Artificial Intelligence for Economics

Prof. Adway Mitra

Artificial Intelligence

Indian Institute of Technology Kharagpur

Week – 02

Lecture - 10

Lecture 10 : Clustering and Segmentation

Hello everyone, welcome to this NPTEL course on artificial intelligence for economics. I am Adway Mitra assistant professor at Indian Institute of Technology Kharagpur and I am today we are going to start with our tenth lecture in this course the topic of which is clustering and segmentation. So, in our previous lectures we have discussed certain methods related to heuristic search optimization and so on. In this and in the next few lectures our aim will be like how we can deal with data and extract knowledge from data from past data and using which we can make predictions on new data. So, as so this is relates to the broad technique set of techniques which are known as machine learning. So, like machine learning as many of you know has two main parts supervised learning and unsupervised learning.

Today we are going to start with unsupervised learning and in the coming few lectures we will be dealing with supervised learning. So, the concepts which we are going to cover today is we will first talk about ah feature based representation of data ah next ah we will talk about what clustering actually means or what clusters mean and then we will discuss two ah commonly used algorithms for clustering namely agglomerative clustering and k means clustering. Apart from these also there are many other techniques of clustering. However, for the lack of time in this course we will not go into the details of those techniques, but you are welcome to read more on those clustering techniques from other courses related to machine learning.

So, first of all what is clustering? So, like as I have already said this is a method known as based on unsupervised learning which means that you are provided with observations, but no additional labels with each observation. So, let us see let us say that a child is shown these kinds of pictures of different animals. So, these are the examples. Now, they are not told which animal is or which picture belongs to which animal. So, with that kid does not know that this is a tiger, this is a cat and so on and so forth, it just has access to these images.

Now, even if the kid does not know the names of any of these animals, it may still be able to say that these some of these animals are similar to each other. Like they can say that these two animals they have certain features in common like they both of them have these kind of stripes on their body or that these two or these three are similar animals because they are special structures have certain similarity maybe their eyes have certain similarity same with these few animals or with these few animals they have they may have horn or their faces have a particular shape. these animals again the general overall size and shape of their bodies are similar and so on and so forth. So, that way they can try to group these examples together. So, they can find out which sets of examples are similar to each other and group them together and each group which we are talking about this is what we call as cluster.

So, if we consider some numerical data, so let us say that we are trying like we have observations related to the income and expenditure of different people. So, there are every person in this case is represented by two values, one value for income and one value for spending or expenditure. So, like we can represent every person by a two dimensional vector in this case. So, since there are only two dimensions we can plot these there these vectors in the feed in the this two dimensional space as is seen in this image. So, like here for every observation that is for every user we identify the their positional coordinates by using on the basis of these two features that is the income and the expenditure.

So, as you can see along the this horizontal axis we have plotted their income and along the vertical axis we have plotted their expenditure. So, now the each of these dots then represent the different persons. Now, like if you look at these dots. So, we may see certain patterns emerging here. So, we can say that there is like there are some observations are broadly in this region.

So, in general like if you see it at a very crude level we will see that all the observations are all over the place. that is there, but if you zoom in a little bit then you can see that most of the observations are focused mostly in this kind of a region like this kind of a region and there are some observations like these or like these which are a bit away from it. Now, if you zoom in even further then you can say that even within this region which has high concentration there are certain other region certain sub regions of even higher concentration like this is one sub region of high concentration, this is another sub region of high concentration, this is one more sub region, this is yet another sub region and so on and so forth. So, each of these sub regions we can call them as different clusters. So, like the we have the entire set of data which are like we can say that they are present in the feature space every observation is represented as a point in the feature space and then we are trying to see or explore the feature space and try to partition it into high density regions of points.

So, each of those high density regions can be all the points which are present in each high density region we can say that those are forming one cluster. So, this is a like an intuitive definition of cluster, but let us try to come to a more formal definition of cluster. So, to have a formal definition of cluster we need the concept of distances between any pairs of examples. So, we have already mentioned that each example or every each observation is represented by a feature vector in the suitable feature space like in this case it was the two dimensional feature space. Now, if that is the case that is every point or every example is to be represented as a point in the feature space then if we want to identify which At the high density regions as we are saying earlier, we need a notion of closeness.

We want to know which points or which observations are close to each other. And if we are talking about closeness, there must be some kind of measure of distance. Now, there are many distance measures in which are defined in mathematics the most common one is of course, the Euclidean distance. So, if we are using this the feature space as the let us say in this case it was the 2 dimensional real space that was the feature space. So, if this is the feature space then we can use the Euclidean distance itself as the distance measure in not only 2 dimensional if it is any high dimensional real space also we can consider But, it might happen that the feature space which we are considering is not two dimensional real, but it is something else.

So, let us say or like they are all binary vectors every feature vector is a binary vector. So, then we need to compare two binary vectors and we know that hamming distance is one such measure. Apart from that there are other measures like Mahalanobis distance or Manhattan distance and other like in case the space is not Euclidean we can we have to think of other distances also. But as of now let us keep things simple for the sake of this particular course it will mostly suffice to consider Euclidean distances. So, we will stick to that.

So, now coming to the formal definition of clusters. So, here let us consider a set of points here. So, now you can see that if we can consider these subset of points, then the mean distance between every pair of points in this subset is turns out to be 2 let us say. And, if you similarly if you consider this subset of another 4 points and you consider the pair wise distance between every pair like each pair of these points. So, obviously, 4 points are there.

So, 4 c 2 equal to 6 pairs can be calculated for every pair we can calculate the distance let us say the Euclidean distance and then take the mean across the 6 pairs and let us say that mean turns out to be 1.5. So, which means that the like if you consider any two of these any two points in this set their mean distance is 2 any two points you consider in

this set their mean distance is 1.5. But now if you consider.

point from this subset another point from this subset and you calculate the distance between them. Now, again so you can like there are 4 of them here and there are 4 of them here. So, total 16 pairs you can choose. Now, if you like let us say we calculate these 16 distances and calculate their mean and then this mean works out to be 6. So, like we can say the average distance between a point in this subset and another point in this subset is 6.

So, we can see that like these two subsets have this property that the points within each subset are quite close to each other their mean distances are like 1.5 or 2, but if you consider two points which are across the two subsets then their mean distance is quite far from each other like 6. So, we can in this case we will say that these two are clusters. So, then what is a cluster? So, cluster is like basically it is a subset of the observations such that the intra cluster the mean intra cluster distance is low, but the mean inter cluster distance is high. That is if you like I will I can given a set in set of n points of course, I can divide them into or partition them into any number of clusters in like many different ways it is a combinatorial problem how many clusters can be defined.

But like not all of them we will accept as a valid cluster clustering. Clustering will be considered to be valid only if this property is satisfied. That is so what is a valid clustering? A valid I will say that a particular clustering or a partitioning of the points is valid if the intra partition or intra cluster distance is low and the inter cluster or inter partition distance is high. So, that is the general criteria of defining the clusters. So, now, the we must look for algorithms using which we can find this kind of clusters.

So, we are provided with the data with the raw data like this we see that all the observations like all over the place. So, then how do you partition them into clusters, so that the property or the criteria which we mentioned is satisfied. So, these are the clustering algorithms. So, our first clustering algorithms is what we call as agglomerative clustering very simple heuristic idea. The idea is that let us say that we consider the different points are all like all the different points are they are own clusters that is there is no as of now no two points are part of the same partition some somewhat like this.

Now, what is our aim? Our aim will be to identify pairs of points which are quite close to each other with respect to some threshold and we put them together like this. Like these two points may be close to each other, these two points are close to each other, these two points are close to each other. So, we put them within the same cluster. But, note that there are still there are some many points which have not been covered in these clusters and so, what do we do about them right. And so, for this purpose we now to do anything further we must consider ways of computing distances between a point and a cluster or

between two clusters.

So, far we know how to calculate distance between two points like two Euclidean distances, but how do we calculate the like let us say I want to see which cluster this point should define it should join should it join this cluster or should it join this cluster or should it not join any cluster at all. Similarly, the same question may arise for this point. So, we need to see how to compare every point with a cluster. So, there are again certain criteria which are known as the linkage criteria. So, there are multiple linkage criteria, one is a single linkage which is basically the minimum distance between two points from the two sides.

So, let us for generalization let us just consider that there are two clusters. The first one has these n points which we are calling as a_1, a_2 etcetera. So, this is we can call this as cluster A and similarly there is a cluster B which also contains another m number of points. So, then we are interested in calculating the distance between these two clusters. Note that we may like early like when we are talking about calculating distance from one point to a cluster as we are just doing, we can just assume that one of let us say cluster A has two points and cluster B has one point and so on and so forth.

So, the points which are like that is so, we are trying to define the distance between any two clusters and as I said we have certain criteria known as the linkage criteria. So, first is the single linkage criteria that is the minimum distance between two points from the two sets. So, there are m points here there are n points here. So, total m cross n distances can be calculated by considering any one point here from here and another point from here and calculating their this Euclidean or whatever distance. So, total m into n pairs of such points we can form calculate their distances and then among those distances whichever is minimumI consider that as the distance between these two clusters.

So, that is called the single linkage criteria. Multiple linkage criteria is just the opposite instead of the minimum we just consider the maximum distance between any two points in the two sets. So, like in case of single linkage criteria Even if the points of this cluster are more or less far away from the points of this cluster, if there exists one pair of points such that they like one is in this cluster, the other is in this cluster and they are quite close to each other, then also I will consider these clusters are quite close to each other using the single linkage criteria. But using the multiple linkage criteria like it is just the opposite, even if all the other points of the two clusters are relatively close to each other. If there is one pair of points from one from this cluster and another from this cluster which are quite far away from each other, I will say that the these two clusters they are quite far apart from each other.

So, we can say that single linkage is a lenient criteria for calculating distance from each between two clusters. Even if there is only one pair of examples which are quite close to each other we will say the entire clusters are quite close to each other, while multiple linkage is like the more strict criteria. So, even if all the other points are quite close to each other if there is one pair of points which are far apart from each other. we will say that the entire clusters are far apart from each other. And then between them there is the average linkage criteria which is the mean distance between the any two points from the two sets.

So, now as we said earlier we initially are considering each point as a separate cluster, we are going to merge a pair of clusters if they are closer than a threshold and we keep on repeating this till no more mergers are possible. So, this is the initial situation all points are their own clusters, then we merge those points which are close to each other. Next we calculate the distances between like of points from sets as according to the linkage criteria. And, we say like let us say in this case this point is close in like is I mean the distance between this point and this point is 4. Now, the distance from this point and this point may be higher, but we because we are using the single linkage criteria we will say that the distance of this point to this cluster is 4.

Now, we are considering a threshold of 5. So, since 4 is less than 5, I will say that this point is close enough to this cluster. So, I will merge them. Similarly, these two points their distance is 3 which is less than the threshold 5. So, which we will merge them.

Now, if you consider these two clusters. So, like both of them have two points each. Now, from here to here from this point to this this point the distance can be quite large. Let us say 8 or 9 which is more than the threshold. But there is also this pair which whose distance is only 2 which is less than the threshold 5. So, if you are using the single linkage criteria we will focus on these 2 and not on the 5I mean not on the 8.

So, since 2 is less than 5 I will say that these 2 clusters are close enough to each other and I will merge them. But if I were considering the complete linkage then I would or multiple linkage then since this is the like 8. So, I will consider this 8 and not the 2. So, since 8 is greater than the threshold I will not merge them together. But assuming that we are using the single linkage criteria like we merge them and this is what the clustering looks like.

Now, when does this process terminate? Now, the process terminates if no further mergers are possible like let us say we have reached this situation. Now, this one point is sitting alone it is not joined any cluster so far. So, now, I calculate the distance from this to all the other clusters and let us say that this distance that distance from this point to this cluster according to the single linkage criteria is 6. Let the distance from this to the this

cluster may be even higher let us say 8 or 9 and from this point to this cluster is also quite high let us say 7 or 8. So, none of these point to cluster distances are within the threshold of 5.

So, we cannot link any further. So, we just stop it stop here. Similarly, we can also we will also have to see whether this cluster and this cluster can be joined. So, for that we will have to consider the pair wise distances between these points. So, it may turn out that the closest distance between these two clusters is the distance between this point and this point we are between this point and this point, but those distances are also more than this threshold of 5.

So, we do not join these two clusters. The same situation arises between these two clusters also the distance is closest between these two pair, but this distance also turns out to be let us say slightly more than the threshold of 5. So, we do not merge them any further. So, we say that that algorithm has terminated. So, this is the idea ofagglomerative clustering. The other another clustering technique which we will discuss is the prototype based clustering.

So, here given k which is the number of partitions which which you want to create by the way I must also say point out here that in this casein case of agglomerative clustering you do not beforehand specify how many partitions you are going to create. You just specify a threshold based on which the pairs of clusters or points are going to be linked or not. In this in the case of prototype based clustering you do not specify this threshold of distance, but instead you specify how many point or how many clusters you are going to create. So, now, the your task is to construct a partition of these m objects in or m observations into these k number of partitions or clusters which we can call as $C_1, C_2, \ldots, C_k$. So, now, the interesting thing is that each of the clusters will be represented by what we call as a cluster mean denoted by mu 1, mu 2 up to mu k.

So, these are the cluster means is what we call as the prototypes and our like when we are considering prototype based clustering algorithms our target will be to calculate not only the clusters, but also these cluster means or the prototypes themselves. Once we identify the prototypes it will be easy to assign every point to observation to one of these prototypes. So, the this here comes the famous K-means clustering algorithm which was proposed as back as late as I mean as early as 1967 which is like more than 50 years earlier. So, the given K the initialization is you randomly choose K data points and assign them to be the initial cluster centers or the prototypes. That is there is now in general there is no rule that the observations themselves will have to be the prototypes, but in this case we initialize the some of the prototypes with some these any k of the observations.

I mean in general we will have much more observations than k. So, we just randomly pick some of them and call them as the prototypes. Next, comes the second step which is the cluster assignment. Now, each of the other observations which have not been called as the prototypes, we like we try to assign each of those data points to the closest cluster center or prototype. So, the closeness in this case will be determined based on the Euclidean distance.

And once this cluster assignment has been done, then we recalculate the cluster centers or the prototypes using the current cluster memberships. And we keep on doing this process till we reach some kind of a converging criteria. So, let us see a run of this algorithm. So, let us say that these are all the points and we have selected these three points as the initial set of clusters. So, I have decided that three clusters or three partitions are going to be formed and the all the points are represented once again in the 2D feature space.

So, these are the initial clusters using which we are which we have selected for the initializations. So, the second step is to assign the remaining points to these clusters. So, I take this point for example, and I calculate its distance to all the three clusters this, this and this. And among them this cluster I mean this point this cluster mean prototype turns out to be the closest.

So, I just assign it to this cluster. Similarly, this point again I calculate its distance to this center, this center and this Now, obviously, this one turns out to be the closest. So, I add like link it to this point. So, these two are now part of a cluster. Similarly, this point I see that it is closest to this prototype compared to this or this.

So, this is also joined to this point. Same case with this, this and this. On the other hand, these three points like I find that they are closest to this particular centroid. So, they are all assigned to the like or linked to this centroid. So, we can say that these two they form one cluster for the time being these all of these point they form another cluster for the time being and these also form another cluster for the time being four of these ok. So, there are three clusters of varying sizes this has only two members, this has four members and this has six members.

Now, like we have now that we have identified these temporary clusters, now we have to see suitable prototype for these clusters. So, these I called as prototype, but that was done without any basis that was just for initialization, but now that these all these points have been selected as part of one cluster, maybe we should redefine the prototype. So, the prototype is defined as the mean vector of all of these points in which have been assigned to that cluster. So, maybe like so, since these are all part of the same cluster.

So, like I calculate the mean feature vector of all of these 6 points. So, which will be some may be somewhere in the middle like this. So, note that this mean the or this prototype this is not part of the initial observations or initial data points. So, remember there were 6 data points in this cluster I have the the new prototype that has been created as the mean of their feature vectors. is not any of the data points, but a new one. Similarly, like in this cluster of two points I have calculated their mean which of course, lies in the middle.

Similarly, here also for these four data points I have calculated their mean which also lies in the like somewhat equidistance from all of these four. So, these are the different prototypes which we have found like this right. So, now these are the new clusters. So, now that we have identified or a re estimated the different prototypes now we may ask that ok.

So, the prototypes have shifted. So, earlier I was trying to assign every point to one of the prototypes, but the prototypes was at that time were these, these and these, but now the prototypes have shifted the prototypes is here, here and here. So, now the question is like these point the remaining points should they still be part of the same clusters where they are currently placed or should we now move them to a different cluster. So, I repeat this process I again calculate the distance of every point to the new clusters to the new prototypes and whichever is the closest I put them there I will assign them there. And accordingly once again I will have to recompute the cluster centers or the prototypes.

And this process go on till a certain criteria is met. So, before we come to the so, ok. So, what are these criteria? These criteria known as the stopping criteria when they are satisfied then this process stops. So, what can be stopping criteria? First of all if we see that there is no further change of cluster assignment of any data point. That is even after you have re estimated the cluster centers or the prototypes you see that there is no further change to be done. The all the points are like all the other points are remaining in the clusters where they currently are like in this case for example.

That is even if you calculate these three new prototypes, you will see that each of these prototypes are still the nearest to the points which were originally part of this cluster. Like for example, if you consider this point, it was earlier part of this yellow cluster when the cluster center was here. Now, even though the cluster center or prototype has moved here, still this yellow point is closer to this prototype than say this prototype or this prototype. So, this continues to be part of the yellow cluster. Similarly, the same with this point or with this point, this point was earlier part of the brown cluster when the cluster center was somewhere here, but even after the cluster center or prototype has moved to hereit still continues to be the closest prototype.

So, this continues to be part of thesay brown cluster same with this orange cluster also. So, we can say that the cluster assignments of any data points are not changing. Similarly, the like if we talk about the cluster centers or prototypes they are also not changing. I mean naturally if the clusters themselves do not change then of course, the cluster I mean the cluster centers or prototypes also will not be changing.

Also, another thing which we can consider is the intracluster distance. So, as we earlier were talking about in case of clustering there are two major criteria the intercluster distance and intracluster distance. So, we can measure the mean intracluster distance. So, now, at any point if we see that the mean intracluster distance has reached something like a local minima, it is not reducing any further then also we can consider it to be a stopping criteria. So, like like initially the let us say the points were like this only the three all the points were placed like this and the these three were the initial three cluster centers or prototypes selected randomly. Then in the second iteration we see that some of the points are assigned to the blue cluster, some of the points are assigned to the red cluster and the rest are assigned to the green cluster all based on proximity.

Now, accordingly we re-estimate the blue cluster center, the red cluster center, green cluster center and so on. And this process goes on till no further change is possible that is the algorithm has converged and then this is what the clustering looks like. We see that all these points they have they are forming one the blue cluster, these points they are for the red points they are forming as another cluster and these green points they are following another cluster right. So, this is how the clustering takes place. Now, this k-means clustering it is a simple algorithm, but it has certain problems which have to be I mean it is I mean it is not guaranteed to give you the clustering that you are looking for.

So, it provides you only the k-means for only a local optima that is it will give you a clustering, but you can it does not claim that that is the best possible clustering. Similarly, the success of this k means clustering also depends on thechoice of the initial clusters where you chose the initial prototypes to be. Similarly, the choice of k or the number of clusters that is also empirical I mean generally there is no particular way of knowing what value of or how many clusters you want to group these points into. So, like then there are heuristics. So, you typically do it for multiple values of k and you calculate the mean intra cluster distance in each of the cases after running the k means clustering.

That is you first carry out the k means clustering with k equal to 1. That means, all the points are to be placed in the same cluster. Then you do k means clustering with k equal to 2. So, it will convert somewhere. So, 2 clusters will be formed.

So, in both of them you calculate the intracluster distance and take the mean of them. Then you consider k equal to 3 again 3 clusters will be formed you consider the mean of

mean intracluster distance and so on and so forth. and then you will you plot all of these and you will find some like at some value it has converged. So, let us say like you see that here in this graph at k there is no particular distance or the no particular difference in the results for k equal to 4, k equal to 5, k equal to 6 and so on. So, maybe we can go with k equal to 4 or maximum k equal to 6 and we can stop here.

So, now, so this is the story of K-means clustering. So, how do we apply this to economics? So, there are many problems. So, first example market segmentation the aim here is we there are lots of consumers who are all like which may who may be behaving as some collective or collective traits may be visible. So, that aim is to group the customers into these clusters based on certain attributes of these clusters such as their age, gender, demographics etcetera. So, why do we want to do it because we may the any company may try to like instead of trying to attract individual customers which is more difficult they may try to attract or try to sell certain products which to with the purpose of attracting entire groups of customers.

Now, how do you identify such groups of customers by using clustering. Similarly, in case of economic policy and design or analysis, if you want to understand the impacts of different policies at local, regional or national levels, it may be predicted from experience of similar context such as GDP, employment rates, income level etcetera. So, also like many of these characteristics some of them are like are national level characteristics, some of them might be regional characteristics, some of them may be local characteristics like infrastructure and so on. Now, we can actually try to cluster different countries like as the let us say as the developed countries, as the developing countries and so on and so forth. Similarly, the regions or localities we can cluster them into let us say one category of cities, another category of cities or semi rural areas, semi urban areas, fully rural areas etcetera based on the let us say the different types of infrastructure or socio economic characteristics of them. And then like we may design policies which are targeted towards specific types of countries or specific types of cities and so on and so forth.

Another similar application lies in labor market analysis where we are trying to cluster the different groups of workers who are present in the labor force based on their skills, educational experience and other relevant factors. Let us say that like I have some jobs which I want to advertise to the labor market, but the labor market has so many different peoples. So, now it does not now not all the jobs are going to be suitable for all the people. So, I will identify certain sets of people who have similar characteristics and I will advertise my jobs to them. So, just to conclude the clustering is basically the task is to partition the examples into coherent and distinct groups requires the distance measure to compute the these kinds of similarities or to understand which individuals are similar to each other.

There are we discussed two algorithms, one is based on agglomerative clustering where we start with initial all the points are separate clusters and then we gradually merge them together. The other is prototype based k-means clustering like which is like highly popular and is based on Euclidean distance. However, it is a problem with k-means clustering is that it is sensitive to initialization and the number of clusters need to be pre specified. So, the with this we come to the end of this lecture. Thank you and in the coming lectures we will discuss more methods of how to extract knowledge from data and use them for making further predictions. Thank you.