NPTEL

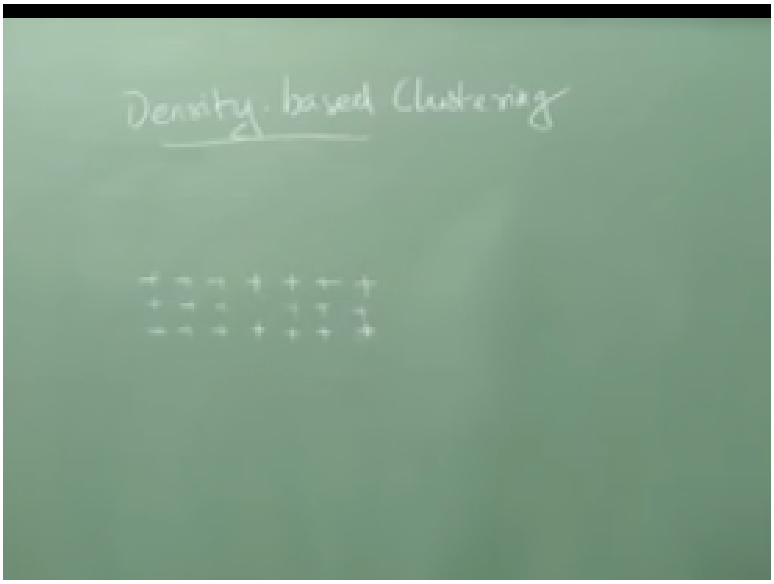
NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture-75
Density Based Clustering

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

(Refer Slide Time: 00:17)



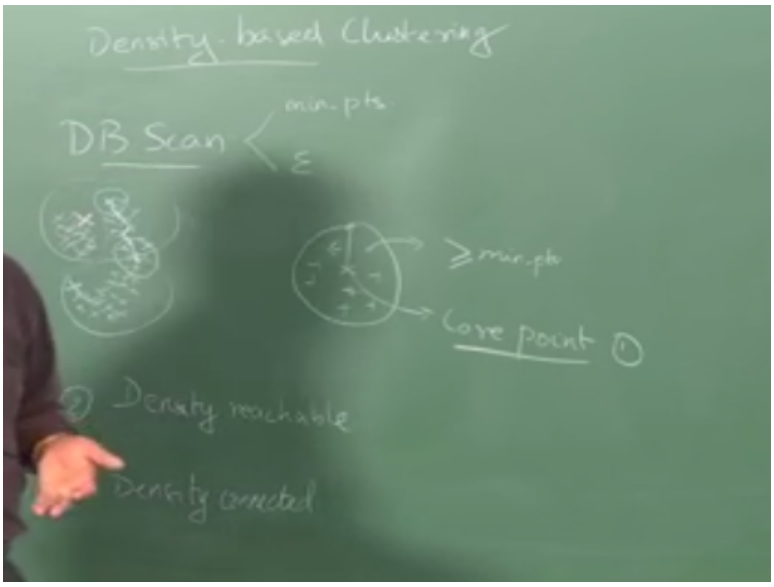
Can you see two clusters on the these points, kind of sort of what are the two classes as a row of pluses there in the row of pluses at the bottom so there are two clusters here right. So I will not touch these data points right I just do the following, now do you see two clusters side by side so what essentially defines clusters is not the distance between the data points that is more like the density of the data points right.

Naturally when we think of clusters it is really thinks that where data points are really dense is one cluster and then we tend to draw the boundary between clusters where the density is lower right so I did not change the data points in the initial set of data points you are very happy to say

that okay the cluster or the top and at the bottom right so what cost you to change your clustering when added the additional data points.

The density went up right the density went up in a different way therefore you said oh okay the least dense point is no longer between the two but vertically right no longer horizontal but vertical was the low density points right if you run k-means you might get anything I do not know and depending on where you start off with right so the question is so if you have your intuitive notion of clustering is to do with density so why do not you try to come up with a clustering algorithm that captures this notion of density right and trying to come up with a clustering algorithm that captures this notion of density.

(Refer Slide Time: 02:45)



So there is a very popular clustering algorithm called DB scan, it very popular algorithm called DB scan that does density based clustering okay so Db scan actually has a lot of terminology that they define right once all the terminology is defined then the clustering algorithm itself becomes nearly trivial okay but the terminology takes a while to get through so the basic idea is very simple right suppose I have a suppose I have data points like this right.

You see two clusters here clearly two clusters right incredibly hard to get k-means return these two classes and if you run k-means what will happen is you will end up with the centroid somewhere here right another centroid somewhere here right and it will say these data points are

one cluster and these data points another cluster is the biggest drawback with k-means right, so what DB scan say is that two points belong to the same cluster okay.

If I can get from one point to another right by moving only through only through dense regions only through points that are close by right two points belong to the same cluster right suppose I take let us take let us say I take this point let me take this point right in fact they look pretty far away and if you look at the direct distance between them they are pretty far away infact the points of other cluster which are closer to this than this point right but when you look at it you think that they this thing is one cluster this thing is another cluster.

So what is the intuition here is that can keep hopping at no point do I take a very big hop right I can keep hopping to things that are nearby and I can go from here to here right, so if you take these two points the blue and the brown one right so if I take these two points the no way I can hop from here to here right because there is a this nice gap here right there is no way I can get from here to here only by going through dense regions right is that clear.

So that is the intuition that we are trying to capture here so what is it that we should define now first what I mean by a dense region right so that is essentially what we have to define so we will start off by defining something called there are two parameters that DB scan uses one is called pinpoints other one is epsilon there are two parameters so min points essentially gives you some kind of a threshold on how many points would you consider as being dense right.

And epsilon gives you the area over which you will perform the count is a min point says okay if you have five points okay then you are in a dense neighborhood but where do I count these five points okay in a radius of ε around me you count the five points okay count in the area epsilon around me if you find five points then you are in a dense region and if you make the ε very large then it might encompass my entire input space then everybody will be dense so it does not make sense.

The epsilon has to be small likewise if I make my min points 1 point 1 not 2 points that means everything will look dense unless make me epsilon very small so these are actually complimentary things I can control my min points and I can make them in points very small right and then make the density high right or I can make my min points large with the larger epsilon

and that it then also I can make me density high the effects that you will see are different for both right so I let you think about it.

Which one I mean what is the effect of increasing min points versus decreasing epsilon okay so essentially what I am saying is okay take a data point right take a radius epsilon around it okay that that is that is a circle so take a radius epsilon around it and count the number of points okay if this count is greater than min points okay then you call this a core point take a data point take a ready ball of radius epsilon around the data point right.

Count the number of points right if the number of points is greater than or equal to min points right number of points is greater than or equal to min points in that ball of epsilon then you call this a core point right a core point is a point that lies in a high density region that is the definition we have right. So we say a point is you see that the point is density reachable okay.

A point is density reachable if there is a core point from which you can reach this point it went by traversing only through core points so this might not be a core point because is it the border rate I draw in the radius around it I get only one point here okay so this might not be a core point right but then if I start here let us say let us say this is a core point for sure a lot of points in the anything from here i can basically move two points within epsilon of itself.

Which are internal core points right so I can move to you can move to core points right and finally reach this that no point I should be making a jump greater than epsilon right because it is starting a core point I will have enough points in the neighborhood that I can actually jump to something within epsilon right so I make steps of size epsilon and I actually go through core points every time then we call them density reachable say point I is density reachable if there exists a core point from which I can reach here.

By jumping only from a core point to core point until the last step obviously every core point is density reachable because it Is reachable from itself right every core point is density reachable and then there will be these border points right which are density reachable from core points there might be other points which are not density reachable from core points which are essentially out layers okay right.

So these are the definitions we have so this is the first right these are the two quantities we need this is these are not definitions this is the first definition what is a core point second definition

third definition is density connected case I will say two points I and J are density connected if that exists a core point k from which both of them or density reachable that makes sense right so what is density reachable I start from a core point and only move to core points right until I make the last hop to this case so no point I will be making a move greater than ε.

And all the points I visit in the on the way will be core points ok that is density reachable density connected is if I and J that exist one core point from which both I and J are density reachable then the I and J are density connected okay here is the next thing. So I and J are in the same cluster if and only if they are density connected this is the definition of a cluster two points I and J belong to the same cluster if and only if they are density connected make sense.

Sorry how do I implement this so I start off with YS any point right I pick up a random point right I figure out whether t is a core point or not right then okay it was a core point great so I will keep that as my starting point for the cluster how do I determines the core point I pick up a point look in the neighborhood figure out if there are epsilon M if there are min points within a neighborhood of Epsilon if that is the case then I will keep it right.

Then what I do is I look at all the neighbors of that point and look at all the neighbors of the point and each point in turn I will check whether it is a core point or not right, so each point in turn I will check if it is a core point or not so any additional points I encounter when I do this check I throw it into my Q so I will keep going right if I reach a point which is not a core point okay so I will not insert the neighbors of that new neighbors of that point I will just stop there if we reach a point is not a core point I will just leave that exploration go back to my Q to see if anything else is still there.

So I keep doing this until my Q becomes empty so all the points I have examined from the time I started till my Q became empty go into a single cluster sorry like it that first search moves like a deficit you do that all these data points go to a single cluster now what I do I go and start at a random point which has not been assigned a cluster so far and then do it if first search again till I find the whole clusters I do this and I am done.

So the nice thing about this is I am really doing only one scan through the data right so every data point I will actually look at it once right I will examine the neighborhood right and then I will go on but then the number of computation I will do will be still significant so EV scan is a

slow algorithm even though I examine each data point only ones but the amount of computation I do an examiner data point is significant.

Because I am looking at the radius epsilon and then we have to find all the neighbors within the radius so unless I have a very efficient data structure that will return to me the nearest neighbors very quickly right so this can take a significant amount of time in running so there are some efficient implementations of DB Scan out there it is really cool in that it gives you all kinds of arbitrary clusters right.

And so these kinds of things today whatever I drew that that you would not be able to recover using k-means or even hierarchical clustering depending on the kind of cluster measures that you choose right cure might be able to give you or might not be able to give you this kind of clusters depends on again how your sampling that you do and what you start off with and so on so far so whole bunch of imponderables.

But again the same thing with DB scan depending on which is of min points in epsilon right you might get very weird results, right yeah so if you look at the data mining text book by Han and camber right Michelin Cameron book right so they have actually horrendous examples of DB scan in fact I think they did a significant optimization to find out which are the worst parameters possible form in points in epsilon and they give you results.

Because they wanted to write a paper that said hey we have an algorithm that does better than DB scan so they said oh DB scan can perform really badly if you give it back parameters so let us give it back parameters and then we will beat it right. So I on this well this is assuming that they were actually being more fair than that but the way the way the way look at the results it looks like that I mean TV scan looks really horrendous and they are a method which is called chameleon.

I is an acronym yeah so I think C stands for clustering I am not sure good I sure about that either so yeah so yeah it is a nick name DB scan right I just start off with some arbitrary point and then I look at look at the exam examine the clusters and soon so forth, so what happens if I happen to start off at a boundary point instead of a core point I will examine it I will say okay it is not a code point I will throw it away right so it will never get assigned to any cluster.

So it will kind of be all by itself as an outlier even though it should belong to a particular cluster that is one thing the second thing is suppose I want to vary ε and I want to run the clustering again I basically have to do it all over from the beginning right so what optics does this gives you a clever way of ordering your data points such that right for different values of epsilon you can recover the clusters very quickly for the same value of min points min points is fixed but epsilon changes right.

So it gives you a way of ordering your scan through the data such that for different values of epsilon I can recover the cluster is very quickly and is a really cool idea right.

IIT Madras Production

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

www.nptel.ac.in