Learning Analytics Tools

Ramkumar Rajendran

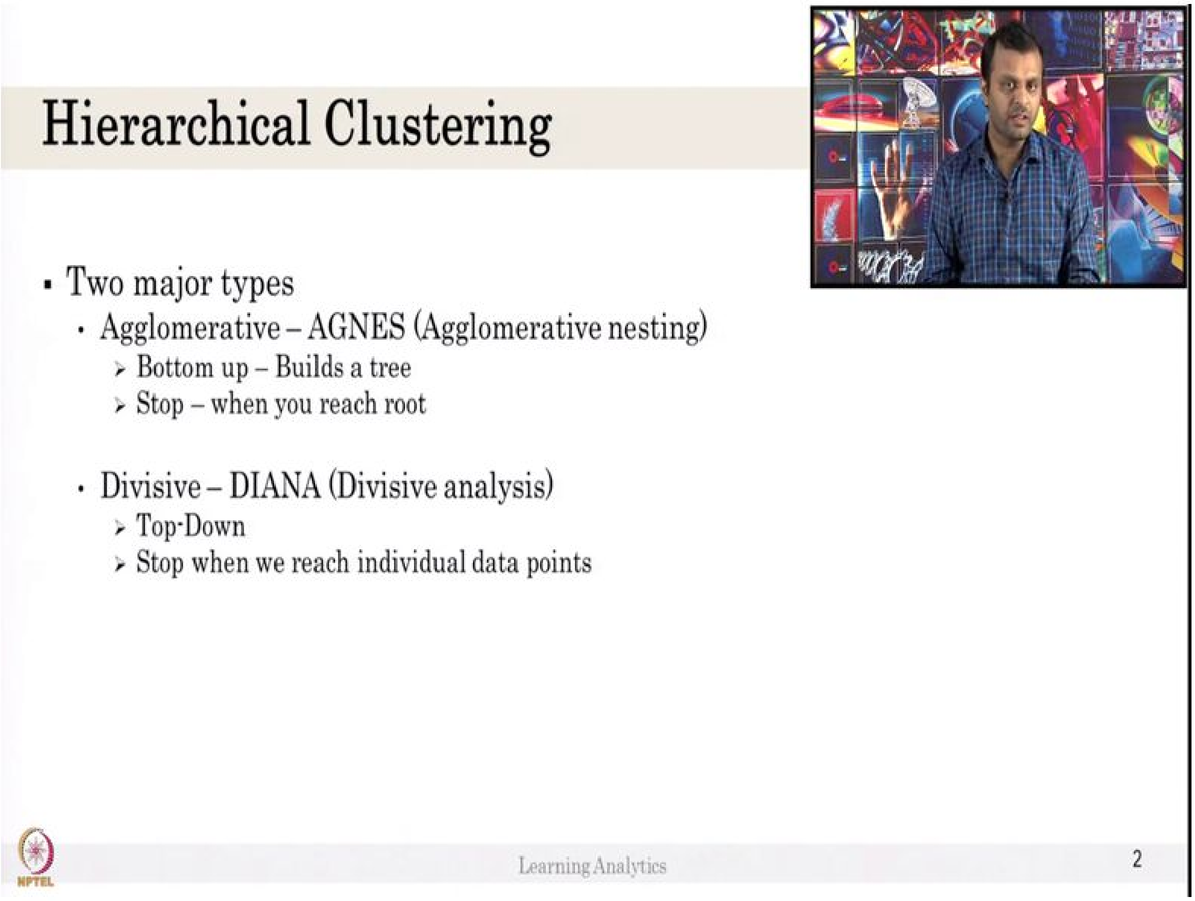
Educational Technology

Indian Institute of Technology, Bombay

Lecture 7.3 - Hierarchical Clustering

In this video, we will discuss what is hierarchical clustering. So, in the last video, we discussed what is K-means clustering that is centroid based clustering, hierarchical clustering is a connectivity-based clustering.

(Refer Slide Time: 0:46)



Hierarchical Clustering

- Two major types
  - Agglomerative – AGNES (Agglomerative nesting)
    - Bottom up – Builds a tree
    - Stop – when you reach root
  - Divisive – DIANA (Divisive analysis)
    - Top-Down
    - Stop when we reach individual data points

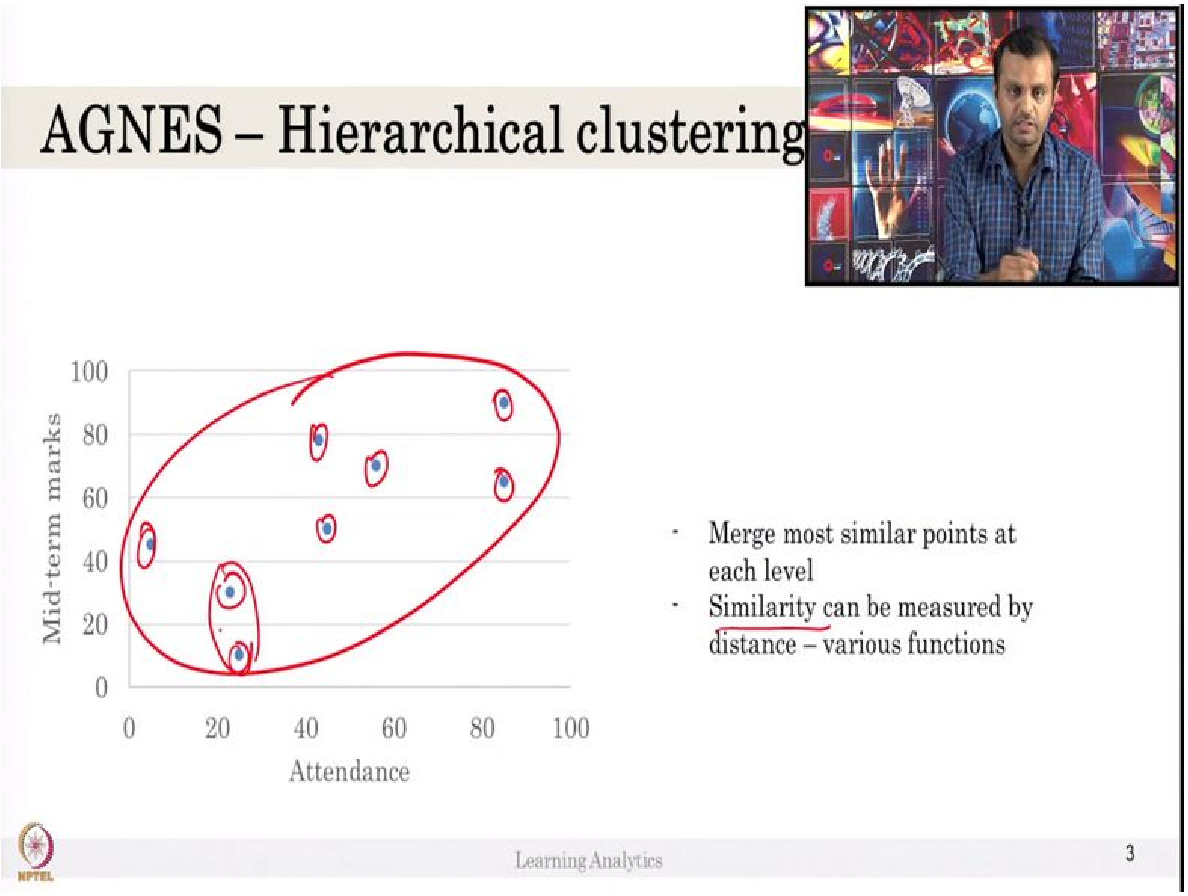
Learning Analytics

2

There are two major types within hierarchical clustering one is agglomerative that is called agglomerative nesting, AGNES. We will see this in detail in this class and the other one is divisive analysis like DIANA (D-I-A-N-A). So, divisive analysis, DIANA analysis. So, divisive hierarchical clustering or agglomerative hierarchical clustering. So, for the AGNES we build a tree from a bottom-up, we will see what is a tree, what is the bottom up? (we will stop when we reach the root).

So, for DIANA, it is just opposite. We will start from the top and we will stop when we reach all the data points, all the children's in the given sample. So we will see what is exactly this bottom-up and top-down approach, what is the difference between AGNES and DIANA? So, we will discuss only the AGNES in this course. So, it is simple, you can understand how the other clustering algorithm works.

(Refer Slide Time: 01:36)



AGNES – Hierarchical clustering

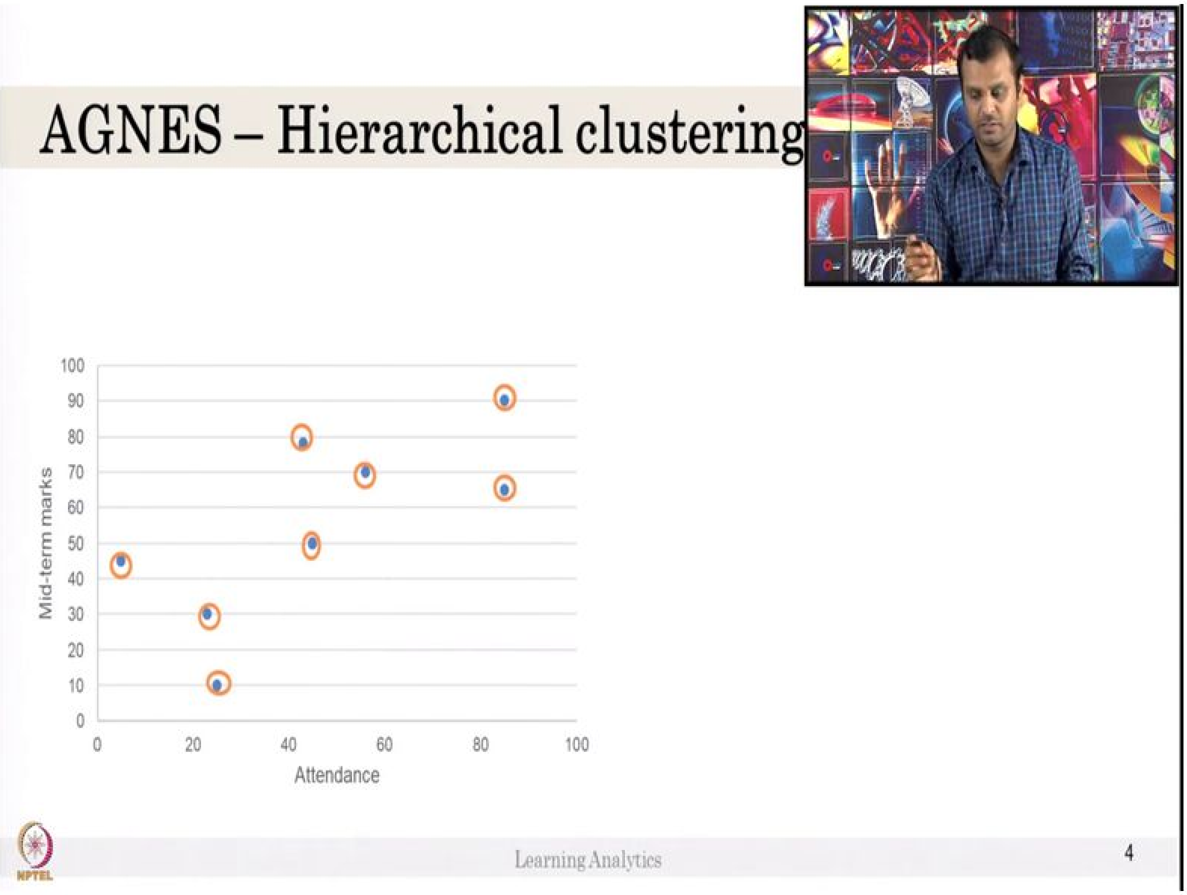
- Merge most similar points at each level
- Similarity can be measured by distance – various functions

Let us see hierarchical clustering, the same dataset, but I think some points were removed to just to make it simple. The idea of hierarchical clustering is that you have to pick each and every data point as a cluster because it is bottom-up. So, the number of clusters here is eight. Each data sample is a cluster. So, each and every data point will act as a cluster, i.e. everything is cluster then you merge the most similar points. So, these are very similar, maybe this can be another cluster, something like that. So, you merge most similar points in the next step, then merge more similar points in the next step.

So, till all of this belongs to one cluster. So, you make multiple levels. So, how would we merge the most similar points? There should be similarity measure like last time we saw the Euclidean distance to measure the distance between the centroid and the data points to pick which cluster the point should go. Similarly, here we also have to ask similarity measures. Similarity measure makes use of various formulas, a lot of formulas.

It is not important to know all the formulas here. But understand there is some similarity function, which measures the similarity between the two points. If it is similar, if this point is more similar to this point compared to this point, these two will be merged first as a step one. That is how the similarity measures are used.

(Refer Slide Time: 03:25)



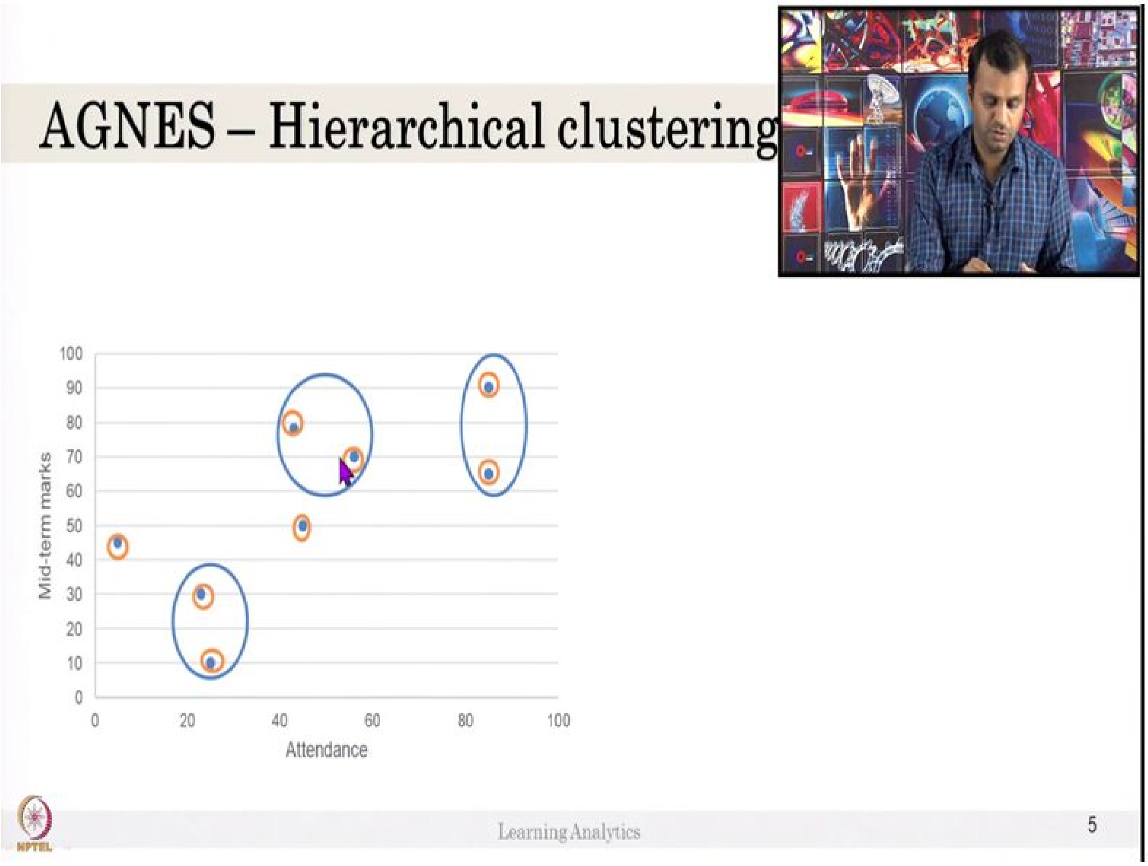
AGNES – Hierarchical clustering

Learning Analytics

4

Let us see how this works in multiple steps. So, the first step, what I did I considered every data point as a cluster. So, we start with eight. So, what I say, it is all the eight of them here. So, all eight of them are here clusters.

(Refer Slide Time: 03:44)



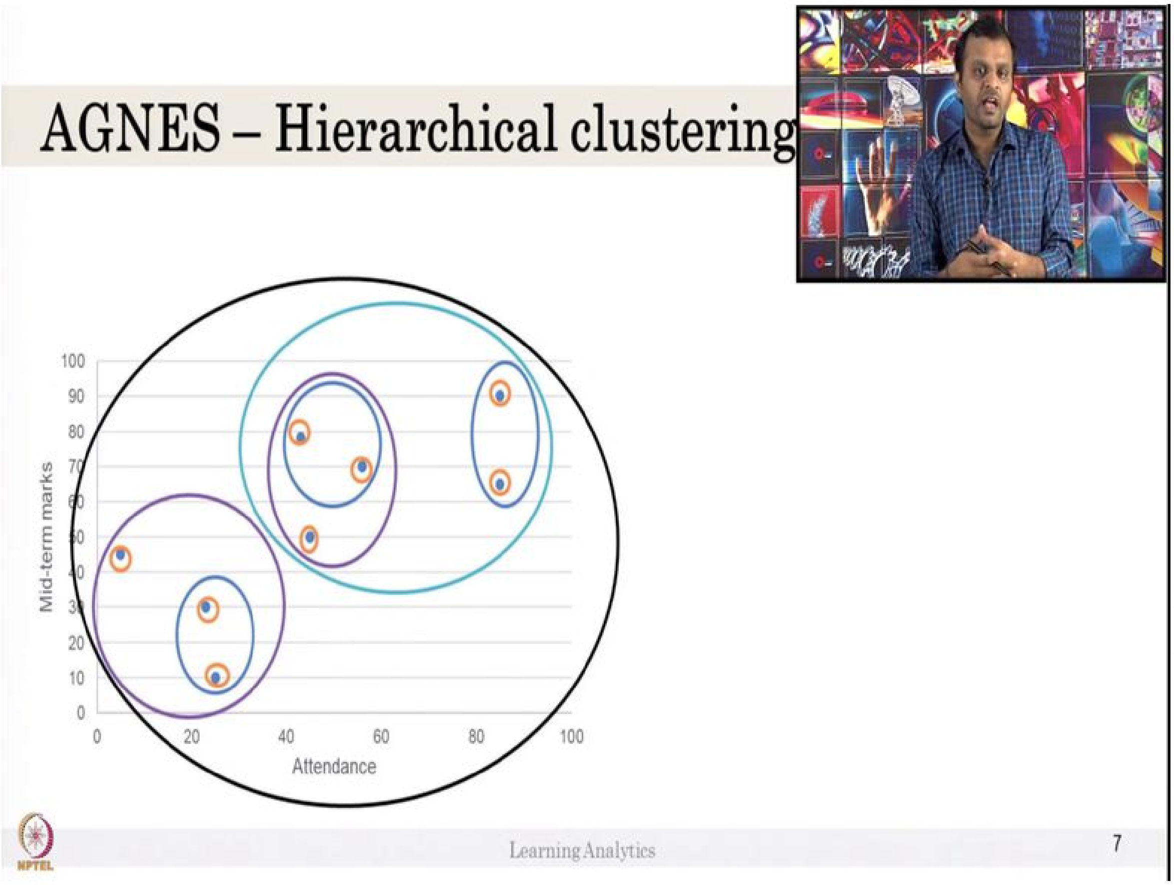
AGNES – Hierarchical clustering

Learning Analytics

5

So, in step two, we merge the most similar data points. By some measure, these two are more similar compared to these two. So, this is a one clusters, it is another cluster, so three clusters. So now we have five clusters, one, two, three, four, five, we have five clusters after step two.

(Refer Slide Time: 04:07)



AGNES – Hierarchical clustering

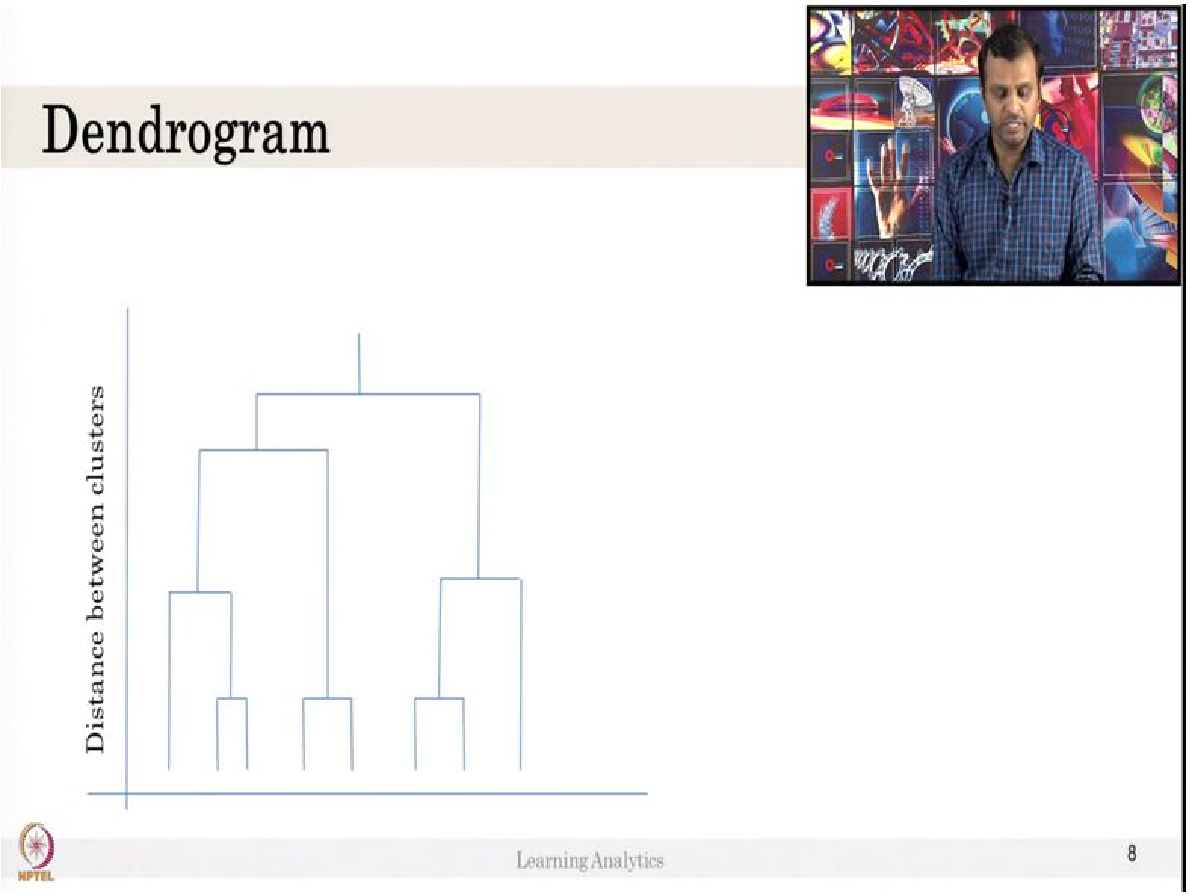
Learning Analytics

7

Now, this cluster and this cluster is similar compared to the other clusters. So, this form another cluster, so another cluster, this one, these two. So, by some similarity measures, okay. Consider now we have three clusters, one, two and this one, the three clusters at step three. At Step four, we might have another cluster combining these two. So, now we have two clusters.

And at Step five, we have one big cluster. So, the number of steps is not dependent on the number of data samples instead is it depends on the behaviour of the data sample. It can converge in six or seven steps or three steps or based on the behaviour of the data. So, this is what hierarchical clustering, this is called actually agglomerative clustering, agglomerative clustering starts with a one data point and tries to create clusters to neighbours using similar measures.

(Refer Slide Time: 05:11)



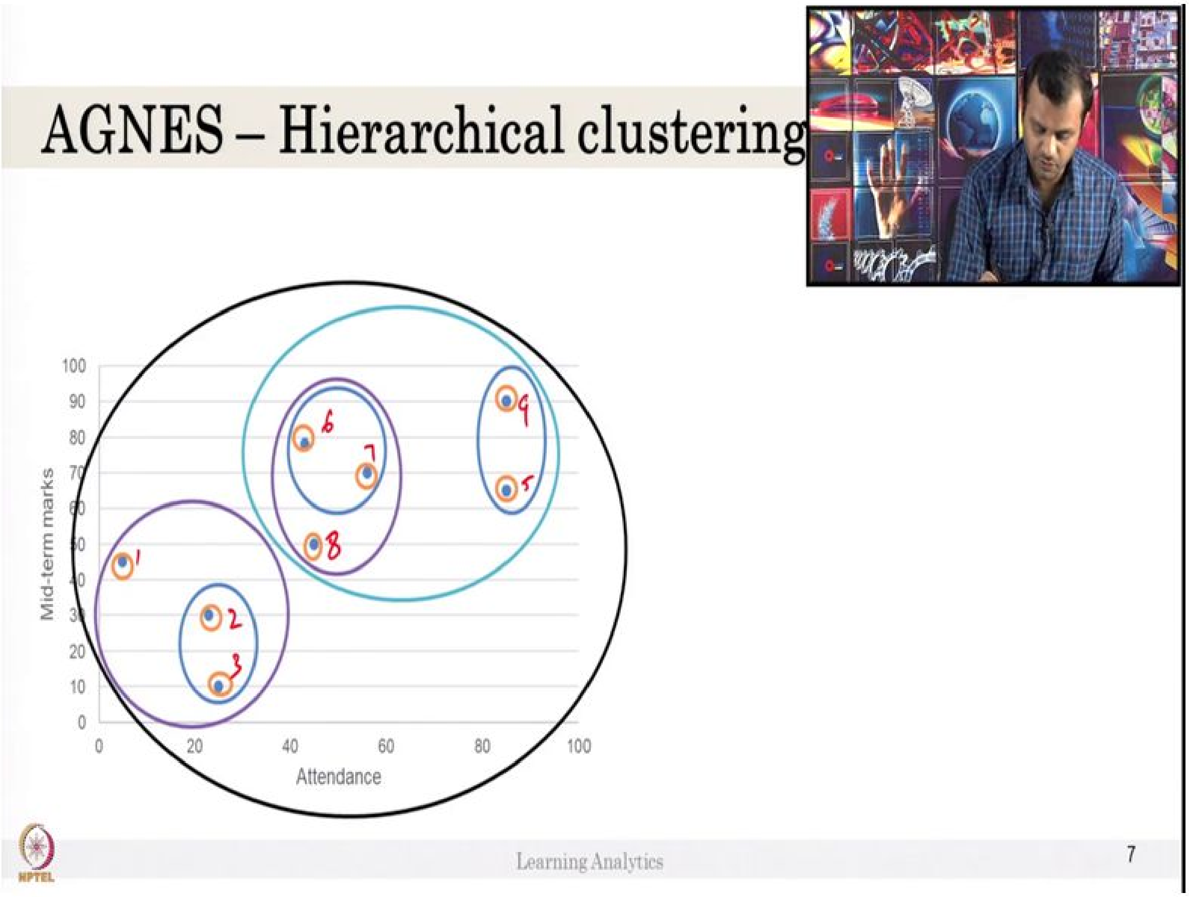
Dendrogram

Learning Analytics

8

So, this cluster will be described very easily with dendrogram diagrams. So, a dendrogram is another type of visual representation which we did not see in descript analytics because we will see this, so I thought not to discuss that multiple times. (This is not actual dendrogram diagram plotted from any library, instead, I just drew this dendrogram in the PowerPoint). So, let us see how this dendrogram works. Let us see how it works.

(Refer Slide Time: 05:43)



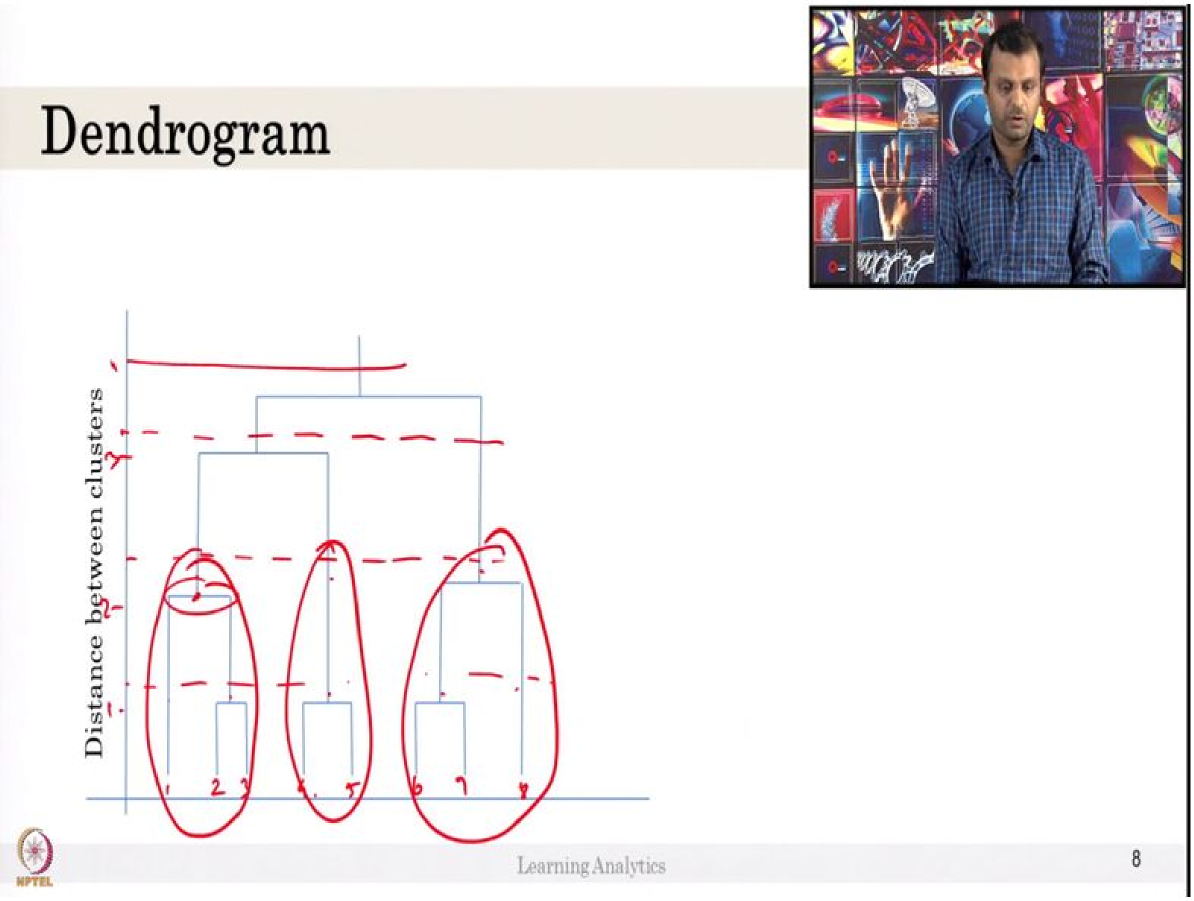
AGNES – Hierarchical clustering

Learning Analytics

7

If we give a number to some of the values, let us assume some numbers, one, two, three, four, five, six, seven and eight. Okay, that is kind of some changes here, this is not exact dendrogram but let us see how it goes.

(Refer Slide Time: 06:13)



Dendrogram

Learning Analytics

8

These two are similar measure so, this will be combined as one cluster. So, this is a step one. At step one these two will be combined and these two will be combined and these two will be combined. And at step two this point will form a new cluster, it is a node, this particular node there is one cluster.

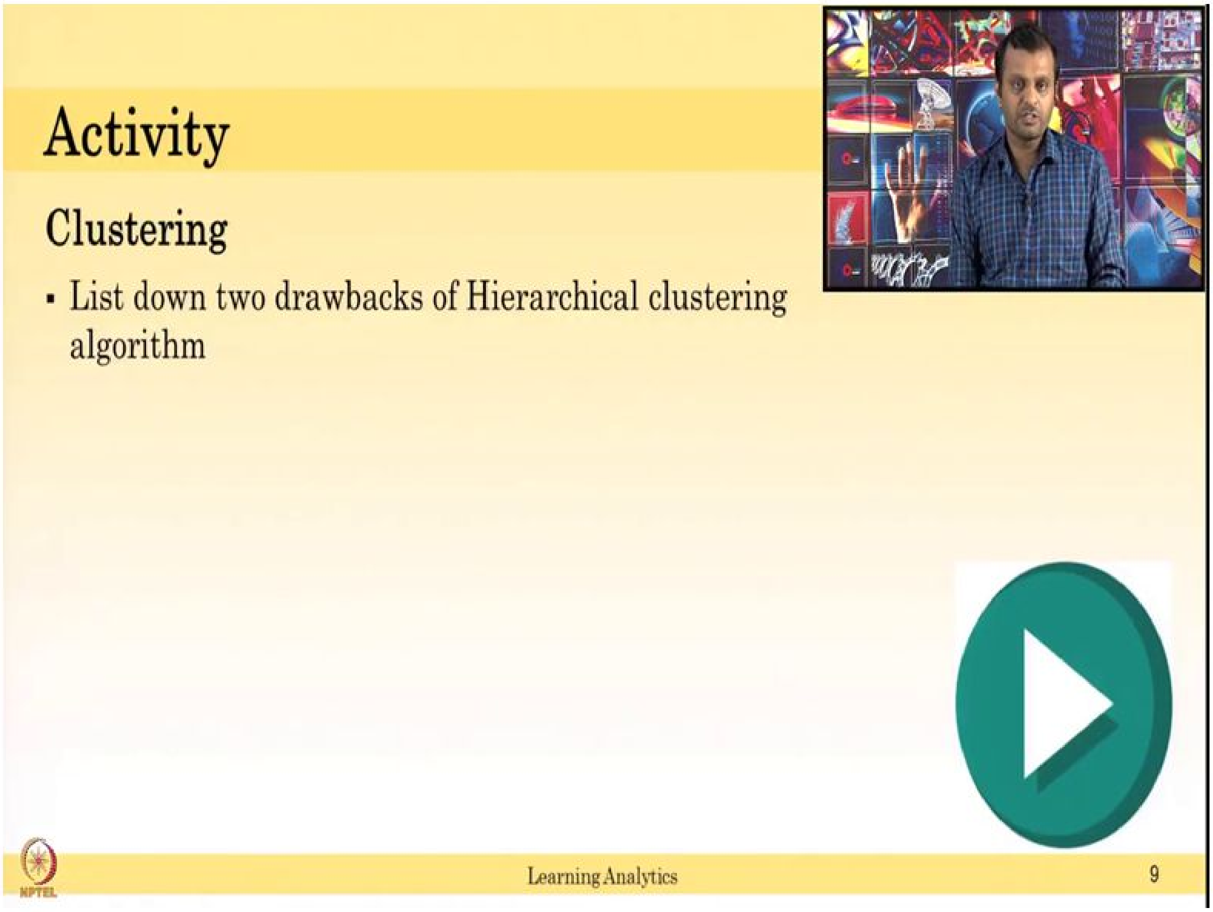
So, if I draw a line here, after step two, I should have three clusters. If we draw the next steps, these two are combined to a new cluster at step three. Then if I draw a line here, I should have one cluster, two cluster then this one more cluster. There is only one cluster after step four.

So, this dendrogram tells that how to represent agglomerative clustering in the visual format. So, if you have data, if you run the AGNES algorithm, you might get the dendrogram diagram on your data points. And that tells you how these clusters are formed and you can choose which level you want to go.

So, I want to say if I want to have three clusters, then these data points will be considered as one cluster, this is another cluster and this is another cluster. So you can choose as many clusters you want. Each data point is each student ID so you use each student's behaviour(student one’s mark in midterm), so each data point is each student's ID. So, I just wrote it in sequential for like one, two, three. But if you really look at the data point, this might be different.

So, this is how the hierarchical clustering can be used. So, to implement hierarchical clustering there are a lot of tools available, we will also discuss some tools for that. So, given the data compute the hierarchical clustering and you choose your level based on your data and your need. So, hope you understood what is dendrogram and what is hierarchical clustering.

(Refer Slide Time: 08:51)



Activity

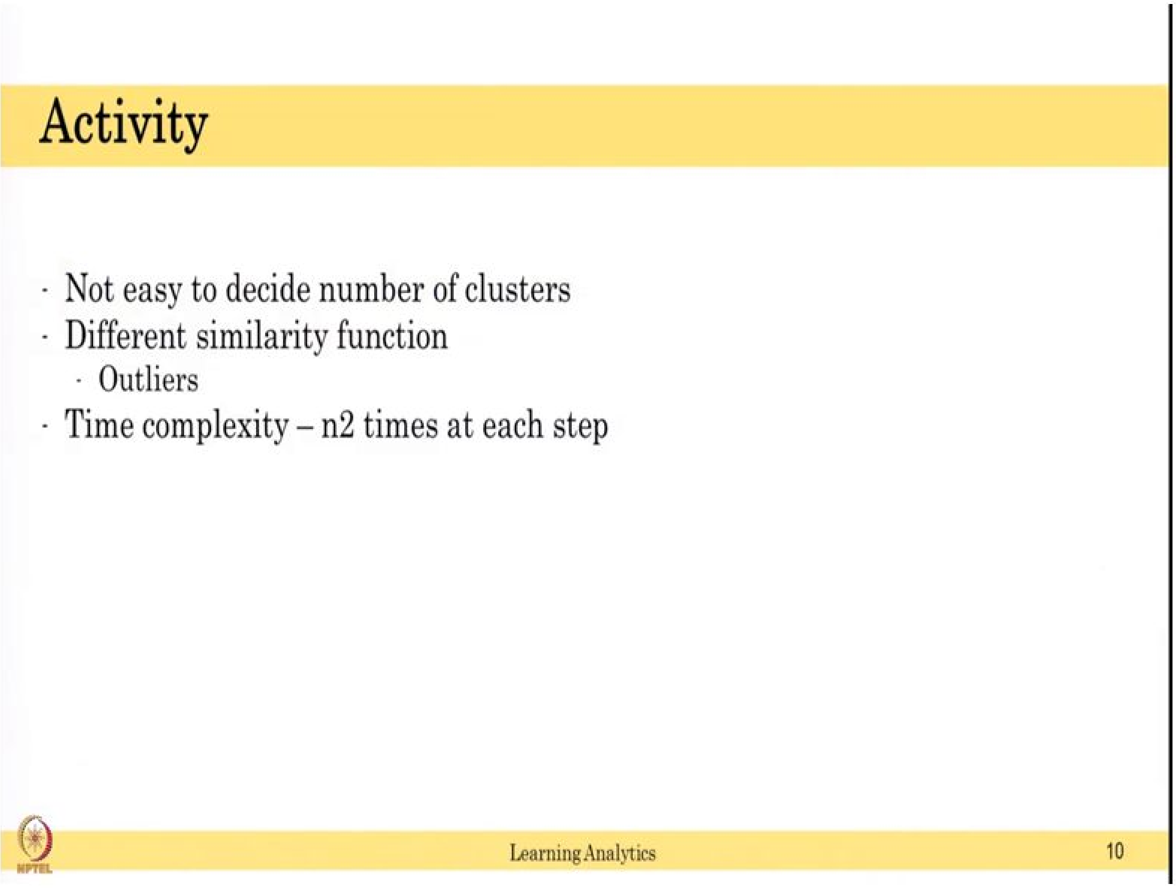
Clustering

- List down two drawbacks of Hierarchical clustering algorithm

Learning Analytics 9

So, can you list down two drawbacks of the hierarchical clustering algorithm? After listing it down rush to continue.

(Refer Slide Time: 09:01)



Activity

- Not easy to decide number of clusters
- Different similarity function
    - Outliers
- Time complexity – n2 times at each step

Learning Analytics 10

So it is not easy to decide number of clusters, but you have a complete picture like what if cluster value equal to one, what if cluster equal to two. So, you have complete picture, there is no need to have elbow method or compute the error function. There is no error function here but you have the complete figure, so you know where to choose. And the distance of the lines also important, I will tell about what is that lines in next video.

There are different similarity functions. One is single connection or single link or average, there are a lot of similarity functions. You should choose right similarity function, and some similarity functions are sensitive to outliers. So, if there is outlier, some similarity function may not work. So, you have to understand which similarity function to choose. So that is one drawback. I did not discuss that, but you can check what are the similarity functions used in hierarchical clustering.

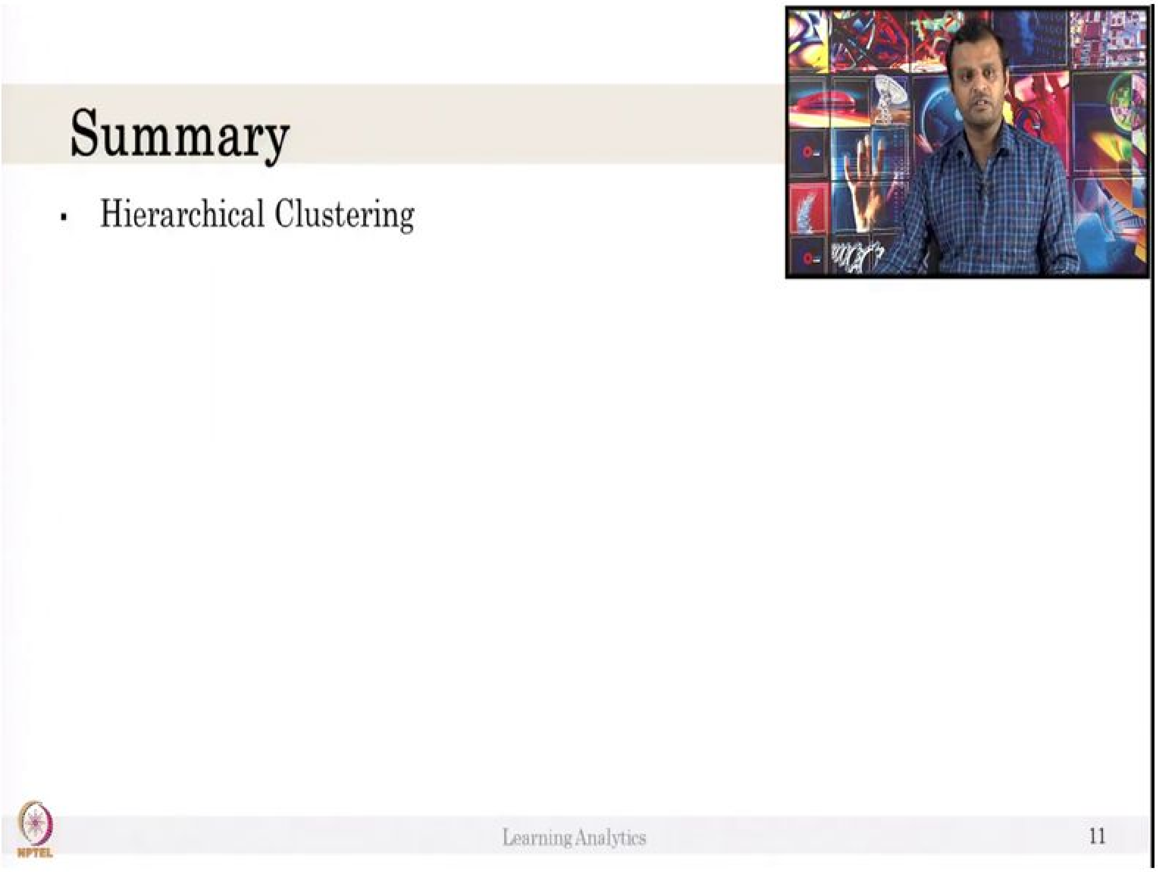
The main problem is time complexity. So, if O(n^2) and if we have say four steps or eight steps, then its computationally costly. The logarithmic scale of complexity, which is computationally feasible than the previous one is more desirable to have.

If you have a very small dataset, which you collect from 60 students with four or five features, do not worry about these things. So, for this 60 students there are five variables or 10 variables, this time is not really important.

For very large datasets, you want to do a cluster in real-time and you want to give feedback in a real-time, there the time is important. So, when you talk about time, you also have to consider where you will be using this? In offline or in a real-time. So, consider that. So, yeah, hierarchical clustering has couple of drawbacks. But it is easy and visually appealing, but you do not know how to choose the clusters.

So, some places in research we use K-means clustering and in other places, we use hierarchical clustering. We will see the example of both K-means clustering, hierarchical clustering in a research paper in the next video.

(Refer Slide Time: 11:43)



Summary

- Hierarchical Clustering

Hope you understood what is hierarchical clustering and thank you.