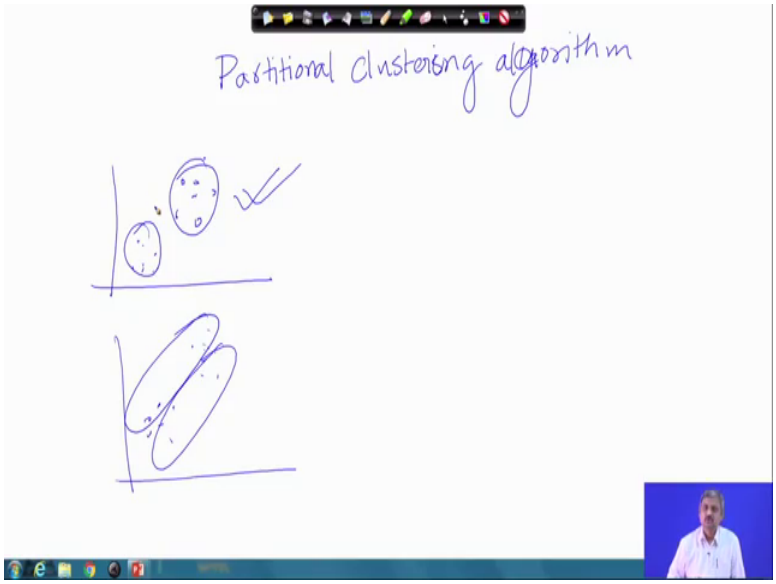
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture - 34
Clustering – III

Welcome to the lecture on the partitional clustering algorithms.

(Refer Slide Time: 00:25)

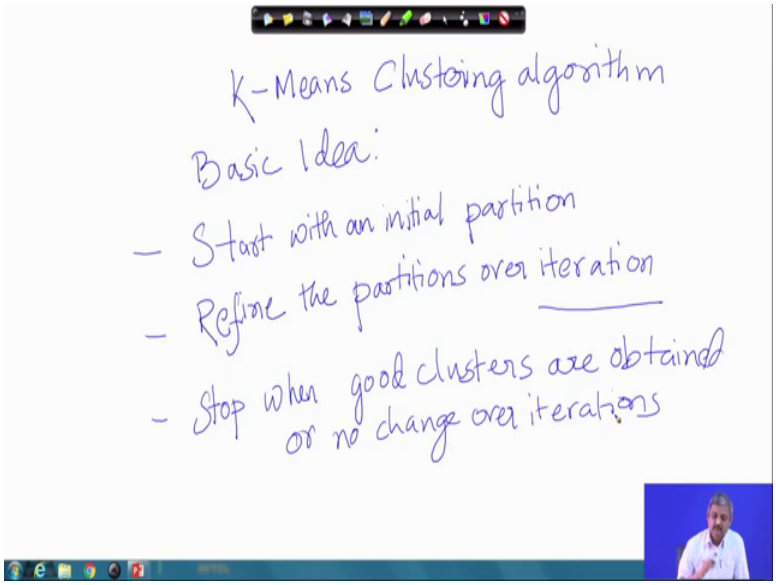


So, as we had told earlier we will what partitional algorithm does used to group points into clusters, clustering algorithms into clusters and then evaluate how good are they and so on.

So as an example let me explain this suppose there are points like this and one possible into two cluster is this another possible is this of course, this is preferred over this because this has a better inter cluster separation and inter cluster homogeneity ok. So, that is what it will try to achieve.

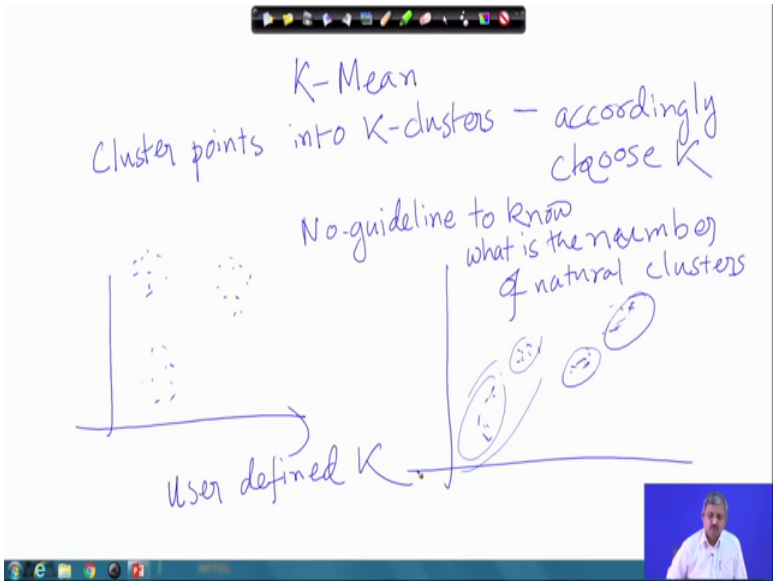
So, there is one greedy algorithm what it does instead of trying out all possible partitions and finding which one is the best it kind of starts with some initial guess of the partition and then refines it.

(Refer Slide Time: 01:42)



This algorithm is called the K- Means clustering algorithm. So, the basic idea is start initial partition; then, refine the partition over iteration. So, this is also a iterative algorithm whereas, the hierarchical one sort singles out this is a iterative algorithm or over iterations this is the idea.

(Refer Slide Time: 03:25)



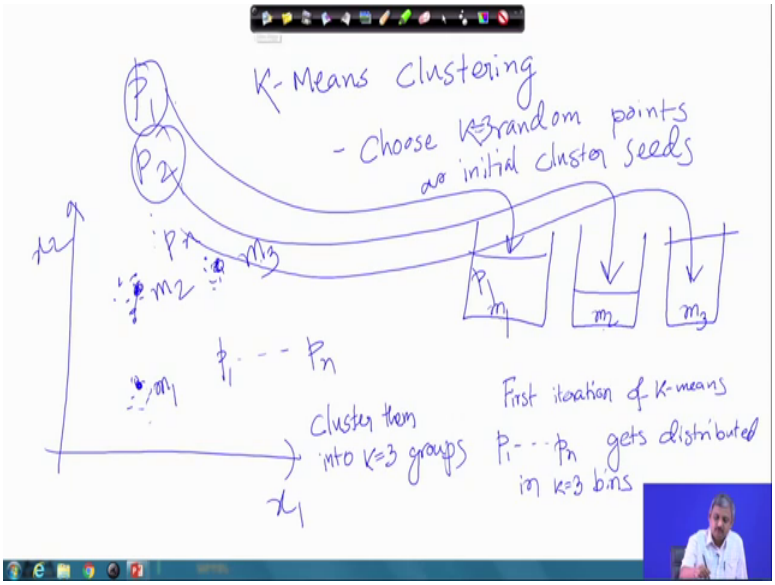
So, you start with some initial partition and then refine it let me tell you. So, the name K means is because it work with mean value the average value of a cluster and there are K such means that determine that. So, if you want to into K groups, K clusters, you have to

accordingly choose value of K. Choose K and note that one very important problem is in clustering is that nobody really knows what the number of clusters are what is the natural number of clusters given the data nobody really knows.

So, maybe I ask you a question suppose I have I have say points like this say. So, what is the number of cluster is it 2 or is it 4 ok. So, that is a open I mean no guideline to know what is the natural I mean what is the number of cluster written naturally there in the data

So, what we do instead is that we depending on the domain we choose a user defined K all right. So, for the K means algorithm first how many clusters you have to know K value you have to know.

(Refer Slide Time: 05:31)



Then I will explain you to do the following. Let me draw the picture again. So, you are given p 1 to p n points. Say cluster them into K equal to say 3 groups, choose K random points as initial cluster seeds. Just out of this n point just pick up K points randomly pick up K points.

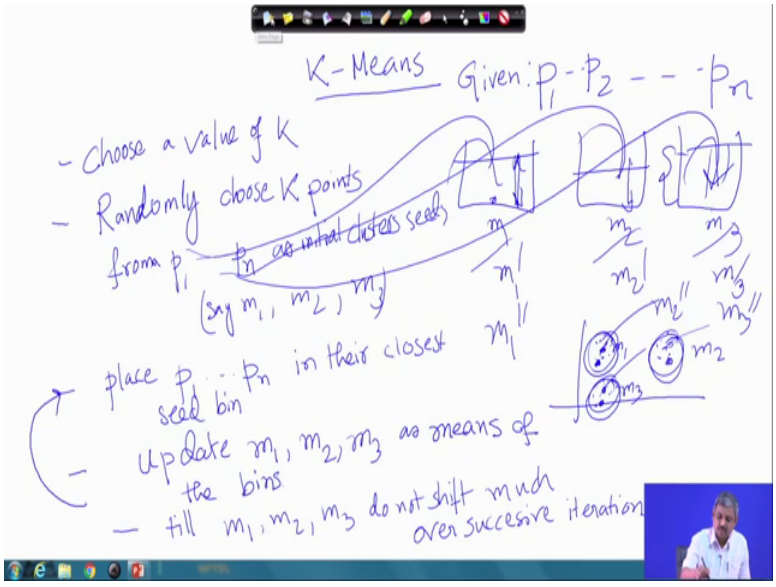
So, suppose this is one point, this is one point, and this is one point this I have actually I am lucky that I have got the 3 points in 3 cluster they may not happen all the 3 points may come from the same cluster if you randomly choose them hm. So, let us be optimistic. So, I have 3 points initially chosen.

Now I do the following I make 3 bins, 3 containers and this initial points that I have randomly chosen let me call them as m 1, m 2 m 3 ok. So, K equal to 3 here place these seeds initial seeds in each of these bins place the seeds in this bin. Now you go through your data points take the first point see which of this seed it is closest to you know the distance between 2 point that is already defined you see which seed it is closest to.

So, for example, if your point is this it is closest to this rather than this and this place it there in that bin your first point goes into this bin. Now look at the second point again see which m 1, m 2, m 3 it is closest to we actually place it in that bin. So, this way go through all the points and place them in the closest mean value closest seed m 1, m 2, m 3.

So, this is the so, what will happen after first iterations your n points will get distributed in these 3 bins. Now we will start the second iteration ok. Let me properly draw it so, that I have some space to write.

(Refer Slide Time: 09:42)



What happened wait a sorry I am just clusters say m 3, are the seeds. So, place the p 1, p 2, p 3 in their bins which every m 1, m 2, m 3 is closest.

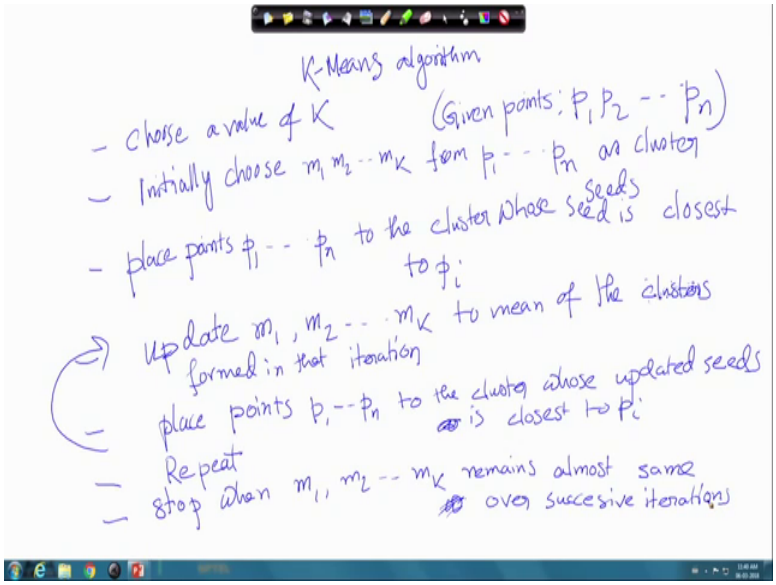
Now update the seeds. So, after first iteration you have distributed the points. Now throw away m 1, m 2, m 3 take the mean data point central point of all the points which in first

bin all the points in second bin, all the points and those are your new m 1 prime, m 2 prime, m 3 prime, the average point of all this.

So, for example, m 3 and say after first iteration this is 1 bin this is 1 bin, this is 1 bin, you throw away m 2, m 1 into m 3 take the centre point of all the bin each of each bin take that as your m 1 prime, m 2 prime, m 3 prime bins ok. Now, repeat this step ok.

Now with this new bin again go through p 1, p 2, p n. Redistribute them again update m 1 double prime, m 2 double prime, m 3 double prime. Keep on doing this till over 2 iteration m 1, m 2, m 3 do not change much. They become fixed, they are converged, shift, successive iterations. Let me properly write down.

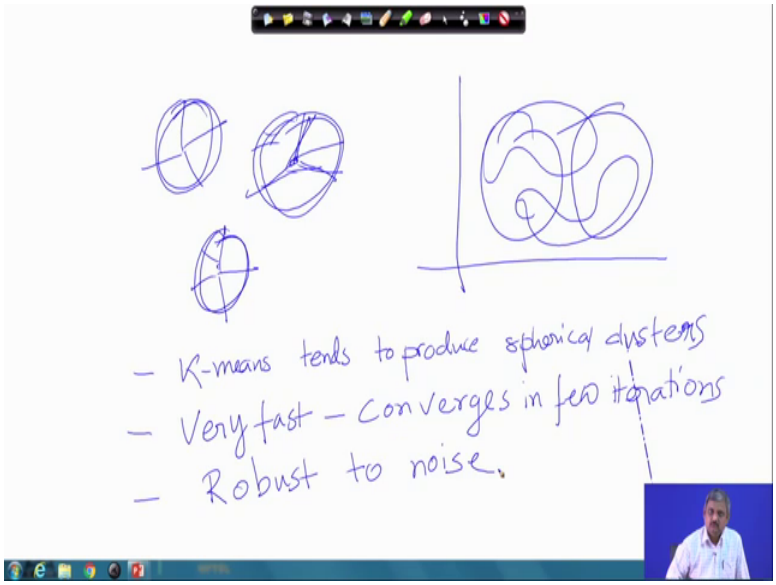
(Refer Slide Time: 14:52)



Choose seeds. Place p n to the cluster whose seed is closest m K intermediate whose seeds are is closest to p i, the point which are placing repeat.

So, repeat this two when m K successive iterations ok. So, this is the K means algorithm so what it does actually is it produces circular clusters, spherical cluster.

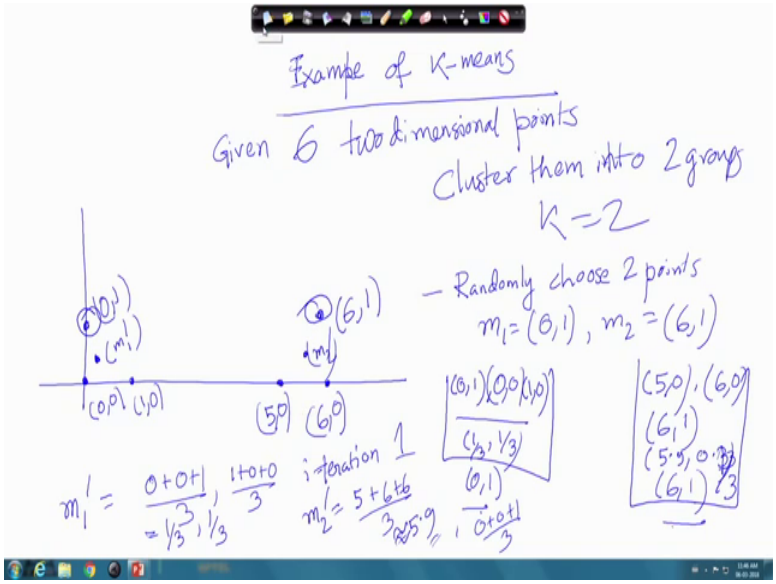
(Refer Slide Time: 18:40)



Because see if you are placing to the closest so it tends to produce. So, if you have natural clusters like this it will actually produce clusters like this. Actually it can be shown that the K means algorithm will converge to a point whose distance from the centre if you add them up that is the minimum ok.

So you take the cluster centre this is the minimum clusters, but it is very fast iterations. And it takes care of noise robust to noise ok. So, very simple algorithm, very simple.

(Refer Slide Time: 20:38)

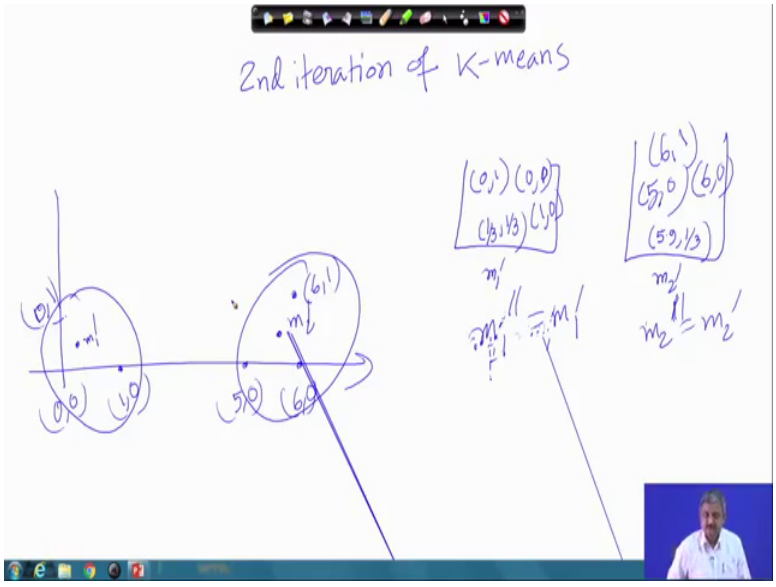


Basically let me illustrate with an example say 6. So, let us say I have this 6 points, I have this 6 points ok. And then what I want to do is to cluster them into two groups alright. So, K is 2 so randomly choose 2 points. Let us see suppose I choose this point and this point. So, m 1 is 0, 1, m 2, 6, 1. This is 1 bin, 6, 1 is another.

So, now let us iterate this point will come here this point will go over 0, 0 this or this which is closest this is closest. So, 0, 0 goes here 1, 0 this is closest or this is closest this is closest so 1, 0 goes here. Now 5, 0 this is closest or this is closest goes here, 6, 0 goes here, goes here, this is 1 ok. So, what is m 1 prime after first iteration average of these 3..

So, it is 0 plus 0 plus 1 by 3, and 1 plus 0, plus 0 by 3. So, it is one third and one third this is m 1, m 1 prime ok. So, somewhere here look the class seed has shifted this one, it is how much 5.9 something like that approximately. And it y axis is 0 plus 0 plus 1 by 3. So, m 2 prime is 3 3 3 here one-third, 5, 5.9 and one-third. So, it is something like here ok.

(Refer Slide Time: 25:16)



So, now, start the second iteration to 0, 1 goes here, 0, 0 goes here, 1, 0 goes here, 6, 1 goes here, 5, 0 goes here, 6 0 goes here. So, what is m 2 prime same it will be same as m 1 prime, sorry m 1 double prime, I made a mistake m 1 double prime is same as m 1 prime. Because these points are same again if you calculate mean same you will get m 2 double prime, is m 2 prime.

So, over two iteration seeds are not shifting now stop. So, your final cluster is this is one cluster this is one cluster so that is how it works alright. So, thank you for today in the next class we will discuss about a density based clustering algorithm known as DB scan.

Thank you.