. In clustering also, you can control the number of clusters by specifying the number of clusters.

In certain algorithms you can directly specify. For example, there is an algorithm called K-means algorithm. K-means is widely used. The K is nothing but the number of clusters. You specify it a priority.

Starting itself you have to specify. Then describe the clusters. So cluster profiling. So basis for choosing base variables, I think I talked about it already. It depends on the purpose in business.



It depends on the purpose of business. If you are marketing a product and your aim is positioning, so you may actually cluster based on perception of brands. Whereas if you are clustering for customer loyalty, customer retention widely used variables are the RFM variables. Recency, frequency and monetary value. In marketing literature, this is widely discussed. Your classification of customers into platinum, gold etc is based on RFM kind of variables. Not just one variable but multiple attributes. I have given you a reading in supplementary readings. You can read that. That is on RFM.

Delta Airlines continues to use RFM as a method for customer classification or customer loyalty program. These are general principles. Let us move on.

Let me actually start with an example. Clustering algorithm. I will discuss the details in the next class. Just to get a sense of how we are going to proceed. There is something called loyalty in business. You actually want to build loyalty in customers and organizations invest hugely in loyalty programs. Investment hugely happens in customer loyalty, customer relationship management.

These are areas of huge investment. Some companies invest more in CRM than in R&D. Many companies actually. Why so? There is of course, evidence for it. Loyal customers, they are buying behavior or they buy more as the years go by.

They bring more customers. They bring referrals and they need much less servicing because you get used to the business etc. So, loyalty is an important aspect which companies build.

Now here is a small survey where two variables V1 and V2 are used. Two variables. So, two variables V1, V2. V1 represent store loyalty. V2 is brand loyalty. Now what is done is, seven respondents were given this instrument. So, you can actually classify seven

respondents or put seven respondents as A, B, C, D, E, F, G. Seven respondents. You get some data from each respondent.

Respondent A will give some value say 7, 3. B will give some value 4, 2. C gives some value 7, 2. D gives some value 3, 1. 6, 2. 6, 1. 3, 2. I am just putting numbers here. Just look at this data, with that the session gets over. Just look at this data, who are the respondents who are sort of having similar response or similar perception about loyalty? A and C. 7, 3, 7, 2. They their perception is very close. And yeah, maybe 6, 2 also you may write. They are one club. By looking at the data itself, we will try to suggest, well they feel their experience is similar.

And look at D, D and G. 3, 1 and 3, 2. They are also very similar. Are you getting a sense of what clustering will be doing? Next class, we will see how this. Essentially we are making certain sense based on proximity. They are close in terms of the values, in terms of the responses they are close. So, they feel similarly, they experience similarly, they must be put into same group.

That is what clustering is going to do. Practically in clustering, you have variables, their values and then you look at each pair of value and see how close some are in terms of these values. So, we talked about similarity. Similarity is a measure now we need to develop, to sort of shape this idea of bringing objects together based on certain measure of similarity. And that is what we are going to discuss in the next class.

There are different measures of similarity for different data types. All that, next class and we will work on a problem in the next class. Thank you very much. See you again very soon. Thank you. .