NPTEL

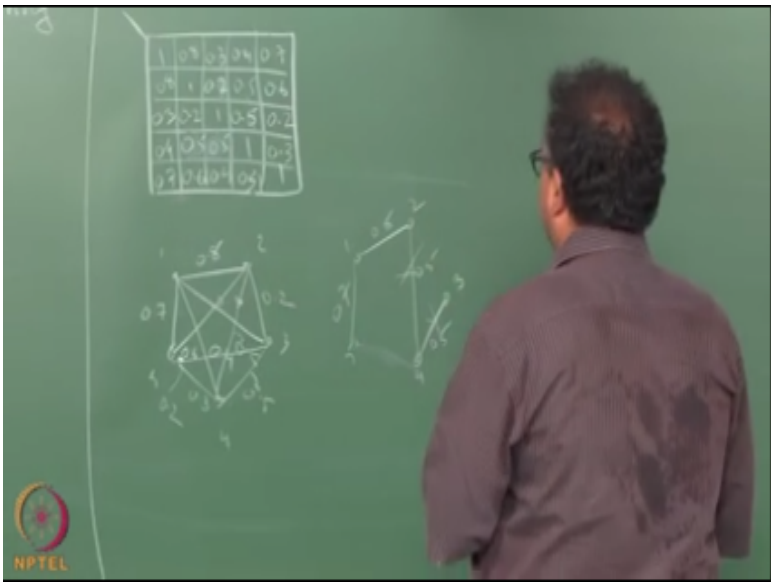
NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture-72
Threshold Graphs

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

(Refer Slide Time: 00:14)



When you looked at many such things so, so I am going to look at a specific setting now which essentially looks at how do you do clustering? Right when I do not really give you the data points right but I give you a similarity matrix between them right. So I do not give you the data points right but I tell you that okay, here are the similarity like similarity matrix between the data points, so this is like say point eight similar is what are the things I can fill in now, I am trying to give you something consistent, when this cooking this numbers up on the fly. I cannot cook everything up on the fly can I something like these rates.

I will give you a matrix like this okay can you do clustering? sorry similarity is the inverse of distance right or I can of course take the inverse of this and give you the distance between the

points as well right, so I give you a similarity matrix right. So the reason I am stating this is sometimes it makes it really convenient to reduce the data that is given to you and even if i give you let us say i give you a huge collection of documents right, so instead of computing the distance between every document again and again when you are doing clustering right.

I can just basically do a n x n I can construct the n cross n matrix like this right. In fact I am assuming this tinsel symmetric right I am assuming the distance is symmetric so it is not really n cross n is only half of that right and so you can construct this matrix you can keep this with you and then you can do clustering based on this right. Suppose I want to do something like Kevin's and how will it work in this case? Little tricky right I want to k-means is a little tricky sorry yeah but then the first you have to find embedding right.

So it is not so that is a is called an embedding into a space right, first you have to find them adding a be the embedding, B might not be sufficient right may not be sufficiently accurate you have to, first figure out what dimensional space you are going to do this embedding in, this is 2d 3d, 4d and finding the embedding itself is a hard problem right and then you want to do clustering on top of that. It is you are going to actually solve a harder problem before you are going to solve clustering. So you do not want to do the embedding right this has some other mechanism which you can do this right. So one way to think about give data like that is to think of it as a graph right, think of it as a graph and think of it as this some kind of weights between the nodes right.

So I have how many nodes I have five nodes right so I have five notes right, so I will give them numbers right, so 1 to 2 okay the weight is 0.8 that is a complete graph by the way. This beginning then 1 to 3 right, the weight is and then also 2 to 3 and becomes more and more at this 0.5 belong to. Now I am confused photo files ok really I mean you can make out the weights know right. So that is a graph right and I want to look at a partition clustering on this graph right. So what I can do is I can solve what is known as a min cut problem on the graph right.

So what is the min cut a cut on a graph is a set of you can do to two things, you can either cut on the edges, and you can cut on the vertices? We will worry about cutting on the edges that cut set of edges on a graph is a set of edges such that if I remove the edges in the cut-set the graph gets split into two components right. So I take a connected graph I remove a set of edges from the graph ok and the graph becomes two separate components ok, so that is called the cut-set right

and the min cut is a set that has the least weight right. In an unweighted graph it is a set that has the fewer stages right.

In the weighted graph it is a set that has the least weight it could have more edges but of all the weight edges could have less weight then that becomes a min cut okay. So you could try and do a min cut on this graph right, so that is one way of solving it and so in the next clustering class that we will have which will be like not next week the week after right the next clustering class will have I will talk about spectral approaches to clustering right. Which essentially talks about different ways of solving this min cut problem, talks about a completely different way of solving the min cut problem, so we look at spectral clustering later right, there are a couple of other things that you can do right.

So especially all of you have done graph theory some point you must have done graph theory all of you have done some graph theory, basic graph theory data structures in okay. People understand what is the meaning by minimum spanning tree very one understands what a minimum spanning tree? Is so what is a minimum spanning tree, a tree okay, b it spans all the vertices it connects all the vertices and three it has the least weight among all those trees that connects all the vertices right. These are the three things I so minimum spanning tree you can just basically take each term and define it and we get the think.

So in this case if you can think of a minimum spanning tree, what would it be I am making people run the extra or something now come on crystal prim what do you want to run crew skill okay give me a minimum spanning tree now right. So that is a minimum spanning tree so I started off by inserting the edges with 0.2 write and then I looked at things that are outside and figured out which is the least cost eight so both of these had the same cost I so now I have a minimum spanning tree. Now I can once I have the minimum spanning tree I can use this to produce clusters. I can start off by saying in this case it is pretty trivial.

I can start off by saying remove the highest weighted edge in the spanning tree, remove the highest weighted edge in the spanning tree right or should it be the lowest fit we are doing similarities right by doing something okay remove the lowest weighted edge in the spanning tree now I could do this either way right if I had add distances instead of similarities also I could do this right, remove the lowest weighted edge. No wait now I think have to the other way around I

am doing similarities right. So I should do a Mac spanning tree may not have min spanning tree. This is it easy to do a Mac spanning tree is the same complexity as a min spanning tree okay.

So you not tell me what is a Mac spanning tree here 5 and 6 is 0.2, 5 and 4 right not is that yeah okay if you want to do it that way sure 5 to 3 is a 0.4, 5 to 3 is 0.2, so 5 to 4 is a 0.3, 2 to 4 the 0.5, yeah that could work yeah so that is a point okay right. So now what I do is I remove a an edge that is got the least weight right, so I will remove this guy, so I am a left with two clusters it is a if i remove an edge from a tree it becomes disconnected right. So get the Mac spanning tree in this case I remove the edge with the least cost right. So I can think of doing clustering by doing it this way right.

In stuff if I had been given distances instead of similarity ever done the minimum spanning tree right removed the edge with the max cost right. Now I did the Mac spanning tree and remove the edge with a min cost to give me two clusters, if i wanted to do work with distances instead of similarity I will do the minimum spanning tree and remove the edge with the max cost right okay this gives me 2 class suppose I want 3 clusters what do I do? Remove another edge, so now that two three classes will be 125 will be one cluster and three will be one cluster by itself or will be another cluster by itself right.

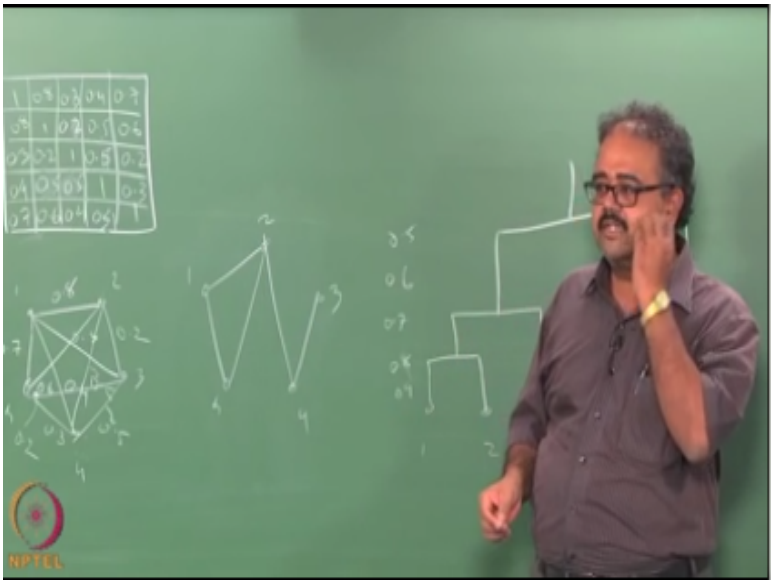
So I do not really need to do the embedding right I can treat it as a graph right and I can still do useful clustering with this. So one thing is do the min cut which we will come back to later other one is to just first do the minimum or maximum spanning tree right depending on what data are given and then do this ok cool. I am going to look at something else I will erase that okay, so that is a good question. So take it pick you know so you have to use some other heuristic even here also there was two possible choices for my first stage itself right I could have when I wanted to cut a 0.5 that I could either cut the one between 3 & 4 are the one between 2 and 4.

That I chose to cut the one between two and four because it gave me more or less equal sized clusters that could be here to strictly use right. So you can just say that ok if I cut this 0.5 I get an isolated node and all the other nodes are in one cluster, look at the other 0.5 I get two nodes 2 and 3 so maybe that is a better division, so you can use additional heuristics like this there are multiple things that are possible. In fact it is even more complex than that there could be minimal many minimum spanning tree is possible. I just showed you one tree it luckily it turns out that

this particular graph there is only one minimum maximum spanning tree that could be minimum many maximum spanning tree is possible. What do you do in that case?

You could only just pick one that is it just pick one and then go ahead and do the clustering it is like yeah there is no single answer for this remember me telling you clustering is a ill-defined problem yeah, so there is no single answer for this right there could be multiple different answers. So let us look more interesting okay, so I am going to introduce you to this concept called threshold graphs right. So I have graphs like this what is the maximum similarity I can have one right I will start off by saying I will connect all the nodes in the graph okay, says that the similarity is 1 or > 1 okay.

(Refer Slide Time: 15:29)



So I will basically end up with that is my craft right so it is essentially the MMT set right now I will say that ok great I have this graph and I am going to treat all the connected components in

this graph as a cluster. All the connected components in this graph I will treat as a cluster, so what do I get five clusters, so remains you of something hierarchical clustering this how we start off in the hierarchical clustering rates I will say that I will start off with each data point of being a cluster of its own right. Now what I do ok I will start decreasing my threshold right so what is the first step that I can do? I will make my threshold 0.8.

Now I will do all my connected components right I will take them as clusters, so how many clusters I have 4 clusters one and two are right. Next what is the best of sorry point it next I say oh 0.7 right that is my graph 0.7 that is my graph so what does it do, so I change the labels here if people are wondering what happen? So and then what is the next level I can do 0.6 is it and what is 0.6 is 1225 0.6 okay sorry does not matter this is still connected right nothing changes I do not hit use any new or connect components it is same set okay. No new clusters have been found right then I will go to 0.5 so what happens to the 0.5 I get that anything else is little tricky.

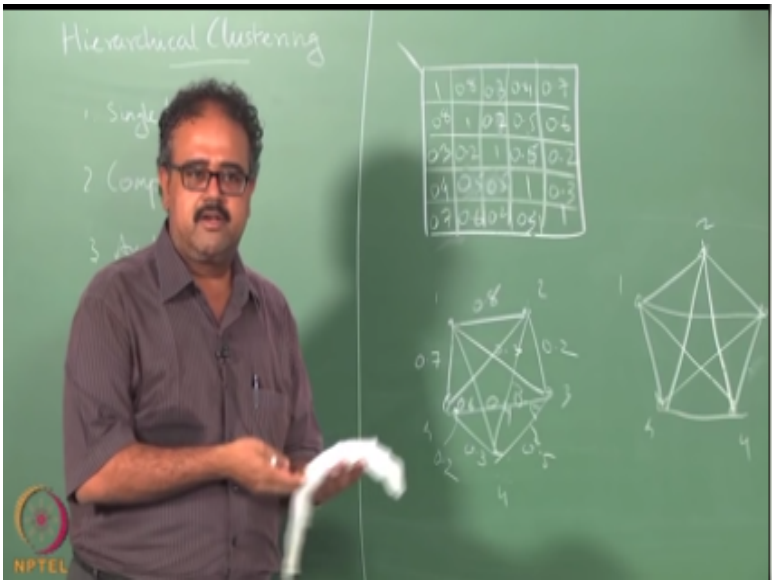
So two and four as well right so what happens now essentially everything is a cluster right, so everything has been boys I just stopped here, so this is my den program correct. So now i have done hierarchical clustering by using just simple graph theoretic concepts, what did I do? I just kept taking threshold graphs and I look at connected components in that graph and I got a hierarchical clusters right. So what is it equal to any one of those things I written down there is it equivalent any one of those distance measures, I wrote down there.

Single link, how many of you think it is single link? One how many of you think half five thank you how many of you think it is complete link okay, one how many of you think is Average link? One and a half so I am taking the average and somebody put their hand up like this somebody who did this so anyway. Single link right so the majority hope it was for single link 5 vs 1vs 1/2 right so single-link clustering right, so if we think about it I so what is the distance between the cluster 1 to 5and the cluster three four things, that are closest right the pair of points that are closer so between four and two.

So that is essentially in the distance I am using right when I have the distances for and to this edge appears in right and then it becomes connected right, that means that the closest points right these are the most similar so similar most similarly means closest distance is smallest right. So they appear in and therefore this is a single link clustering is equal not exactly equal into single-link clustering okay is it fine so can I erase that and I want to do this again. Except that I change

the definition of cluster okay. Threshold graph will start off with this right and I will take all the clicks in this graph as a cluster right, so what are the maximal cliques in this graph all of them right each one of them, so that is it that does not change so the same thing.

(Refer Slide Time: 22:02)



So I start off with sort of with five clusters right then I do the threshold right, so the first page that appears will be this guy right now what are the maximal clicks in this graph one to write everything else is all by themselves, so again I get that I emerge this and the level is 0.8. Now I do the 0.7 level what do I get I get that okay so what are the maximal cliques in this graph is 12 and 15 but we are already inserted 12 as a cluster, so since we have not allowing overlapping clusters or anything here we are thinking of partitions here right.

So 5 will still be left alone right, so I have two possible cluster clicks here 12 or 15 right but since I do not consider overlapping clusters I cannot assign one to two clusters so I just leave it like that so at the 0.7 level I do not do anything, so point name we do not know anything, also I do not do anything. Next what do i do I go to 0.6 which is what 25 a okay, now what are we have

a maximal cliques here 125 right. So earlier at the 0.6 level nothing happened this cluster got formed at the 0.7 level itself here you have to go down 2.6 before the cluster gets formed okay.

Next what we do raise it 2.5 I do not need to put the right 0.5 this is what I get, so there is a new cliques that is formed like 34 right, this is what this is 0.5 right then you click that gets formed a 2.5 then what I do 0.4 that is a point for which is what one and 4 okay let us say change any rate no rate I do not want to disturb any of the clicks that has already been formed unless a new clique is forming I do not want to break this and put it there or anything right so i will leave it like that. Then i go to a 0.3 what is the 0.3, 1 and 3 is it 1 & 34 & 5 what about three and five know what about two and three no right, so at 0.3 nothing happens but then I go to 0.2 ha.

Now I finally have everything down right then basically 0.3,0.4 nothing happens at 0.2 I get the final merging okay it is fine, so there are two ways in which I can do is I can just think of connected components or I can so also think of clicks right and so what is the difference here if I choose to cut, remember I was telling you can cut that tree at some point and retrieve your clusters right. So now I can set myself a threshold okay I want to cut the tree says that the similarity between the data points in a cluster is at least 0.5 right, so then what do I do so I cut the tree just below that right like that and here I will cut it just here right oh well in this case it turns out to be the same sorry yeah.

So I said at least 0.5 right so 0.5 is a bad idea let us let us take let us take 0.6 right I want at least 0.6, so what happens is right so I will cut it off here just below 0.6 I want at least 0.6 means I want it to be at least point 600001 which I do not want it to be exactly points it going to be slightly more than 0.6 I can't it just below 0.6 level here, so what do I get I get 12 as a cluster five as a cluster four as a cluster 3 as a cluster but if you do the same thing here just below 0.6 right I get 125 as a cluster four and three as a cluster.

So depending on how I did the clustering and how I built the tender gram given the same tolerance level my teal different clusters for me right so is this Plus this tie up with any of the clustering technique that we already saw its complete link like why is it complete link? So I consider two clusters as having merged only if all points are connected that means, even the specifically the farthest most points also should get connected right. So the level at which I will merge the cluster now this will be the level at which the to the farthest point slide right so this is essentially this is complete link.

So this is single link okay makes sense so you can always think of your data points lying on a graph and we can do all of these things but a nice thing about data points lying on a graph it can visualize them harm it is 2d that is interesting are all graphs visualize able in 2d, so they were named for it they call planar graph I mean you can still utilize other kinds of grass is just that they will have all kinds of course crossing lines right, so and so it becomes a little harder to visualize right but yeah huh this is planar complete not claim I don't think Theodor gram trainer can visualize anyway come on doesn't have to be a graph right.

You can visualize the dendrogram given any points it seems something even more important and when I start embedding them in some space and start giving you a distance measure it basically has to follow certain properties of that space right typically you end up wanting their distances to follow some kind of a metric right. When I start putting them in a graph right there basically arbitrary numbers I can fill in there right so I do not I can have some similarity measure which is not even a metric right, I do not have to worry about whether it makes sense whether triangle in equality is followed or anything again just take our can assign arbitrary.

Similarities to get a points okay and then I can say do clustering there might be applications where you need this kind of a power, so that is the nice advantage of thinking about this as graphs right. So once you have these as graphs then you can go ahead and do all kinds of your single link complete length clustering or do minimum spanning tree do minutes whatever it is and you can do your clustering. So that is the power of the graph modeling in fact so much so that when nowadays when I think of clustering applications I almost always think of okay what is the graph right that I can construct out of the data and once I construct the graph and then feed it into my clustering algorithm right.

So that is that is typically how many people operate right because there are so many powerful clustering algorithm there are based on graphs okay good any questions on this is not possible to reduce. So the complicity of the intelligence is it possible to reduce it most clustering algorithms are way more than order N^2 in their operations because if you think about it right so if you use something like k-means and what is the problem with k-means? Every time the centroids change and I have to redo the computation to the centroids. Suppose I have case in Troy's and I have n data points okay.

So every time the things change I have to do a for every iteration allowed to NK computation okay and the number of iterations can be fab pretty large. so yes that is the problem right but then k means in if you have effect a fairly large data set a small number of cluster centroids k-means is actually not a bad thing for example k roid right think about the complexity of no it is humongous or Pam right, so when you do Pam basically you take each yeah exactly, so many clustering algorithm very expensive. So N^2 is not too bad right but of course if you have a cheaper way of if you have a way of getting around computing the N^2 distances great.

If you have better way of computer computing that is great. So what typically yeah I mean there are ways of doing that okay but it involves using very clever data structures and trying to reduce the amount of computation that you do right, by doing some cheap computations and then trying to do more expensive computations and so on so forth. We are depending on the size of the depending on the volume of data that you are handling right and a amount of resources that is available to you and so on so forth you might want to choose something over them it turns out that.

The overhead in doing this N^2 computation is lot lesser then in some of those techniques is try to avoid the N^2 computation right so one of the things which way should realize this big bow notation is very deceptive right. Suppose you have ten elements in an array what is the best way to sort it likewise that might be instances where even though it looks N^2 it might be cheaper to do that rather than try to set up something that is more clever okay, so that that is the thing you will not to think about but clustering is inherently an expensive operation there is no way around.

IIT Madras Production

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

www.nptel.ac.in