

Learning Analytics Tools

Ramkumar Rajendran

Educational Technology

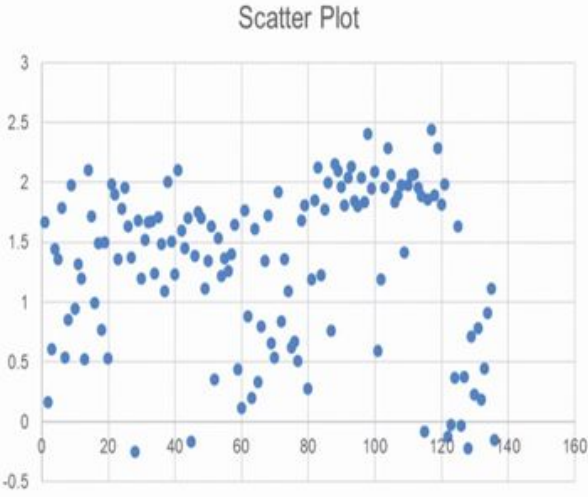
Indian Institute of Technology, Bombay

Lecture 7.2: K-means Clustering

(Refer Slide Time: 0:24)


K-Means Clustering

- Given data of students average rating in a survey.
- Three point Likert survey with 10 questions
- Scatter plot is shown
- How many clusters we can make from this data



Scatter Plot

The scatter plot displays a distribution of data points on a 2D plane. The x-axis ranges from 0 to 160 with major ticks every 20 units. The y-axis ranges from -0.5 to 3 with major ticks every 0.5 units. The data points are represented by blue dots and are scattered across the plot area, with a higher density between x=20 and x=120, and y=0.5 and y=2.5. There are no obvious distinct clusters visible in this scatter plot.



Learning Analytics

2

In this video, we will describe what is K-means clustering. K-means clustering is one of the centroid-based clustering technique. We saw two types of the major category of clustering techniques, one is centroid based or quantity based. K-means is actually from the centroid-based clustering algorithms.

Let us look at this data. It is a response of three-point Likert survey having 10 questions. Ignore the x-axis, the numbers or y-axis. Consider this as a scatter plot you have and this is plotting. So there are say 200 students three-point Likert survey questions you are plotting in a scatter plot. You want to see whether these students' survey has some clusters.


For example, You want to see among the students, Is there a cluster? Is there a pattern? again you group them into something. If we ignore the x-axis, consider this is the scatter plot you have, you can form clusters. So how many clusters can we make from this data? If you want to group them into say n groups, two groups or three groups, four groups, how many clusters can you form on this data?

(Refer Slide Time: 1:50)

K-Means Clustering

- Two Random points as Centroids
- Euclidian Distance of each data from centroid is measured. The points near to the centroid are formed as cluster.
- Create new centroid of cluster

Scatter Plot



Learning Analytics

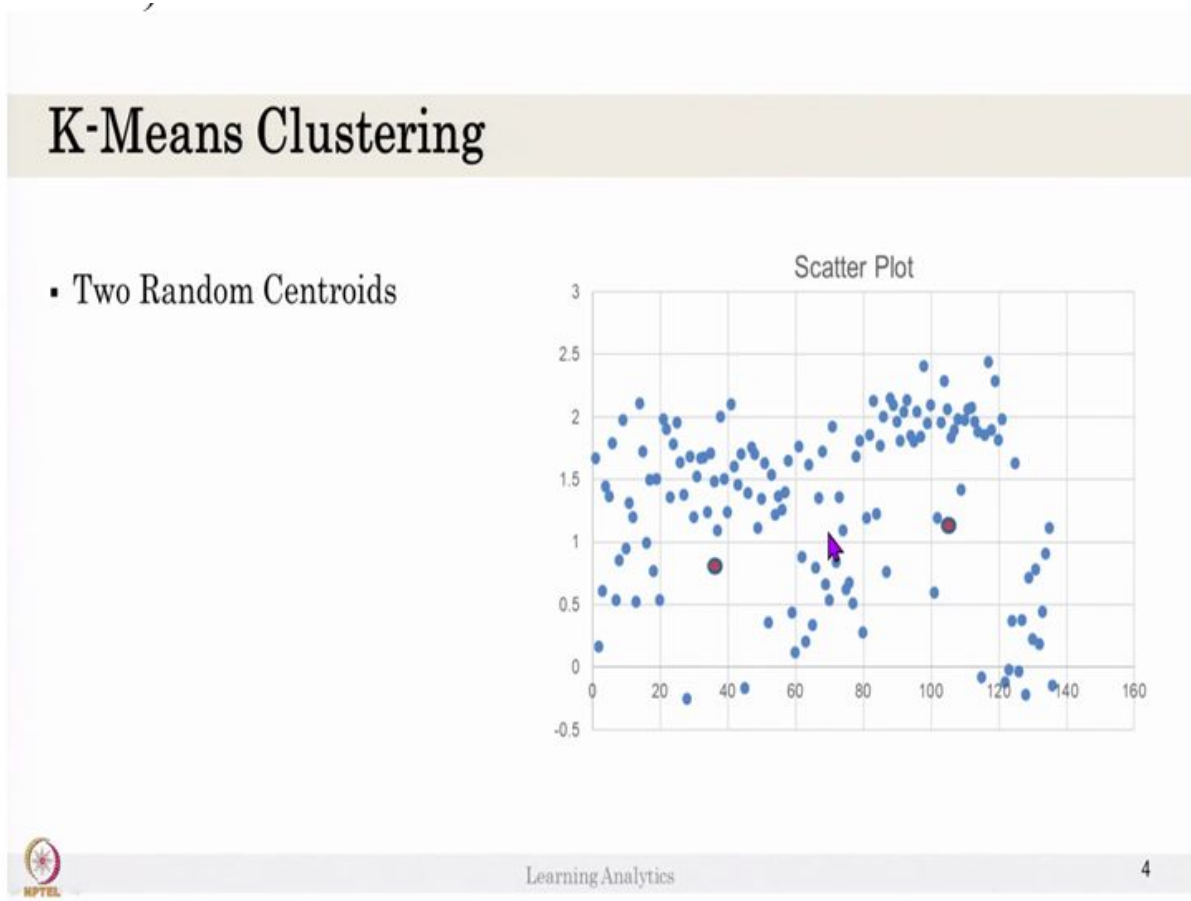
3

So let us say we have two random points, one point here, the second point here, two random points. And so we can create a clustering moving these random points over this figure means I am trying to create two clusters from this plot. So what happens is it tries to identify the nearest point to the centre and all the

point nearest to the centre will be considered as a cluster and all these points near this centre will be considered to another cluster.

After creating two clusters it identifies again the mean of all these values, the mean value will be the centroid value for the new cluster. So then it iterates further, further until it reaches the convergent point, then it creates a two cluster.

(Refer Slide Time: 2:40)

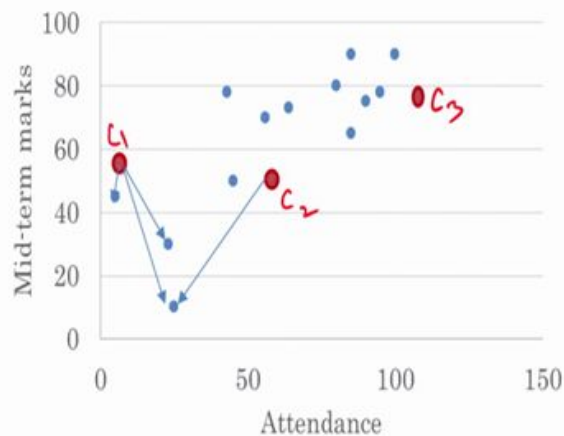


So let us see how these two random points vary. So consider two points have been moved here and this can be one cluster, this can be the second cluster. Let us see what is this distance and how this distance is calculated.

(Refer Slide Time: 2:54)

K-means Clustering

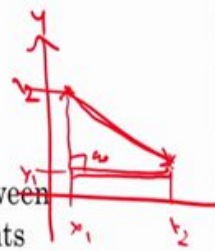
Another Example



Euclidian distance between centroids and data points

$$= (||X_i - C_j||)^2$$

(x, y) C_1



Consider these three, this data, okay, consider this data and we have seen this data in previous classes. We have students' attendance and students' midterm marks, we do not know what is the answer mark or anything. So we have students' attendance and midterm marks, we have the scatter plot. We would like to cluster them because it is unsupervised learning, and we do not know what we are trying to predict.

We want to know what is happening with some students who are getting a low score, is there a pattern that exists between attendance versus midterm marks? We have, we do not know but we have some hypothesis, it might be a pattern with attendance and midterm marks, can we check it?

So let us see these are the students' plots, I randomly selected three points, so I want to create three clusters randomly. I selected three cluster points, so point one, point two, point three. Remember that I have only two variables here if I have more than two variables like three, four, five, we are collecting more variables, you cannot plot it easily to see how many clusters to come up with. So let us consider the example of two variables in this plot and see how this clustering works.

So I randomly selected three random points and these random points over here, so what happens is this random point is checking the distance between all the available data points, this random point- this is a centroid, first centroid. So the k is three, there are three clusters, one, two, three. So I put three clusters in this data, so each random point is called cluster centroid, this cluster centroid measures the Euclidean distance between all the data points in the figure.

So when it measures the Euclidean distance between this point and all other points. This is done for other two centroids. From this centroid Euclidean distance is large, but from this centroid Euclidean distance is less, so what happens? This point will be considered as a part of this cluster, not these clusters.

Similarly, this particular data point is very near to this centroid compared to this centroid. So this will be formed as one cluster, for example, this Euclidean distance of this is less compared to this, so this will be another cluster. So how to identify the Euclidean distance? So what is Euclidean distance? Euclidean distance is the point's distance between centroids and data points computed by this formula, so c is the centroid, so this is c1, consider this as c1 and this is c2 and this as c3 and x maybe x_1 , x_2 , x_3 and all the points, one, two, three, four, five, six, seven and other points.

So, x will be from x_1 to x_n , it is the number of data points you have. So you have to compute Euclidean distance between each x_1 to x_n to c1, c2, c3 so this Euclidean distance is computed. So what is Euclidean distance? In a plot, if you want to compute Euclidean distance, this is actually the vector distance between these two. If you know how to find vector distance, it's simple, it is like,

In the Euclidean plane, let point p have Cartesian coordinates (p_1, p_2) and let point q have coordinates (q_1, q_2) . Then the distance between p and q is given by:

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}.$$

So this is squared Euclidean distance.

Let us see how this compute the Euclidean distance between two points, so there is a point here, there is another point here. If you want to find the Euclidean distance, we have to use the Euclidean formula,. So simple Euclidean distance actually considers this as 90 degrees. So we know the distance between these two points along the x-axis and also we know the values of x_1 and x_2 , y_1 and y_2 , this is a right-angle triangle, now we know how to compute the value of this if you know the value of these two because

$|x_1 - x_2|$, is this,

$|y_1 - y_2|$, is this value.

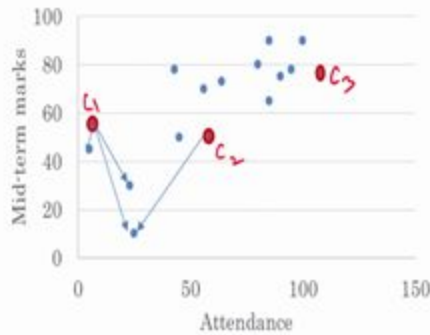
So you know how to compute it, so this is how you compute the Euclidean distance, so that is it. It is simple to do, so yeah, check it out. If you have not come across the system called Euclidean distance, it's very simple to compute and it's logical, intuitive also to compute distance.

There are other distances to compute not just Euclidean distance in clustering, Manhattan distance or some other distances but let us use Euclidean distance for this clustering technique. So what happens is the centroid computes the Euclidean distance of all the points, then whichever point which is near to the cluster it will be considered to be part of that cluster. So let us see the next point.

(Refer Slide Time: 7:49)

K-means Clustering

Another Example



Euclidian distance between centroids and data points

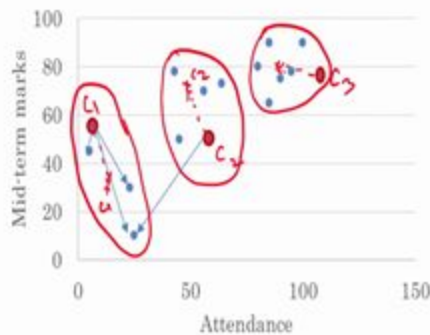
$$= (||X_i - C_j||)^2$$

$\{ \dots \} C_1$



K-means Clustering

Another Example



Euclidian distance between centroids and data points

$$= (||X_i - C_j||)^2$$

$\{ \dots \} C_1$



So after it picked three points what happens is let us look back. So consider it, these three points are considered to be one cluster and this is part of another cluster and these are all near to this cluster. So this

particular centroid, these three points are near to this centroid, these points are near to this centroid and these points are near to this centroid. So based on that value the three clusters are formed.

After you do that, the new centroid will be computed by finding the average value of these four points. So there will be the average value of these four points. Using the four points average the new centroid will be computed, new centroid might come here, the c2 will move here.

Similarly, these three points if you compute the average of these variables, it might come here, c1 might come here. Similarly, this if you compute for these points and the new centroid might be somewhere here, so this will be moved here. So the new centroid of this cluster will be c1 here and c2 here and c3 here, so that is how the three clusters are formed. So the first step is you find/compute the Euclidean distance of centroid to all the data points. After creating, after finding the data points which is near to the centroid you compute the new centroids based on the cluster's values, that is the mean value.

So now the centroids are moved here. After this again compute the Euclidean distance of centroids to all the data points. Obviously, these three points, these three points are very close to this, so this is one cluster. And these points are very close to this cluster and maybe, so this is close to this cluster, so there are three clusters formed as of now.

Again, if you compute the centroid if you try to compute the mean of these three points and move, it is already the mean computed. So there is nowhere to move, it is already in the center place. So if the centroid is not moving, then you stop iterating, that is where you have to stop. Saying that hey, there is no change in centroid movement or there is no change in the one data point, one cluster to another cluster, then you consider stopping.

If that does not happen, so you set a small threshold value. If the centroid is moving with this limited threshold, it is okay, keep it out because some places we do not get the exact clusters from the K-means clustering algorithms. So you can have a threshold value, it is okay, one or two points can move and I want to stop it or you can say I will not iterate to over 100 iterations, I do not want to do beyond 100 iterations, that is fine. So you can stop based on number of iterations or with some threshold or no change in movement of centroids.

So this is cluster one, cluster two, cluster three. Now I hope you know what is K-means clustering, it is actually trying to find the means of the data points from the three centroids or K centroids.

(Refer Slide Time: 11:08)

Activity

K-means Clustering

- List down steps in the clustering operation seen in previous slide.



Activity



- K-Means algorithm
 - Select the value for K
 - Randomly assign k points
 - Compute Euclidian distance between K point to other points
 - Compute the Centroid, → New K point
- Repeat again until the errors values is within the threshold or no change in centroids



Why cannot you go ahead and list down the steps we saw in the clustering operation in previous slides? List down the steps like, what is the first step? In your own words, it does not matter what you write. So

in your own words, the first point is, randomly assign k clusters, one or two or three, k clusters, then you do the Euclidean distance and do move the centroid, so write down how do you move the centroids, write down those steps.

After writing it down resume the video to continue. I hope you would have written down the K-means algorithm, this is the algorithm of K-means. The first step is to select the value of K , so the K can be two, three or anything, why we have to pick two or three, which number is good? We will talk about that and after selecting the value of K randomly assign the K points. So the centroid, the initial centroid points are randomly assigned, that is the key. Remember this.

I select K value equal to two or three, then I randomly assign this K points (centroids) may be in one corner or it can be randomly split in the chart. After that compute Euclidean distance between the K points to other data points, compute all, then create a cluster by grouping points which are near to the centroid, i.e. all the points which are near to the centres are grouped to create/form a cluster.

After you create a cluster, compute the centroid (the centre value) that is the mean value of all the data points in the cluster, that will be the new K point. That centroid is moving to the average or the centre of, all data points in the cluster, that is a new K point. So you have to compute the centroid again of the data points of the newly formed clusters and that centroid will act as a new centroid, new K point.

So now the number of cluster is still same, it is not reducing, we are just moving the centroid here and there, repeat again, repeat this process again until you see the errors values is very less, there is no change or you have done certain iterations (100 iterations or 50 iterations). So, repeat the points like the computing the Euclidean distance and compute the centroid again and again till there is no change in the centroids or the number of iteration is reached.

This is the K-means algorithm, this is the basics of the K-means algorithm. If you have written down this, that is right, you already understood the K-means algorithm and if not, please go and watch the video again or go to the Internet and check the resources. There are some good simulators available to show how these K points are moved, how they are computed, so check those videos.

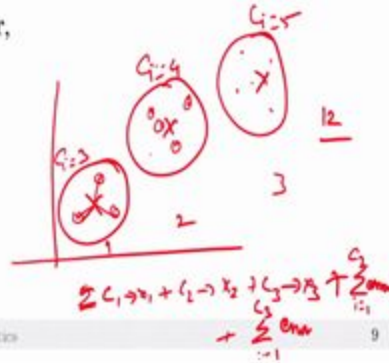
(Refer Slide Time: 14:10)

Selecting Correct K

- Compute the squared error function using below formula

$$J = \sum_{j=1}^k \sum_{i=1}^{C_i} \text{Distance between centroid and data points}$$

- C_i – number of data points in i^{th} cluster,
- Number of clusters $\cdot K$



Then, now we have to see how to select the correct number of K. I said that, can we select two, three, four or what is the number you want to select? So in order to select the correct number of K first, you have to identify the error function or the objective, the objective is to keep the value of error function minimal, and what is this error function? It is actually the distance between the centroid and the data points.

After you complete the iterations and there are three clusters. There are one, two, three centroids. In each centroid, so let us see the three centroids. Within the three centroids there will be a lot of data points. In this cluster, there might be like three data points.

So compute the distance between centroid and data point. So compute the distance between centroid, so summation of all this distance,

$$= \text{dist}(c_1, x_1) + \text{dist}(c_1, x_2) + \text{dist}(c_1, x_3)$$

The distance between centroid and x_1 , x_2 , and x_3 . So there are only three points in cluster one.

C_i = no of data points in cluster i ;

c_i = centroid in i^{th} cluster;

So,

$$C_2 = 4$$

$$C_3 = 5$$

So for each cluster you identify the distance between the centroid and all the points in that cluster i.e.

$$\sum_{j=1}^{j=C_i} dist(c_i, x_j)$$

This value is obtained for each cluster and then added to get error function.

If you have this error computed then our objective is to reduce this error as much as possible. So you can create more clusters. If you create more clusters, the distance might even reduce further. So just a quick question, consider I have twelve data points, what if I choose K equal to twelve, what will happen? If I choose K equal to twelve, which means the centre of the cluster is actually the data point and the distance between the data point and the cluster is zero.

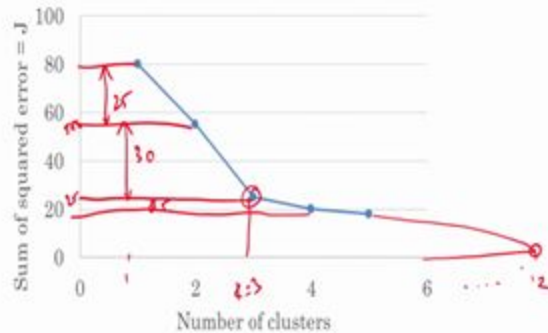
If you add the error of all these values, it will be zero. So if you choose K equal to the number of data samples, the value will be zero. If I choose only one cluster, only one cluster then this complete, this complete will be one centroid, that value will be the maximum error function. We do not know what is the value, it changes based on the data points but that will be the maximum error function. If we choose K equal to the number of data points, it will be zero.

So now, what we have to do? You have to compute this J objective function and for clusters K equal to one and K equal to two, K equal to three, you have to compute for different cases, let us see.

(Refer Slide Time: 18:23)

Selecting Correct K

- Plot J value for different values of K



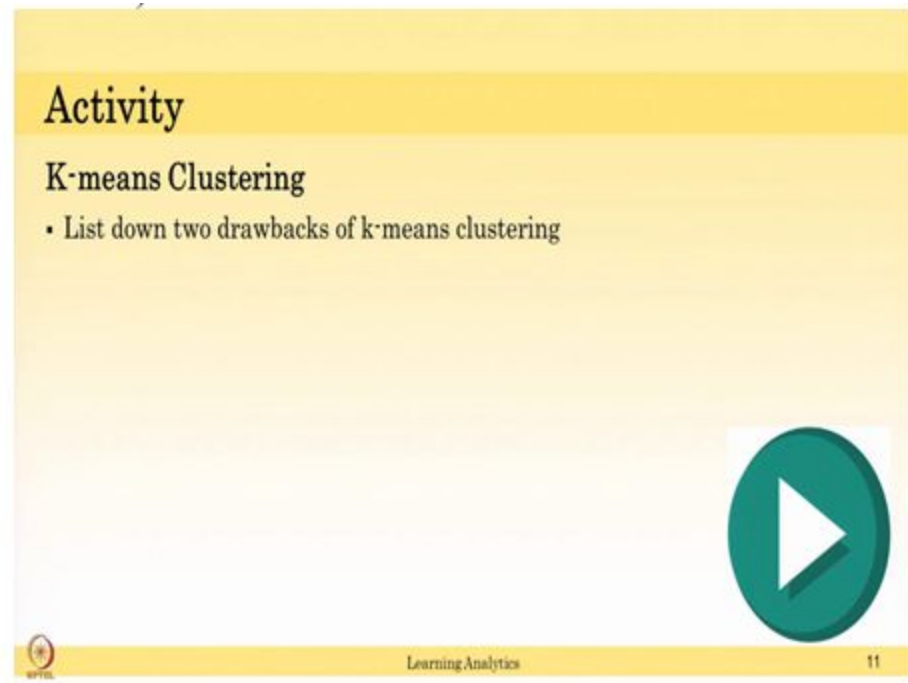
If I plot it for cluster one to clusters three, if I computed that previous, what I was trying was the sum of squared errors is objective function J. If I compute it for K equal to one cluster, this is the maximum value and if you go there like twelve points, if it is twelve here, it might be zero, the point might sit here.

So now if I compute it, I can plot it like that, so how many clusters we should consider, what is the K value, which K value I should pick, that is based on this curve. If you plot this curve, you can say this is elbow curve, so there is a bend here, there is a kind of elbow, just like your elbow. So this is considered to be the optimum K value. So for this particular data you choose K equal to 3, that is the best optimal value, so you can choose K equal to 3, three clusters.

And why we choose that, I will tell you the reason, for example, for K equal to 1 to 2, the difference in the sum of squared errors is reduced to say 55 from 80, say 25 points, here say by 25 points, but this difference it is just a 5.

So the sum of square error reduction is really less from 3 to 4, so that is why we are saying that we will pick the point 3. So the sum of squared error will be really less going forward, so do not pick the cluster equal to 12, which means every data points is the cluster that makes no sense and so pick the cluster which is elbow point, so K equal to 3. It works with this value(elbow point here and check it out). You create your own data and try clustering algorithm and see if this elbow point works or not.

(Refer Slide Time: 20:44)



The slide has a yellow header with the word "Activity" in black. Below it, the title "K-means Clustering" is displayed. A bullet point lists the task: "List down two drawbacks of k-means clustering". A large green play button icon is positioned on the right side of the slide. The footer contains a small logo on the left, the text "Learning Analytics" in the center, and the number "11" on the right.

Activity

K-means Clustering

- List down two drawbacks of k-means clustering

Learning Analytics 11

So now I hope you know what is K-means clustering and you know how to pick the right K value and if you understood, can you list down two drawbacks of K-means clustering? Based on your understanding, what are the drawbacks? Can you list down two of them? After listing it down rush to continue.

(Refer Slide Time: 21:12)

Activity

- Initial centroids
- Categorical data
- Non-linear datasets



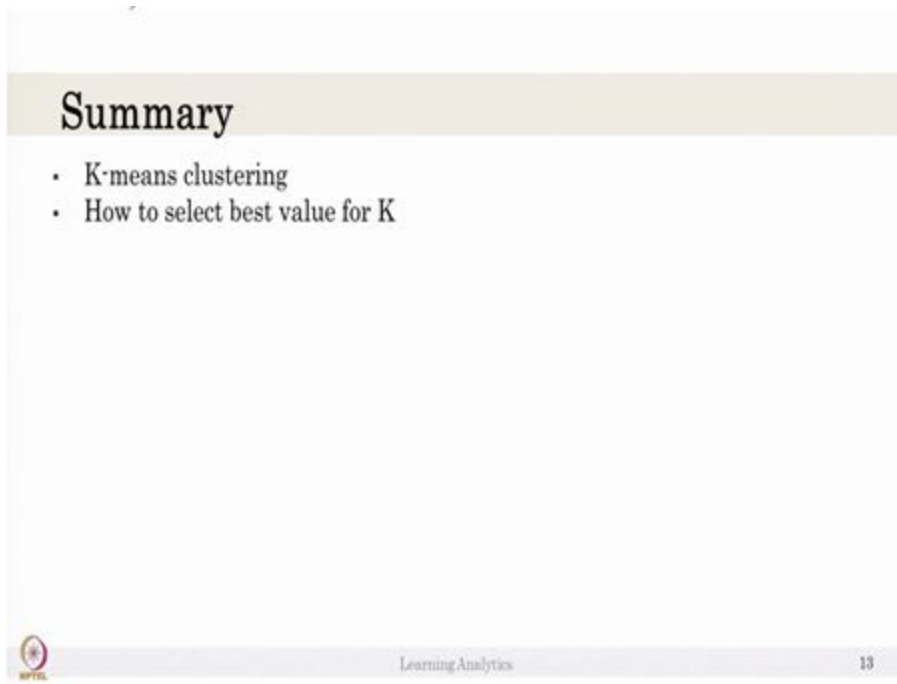
So, initial centroids, I said you pick K equal to a random value, K equal to 2 or 3 and centroid (initial centroids) is assigned randomly, that is a tricky part because where we assign the centroid makes the different type of clusters. In a two-variable it is easy to see but in a three-variable or four variable it is really tough and it would not work for categorical data, also for non-linear datasets.

So to handle that there are techniques, so what is initial? So let us see there are points like this, if my initial centroids are here. So based on how you choose the initial centroid matters a lot. To avoid this problem you can randomly select the initial data points and do it multiple times.

So what I mean is consider for K equal to 2, for K equal to 2, the two random points might be somewhere here when you assign it. So create the K equal to 2 multiple times, check if your K equal to 2 if you do the same clustering same data say multiple times, three times, four times or five times, check the number of clusters or the J value is same then you stop it. Simply for K equal to 3 iterate for ten times, K equal to five iterate for ten times, in each time you have to iterate multiple times to find right clusters, right centroids. What I am saying is run the K -means clustering for K equal to n , multiple times.

So take K equal to 2 and run the K -means clustering for say 10 times, similarly K equal to 3 for 10 times and pick the minimum value of J from these iterations or pick the mean value, something like that, then you plot the elbow curve or the bend, then you pick the right K . This is the best option we have, to avoid the initial centroids issue in the K -means clustering.

(Refer Slide Time: 23:45)



Summary

- K-means clustering
- How to select best value for K

NPTEL Learning Analytics 13

So in this video, we saw what is K-means clustering and we discussed how to select the right number of K using the elbow curve. So hope you understood K-means clustering, if you do not get what is K-means clustering from this video, I recommend you to go and check internet videos. There are a lot of simulations to explain K-means clustering, I am not talking about the mathematics behind the centroid computation and everything because that is not needed for this course.

The idea for this course is, if you have a data and if you can apply the K-means algorithm, you have to understand which K to pick, what is K-means algorithm and what this clustering means. If you understand what is K-means algorithm, how to pick the number of K, perfect, that is enough for this course but if you want to know more about K-means algorithm, please refer the data in the website. Thank you.