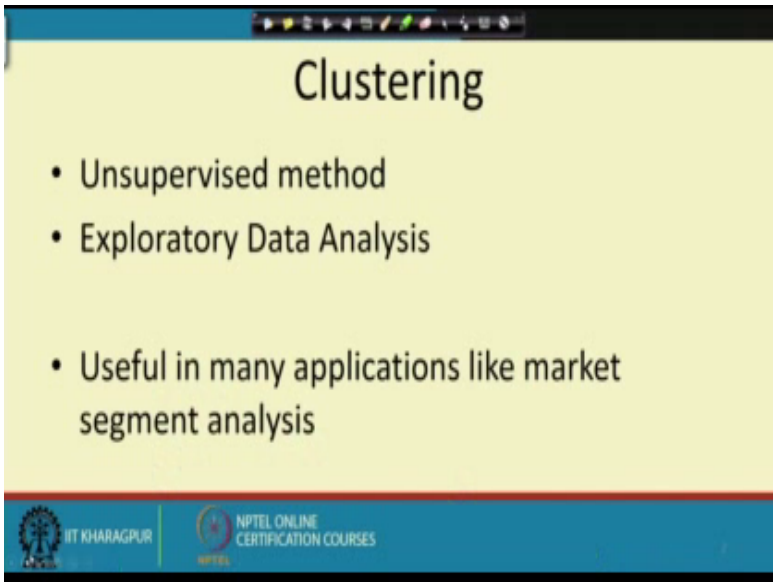
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture - 32
Clustering – I

Welcome to the lecture on Clustering. So far we have studied the algorithms where you had a training set and you have to propose, you have to predict some kind of class level for an unknown example.

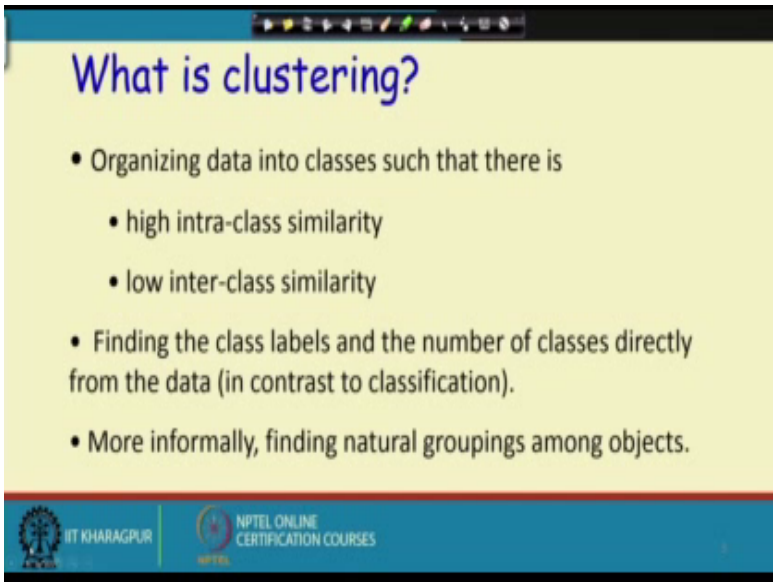
(Refer Slide Time: 00:30)



So, the number of classes are fixed beforehand and you have a training data which for a given input tells you that this would be the class and you train on that. So, this type of learning is called supervised learning.

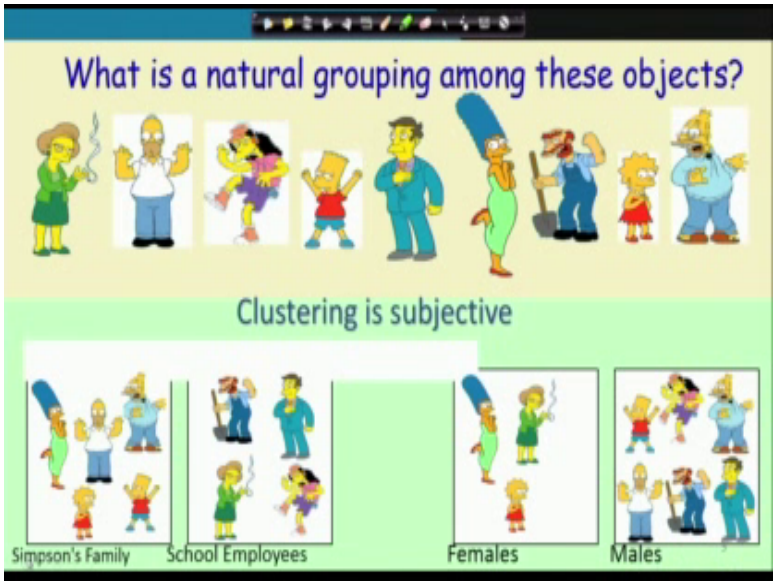
On the other hand in some of the application it may be the case that we need unsupervised learning or exploratory data analysis. For example, suppose I am looking at customers in a supermarket say big bazaar or something and I want to find group of customers I do not know beforehand what are the groups. But looking at the behaviour of the customers I want to find out the groups. And there are many other applications for example, I want to group the images in internet into similar things.

(Refer Slide Time: 01:47)



So, the definition of clustering is you have to organize the data the points, the examples into groups which are homogeneous. Homogeneous means if you take points from the same group they are close whereas, you want to find if you take the points from different groups they are apart. So, you want to maximize intra class similarity and minimize inter class similarity among the points. So, this actually amounts to finding the natural groups in the data, natural groups in the data.

(Refer Slide Time: 02:43)

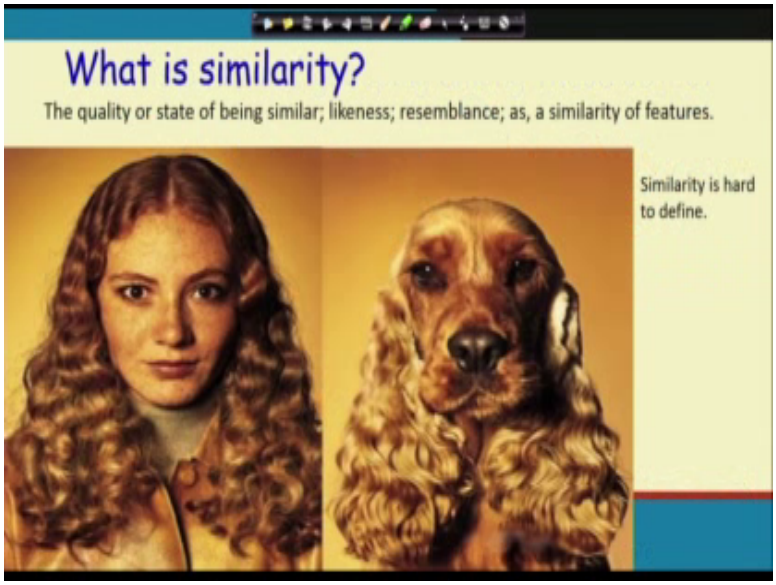


So, what does natural groups mean? So, suppose I have this images, so if you note that the previous definition I said that you group your data into similar objects, similar. So, the terms similar you have to first define before you define what is a group of similar object ok. So, that this similarity measure defines what is a natural group.

For example, what is the natural grouping among these objects that I am shown in the picture. So, I can have several possible groupings. For example, I can have two groups this 9 images I can let us say all members of this family these are people have a family and maybe some employee of the school, that can be a possible grouping one clustering or I can say I group into male and female that is another possible clustering.

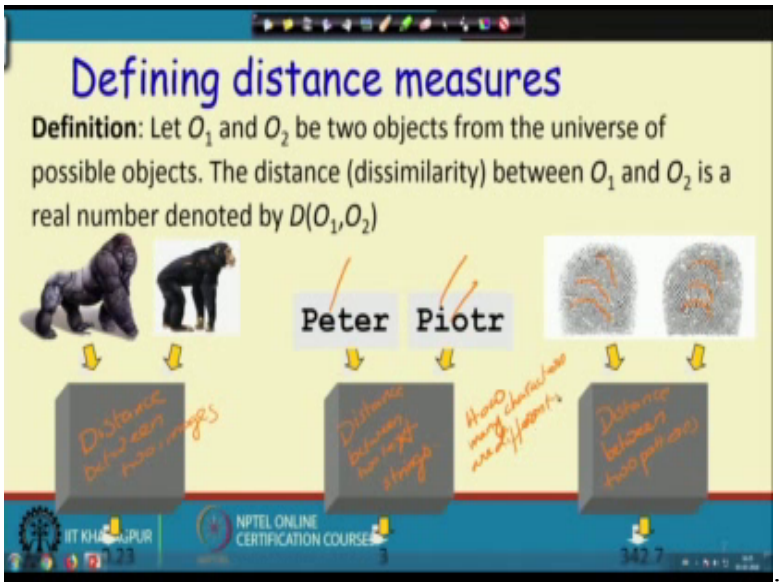
So, what is the natural grouping, what is a similarity is subjective ok.

(Refer Slide Time: 04:08)



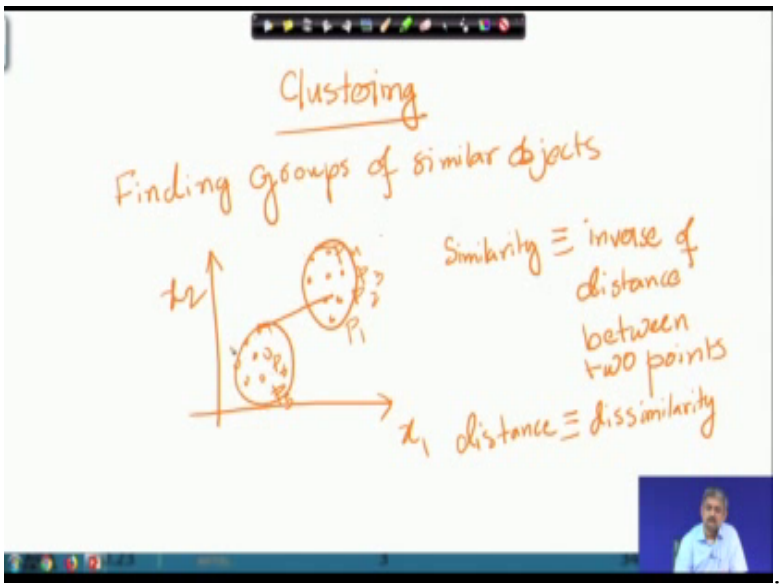
So, first you have to define what is similarity. So, naturally it will depend on the attributes of the or the features of the objects between which you find in similarity. For example, here if you are looking at the texture of the images or the colour of the images or the position of the eyes in the images they are similar ok. But actually in there is some other aspect in which they are not similar. So, it is a difficult task to define similarity, difficult task to define similarity.

(Refer Slide Time: 04:50)



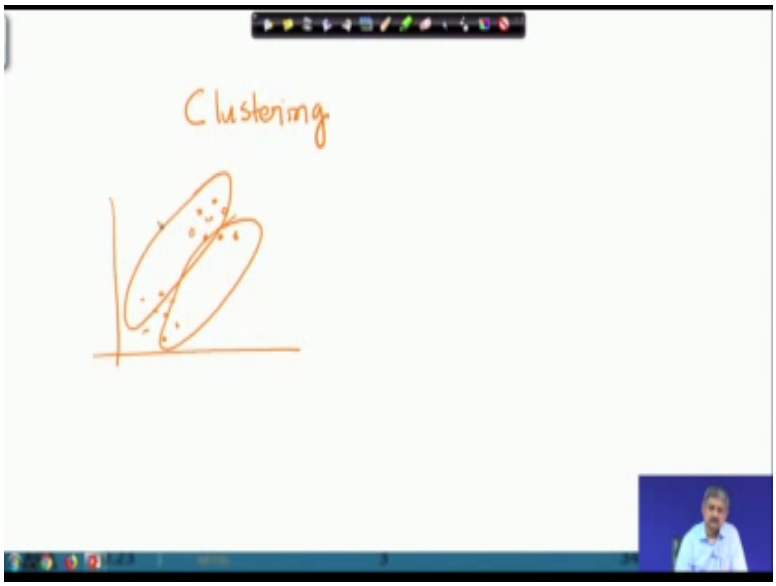
Usually, and once similarity is defined that is clear. So, once similarity is defined what we will do is the following I will explain you with an example.

(Refer Slide Time: 05:10)



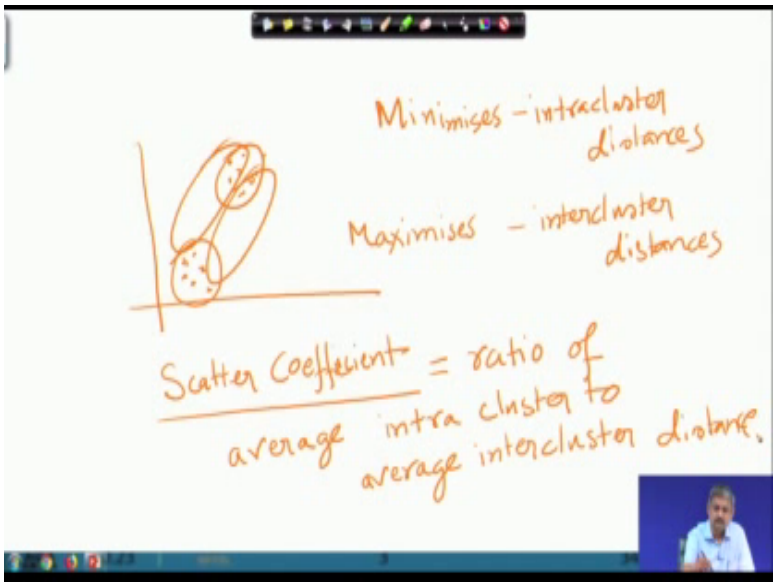
So for example, if I have two features and my data points are say and so on and maybe that is another set of points, note that there are no more class levels and suppose two points ok. So, I use similarity between two thing is this distance say; so basically dissimilarity. Then what I am trying to do is to find all groups which are close ok.

(Refer Slide Time: 07:06)



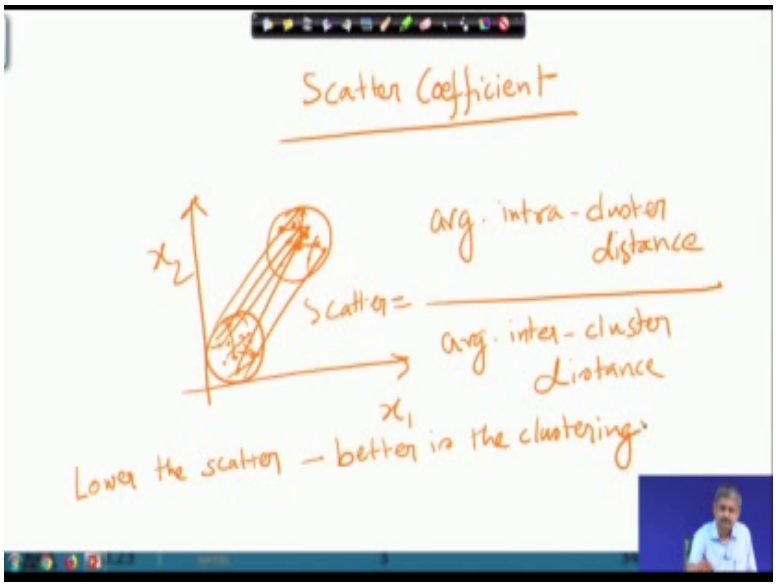
So, to x; sorry to explain again, if I ask you to split up these points into two groups you will never do like this, all right; instead you will do this group. Why? Because it, or this distances ok.

(Refer Slide Time: 07:30)



In fact, there is a measure of how good the clustering is which will be which will prefer this cluster, the these two cluster instead of this is called scatter coefficient, scatter coefficient. This is a very important term ok. It is the average, it is the distance ok. So, how do you calculate it ?

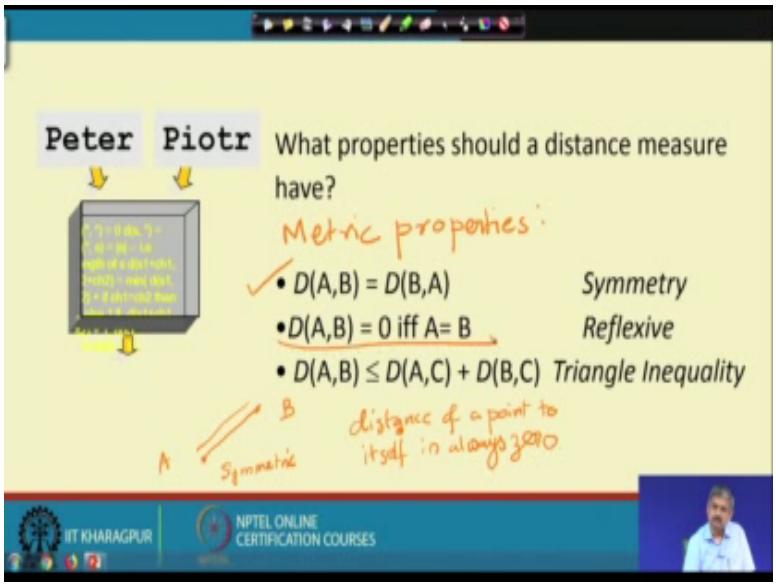
(Refer Slide Time: 09:25)



If I have points like this and this is my clustering you take all these distances. So, the average of all this take all these distances, distance ok. So, this ratio is scatter ok. So, given a set of point, given a partition into two groups you can always measure the scatter and lower the scatter better is the cluster all right. So, let us come back to our discussion.

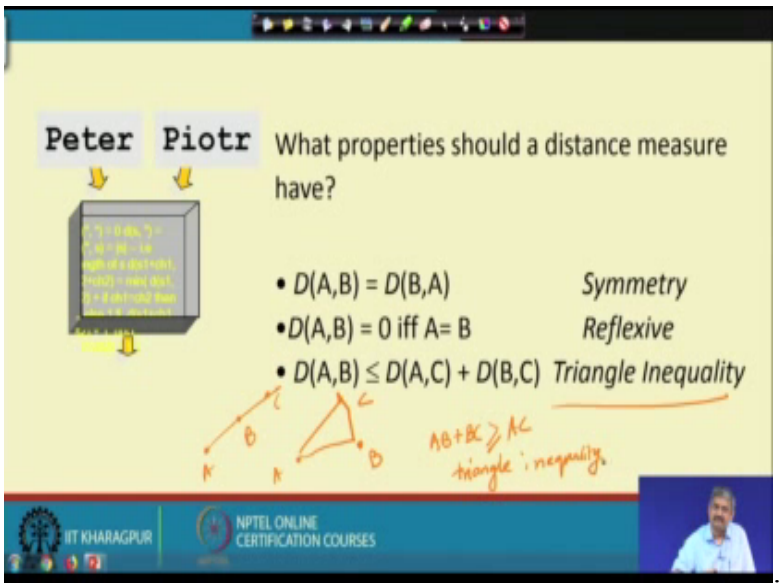
So, basically we need some kind of distance measured between two objects, as I have said dissimilarity. So, in general distances are complex. For example, you how to distance between two images, two images. How do you do? Maybe I look at the pixel values and see how much they are different hm. You can have two patterns maybe their shape, how does they match and so on ok. You can have strings how maybe how many alphabets are different ok, how many alphabets are different, are different ok. Whatever value you take you have to satisfy certain property for it to qualify as a distance measure; however, you define distance. This properties are called.

(Refer Slide Time: 13:09)



So, you can define your own distance function, but you have to ensure that these properties hold. What is the property? Symmetric, that means, if you take two points A and B, the distance between A to B is same as the distance between B to A ok, symmetric. Reflexive, so this is, reflexive means if you take the same point the distance of a point to itself is 0, is always 0, reflexive all right.

(Refer Slide Time: 14:38)

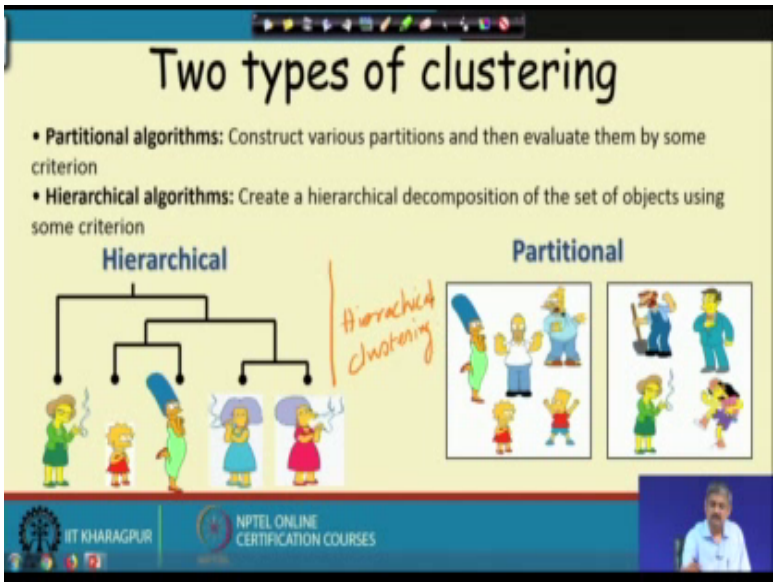


The third one is the triangle inequality that means if you take 3 points if you take the distance A to B plus B to C it is greater than or equal to distance from A to C directly. So,

AB plus BC in some cases they will be equal if a BC are in a straight line, AC, this is the triangle inequality ok.

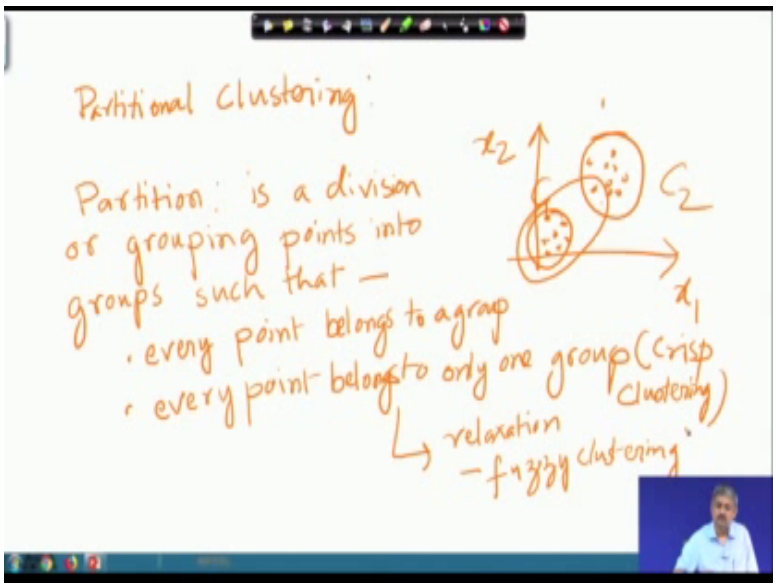
So, this is that three properties you should satisfy.

(Refer Slide Time: 18:38)



So, now, using this properties you have to group them, you have to cluster them ok. So, there are again two types of clustering, one is called hierarchical step by step, one is called partitional right. So, what is partitional ?

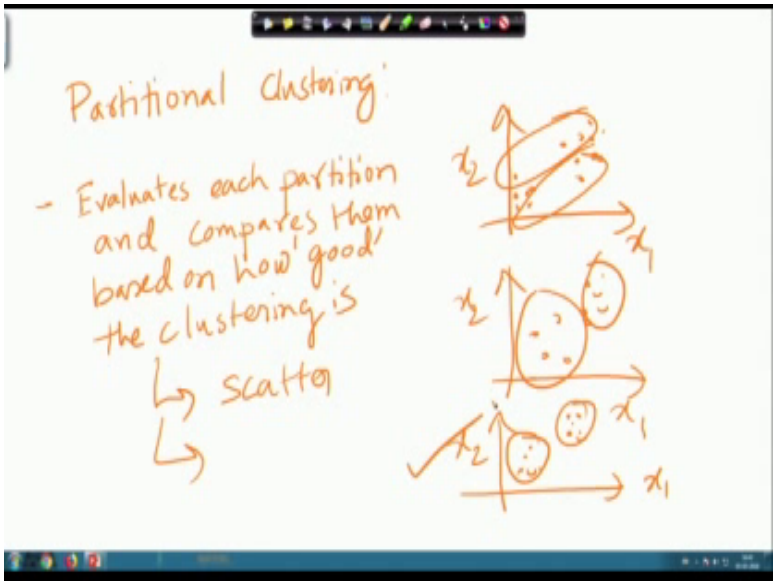
(Refer Slide Time: 16:17)



Now, will I do it. What I will do is given my points to be clustered. By the way what is a partition? Is a division or grouping points into groups such that two things point belongs to a group, every point belongs to only one group.

So, if I group it into two this is group 1, this is group 2, let me call it cluster 1, cluster 2 every point falls in one of this two group ok. So, if your point is left out it is not clustering. Similarly a point belongs to only one group in other words you cannot have clusters like this, overlapping thing cannot have ok. There is an alternate called fuzzy clustering which we have not discussing, where it is allowed. So, relaxation of this only one group is called the fuzzy clustering, whereas this is called a crisp clustering right. So, what does partitional algorithms do ?

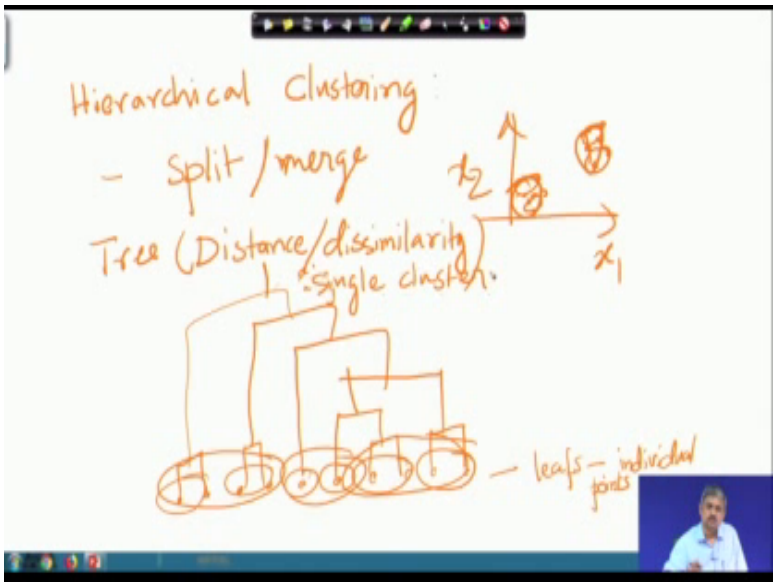
(Refer Slide Time: 18:50)



What it does? You have your points, you make all possible partition, this is a partition, this is another partition, this is another partition. What it does? Evaluates each partition and compare them based on how good the clustering is ok.

So, which will of course, to when you need how good I tell you can measure by scatter or some other measure we will discuss later ok. So, it checks all this, it checks all the possible clustering and chooses the best in this case maybe this is the best all right. So, this is partitional clustering.

(Refer Slide Time: 20:44)

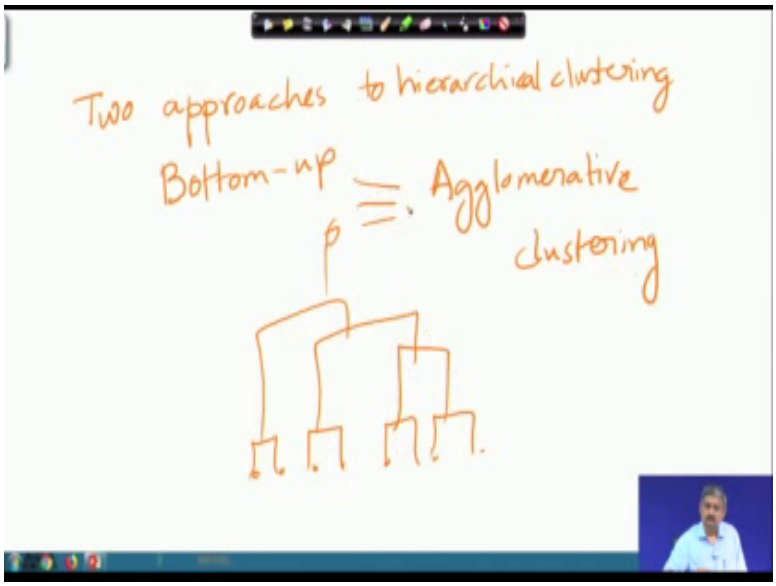


An alternate is hierarchical clustering ok. So, what it does is like this. It follows a policy of either split or merge ok, it does not try out partition. What it does for example, it will first try to make small clusters and say this is 4 cluster, then merge these clusters into larger clusters ok. So, you construct a tree, a dissimilarity tree or a distance tree. What you do in the tree? If a tree looks like, this in the leaf you have the individual points.

Now, the tree merges this leaves some order, some order it merges these leaves, they keep on merging till every. So, so what happens in the mod does merging means? This means this two array cluster, this means this two array cluster, this means these two array cluster, these two, these two, these two. Then at the second level these two array cluster these two array cluster. So, this way at every level you keep on merging till everything gets merged into a single cluster ok.

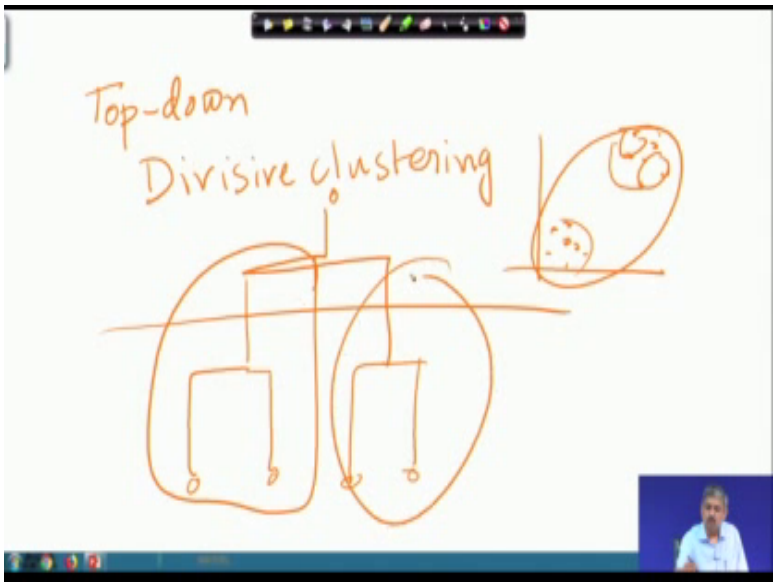
So, let me give an example. Suppose all the students first I look who have maybe in same class a group together, then maybe who are in the same department a group together, then maybe who are in the same college I group together, then maybe who are in the same state group together, then maybe who are in the same country group together. So, this is hierarchical clustering. There can be two approaches. So, see this is what, that is what we may done here hierarchical clustering.

(Refer Slide Time: 23:59)



So, there are two approaches, bottom up where you start from the bottom of that tree and then keep on merging and finally, everything one. So, this is called agglomerative clustering bottom up.

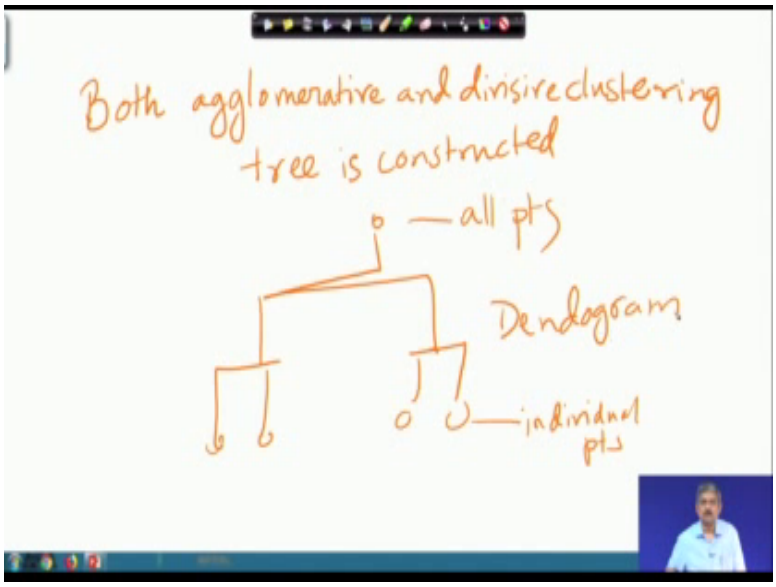
(Refer Slide Time: 24:59)



Similarly you can have top down. So, what you do? Initially everything is a cluster then you break them up into two cluster, then maybe subdivide. So, you construct the tree top down, till you every point is a separate cluster. So, this is called divisive clustering. Both are hierarchical, hierarchical , both way you can construct. And the how do you get the

actual clusters from this? You cut this tree at certain level and say everybody in the left of the tree is one, in the right of the tree is one, so that you can do either in agglomerative, both.

(Refer Slide Time: 26:05)



See in both this clustering tree is constructed ok, either bottom up or top down. So, this is all points and the leaf is individual points. This tree is known as the dendogram, all right.

So, I stop here today. In the next class I will explain how these dendogram is constructed in different algorithms.

Thank you.