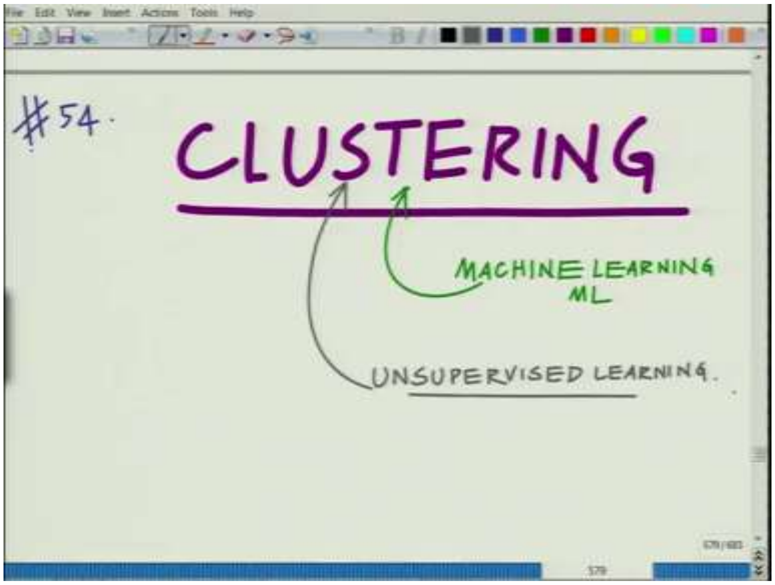
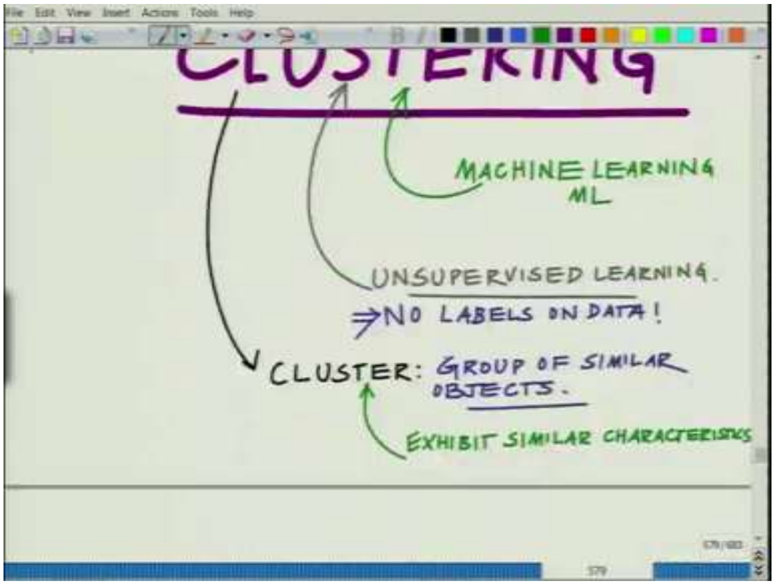
Applied Linear Algebra for Signal Processing, Data Analytics and Machine Learning
Professor. Aaditya K Jagannatham
Department of Electrical Engineering
Indian Institute of Technology, Kanpur
Lecture No. 54
Machine Learning Application: Clustering

Hello, welcome to the another module in this massive open online course. So, in this module let us look at another important application of the principles of linear algebra, that is in the context of clustering which is essentially a very important technique or algorithm in machine learning.

(Refer Slide Time: 00:32)





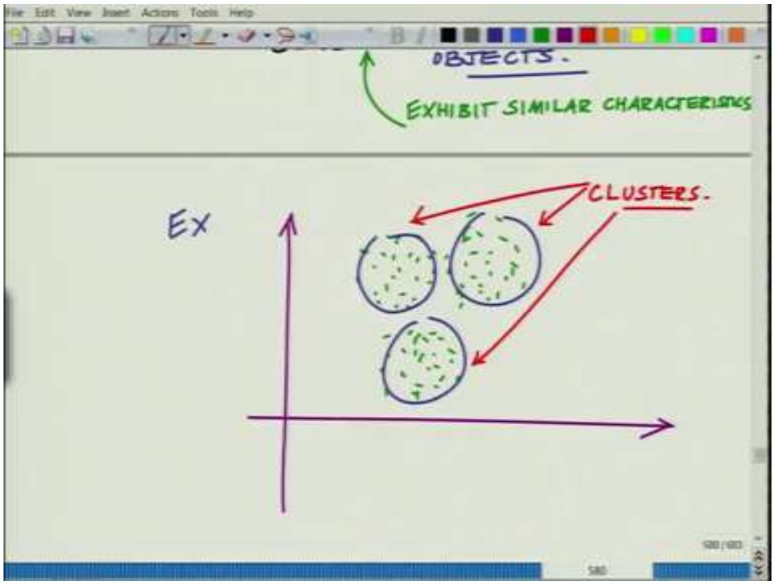
So, let us continue our discussion and look at another novel technique which is essentially clustering and this is very a important concept in the context of as said in the context of machine learning. So ,clustering is a very important machine learning tool or what we now call more often is ML and clustering essentially what it mean is, now what do we, I will come to this in a little bit now clustering belongs to a class of techniques known as unsupervised learning.

What does this notion of unsupervised mean? Unsupervised means you have a large amount of data, and you just have a bunch of data, huge amount of data and this data is not essentially labeled, no one had essentially labeled or put meaningful tags on this data. So, it is essential that one has to now organize this data so as to say into whatever it is, in a hierarchical fashion, clusters or so on, so as to extract relations and meaningful patterns from this data.

So, unsupervised learning is essentially learning or machine learning that is performed when there are no, this implies there are no labels on data or data is not labeled. That is another way of saying the same thing, and now clustering naturally means it begins with the notion of a cluster. So, naturally a cluster, and all of you must heard this them cluster. Which is essentially in English means a group of similar project, product or group of similar objects.

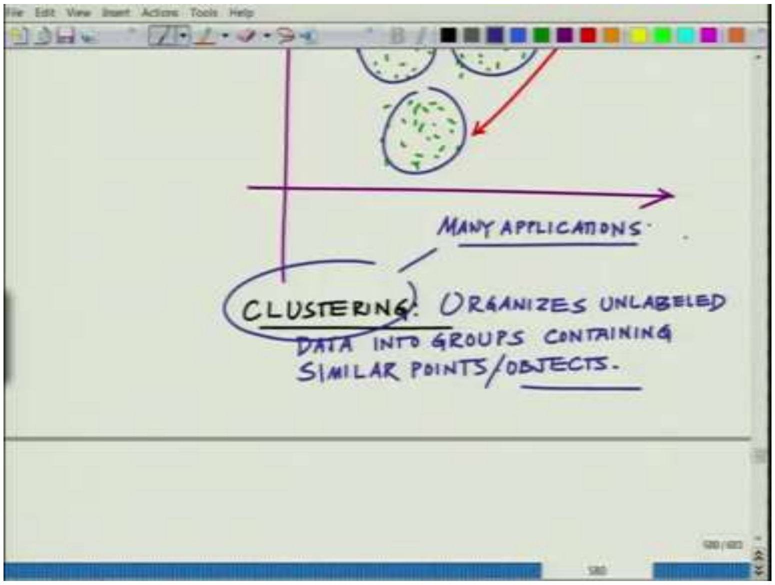
Cluster essentially means a group of similar objects who exhibit some kind of similar characteristics. I mean that is the intention, that is the cluster is tight knitted group exhibiting some similar characterizing features. So, that is the important aspect here. This exhibit similar characteristics, that is essentially, now for instance example you can visualize this as a following diagram.

(Refer Slide Time: 04:21)



OBJECTS.

EXHIBIT SIMILAR CHARACTERISTICS

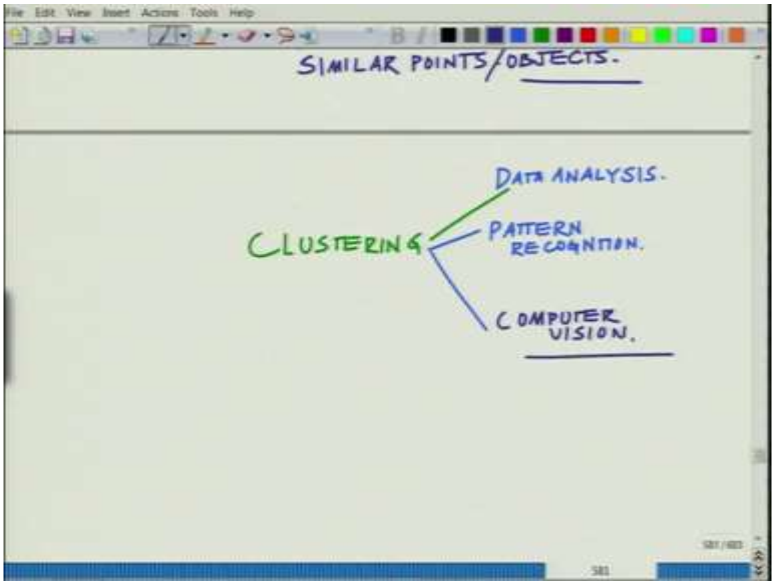


CLUSTERING: ORGANIZES UNLABELED DATA INTO GROUPS CONTAINING SIMILAR POINTS/OBJECTS.

For instance let us say you have space and then you have the points that are lying as such and they are let us say your points and then you can see you can extract some kind of meaningful clusters from this. You can form groups closely related data points and these are essentially what are, these are essentially your clusters. I mean closely related points and clustering is nothing but essentially organizing your data into such clusters, groups of objects or data that exhibits similar characteristics.

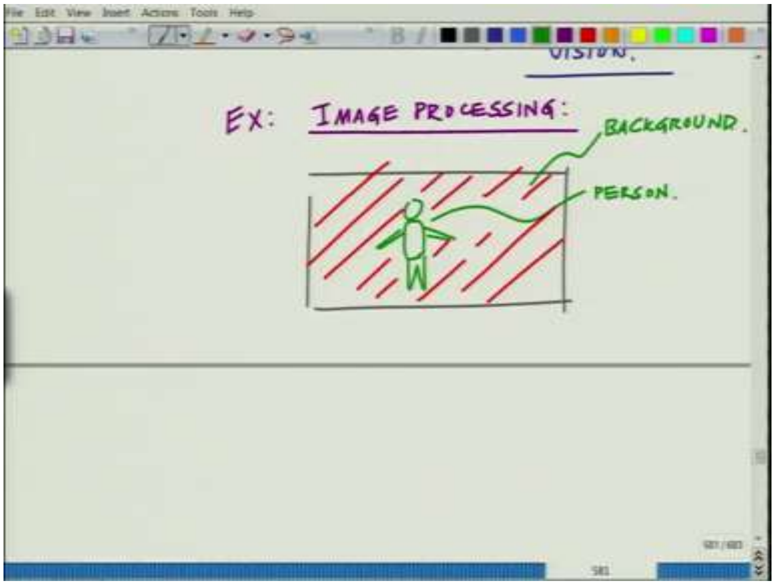
So, the clustering algorithm is, which is very important machine learning performs this task. So, what clustering does is, the clustering algorithm, it organizes unlabeled data. Remember, we said this is unsupervised, so this organizes unlabeled data into groups containing logically similar points, in to groups containing, you can say similar points or you can say objects, and this particular aspect of clustering, this has many applications. It is one of the, it has many applications such as what are the applications of clustering let us look at some.

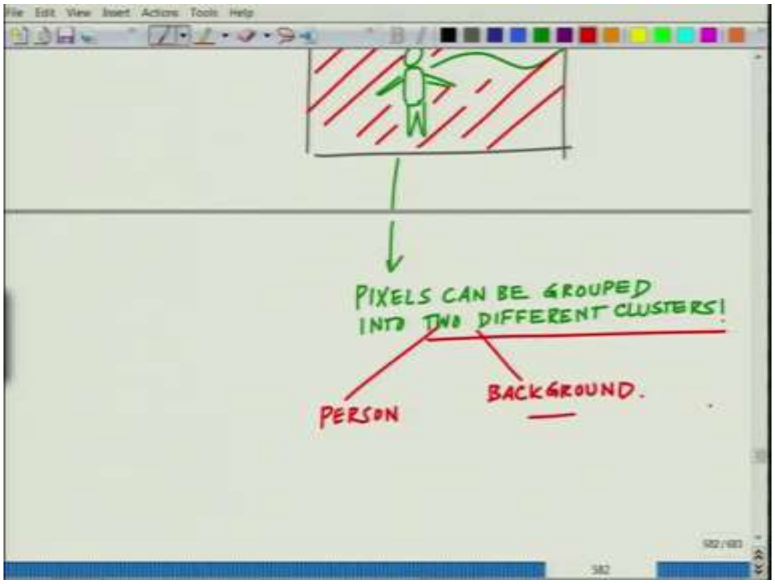
(Refer Slide Time: 07:09)



What are the applications of clustering? For instance this is applications in data analysis, pattern recognition, recognizing patterns. You can also think of, for instance such as computer vision and so on and so forth.

(Refer Slide Time: 08:03)



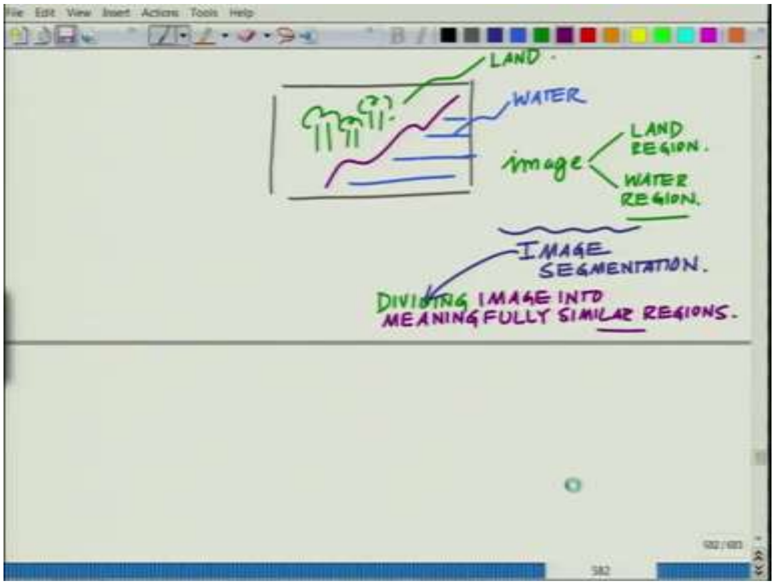


For instance let us take a simple example from this area of computer vision let us say or image processing. Let us take simple example from the area of, so given an image one would have perhaps like a segment that image into two parts. So, let us say you have a certain person or an animal in that image, you would like to segment the pixels of that image into two clusters. One, corresponding to the pixels of the person and the other corresponding to the pixels of the background so that is a classical example.

So, for instance let us say you have an image and you have let us say my, cartoon figure of a person. So, this is your person and the rest is essentially your background. So, you have the person and you have the background and therefore we can. The pixels can be grouped into two different clusters, so the pixels can be grouped into two different clusters and what are these two different clusters? These can be essentially your person and then you have the background. So, these are the two different clusters.

Or for example, you have an image, image containing for instance image of a large, a physical image of a large area containing geographical features such as, for instance large land body, large areas of land and then adjuring water bodies, so you would like to probably cluster the pixels into the pixels corresponding to a land and water.

(Refer Slide Time: 11:17)


File Edit View Insert Actions Tools Help
LAND
WATER
image
LAND REGION.
WATER REGION.
IMAGE SEGMENTATION.
DIVIDING IMAGE INTO MEANINGFULLY SIMILAR REGIONS.

So, you can have an image or for example, an another example you have an image and then let us say you have, over here you have the water and over here let us say you have the land, then you would like to take the image and divide that into clusters or pixels into land and water portion, water region. So, you have the land region and you have the water region. So, essentially what you are doing, this is also termed as by the way image segmentation.

So, you are segmenting the image so this basically is also known as image segmentation. What is the meaning of image segmentation? This is essentially segmenting or dividing the image into meaningful visually similar regions. So, meaningful perceptually similar, some regions that have some similar characteristics for instance in this case land and water. So, you are taking the regions of the image, dividing the image into meaningfully similar regions

(Refer Slide Time: 13:32)

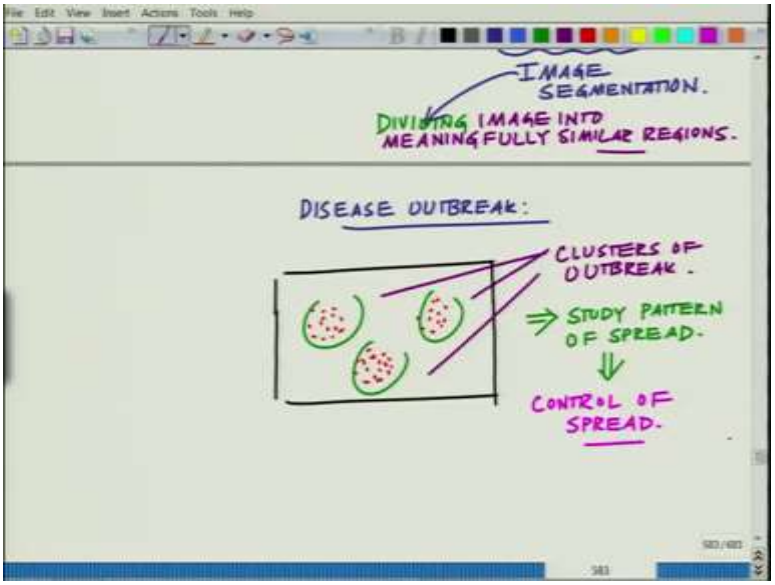


IMAGE SEGMENTATION.

DIVIDING IMAGE INTO MEANINGFULLY SIMILAR REGIONS.

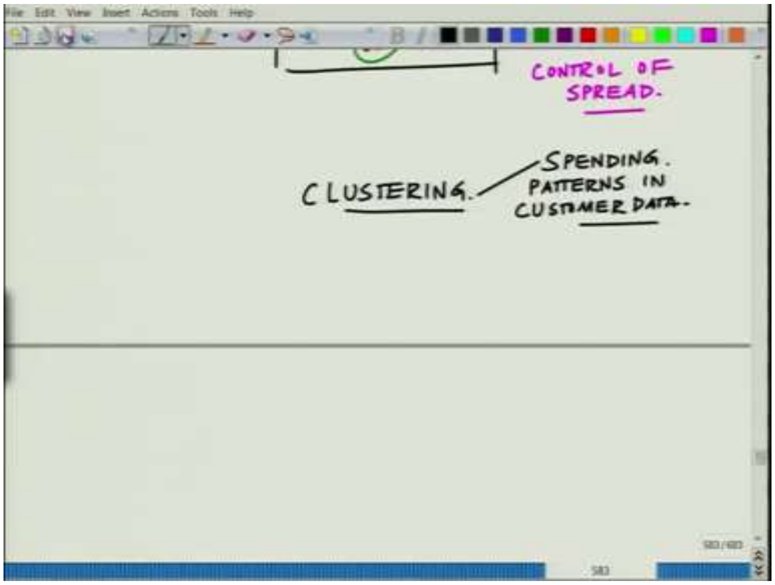
DISEASE OUTBREAK:

CLUSTERS OF OUTBREAK.

⇒ STUDY PATTERN OF SPREAD.

⇓

CONTROL OF SPREAD.



CONTROL OF SPREAD.

CLUSTERING. SPENDING. PATTERNS IN CUSTOMER DATA.

And for instance similarly clustering can used to also, you can have a large city and you have for instance when you studying a disease outbreak, you have clusters and this is of course shows you directly which parts are, which patterns of, where the disease is growing fastest.

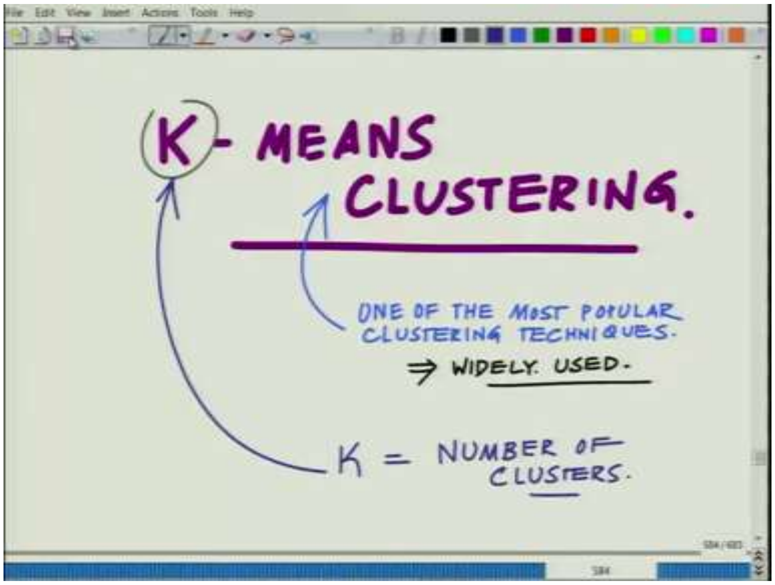
So, you have for instance, so consider the example of an epidemic or a disease outbreak organizing them into clusters of affected, clusters or outbreak that helps us understand how it is evolving, where it is spreading fastest and their implies basically helps us in able to control the spread. So, this implies one can study the pattern of spread plus then followed by, this can be followed naturally by the control of such spread.

And clustering can also be used for instance to study other, for instance in finance and marketing you can have applications, which can, for instance you have a consumer database. This can be organized to study spending patterns in your, it can also be used to study for instance the spending patterns in your customer data so on and so forth.

So, clustering has essentially a very large number of applications and as you will see because of its rather, I would say unsupervised nature because it does not require labeling and labeling can sometimes be a problem, especially if you have a large amount of data and especially because given, because as you going to see, because of the simplistic, rather simplistic nature of the algorithm, it is very, very appealing for implementation in practice.

In fact one of the algorithms that is used very frequently in practice in machine learning. So, let us learn look at one of the simplest algorithms for clustering this is known as K-Means Clustering.

(Refer Slide Time: 16:34)



K - MEANS CLUSTERING.

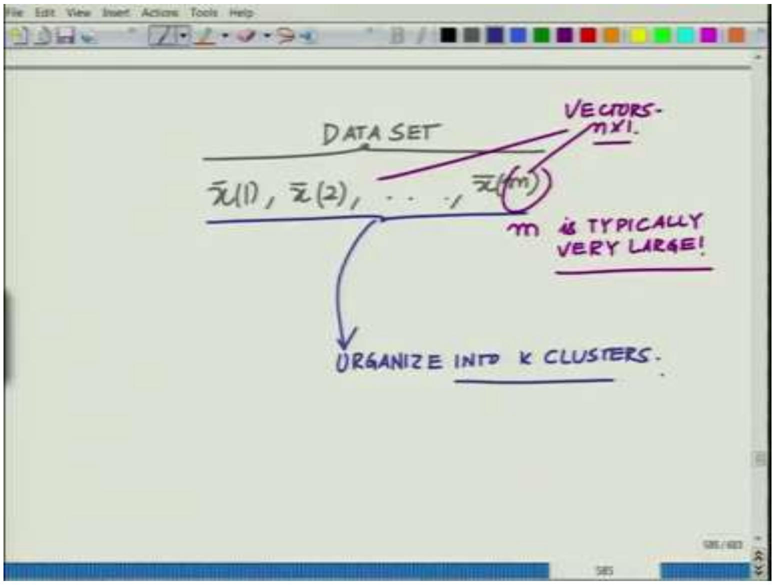
ONE OF THE MOST POPULAR CLUSTERING TECHNIQUES.

⇒ WIDELY USED.

K = NUMBER OF CLUSTERS.

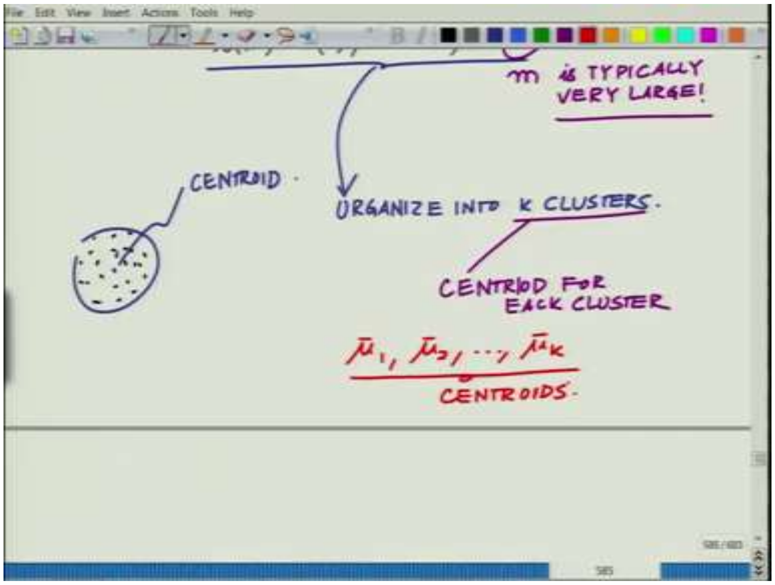
So, the algorithm that we want to look at, this is known as the K-Means Clustering, which is a one of the most popular, this is one of the most popular clustering techniques, and also one of the most, which implies that basically probably it is also one of the most widely embraced and widely used techniques for clustering. The name is simple. The K essentially denotes the number of clusters and these are, so K denotes, K equals to the number of clusters and if K equal to 2 then it will be binary clustering and so on. So, K denotes the number of clusters. Now we have these data points how does this K-Means algorithm proceed?

(Refer Slide Time: 18:21)


File Edit View Insert Actions Tools Help
DATA SET
VECTORS - n×1.
$\bar{x}(1), \bar{x}(2), \ldots, \bar{x}(m)$
m is TYPICALLY VERY LARGE!
ORGANIZE INTO K CLUSTERS.

So, you have these data points x 1 bar, x 2 bar, x m bar. So, these are your data set, these are vectors, these are your data vectors. Let us say these are n dimensional vectors. Now typically m is a very large number and now we wish to organize this, the aim is naturally, organize data into, the aim is to organize this into K clusters.

(Refer Slide Time: 19:39)

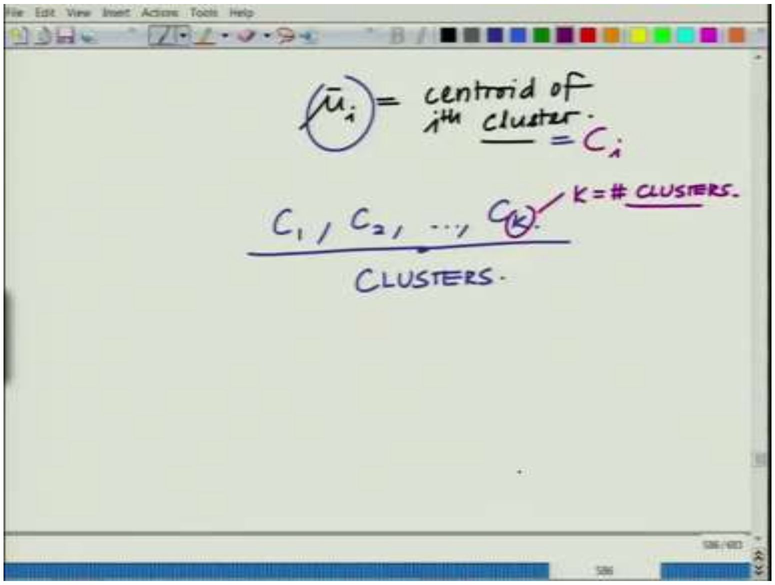

File Edit View Insert Actions Tools Help
m is TYPICALLY VERY LARGE!
CENTROID.
ORGANIZE INTO K CLUSTERS.
CENTROID FOR EACH CLUSTER
$\bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_K$
CENTROIDS.

Let us say these K clusters, we denote these K clusters. We have this notion also; we have a cluster, let us say whenever we have a cluster. One way to characterize this cluster is by what is known as a centroid, something that is a central point, sort of that captures the location of each cluster in an n dimensional space. It might not be simply based on distance; it might be based on some notion of distance. So, this is something that is closest in some sort of metric to these different points belonging to the cluster.

So, this is what we call as each centroid, so this cluster can be characterized by the centroid and each cluster has a centroid, centroid for each cluster. So, we will determine the centroid for each cluster, let us denote these by mu bar 1, mu bar 2, mu bar k these are essentially the centroids or center of gravity or you can simply call them as the centers and so on and so forth.

So, centroid similar to a large object where the centroid is essentially the center of gravity or the center of mass, so that essentially you can think of that as a point where the mass of the object is concentrated. So, the notion is something similar, so that is a point which sort of best describes this cluster.

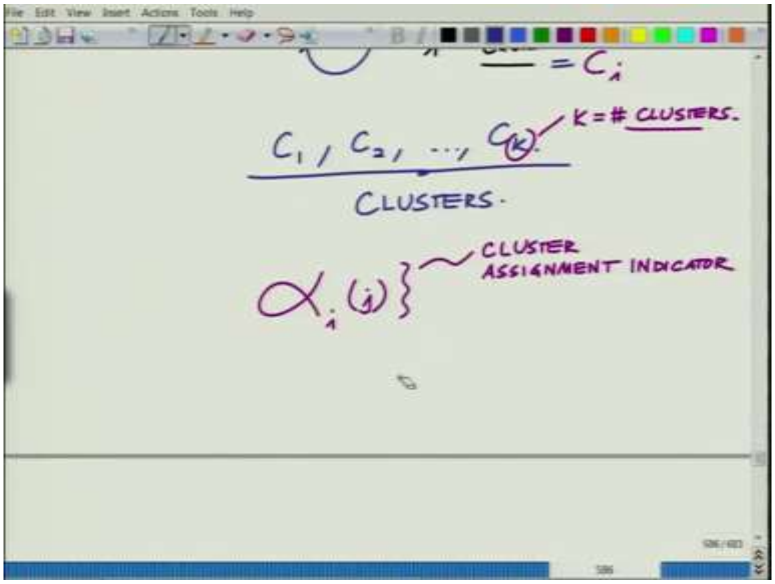
(Refer Slide Time: 21:41)



$\bar{\mu}_i$ = centroid of i^{th} cluster = C_i

$C_1, C_2, \ldots, C_k$ — k = # CLUSTERS.

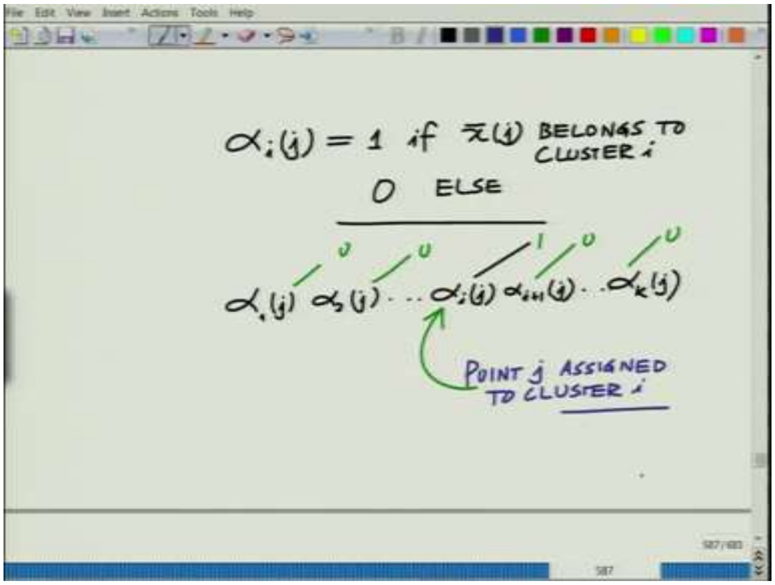
CLUSTERS.


File Edit View Insert Actions Tools Help
$= C_i$
$C_1, C_2, \ldots, C_K$
K = # CLUSTERS.
CLUSTERS.
$\alpha_i(j)$
CLUSTER ASSIGNMENT INDICATOR

So, these are the centroids and naturally this mu i, mu bar i this is essentially centroid of the i th, this is the centroid of i th cluster, and we denote the clusters themselves by C 1, remember we have K cluster. So, you have K centroids and you have K clusters. So, these are the clusters. So, C i bar, so mu i bar this is the centroid of the i th cluster which essentially what we are denoting this by is, we are denoting this by the set C i that is essentially your i th cluster. Then you have the K cluster. So, this K not to forget K equals, once again the K equals, clusters.

Now the key parameter here, this cluster assignment parameter which we will denote by this term, what we call as alpha i of j. This is essentially your cluster assignment indicator or parameter and what this, so this alpha i j, as we have alpha i j equal to 1.

(Refer Slide Time: 23:40)



$\alpha_i(j) = 1$ if $\bar{x}(j)$ BELONGS TO CLUSTER i

0 ELSE

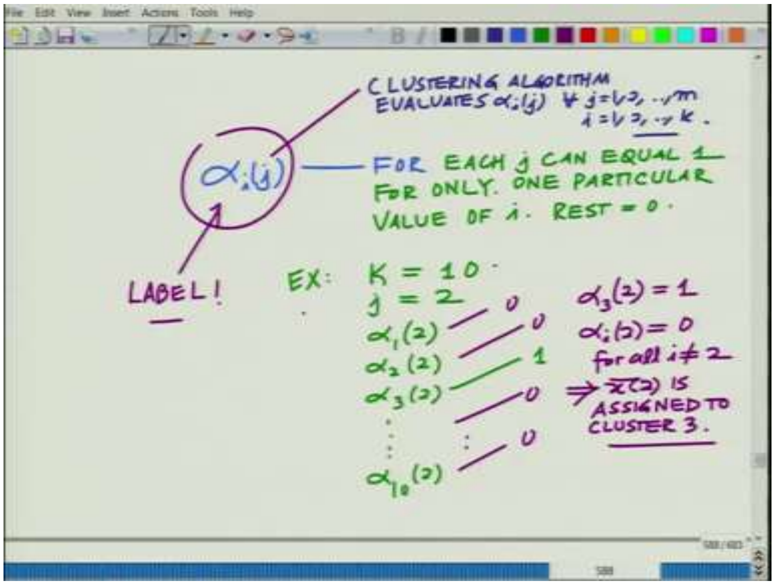
$\overset{0}{\alpha_1(j)}\ \overset{0}{\alpha_2(j)} \cdots \overset{1}{\alpha_i(j)}\ \overset{0}{\alpha_{i+1}(j)} \cdots \overset{0}{\alpha_k(j)}$

POINT j ASSIGNED TO CLUSTER i

What is the role this thing plays? This alpha i j is naturally defined as follows. This alpha i j equal to 1 if x bar j belongs to cluster i equal to 0 else. So, we have alpha 1, so we will basically have alpha 1 j, alpha 2 j so on alpha i j, alpha i plus 1 j and so on and then we will have alpha k j. One for each cluster, for each point j and this point if it belongs to cluster i this will be equal to 1 rest all will be equal to 0.

So, this implies that point j that is assigned to, point j that is x bar j, point j is assigned to cluster i. So, for a given j only one of the alpha i j's can be non 0, because each point, remember the fundamental assumption is each point can be assigned only to the 1 cluster. So, only 1 alpha i j, for each j can be equal to 0 for only one particular value of i.

(Refer Slide Time: 25:30)



CLUSTERING ALGORITHM EVALUATES $\alpha_i(j)$ $\forall$ $j = 1, 2, \ldots, m$
$i = 1, 2, \ldots, k$.

$\alpha_i(j)$

FOR EACH j CAN EQUAL 1 FOR ONLY ONE PARTICULAR VALUE OF i. REST = 0.

LABEL!

EX: $K = 10$.
$j = 2$
$\alpha_1(2)$ — 0
$\alpha_2(2)$ — 0
$\alpha_3(2)$ — 1
— 0
$\vdots$
$\alpha_{10}(2)$ — 0

$\alpha_3(2) = 1$
$\alpha_i(2) = 0$ for all $i \neq 2$
$\Rightarrow$ $\bar{x}(2)$ IS ASSIGNED TO CLUSTER 3.

So, alpha i for each j, so if you look at alpha i j, for each j can equal 1, for only one particular value of i, rest equal to 0. Example let us take for instance, 10 clusters, example let us say k equals 10 and then we look at alpha, let us look at j equal to, let us say 2, the point number 2, then we have alpha 1 of 2, alpha 2 of 2, alpha 3 of 2 so on up to alpha 10 of 2, and let us say point number 2 is assigned to cluster 3 this implies this is equal to 1 rest all must be equal to 0.

So, alpha 3 2, alpha 3 equal to 1, and alpha i of j equal to 0 or alpha i of 2 is equal to 0, for all i not equal to 2. So, this implies x bar 2 is assigned to cluster number, is assigned, this implies x bar 2 is basically assigned to this cluster number 3. So, essentially the clustering algorithm is nothing but computing, so what the clustering algorithm does is essentially computes these alpha i j's remember that is the end of the, that is what you are going to have at the end of this algorithm. You want to have, for each point you want to have a cluster assignment.

So, clustering algorithm, so that is, for all j equal to 1 to up to m and i equal to 1 to up to k. That is what your clustering algorithm is essentially giving you. The clustering algorithm is essentially giving you these labels alpha i's. So, you are now taking the label and you can think of the cluster as nothing but assigning a label, so you are taking unlabeled data assigning the labels. So, that is, so you can think of these also fundamentally as a label.

So, the clustering algorithm what it is doing is essentially it is taking these chaotic, large amount of chaotic data assigned, clustering them assigning these labels and the assignment to a cluster itself is nothing but the assignment of a label. So, that is essentially is basically a high level description of the clustering algorithm. We will specifically look at the clustering algorithm starting with the next module. So, let us stop here and we will continue this discussion subsequently. Thank you very much.