Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 02
Lecture - 10

Lecture 10: Introduction to machine learning-III

Hello everyone. Once again come to the course. Today we will be covering the lecture 10. Which is Introduction to Machine Learning, the last part. In last two lectures we have introduced this topic machine learning which is the foundation of recommender system. So, in the first two of this we have talked about supervised learning, then today we are going to talk about unsupervised learning.

So, today we are going to talk about unsupervised learning and here we will be talking about clustering and PCA which we discussed at the time of dimensionality reduction we will consider it once again as a topic of unsupervised approach. Let us first try to understand what this unsupervised approach is. So, in case of unsupervised learning. Is it ok now? So, today we are going to talk about unsupervised learning and let us first see that how it is different from supervised learning.

Supervised learning is very well understood. In case of supervised learning we have a number of data points let us say 1 to n with p number of features and we also have. . Is it ok? Yes sir. You will restart this particular slide.

Yes, yes I will restart. So, let us discuss about unsupervised learning. So, far we have covered supervised learning and there we saw both regression as well as classification approach. Now, in both these settings supervised settings in the data set we have that attributes as well as one response variable. So, data if we consider if we consider the data which was a matrix.

And we had n number of observations with attributes 1 to let us say p and we had a response variable. In case of regression problem this response variable in case of regression problem this response variable can take any value any numeric value. But in case of classification problem we had this variable to be qualitative which was represented in terms of certain numeric, but in nature it was qualitative it was called class variable either it will be binary in nature I presence or absence of a response or it can take certain categorical value like 1, 2, 3, 4 etcetera in let us say in the Likert scale. Now, in case of unsupervised learning the situation is a bit different we really do not have this variable y what we have is this set of records and all these elements that we have has to be now we have we will be applying techniques on those features only. So, what are those features? We have this x p number of features.

So, x 1 to x p. So, these are the features which are measured for n number of observations. Now, the prediction is not them here because there is no associated response variable that I told you

just now. So, rather the goal is to discover the interesting things about the measurements on x 1 to x p. So, based on all these n observations we are supposed to find out certain interesting things.

So, what could be those interesting things? Is there any informative way to visualize the data? Can we discover the subgroups among the variables or among the observations? Now, when it comes to the first one to visualize the data in two dimension we can conveniently visualize the data, but if the data consists of multiple dimensions how are we going to visualize it? Up to three dimensions probability is beyond that it is extremely difficult. So, then when in recommender system we deal with very large data sets running the algorithm on this entire data set first of all will be very computationally expensive besides in this entire data set if it is possible to have certain groups which are very similar to each other then probably if we run the algorithms on these groups individually then our performance measures are going to be improved. So, in unsupervised learning we can apply techniques such as clustering and principal component analysis. So, let us go ahead. Coming to this cluster analysis this cluster analysis is the process of partitioning a set of data objects or observations into subsets.

Now, these subsets which may or may not be overlapping have similar objects. So, when we compare objects of one group with another within group similarity will be more and similarity with the other group will be fairly less. So, clustering can lead to discovery of previously unknown groups within the data. So, there are many types of clustering algorithms to start with the basic one is the partitioning method. In this partitioning method we mostly are aware of one which is k means besides that there are many more approaches.

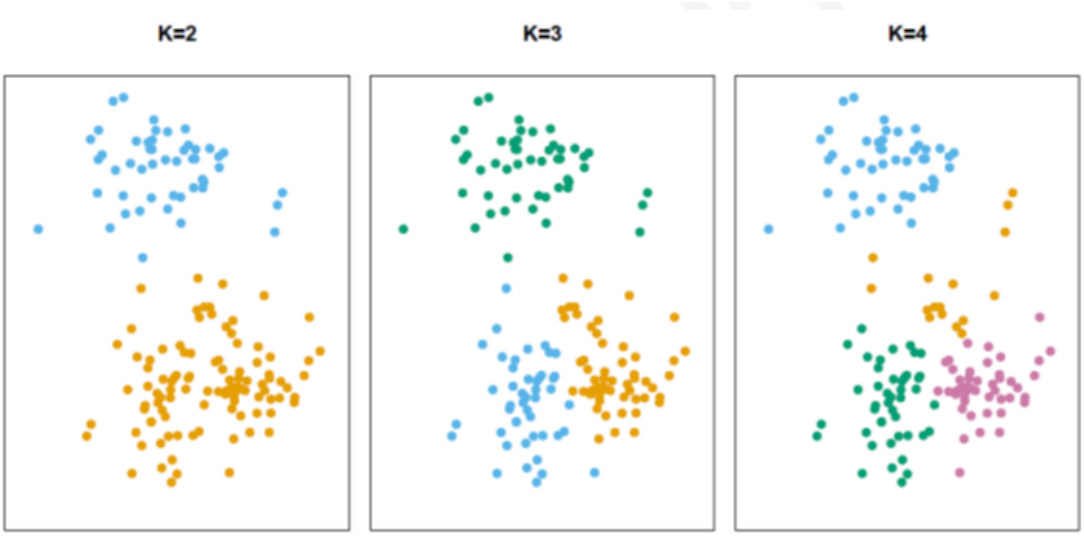
And coming to this partitioning based method if we have a number of let us say data points. So, then we will be able to group them let us say into 3 groups. So, in this particular context all the data points which are here will belong to either any one of this cluster. So, all this will be mutually exclusive this partitions are mutually exclusive and they can take any shape typically spherical in shape because they will be starting from something called a centroid. That centroid can be either min or medoid if the centroid is min it is a called k means otherwise it is k medoid.

They are very simple algorithms yet they are very computationally expensive. So, they are typically suitable for small or medium data sets. Next comes your hierarchical method. In case of hierarchical method you start with you can have 2 approaches may be you start with individual data elements and keep combining them based on certain measure. Or you can start from the entire data set and keep dividing them into various groups.

So, there are again many methods in this and the main advantage of this is in this tree based structure you can stop at any point. Suppose you stop at any point at this point you will have 2 clusters one is this second is this. Here suppose you stop at let us say some other point. So, at this you will have let us say 3 clusters at this point you will have 2 clusters and so on. So, you can either start with the entire group and keep on breaking there are different approaches for doing that or you can start from individual elements and keep on clubbing.

So, you have to link the objects with each other and what just now you saw it called a dendogram. But anyway we are not going to talk about hierarchical methods we will be focusing only on k-means later at a later point. So, the third one is the density based method. In case of

density based method we can find arbitrary shaped clusters and looking at the region where the objects are very densely populated. Let us say this is our space and so, we may get some cluster like this because this is densely populated here another cluster could be this another cluster could be this.



With this now you can see the points which are away from any densely populated region can be termed as outliers. So, these techniques can also help in filtering out the outliers. So, to tell once this can have arbitrary shaped clusters this clusters are in dense regions and this cluster density represents each point must have a minimum number of points within its neighborhood. So, that which means this one cannot be called as a cluster this one cannot be called as a cluster they are rather outliers. So, db scan optics denclu are some of this type of clustering algorithm.

Then the third one is your grid based approach. In case of grid based approach you can have the data with a higher resolution in the sense you can think of processing it with a higher dimensional space. So, multi space clustering is possible under this grid based method. So, these algorithms are typically very fast and independent of the number of data objects, but they depend on the grid size. So, how many grids you make matters here.

Then the third one is your distribution based method. In this distribution based method which are also called model based approach we need to a priori before clustering we need to know the distribution of need to have some idea about the distribution of the data. Typically Gaussian distribution is assumed and let us say we know the distribution and we have some data points let us say these are the data points. Now, look at this if we have this some point this may belong to this distribution as well as we go away from the mean may be probability of being included in this is less, but there is some probability. Similarly, this can also be part of this one and not only this a point like let us say this one can be part of this cluster as well because if we look at the spread of this and if this is Gaussian it can go to any extent.

So, therefore, probability may be less, but this particular point which is very much within this particular cluster can also be part of the second cluster with some probability. So, anyway this

distribution based methods can give more explanation why certain points belongs to some cluster. Now, with this let us move ahead and look at one specific partition based clustering algorithm. This is a very well known algorithm and most of the courses that will you will be taking on may be you will be may be taking on machine learning. This is the first algorithm probably in clustering you might be studying.

1. $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \ldots, n\}$ — each observation belongs to at least one of the K clusters
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$ — clusters are nonoverlapping

So, it is a very simple and elegant approach. It partitions the data into k distinct non-overlapping cluster. So, just like I discussed in case of density based the partition can be overlapping. So, which means a particular point can belong to any of the cluster with certain probability. So, such kind of thing is not present here.

You will be a particular point will be either in a group in a specific cluster or it is not and this is very very as a result this becomes very very sensitive to the outliers. Let us say in this you have a point this one. It appears that this does not belong to any cluster. However, the algorithm somehow will put it either in this cluster or this cluster depending on to whichever it is the distance is less. Let us say for this one the distance is probably to the let us say this is the cluster center here, this is the cluster center here and and the point is become little let us say this is the point.

So, this is the distance and visually we can see this distance is more this is less. So, even if this is not a part of this appears to be an outlier this will become part of this yellow cluster. And as we move ahead we can see as we increase the value of k which is a user input here we will be accordingly getting more number of clusters. So, here 3 this is here 4. So, you can keep on moving with you can keep on increasing the value of k and get more number of clusters.

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\} \quad W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

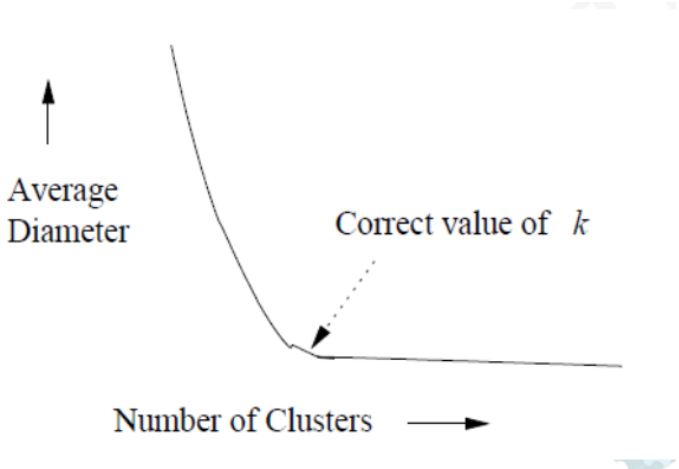
$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

Now, what is the foundation for this K-means? This K-means clustering procedure results from a very simple intuitive mathematical problem. Let C 1 to C k denote the set sets containing indices of the observations in each of the cluster. Now, this set satisfies two properties what are the C 1 and C 2 up to C k? They are the clusters containing the points. Now, if we take the union of all these clusters this has to be n and intersection of any two cluster is phi. So, which means a

particular point cannot belong to two clusters or in other words the clusters are non-overlapping and that is what we saw here the clusters are non-overlapping they are distinct.

So, this particular point you cannot say this belongs to green or there is some with some probability this belongs to blue with some probability this belongs to yellow to green. No, it is distinctly told here by this algorithm that it belongs to this blue cluster. Now, these clusters are non-overlapping. Now, what is the mathematical intuition behind this? The idea behind K-means clustering is that a good clustering is one for which the within cluster variation is as small as possible and most common choice for within cluster variation is actually Euclidean distance. So, our idea is you have to minimize this.

So, which means you have to distribute the points in a manner. So, that this within cluster sum of the within cluster variation among all the clusters when summed up this should be minimum. Now, extending it little further we can see that K which is a variable here we can find out this sum for different values of K and definitely for certain value of the K this is going to be the minimum among all. So, that may be the right value of K. Now, what is as it has been told the the most common choice so far is the Euclidean distance.



In that sense, other distance functions as we move move ahead to the next lecture we will be talking more about the distance functions, but Euclidean distance is widely adopted choice, but other distance functions cannot be ignored as well you can take other distance function as well. So, this c k can be computed considering the distance of the points within the cluster. So, in essence this distance based function has to be now minimized. This is the iterative algorithm for realizing this optimization problem. This optimization problem if you try to solve it numerically probably it is becoming it will become complex, but there is a very efficient algorithm which solves this in an iterative manner.

And here we randomly assign a number from 1 to K to each of the observations. So, these observations serve as the initial cluster assignments for each. Now, out of this sum will be getting value 1, sum will be getting value 2 and so on. So, the points which has certain cluster value they will be sum together and you can take the cluster centroid. So, you come to iterate until the cluster assignment stops changing.

So, what you have to do in each step? For each of the k clusters compute the cluster centroid, the kth cluster centroid is the vector of the p feature means of the observation in the kth cluster. So, what does it say? What does it mean? So, when we talk about a data point, data point consists of many features. The example that we saw we were representing we saw here we represent the data points are in two dimension, but this need not be the case. Because we are considering Euclidean distance we can consider the distance between two points even if their dimensions are more. So, let us say x 1 is one point, x 2 is the second point and both of them will have p dimensions.

So, x 1 1, x 1 2, x 1 p, x 2 1, x 2 2, x 2 p. So, which means now individually attribute wise we can compute this distance function. We can compute this distance function for taking each attribute, comparing each attribute at a time and we can find out the Euclidean distance. So, taking all the attributes together if it has let us say some l number of observations in this let us say this is cluster 1, then you are supposed to take the sum of each x i 1, x i 2 and find out the mean by dividing l. Okay? So, once you find out this then from this cluster center you find out the distance of all the points start from this centroid. So, there how many centroids you are going to get? p such k such centroid you are going to get because your input was k.

So, find out and once again do reassignment. Again repeat the procedure find this out do reassignment and so on. So, now the question is how many clusters we must consider? What is the right value of k? As I have told you in this particular setting we are supposed to minimize this for a particular k. So, which means there might be existing some k for which this is this will be minimum among the all besides that there is something approach which is popularly called the elbow approach. To understand this we have to know what is the two different things. One is the radius of a cluster and diameter of a cluster.

Now, radius of a cluster is the maximum distance between all the points and the centroid. Now, diameter of a cluster is the maximum distance between any two points of the cluster. So, now for all the clusters you start with let us say minimum number of clusters that is to find out the diameter make it three find out the diameter and so on. As the number of clusters increase this diameter value is going to decrease. And at certain point of time there will not be any substantial change.

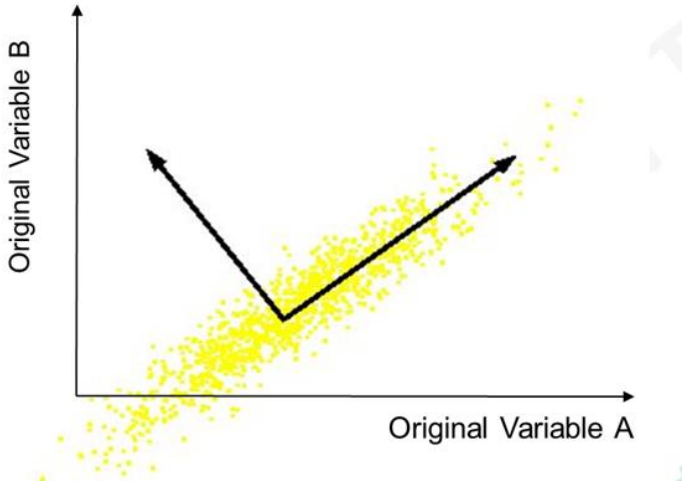
So, this may be considered as the correct value of k, but this anyway there are many other heuristics which also exist. And seldom people will actually be using because it is a very expensive algorithm people would not be actually changing all the time the value of k and testing this elbow. The probably they will be using some kind of intuition, but this is the right approach. Now, what is the role of this clustering in recommender system? So, mostly in recommender system this clustering is used as a preprocessing step. Where exactly it you will be doing this preprocessing? See, when we are talk about recommender system it is a it consists of the basic recommender system consists of three matrices what are they? User matrix, item matrix and one preference the rating matrix.

So, all these three matrices are large matrices. Suppose, there are some recommender systems which work only on only on the rating matrix. In that case is it possible that we have to deal with this entire matrix at a time? It we may even think of certain hybrid approach in which will be categorized the items based on where the items are to get items are rated together. And the small

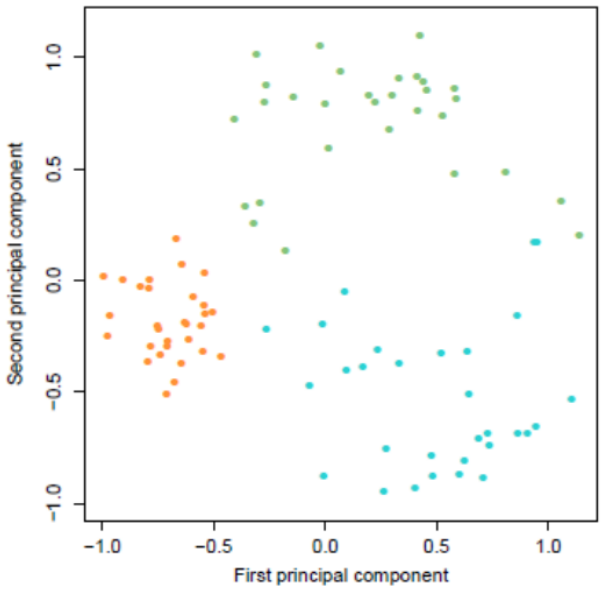
clusters we can now focus on. Similarly, items based on the item feature we can cluster users based on the user feature we can also cluster depending on the technique that we are going to use. The second one under this unsupervised approach which we have already studied is principal component analysis.

We have studied principal component analysis in somewhat little bit in depth in some earlier lecture, but try to remember how it was working. We had some data matrix M with n objects and p number of features. And what we did? We transformed it to a new matrix say M dash with n observations, but some r features where this r was less than p. And this all the features that you had here are called principal components.

So, you had let us say you decided two features. So, you will have PC 1 and principal component 2. So, the data which was in p dimension now you represented it in two dimension. Now, remembering little bit more actually what we did? How did we discover this? By rotating the original axis in the direction of maximum variance for this was just for your just to remember let us say in two dimension this is the distribution of the data points these are the distribution of the data points and we tilted this. And how did we tilted this? We this new axis where on the on one of the axis the variance was maximized and it kept on decreasing as we move ahead with PC 1, PC 2 etcetera was actually the eigenvectors. And we also studied how to I mean of course, there are many efficient algorithm, but in the basic form we also studied how to find out the eigenvalues then from the using characteristics equation then finding the eigenvectors.



Once we find the eigenvectors using those eigenvectors we multiply with the original data matrix and we get the principal components. Now, what is this use of this principal component? One of the basic question that we asked in the beginning of this lecture is unsupervised methods work only on the attribute part. So, in case of principal component also we used only this p attributes not the variable. Of course, this principal components we can use in case of supervised approach where we have this latent features PC 1 and PC 2 and we also use the response variable y and put together we can have our data set with latent features and response variable. So, this we can use our training data and build our supervised model, but in case of unsupervised approach, but this is PCA itself is an unsupervised approach because this works only on this data matrix m on the p features.



Now, what is the use of this in case of say recommender system? One of the benefit of this PCA of course, we have already seen that using PCA we can find out the latent features specifically when we are dealing with let us say item feature. For example, let us say the item is that of news item or the items are the news articles. So, we have text features and we will have thousands and thousands of text features. So, out of that finding the major directions can be possible using principal component analysis. Besides that, if we would like to visualize the data which is in higher dimension in two dimension, we can simply plot the first and second principal component.

And looking at how these points are distributed, we can in fact find out the clusters based on the variability that is observed within the data and that is captured through your principal components. So, we can use this as a data visualization tool visualization tool, identify the cluster and probably based on this cluster you can partition the data and carry on your activity. So, these are the two sources from which I tried following. Of course, I have not mentioned that statistical learning also used I mean the as one of my resource, but anyway that book is already mentioned in some other lecture, but I have taken some examples from that one as well.

So, this is my concluding remark. So, we have studied about the unsupervised approaches. We saw that the approaches which work only on the attribute order picture vector and do not utilize response variable in any sense are called clustering are called unsupervised approaches. Clustering PCA are two such approaches there can be many more. Typically these two are quite prominent approaches. We studied we had the understanding of the basics of different types of clustering algorithms and we tried looking at K-means in a little bit more depth and we understood that PCA which typically is a dimensionality reduction approach is a unsupervised approach as well and we can use PCA to visualize higher dimensional data in two dimension using principal component 1 and 2. Thank you everyone.