NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

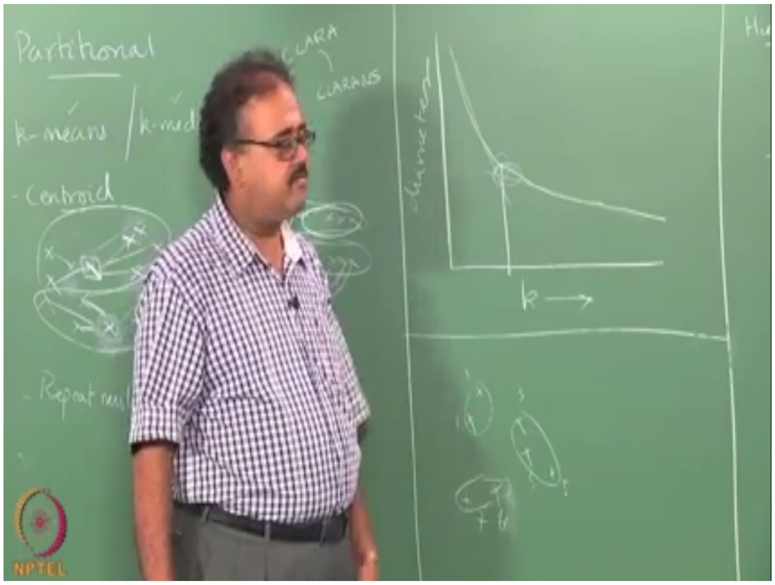
Lecture-71
Hierarchical Clustering

Prof. Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

Hierarchical Clustering

Go on to Hierarchical Clustering.

(Refer Slide Time: 00:17)



So in hierarchical clustering what I will do is I will start off with there are many ways in which to do this right, so one way to do this is to say that I will start off with each data point being a cluster off its own right. Each data point is a cluster of its own and then what do I do I try to merge them into larger clusters right each data, oh! this is not a special distribution of data points okay just put these things here that is I mean individual clusters right data points are individual clusters then what I do is I compare distances between the data points and say maybe merge these merge these merge these merge these right.

So when I draw a line like this it means these two have been merged then likewise these two are merged and then the third data point was merged with that and then these two got merged right. So how could this things look like? Yeah that is what, so I am not told doing what the data points are right so we will call them I should have one two three four five six seven one two three four five six right.

So I initially start looking at this way okay one and two are close together let us merge them right three and four are close together let us merge them five and six or six and seven are close together let us merge them after I have done these things okay next thing I look at okay five is close to three and four okay, let me merge them and so on so forth right and next what we would want to merge?

Which looks closer? That should bring you to the question how do you measure the distance between clusters you know how to measure the distance between data points and how do you measure distances between clusters even how do you know that five is close to three and four you could do variety of things right there are many different measures that you could use right.

So in this case so one thing is the right, so centroid based sometimes how that another thing that people use is called single link distance you know what a single link distance is? So I look at two clusters I look at the pair wise distance of taking one point from this cluster another point from that cluster right. So I look at all possible pairs right I look at the closest to such pair right look at the closest such pair and then I use that as my distance between the two clusters what do you mean by not equal every possible pair.

So there are five here to here I do ten pairs okay, why not it you could, so what do you think that is called so that is another decision we should call average link clustering okay so the single link clustering single link clustering essentially takes the closest data points right so if we take the closest data points here so which is closer so I already merge this right so this is one cluster of now that is another cluster this is in the cluster so which is closer these two are closer at least or closer the question now boils down to is one closer to three then four to six.

So here is not very clear to me but I will take your word for it okay so now that is basically done so all the data points are merged at this point right so I did four and six because there so that a single link right so more I mean single link is by far the most popular hierarchical clustering distance measure right and then there is another one called so what do you think complete link would mean summation farthest max.

So I look at pair wise distances and the max of these pair wise distances is assigned as the distance between the two clusters so in single link clustering so distance between these two clusters was given by the distance between one and three and the distance between these two clusters was given by the distance between four and six that is single link clustering right incomplete link clustering the distance between these two will be given by two and five or one and five where okay.

Whatever two and five and the distance between these two will be given by seven and five right, so what you think is larger? Oh my god! It is hard to make it how to make two and five smaller is it and make it easier for us there you go, yeah so two and five is moving two and five smaller than 7 and 5 right so if I am doing complete link clustering then I would have merged these first right and then I would have merged right.

So I could do a centroid based distance equal to a single lane complete link I could do average link anything else you can think of yeah, so I could Lu radius based. So what do I do in that is that? So I will take the two clusters I want to find the distance between these two clusters I essentially merge the two clusters and find their radius if I want to find the distance between these two clusters I will merge these two clusters and find their radius.

So the smaller the radius the closer the two clusters are so no centroid I am looking at the distance of the centroids of the two rights here, I am merging the two and then finding the centroid of the merged cluster. So it is different right I like and similarly I can do that you know diameter I can merge the two clusters take the diameter, so the smaller of the two is the better so these are essentially more useful for comparison purposes right I want to know whether cluster one is closer to cluster two or two cluster three then I can merge the two find the diameter and then make a decision right it is not really a true distance measure in the sense that I cannot work by say what is a distance of cluster one from cluster 2 okay diameter does not make sense right.

But then if you want to say it is Cluster one closer to cluster two then to cluster three then I can use the merge diameter and I can make those decisions okay, so all of these are valid ways of doing hierarchical clustering yeah, you could I mean define whatever you want right so these are popular ones yeah and yeah they do use other distance measures as well for doing hierarchical clustering okay.

So well now it is a nice thing about hierarchical clustering once I choose this distance measure right, so think one minute stop and think right so these are meta distance measures right I still need to decide on point-wise distance measure right. So that could be an equilibrium measure so when I say single-link I said that distance between the two closest points right but what is that distance? That could be Euclidian that could be Jaccard similarity that could be whatever you want absolute deviation whatever then you can look at whatever distance measures you want.

So that I still have another distance measure, so that is dependent on the data type that distance measure typically depends on the data type well this distance measure typically depends on the kind of clustering that you are looking at right. So once I have a tree like this right what you did not have to choose here? K I did not have to choose K here, so once I have a tree like this how do I recover clusters if you think about it when you completed my clustering all I was left with was one cluster right?

How do I recover the cluster? Traverse the tree, how would you traverse the tree? Now when they start here at a single cluster if I come here I have individual data points yeah, so I basically have to figure out some point okay I will break here if I break here how many clusters do I end up with three clusters if I break here how many clusters I end up with two okay I could choose to break at some point in between and then say that I will take that many clusters I get right.

How do you choose which point to break it at? I can do the need method again right so I can do a K versus evaluation and I can get this thing, so what is the advantage of doing hierarchical clustering is I get that entire graph generated to me in one go right. So I am not actually having to rerun everything for different choices of K I get the entire K graph that nee graphs are generated for me in one shot you do not, you know well yeah you do not so as many as you can get see the point is the goals that you do not get or the ones where it was very hard for you to find a breaking point right.

So essentially if you choose a threshold at which you are merging their points right, so all these things get merged there is no real reason for you to choose three over seven or something if you don't get anything for five six in between so after three you move to seven, so maybe we did not make any difference to look at four five six usually that is what we will end up great. So depending on what kind of a measure you choose you end up with different kinds of clusters for example if we choose single link?

So what is single link say? That closed the distance between two clusters is the closest data points right so I could have a cluster that like that another cluster that is like that and another cluster that is like this okay which are the okay so we tell the two you should merge now where well give me names leave me numbers one two three. So for you single link clustering I will merge one and two right if you single link I will merge one and two so I will basically get this humongous very long cluster right.

So that is the problem with single link clustering you might end up with very long clusters essentially the points at one end of the cluster to the other end might not be very related at all that is the problem with single link clustering yeah right on the other hand if I had used complete link clustering I would have merged two and three first right and then I would have merged it with one but if I use complete link clustering is highly unlikely that I would actually produce say it elongated clusters as one and two in the first place right, single link clustering tends to produce very tight small clusters right.

And at some point you then merge a lot of clusters but then you will merge them at very high levels in the tree right it is the lower levels in the tree you'll be getting smaller clusters okay ah where is the tree here that thing that figure I drew there is a tree okay but it is not called a tree in the hierarchical cluster in literature it is called something else Dendrogram, you know what is the dendrogram means?

Say tree yeah dendrogram means tree just said they went to a different language and pulled out the one loaded three okay, so dendrogram is a tree right. So that is a dendrogram right and what are these levels I am talking about that yeah no not really iterations they are the levels at distances at which I merged right.

So when I merged one done to right, so I had some threshold I start off with one I say anything that is within point one distance unit of it I will merge into a cluster there's nothing right it stays as one and then I say okay within point two point three point four now at point four great, so now

I have two so at the level of point four I have merged one and two and also turns out that level of point four emerged seven and six and three and four right.

That is why well this all of this should be at the same level but you can think of this being slightly different because three and forests are slightly farther apart then one and two all right, so this these levels are essentially the distances at which you are merging there right. So that is why five gets merged with three and four at the higher level because if I slightly farther away from four then three is right so that is the reason for the levels and then these two get merged at a higher level because we are farther away, so the levels in the tree at which you merge or essentially the distances rate at which you are doing the merging okay.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in