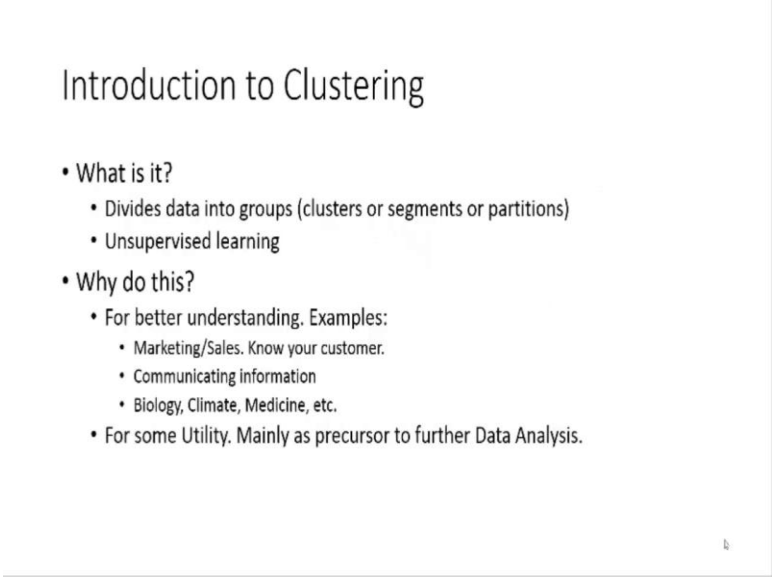
Introduction to Data Analytics
Prof. Nandan Sudarsanam and Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module - 08
Lecture - 41
Clustering Analysis -1

Hello and welcome to our lecture, our first lecture on Clustering Analysis. This is the first of two lectures that we will have on this subject, this one is an introductory lecture. It introduces the basic concepts and algorithms behind clustering analysis and in the next lecture; we will actually go into two specific algorithms and see how those two algorithms work.

(Refer Slide Time: 00:37)



Introduction to Clustering

- What is it?
  - Divides data into groups (clusters or segments or partitions)
  - Unsupervised learning
- Why do this?
  - For better understanding. Examples:
    - Marketing/Sales. Know your customer.
    - Communicating information
    - Biology, Climate, Medicine, etc.
  - For some Utility. Mainly as precursor to further Data Analysis.

So, deriving to the subject, what exactly is clustering? Clustering is the idea of dividing data into groups and we do that, because sometimes there is inherent meaning to doing such an activity. And in other cases, it serves as the first step, it is fairly useful to do this. Now, you might notice that you might have come across clustering, but some other names as well, which such segmentation or partitioning.

So, that is kind of why we have written out all three for you in this first bullet point. But, the core idea is the same, it is fairly interesting, it also partly what terminology wind up using also has to do with the context or the background. So, typically something like segmentation is used more in a marketing sense. So, when you grouping people or customers, sometimes people use our segmentation. Partitioning, again interestingly comes more from the computer science community, where you are breaking a graphs, so you are partitioning graphs.

But, again the core idea is that you have some data and taking the data and based on certain properties of this data, you choose to group it. So, you can think of it as actually grouping data objects based on various attributes associated with this data. The idea behind such an activity is that, the data itself is represented through various attributes or features and some data points are similar to others based on these attributes or features.

And you want to group those that are similar and put them into one cluster or group and differentiate that from other groups or clusters, which are similarly formed based on some measure of similarity. So, the core idea is to put things that are similar together and therefore, as a result the different groups are as dissimilar from each other is possible. So, data points within a group are similar and data points between groups as a result are dissimilar to each other.

Now, the core idea the clustering often relates to and sometimes it is often confused with it is classification. I think an important thing to recognize here is the clustering is primarily seen as an unsupervised learning technique, whereas classification is supervised learning technique. The idea is that you have some, in classification you have some input data and you use the historic input data and the specific output and in the case of classification, the output actually has classes it is a categorical variable.

And, so you used some kind of supervised learning technique to look at the relationship between these input variables and the task there is to make a prediction and assign it a class. In the case of clustering, you are again dividing the input data space in some sense into groups, but the groups are not based on labels that have been explicitly given to you. So, in some sense there is no output variable with clustering and what you have is only the input data space and it is not even clear that, there is another there is a classification scheme like in the sense of the, in the classification there was this output variable and this output variable had 1, 2, 3 or more categorical states.

So, there were the states that are labels that was explicitly given to you and you are now trying to create that relationship between these labels and the input variables. In the case of clustering, not only DA not have the labels or the output variable, there might be no meaning to having such labels. So, in some sense with clustering you are really breaking the data, input data set based on the inherent relationship between data points and how similar they are to each other, you do not have external label that is given to you about the data points.

So, if two data points are very similar across these attributes, then you group them together if they are not you group them apart. So, in the end you might create a few clusters and you can call them cluster A, cluster B, cluster C, but that is not the same thing as classification, which is more of the supervised learning process. So, why really do this and the best way to see that is this brought the two major reasons, one is clustering itself might just be useful to understand the universe of the data that you are dealing with and we are going to look at some examples in that light.

The other reason and sometimes the clustering is used is that, it serves as a precursor to further data analysis. So, it has some utility from a machine learning sense itself to do a clustering. So, now, let us look at the first case, which is that in some sense you get a better understanding of a data when you do clustering. A fairly common example of this is in marketing or sales, your business is for instance collect you know lots of information about a customer.

So, the way you should be thinking about it is, each data object is essentially a customer and the customer has for instance various attributes. They could be things like gender, age and you know it could extend all the way to the kinds of products of the person tensed to buy and so on and so forth. And you have a full data set of a lot of customers or potential customers and you might be interested in grouping that.

So, there is no explicit output variable, but certain types of customers are very similar to certain other types of customers. They could be very dissimilar to the third type. So, creating groups out there could help for instance like a market research initiative into looking into a particular group and coming up with ideas, so how to specifically target our market to them. Another example you know, another very common example is in terms of just communicating information, take a simple example such as Googling or you know searching for a movie on the internet.

Now, there are lots of things related to a movie. For instance, there could be a reviews of a movie, they could be trailers and videos of a movie, they could be a ratings of a movie and they could be information on which theatres the movie is running in or where you can purchase a movie. Now, a search itself, the first search itself across the web might give you a huge large bucket of things that could belong to any of these categories. But, if you have an ability to for instance look to see which pieces of information are similar to each other, you might be interested in representing one of each type.

So, you might have a cluster; that is created, which talks about essentially reviews for movie and another cluster that gives of potential links that talk about, where this movie is playing. And, so in some sense you have you could create this clusters and present the viewer or present the person, who is searching with representative of each cluster or typical value needs cluster. So, that a person, who is searching for a review of the movie does not get 20 links on the first page that give him trailer, give him or her trailers, which should be quite frustrating.

So, in terms of information retrieval, communication of the information it might be fairly useful. Clustering is used a lot in biology mainly in taxonomy, where if you just said the wild universe of mammals or insects or something like that, you kind of want to group animals or mammals that are similar to each other and give them some taxonomy that is different from animals that are different from each other. And here for instance each animal would be the data object and the attributes would be, you know various animal related attribute such as, how they feed, what their family or genes or species and so on and so forth.

It is also used in climate, many times understanding ocean temperature ranging from ocean temperature to hurricanes to various other things, it can be better done you know if you cluster the data on other patterns. Medicine of course, again a certain types of disease similar to other types of disease or certain medicines similar to other medicines, it can simplify the word in some sense and help people understand, how for instance certain medicine interact with certain disease and so on and so forth.

Now, these give you an understanding and these kind of talk about, where clustering is useful or has been used or just to get a better sense of the data. But, there is this completely different and other utilities sometimes to clustering and that is mainly in the form of serving as a precursor to further data analysis. The idea here is that, clustering for instance can be used as a great way to summarize data. Especially, when the algorithm that a person needs to use. Let us say a person is interested in performing some form of regression or another more complex supervised, unsupervised learning tasks.

Sometimes with more and more data the task just becomes computationally harder and harder and it might be sometimes beneficial to perform a clustering analysis on the data which could be in some cases it is computationally easier to do the clustering. And then, just have a representative from each cluster as a data point, so you could have a representative or you could have essentially a cluster center the clusters representative. So, was the data point for all the data points in that cluster?

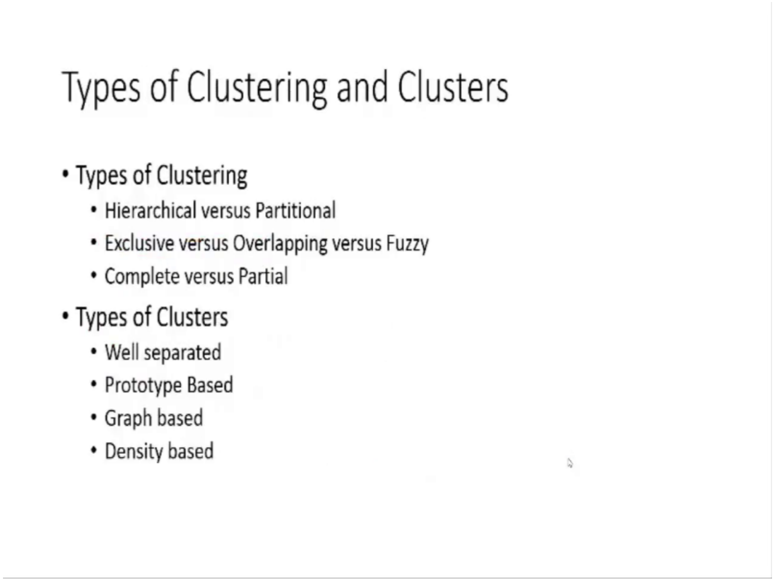
And, so you are now, doing that same machine learning task be it a regression or a factor analysis. So, whatever it is your really doing it on the prototypes on the representatives rather than the full data set. It is also used an extension of that is also how it could be really useful like say in a nearest neighbor task we in an earlier in earlier lectures we spoken about this algorithm called K and n K nearest neighbours when you think about the problem that is involved with that for any given point.

If you need to find it is nearest neighbours you need to actually try and look at the distance between the point under question and every other point in the data set. So, you need to evaluate the distance between the point you are interested in and every other data point and then pick the K closest neighbours. Now, that can become computationally very, very, very hard and you need to keep either the memory and you need to do the computation from scratch a simpler approach could be that you just take the K you do the clustering on the data.

And if you look to see, which cluster center is closest and once you identify cluster center that is closest you now, take your point under question and only evaluate it is distance to all the points in that cluster. And the assumption here is that the points the points under that cluster are the once that are going to be closest to this point, because this cluster center was the closest for instance. It is also used in certain other forms of you know compression of data specifically something called vector quantization, which is used a lot in image or sound or video data, where you typically find that lucky in a particular image, if you break it up into pictures there is the whole host of data points that will look very, very similar to each other.

Let us say you take a photo of a person almost 20 to 30 percent might be the background behind the person and that might have the same color. So, it would be more efficient to just acknowledge that there is a very good chance that the pixel on all four sides are going to be the same color is this pixel. And, so you kind summarize the data you reduce the data to few a number of pixels in a memory values and; obviously, would this the some loss of resolution either some loss of information essentially, but that might be acceptable then the data size itself get substantially reduced.

(Refer Slide Time: 15:00)



Types of Clustering and Clusters

- Types of Clustering
  - Hierarchical versus Partitional
  - Exclusive versus Overlapping versus Fuzzy
  - Complete versus Partial
- Types of Clusters
  - Well separated
  - Prototype Based
  - Graph based
  - Density based

So, we spoke about, what clustering is and why used, now let us just briefly talk a little bit about the different types of clustering. The major classifications tend to be one hierarchical versus partitional and the major idea here is that hierarchical clustering is essentially a nested form of cluster. So, think of it this way you can start with this one mega cluster, which is essentially a single cluster of all the data points and that is on one end of the scale and then, you go to the other end of the scale where you have as many clusters as the number of data points and each data point it is a cluster is its own cluster.

Now, some where hierarchical clustering works on creating this tree between a single cluster of all data points to each data point being it is own cluster now either of these extremes are useless. Because, the single cluster of all data point is not really clustering you just created you put all the data points in one group and its called it group one that is not very useful. And neither is it useful to call each data point it is own group.

So, somewhere in between these two extremes is the real value at, but hierarchical clustering works on some form of nested or kind of tree form of clustering when you start with this one you can you can start it either end, but you might start with one large cluster and then, break that into two. And now, you have these two clusters, now you go into either of these two cluster s and break that into created division there.

So, now, you will have three clusters, but the three cluster is strictly a division of the two clusters. So, it is not like you are going to take some points from cluster when you have two clusters, let us say you had some cluster A cluster B it is not like you are going to take some points from cluster A and some points from cluster B and call it cluster C. It is in that sense nested it is in the sense that you make one split and very, very similar to decision trees you keep making further splits and this is contrasted to a approach of partitional clustering.

Partitional clustering is not of this kind of nested approach it is just that you explicitly decide on the number of clusters in some sense and you go and partition the data its it is it is simply a division of the set into non overlapping set, so it is essentially clusters. And, so in that sense there is no tree diagram or anything of that nature with this. So, for instance a partition of cluster that if I decide to do the partition of cluster and create four clusters and I contrast that to a partitional clustering approach, where I had three clusters.

Therefore, need not be a further division of the three they just completely you know the one that said decided to break in into four has a very little or no relationship theoretically to the division the separate exercise of the separate effort into creating a partitional cluster with three clusters, whereas the same cannot be said for hierarchy.

Because, hierarchical went in sequence, it could have started top down or bottom up meaning you could have started with all in hierarchical all the individual clusters, where it is each day you know each cluster is its data point and then started grouping them or the other way around. But, essentially with hierarchical wall is nestled into the other and so on. The other major type of clustering that people talk about is exclusive verses overlapping versus fuzzy.

The idea here is with exclusive clustering each object is a sigh to a single cluster and that object therefore, cannot be assigned to another cluster. And, so there is no there is no notion that one can simultaneously belong to multiple clusters, where as in clustering overlapping in case in clustering algorithm that allow for over lapping. You can basically it is not exclusive a data point can choose to belong to more than one cluster at a given point and decision on which, of these are really depends on the under lying system.

And some cases it might just make sense to have some data points belonging to multiple clusters and in some in some cases that that notion of division is just not sensical. And finally, you have you have the notion of fuzzy clustering fuzzy clustering kind of takes this over lapping even further, where each data point is not really assigned to a cluster, but it basically gets a number between 0 to 1 that that talks about the weight associated with that data point belonging to the different clusters.

So, for a data point each data point gets total weight of 1 and it takes that data that weight of I am going to assign 0.3 in belonging to cluster A I am going to assign 0.7 in belong to cluster B and I am going to assign myself 0 belonging to cluster C, So, the constrained is that the sum of its weights in terms of belonging to the different clusters adds up to 1, but it can use it is weight of one in any way chooses to belong to the different clusters.

So, that is to give you some idea between of exclusive versus overlapping versus fuzzy the last that is what mentioning is you can also have a complete process partial clustering. A complete clustering basically just assigns every objects to a cluster, where as a partial clustering does not it starts with the data points chooses to you know cluster as many points to different clusters and data points that do not really help in terms of belonging, so clearly to given cluster just not cluster they are just left out.

And these might and the and the motivation there is that you are more interested not in kind of pigeon holing each data point into a cluster, but you are more interested in the cluster formation itself. And there you do not want data points are not, so cluster able, that do not really belongs, so clearly into 1 of 2 clusters to kind of ruin the cluster center or ruin that nice division that you created. So, these are set different types of clustering, now the algorithm themselves and these are more descriptive of the type of algorithm that goes into it.

Now, another area focused could be the kind of clusters that your forming and while this has everything to do with the algorithm also here we are just looking at the end product. So, the algorithm that we are using what kind of end product does it can it give you. So, there is the first one is the well separated idea the idea the well separated is that you want to create clusters, where get the objects where each object is similar to every other object in that cluster.

So, the way you defining well separated clusters are is more from the core idea that you are very interested in looking at the relationship between each data point with respect to each other data point and your measuring or your quantifying or the clustering itself by looking at how similar that data point is to that other data points that are in that cluster. And therefore, how dissimilar this data point is to the data points that are in the other clusters and this kind of thinking you know is very useful especially when the data itself can be very nicely separated.

When the distances between the clusters are fairly significant, then this conception becomes very useful. The prototype based approach really talks more about how each data point is close or to it is cluster representative. So, that the commonly used terminology is to the proto type that defines the cluster and when I describe it I can try to think of it as there is a cluster and there is there is the main representative of the cluster the prototype the one that kind of signifies that what the cluster is in the center of that cluster.

So, in the prototype based approaches, because is representative is in some sense. So, central it is also sometimes called central based clusters, but the idea is as following, where each data point is looked at more from the prospective of how close this data point is to the cluster centre to the proto type. And, how far this data point is from the prototypes of the other clusters or the other representatives and the classification is done in this fashion.

We then, have the graph based clustering and this is really useful if the data is represented as a graph and you have nodes, which are represented each data point or object and you have this links or edges, which represent some notion of connection between the data point. And this notion of connection could be the one that talks about how similar a data point is to the other data point you could have some kind of a threshold value or especially when sometimes you are variable itself is not quantitative variables.

So, you could have some notion that two data points are connected and the idea here is to do some graph based clustering, where you really looking for a high density of connection this between the data points that belongs to a cluster and a very low level of connectivity of the data points between two clusters. And for that reason you know this whole language that comes from the graph theory community, where you define things called cleeks, which essentially just means that at the set of nodes in the graph of a completely connected to each other.

And, so a lot of that that kind of clustering ideas that go in to graph based clustering are once that kind of look out for cleeks and say this guys are all connected to each other, so they must be a cluster. Finally, you have density based clustering the idea behind density based clustering is that a cluster is essentially a dense region of objects that are surrounded by regions of lower density.

So, the idea here is that, because there are not such well defined clusters like in the case of the well separated idea that you are not really looking two often do a complete clustering and there is a lot of noise in the data and you know the clusters themselves are irregular or intertwines and there are lots of outliers and so on. So, the idea here is that you acknowledge all of that, but you just try to identify spots of extreme density, where once you identify that spot that dense region becomes a cluster and so that is fairly useful in defining a cluster, where there is a lot of noise and so on.

With that we will conclude our lecture on lecture that introduces clustering and the different types of clusters and where it is useful and so on. In the next lecture we will looking to basic algorithms one is the K means algorithm and that really belongs to that partitional camp and we will look at another algorithm called the hierarchical clustering algorithm which belongs to the hierarchical camp.

Thank you.