

Machine Learning of Soil and Crop Management
Professor Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology, Kharagpur
Lecture 19

Applications of Classification and Clustering Methods in Agriculture (contd.)

Welcome friends to this nineteenth lecture of NPTEL online certification course of Machine Learning for Soil and Crop Management. And in this week, this is week 4 and in this week we are discussing the Application of Classification and Clustering Methods in Agriculture. So, in our first 3 lectures, we have discussed some important classification methods, we have differentiated classification column clustering method.

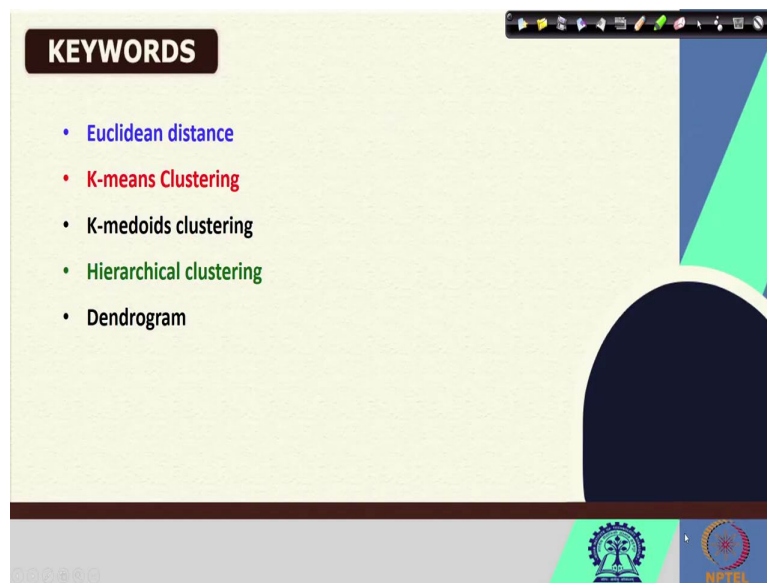
Some important classification methods, we have discussed like linear discriminant analysis, then logistic regression, we have also discussed the classification 1tree and then we have discussed the support vector machine based classification and we have seen some examples of classification algorithm application in agriculture specifically focusing on soil and crop.

We have also discussed the differences between classification and clustering, classification is supervised, whereas clustering is unsupervised. In case of clustering methods, we want to identify some important trend in the feature space itself without the help of any target parameter or target variable or dependent variable. So, we have seen the broad classification of clustering also, one is partitioning method another hierarchical methods.

And remember in case of partitioning methods, we partition the data into non overlapping subset or clusters. Whereas, hierarchical method, we hierarchically just like a tree, we split the data into the subclass or clusters. So, all this classification of in the clustering is based on all these grouping in the clustering is based on some 10 kind of similarity and dissimilarity measures. And we are going to discuss those similarity and dissimilarity measures in today's lecture in details.

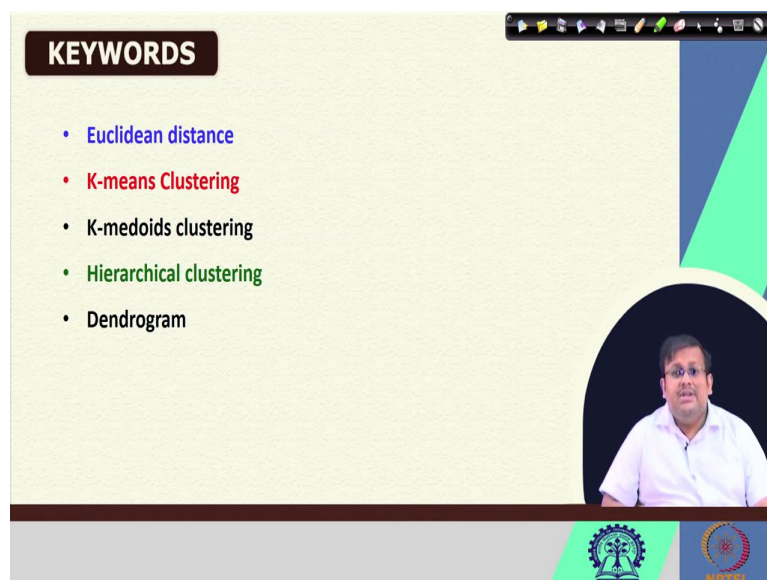
And then we are going to discuss one of the most important clustering method that is K-means clustering, and also we are going to discuss another important clustering method that is K-medoids clustering, and we will see the difference between K-means clustering and K-medoids clustering.

(Refer Slide Time: 3:03)



So, these are the concepts which we are going to discuss today, dissimilarity first of all we are going to discuss the dissimilarity measures in clustering. Also we are going to discuss the K-means clustering and also the next we are going to discuss the K-medoids clustering and finally, we are going to discuss the hierarchical clustering.

(Refer Slide Time: 3:26)



These are the three key words for today's this lecture, number 19 we are going to discuss Euclidean distance, K-means clustering, K-medoids clustering, hierarchical clustering and also we are going to see an example of Dendrogram.

(Refer Slide Time: 3:43)

MEASURING DISSIMILARITY IN CLUSTERING

- ▶ Sometimes the data is represented directly in terms of dissimilarity/similarity. Dissimilarity metric is expressed in terms of a distance function, which is typically metric: $d(i, j) = d(j, i)$ and $d(i, i) = 0$. If not symmetric, can be replaced by $(D + D^T)/2$.
- ▶ Dissimilarities based on attributes X_1, X_2, \dots, X_p is often defined by $D(x_i, x_k) = \sum_j w_j d_j(x_{ij}, x_{kj})$. Common choices of $d_j(\cdot)$ are:
 - ▶ Squared distance: $d_j(x_{ij}, x_{kj}) = (x_{ij} - x_{kj})^2$. More emphasis on large differences than smaller ones.
 - ▶ Correlation $\rho(x_i, x_k)$:

$$\rho(x_i, x_k) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{kj} - \bar{x}_k)^2}}$$
 with $\bar{x}_i = \sum_j x_{ij}/p$ (average over variables). If input is standardized, then $\sum_j (x_{ij} - \bar{x}_i)^2 \propto 2(1 - \rho(x_i, x_k))$. Hence clustering based on squared distance is equivalent to correlation.
 - ▶ Absolute error $d_j(x_{ij}, x_{kj}) = |x_{ij} - x_{kj}|$. Compared to squared distance, it is more robust to outliers.
 - ▶ Discrete variables: ordinal variable uses ranks; categorical variable with M categories needs a $M \times M$ distance matrix.

MEASURING DISSIMILARITY IN CLUSTERING

- ▶ Setting equal weights w_j for each variable does NOT necessarily give all attributes equal influence. If $w_j \propto 1/D_j$, then equal influence.

$$D_j = \frac{1}{N^2} \sum_{i=1}^N \sum_{k=1}^N d_j(x_{ij}, x_{kj})$$

- ▶ In general, setting $w_j = 1/D_j$ is often recommended, however it can be highly counterproductive. Some attribute value differences may reflect greater actual object dissimilarity in the context of the problem domain.
- ▶ Variables that are more relevant in separating the groups should be assigned a higher influence in defining object dissimilarity.
- ▶ Choosing an appropriate dissimilarity measure is FAR MORE IMPORTANT than the choice of clustering method!
- ▶ Miss values:
 - ▶ Omit observations with missing values.
 - ▶ Numeric: impute the missing values by mean or median.
 - ▶ Categorical: treat "missing value" as a new category.
- ▶ Standardize variable to mean zero and unit variance.

Now, let us see what is the measurement how we can measure the similar dissimilarity in case of clustering methods. So, sometimes the data is represented directly in terms of dissimilarity or similarity measures or some indices and these this similarity metric is expressed in terms of a distance function between the data.

Now, which is basically it typically metrics so, you can see here the distance between point i and j is kind of symmetric and so, $d(i, i) = 0$. So, if not symmetric then we can this can be replaced by these d plus d transpose divided by 2. So, this dissimilarity based on attributes which are X_1, X_2 or variables is often defined by this by this index called D capital D X_i by X_k where it is basically a some multiplication of weights and distance between the observations.

So, common choices for these distance measurement are as follows. First, we can calculate the squared distance. So, squared distance here the distance is basically squared, but remember that when we squared the distance more emphasis is given on large differences than the smaller ones because it is a squared term.

Also one another major important dissimilarity measure is correlation based on the Pearson correlation. So, we know the correlation already we have discussed this formula of the correlation where these \bar{X}_i equal to summation of X_{ij} by P which is average of all the variables. So, if inputs is standardized then the summation of $X_{ij} - \bar{X}_i$ squared proportional to this term.



So, we can see mathematically that clustering based on the square distance is equivalent to correlation. So, sometime we use the correlation matrix also the correlation values also our correlation indices also to calculate the dissimilarity. Another metrics it is still another matrix is a dissimilarity matrix is the absolute error which is denoted by these d_{ij} , X_{ij} , X_{kj} . So, it is basically it takes this form where, so, compared to these squared distance it is more robust to outliers.

So, these the squared distance is not very much robust to outliers, because whenever there is an outlier and we are taking the square that influence the results, so, these absolute error d is more robust to the outlier and so, these are about the continuous variable. So, what about the discrete variables? So, discrete variable could be either ordinal variable or it could be categorical variable with M categories. So, ordinal variables generally use the ranks whereas, categorical variable with M categories need say M into M metrics.

(Refer Slide Time: 7:26)



MEASURING DISSIMILARITY IN CLUSTERING

- ▶ Sometimes the data is represented directly in terms of dissimilarity/similarity. Dissimilarity metric is expressed in terms of a distance function, which is typically metric: $d(i,j) = d(j,i)$ and $d(i,i) = 0$. If not symmetric, can be replaced by $(D + D^T)/2$.
- ▶ Dissimilarities based on attributes X_1, X_2, \dots, X_p is often defined by $D(x_i, x_k) = \sum_j w_j d_j(x_{ij}, x_{kj})$. Common choices of $d_j(\cdot)$ are:
 - ▶ Squared distance: $d_j(x_{ij}, x_{kj}) = (x_{ij} - x_{kj})^2$. More emphasis on large differences than smaller ones.
 - ▶ Correlation $\rho(x_i, x_k)$:
$$\rho(x_i, x_k) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{kj} - \bar{x}_k)^2}}$$
with $\bar{x}_i = \sum_j x_{ij}/p$ (average over variables). If input is standardized, then $\sum_j (x_{ij} - \bar{x}_i)^2 \propto 2(1 - \rho(x_i, x_k))$. Hence clustering based on squared distance is equivalent to correlation.
 - ▶ Absolute error $d_j(x_{ij}, x_{kj}) = |x_{ij} - x_{kj}|$. Compared to squared distance, it is more robust to outliers.
 - ▶ Discrete variables: ordinal variable uses ranks; categorical variable with M categories needs a $M \times M$ distance matrix.



MEASURING DISSIMILARITY IN CLUSTERING

- ▶ Setting equal weights w_j for each variable does NOT necessarily give all attributes equal influence. If $w_j \propto 1/D_j$, then equal influence.
$$D_j = \frac{1}{N^2} \sum_{i=1}^N \sum_{k=1}^N d_j(x_{ij}, x_{kj})$$
- ▶ In general, setting $w_j = 1/D_j$ is often recommended, however it can be highly counterproductive. Some attribute value differences may reflect greater actual object dissimilarity in the context of the problem domain.
- ▶ Variables that are more relevant in separating the groups should be assigned a higher influence in defining object dissimilarity.
- ▶ Choosing an appropriate dissimilarity measure is FAR MORE IMPORTANT than the choice of clustering method!
- ▶ Miss values:
 - ▶ Omit observations with missing values.
 - ▶ Numeric: impute the missing values by mean or median.
 - ▶ Categorical: treat "missing value" as a new category.
- ▶ Standardize variable to mean zero and unit variance.



KEYWORDS

- Euclidean distance
- K-means Clustering
- K-medoids clustering
- Hierarchical clustering
- Dendrogram

The slide also features a video feed of a man in a white shirt speaking, and logos for IIT Bombay and NPTEL at the bottom.

Now, so, in case of categorical variables, so, in case of categorical variable we can see that there is it needs an M into M distance metrics. Now, so, we know that the calculation of d we have seen is dependent on some assigning on some weights. So, setting the equal weights how we can assign the weights? So, setting the equal weights W_j for each variable does not necessarily give all attributes equal influence.

So, if W_j is inversely proportional to these capital D_j then we can get the equal inference. Now, this capital D_j is calculated by using this formula. So, in general settings we can see that W_j equal to one by D_j is often recommended. So, however, so, weight is in other words in a very generalized term I would say without considering this it is very generalized term.

If the distance is more then we assign lower weight and if the distance is high then I am sorry if the distance is high or more we assign lower weights and if the distance is low between 2 observations, we assign higher weights. So, in general setting these w_j equal to 1 by D_j is often recommended. However, it can be highly counterproductive also, because some attributes value differences may reflect greater of actual object dissimilarity as in the context of the problem domain.

So, if we use this type of metric sometime it is happening sometime it happens that the dissimilarity notion it calculates maybe much more much more influential than the actual difference between the samples. So, we need to be very, very careful about selecting the most important dissimilarity measure.

So, remember that variables that are most relevant in separating the groups should be assigned the higher influence in defining the objects dissimilarity. Now, choosing an appropriate, so, that is why I am emphasizing again choosing an appropriate dissimilarity measure is far more important than the choice of clustering method.

(Refer Slide Time: 10:31)

MEASURING DISSIMILARITY IN CLUSTERING

- ▶ Setting equal weights w_j for each variable does NOT necessarily give all attributes equal influence. If $w_j \propto 1/D_j$, then equal influence.

$$D_j = \frac{1}{N^2} \sum_{i=1}^N \sum_{k=1}^N d_j(x_{ij}, x_{kj})$$

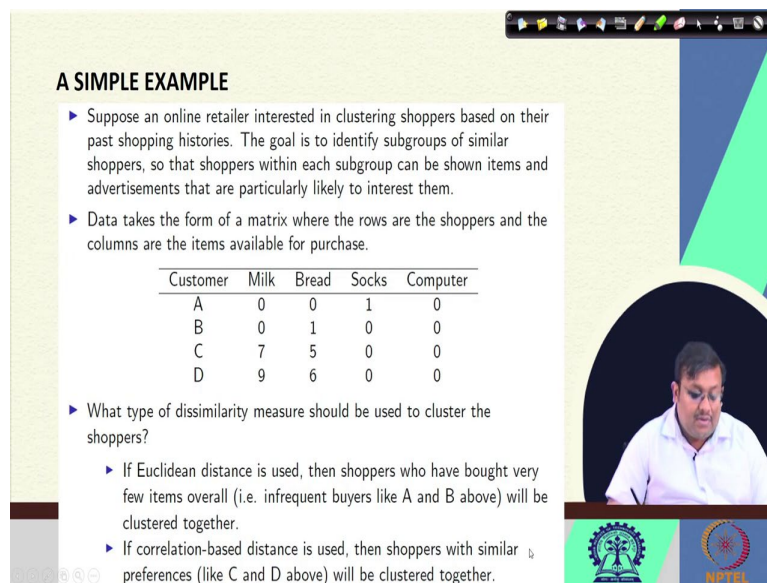
- ▶ In general, setting $w_j = 1/D_j$ is often recommended, however it can be highly counterproductive. Some attribute value differences may reflect greater actual object dissimilarity in the context of the problem domain.
- ▶ Variables that are more relevant in separating the groups should be assigned a higher influence in defining object dissimilarity.
- ▶ Choosing an appropriate dissimilarity measure is FAR MORE IMPORTANT than the choice of clustering method!
- ▶ Miss values:
 - ▶ Omit observations with missing values.
 - ▶ Numeric: impute the missing values by mean or median.
 - ▶ Categorical: treat "missing value" as a new category.
- ▶ Standardize variable to mean zero and unit variance.

So, you give more importance for choosing the appropriate distance measure as compared to choosing the appropriate algorithm for clustering. So, this is very, very important. And if there are some missing values, you should remove those observations before going for clustering, you should or calculating these dissimilarity measures, if there are some missing values, you should omit those value values or observations with the missing values and you can either input the missing values by taking a mean or median value.

So, you can you can either omit the missing value or you can input the median or mean value in that missing values in the place of that missing values or in case of in case of categorical variable, you treat the missing values as a new category. So, these are the 3 ways through which you can manage the missing observations or missing values in observation when you are going for calculating these dissimilarity measure for clustering.

And also as I have already mentioned couple of times you need to standardize the variable to mean 0 and you need variance we have also seen it previously. So, these are different types of dissimilarity measures in place of clustering.

(Refer Slide Time: 12:17)



A SIMPLE EXAMPLE

- ▶ Suppose an online retailer interested in clustering shoppers based on their past shopping histories. The goal is to identify subgroups of similar shoppers, so that shoppers within each subgroup can be shown items and advertisements that are particularly likely to interest them.
- ▶ Data takes the form of a matrix where the rows are the shoppers and the columns are the items available for purchase.

Customer	Milk	Bread	Socks	Computer
A	0	0	1	0
B	0	1	0	0
C	7	5	0	0
D	9	6	0	0

- ▶ What type of dissimilarity measure should be used to cluster the shoppers?
 - ▶ If Euclidean distance is used, then shoppers who have bought very few items overall (i.e. infrequent buyers like A and B above) will be clustered together.
 - ▶ If correlation-based distance is used, then shoppers with similar preferences (like C and D above) will be clustered together.

And a simple example we can see here suppose, an online retailers interested in clustering shoppers based on their past shopping histories. So, the goal is to identify subgroups of similar shoppers, so, that shoppers will each subgroup can be shown items and advertisement that are particularly likely to interest them. So, here you can see the data take the form of a matrix, this is a matrix form where the rows are the shoppers.

So, these rows are the shoppers or customers A, B, C, D customer and the columns are the items available for the purchase some milk, bread, socks and computer. So, what type of dissimilarity measures should be used to cluster this shoppers? So, another dissimilarity measure is called the Euclidean distance which we are going to discuss.

Now, Euclidean distance if we see that Euclidean distance for measuring the dissimilarity measure, then shoppers who have brought very few items overall that is infrequent buyers, like A and B, so A and B just bought only one-one item. So, if Euclidean distance we use, then shoppers who have bought very few items overall will be clustered together and if correlation is distance is used, then shoppers with similar preferences like C and D there they bought these milk and also bread will be clustered together. So, depending on which dissimilarity measures you are using, we can clustered differently.

(Refer Slide Time: 14:12)

EUCLIDEAN AND MANHATTAN DISTANCE

1. Euclidean distance:
$$d_{\text{euc}}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$
2. Manhattan distance:
$$d_{\text{man}}(x, y) = \sum_{i=1}^n |x_i - y_i|$$

So, Euclidean there are the 2 important dissimilarity measure, one is Euclidean distance and Manhattan distance and you can calculate the Euclidean distance in this fashion and Manhattan distance between 2 values using this formula. Now, this Euclidean distance or Manhattan distance are going to use in our different types of clustering methods which we are going to discuss.

(Refer Slide Time: 14:36)

A FIGURE SHOWS DIFFERENCE BETWEEN TWO DISTANCE MEASURES

FIGURE 10.13. Three observations with measurements on 20 variables are shown. Observations 1 and 3 have similar values for each variable and so there is a small Euclidean distance between them. But they are very weakly correlated, so they have a large correlation-based distance. On the other hand, observations 1 and 2 have quite different values for each variable, and so there is a large Euclidean distance between them. But they are highly correlated, so there is a small correlation-based distance between them.

So, if we see the 3 observations suppose there are 3 observation in another example you can see there are 3 observation 1, 2 and 3 with measurement on 20 variables shown, there are 20 variables measurement on 20 variables and also observation. So, we can see here this

observation let me point out this. So, here this observation 1 and these 3 have similar values for each variable. So, you can see they are almost similar values for each variable, the variables are starting from 1, 2, 3, 4, 5 up to 20.

So, we can see both this observation, observation 1 and observation 3 are having similar values for each variable. And so, they are have a large they are there they are having a small Euclidean distance between them. So, they have very small Euclidean distance between them. But they are very weakly correlated, we cannot see any correlation pattern in we are on these 2 patterns of observation 1 and observation 2.

On contrary, if we consider these observation 1 observation 2 although they have quite different values for each variable, observation 2 is having higher values then for each variable but so, they have very large Euclidean distance between them, but they are highly correlated if you see these pattern this orange pattern and this green pattern they are almost similar. So, from there we can see that although the Euclidean distance is less small, the correlation between them is high.

So, it shows the difference between 2 distance measures and this should be carefully observed before selecting that dissimilarity measure in an objectively way in an objective way, otherwise, you will may not get the optimum classification of your data set.

(Refer Slide Time: 17:31)

K-MEANS CLUSTERING

- ▶ The objective is to have a minimal "within-cluster variation" $W(C)$, i.e. the elements within a cluster should be as similar as possible.

$$W(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(i')=k} \|x_i - x_{i'}\|^2 = \sum_{k=1}^K N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2$$

where $\bar{x}_k = (\bar{x}_{1k}, \dots, \bar{x}_{pk})$ is the k th cluster *centroid* and N_k is the k th cluster size.

- ▶ Getting the global optimal solution above requires trying all possible assignments of the N points into K clusters. This number is huge! For example, to assign 25 obs. into 4 clusters has about 5×10^{13} possible assignments.
- ▶ The K-means algorithm:
 1. Randomly assign each observation to one of K clusters.
 2. Iterate until the cluster assignments stop changing.
 - (a): For each of the K clusters, compute the cluster centroid \bar{x}_k .
 - (b): Assign each observation to the cluster whose centroid is closest (in terms of the Euclidean distance).

So, let us start with the K-means clustering. So, K-means clustering is one of the most popular identity decision descent clustering methods and for partitioning a data set into K

distinct non overlapping cluster and to perform clustering, one must first specify the desired number of clusters K . So, here in case of K-means clustering, we first define the number of clusters that is just generally denoted by this K and then the K-means algorithm will assign each observation to exactly one of these clusters.

So, if we start with 2 cluster, three 3 cluster K , we can value we can vary the number of clusters that is K from 2, 3, 4 and so, on. Then these algorithm will assign each observation in the data set to exactly one of this of the one of the clusters. Now, since it is the iterative method, so the objective of this K-means clustering is to have a minimum within class variation, WC , which is denoted by this WC , and WC can be calculated by using this formula.

So, remember, in each iteration I will show you in a with a very good graph, the things will be clearer to you. So, in each iteration, we assign the sample each of the sample to one of these 2 clusters and we calculate these within cluster variation. So, these within cluster variation, so, basically we calculate the centroid for each cluster. And then, we see the distance between the individual data point to this cluster centroid to which it has been assigned and then we sum up the distance.

When we sum up the same for all the observation for all the clusters then we can get this value of WC where these \bar{X}_k is the k th cluster centroid and in case that k th cluster size, so, we do it for all the clusters from one to k and for each of these clusters, we get this value of the sum of the difference between individual observation to the cluster center corresponding cluster centroid. So, getting the global optimal solution above requests trying all possible assignments of endpoints into K clusters.

So, we have to do it K number of we have to assign all the N samples into K clusters. So, this number is really huge for example, to assign 25 observation in 4 clusters, they have about 5 into 10 to the power 13 possible assignments. So, what is the method of K-means clustering? So, in the method of K-means clustering, we randomly assign each observation to one of the K clusters and iterate until the cluster assignment stop changing.

So, in each iteration for each of the cluster compute the cluster center at which I have already told you and assign each observation to this cluster whose center is closest. So, in terms of Euclidean distance, so, here for K-means clustering we do the Euclidean distance this cluster calculation.

(Refer Slide Time: 21:54)

K-MEANS CLUSTERING

- ▶ *K*-means clustering derives its name from the fact that in Step 2(a), the cluster centroids are computed as the mean of the observations assigned to each cluster.
- ▶ Each of the Step 2(a) and 2(b) reduces the value of the within-cluster variation $W(C)$, so that convergence is assured.
- ▶ However, the result may represent a suboptimal local minimum.
- ▶ Because the *K*-means algorithm finds a local rather than a global optimum, the results obtained will depend on the initial random cluster assignment of each observation in Step 1. Hence, it is important to run the algorithm multiple times from different random initial assignment of clusters.
- ▶ Then one selects the best solution, i.e. that for which the objective $W(C)$ is smallest.

K-MEANS CLUSTERING

- ▶ The objective is to have a minimal "within-cluster variation" $W(C)$, i.e. the elements within a cluster should be as similar as possible.

$$W(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(i')=k} \|x_i - x_{i'}\|^2 = \sum_{k=1}^K N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2$$

where $\bar{x}_k = (\bar{x}_{1k}, \dots, \bar{x}_{pk})$ is the *k*th cluster centroid and N_k is the *k*th cluster size.

- ▶ Getting the global optimal solution above requires trying all possible assignments of the N points into K clusters. This number is huge! For example, to assign 25 obs. into 4 clusters has about 5×10^{13} possible assignments.
- ▶ The *K*-means algorithm:
 1. Randomly assign each observation to one of K clusters.
 2. Iterate until the cluster assignments stop changing.
 - (a): For each of the K clusters, compute the cluster centroid \bar{x}_k .
 - (b): Assign each observation to the cluster whose centroid is closest (in terms of the Euclidean distance).

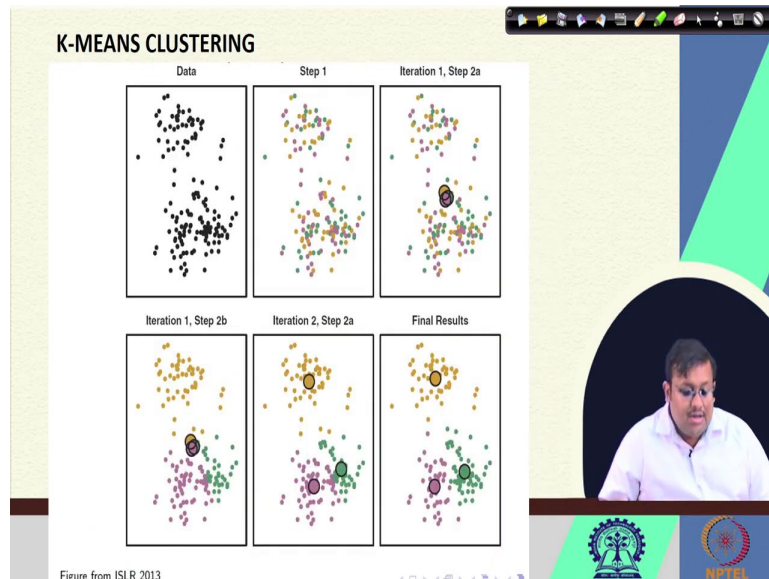
So, K-means clustering derives its name from the fact that in step 2a that is this step 2a for each of these K cluster, compute the cluster center at \bar{x}_k . So, for each of these 2 cluster, the cluster centroids are computed at a mean of the observation assigned to the to each of the cluster. So, we calculate the mean, so, that is the cluster centroid. So, that is why it is called K-means clustering. So, each of these steps 2a and 2b reduces the value of within cluster variation. So, that convergence is assured.

So, however, this result may represent suboptimal local minimum, so, because K-means algorithm finds a local rather than a global optimum, the results obtained will depend on the initial random clustering assignment of each observation in step one and remember that it is important to run the algorithm multiple times from different random initial assignment of

clusters. So, then one select the best solution that for the objective, these WC, which I have calculated is minimum.

So, our idea is to by iterating the same thing and by assigning the N number of observation to each of 1 of these K clusters, we will calculate this WC and each iteration will calculate the cluster centroid based on the Euclidean distance.

(Refer Slide Time: 23:31)



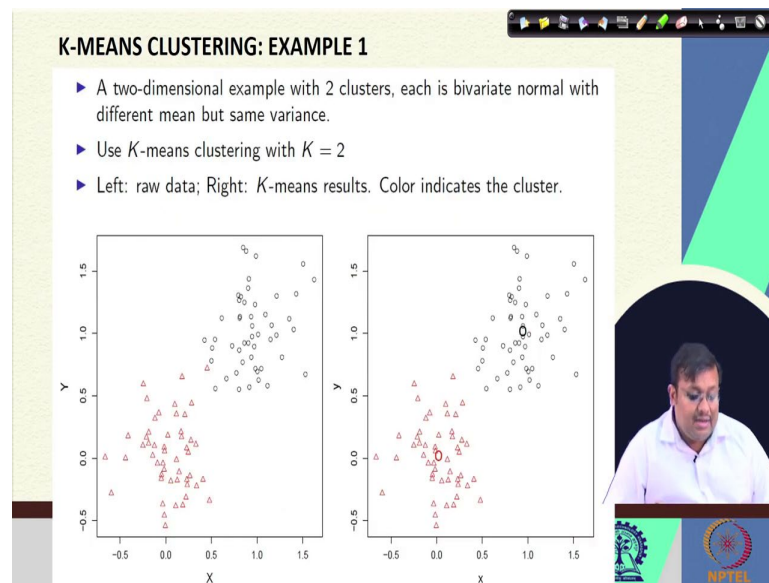
So, this is a very good example. So, let us see this for example this is the data and in the first step, first step we assigned the data into suppose there are 3 here we can see how many groups are there? 1, 2, 3 groups are there denoted by these orange, pink and green color. So, we have assigned randomly the samples into 3 clusters because we have here fix the K value as 3. So, we have assigned all the samples to one of these 3 clusters in first step iteration one and once we do that, we calculate their centroids by calculating the mean.

So that is why they are called the K-means clustering. So, in the iteration one, we can see this one and then we calculate the difference. So, in the iteration one, we can see the assignment of the cluster centroid here. So, when the cluster centroid are here we can see the different distance Euclidean distance from the samples to the centroids our maximum or high but as we go on for multiple iteration, we can see that this centroids will ultimately move to their respective clusters.

So, we can see the convergence of finally, we can get these final results where this distance from this cluster centers for these individual samples from this cluster is minimum, and in

this condition, we will get the lowest value of WC. So, this is the final optimum results, this is the K-means clustering, where we have assigned the sample based on the Euclidean distance from each of these observation to their corresponding centroid, and we have assigned them we have classify them into 3 clusters, because here the K value was 3. So, this is how we do the K-means clustering, I hope this is clear.

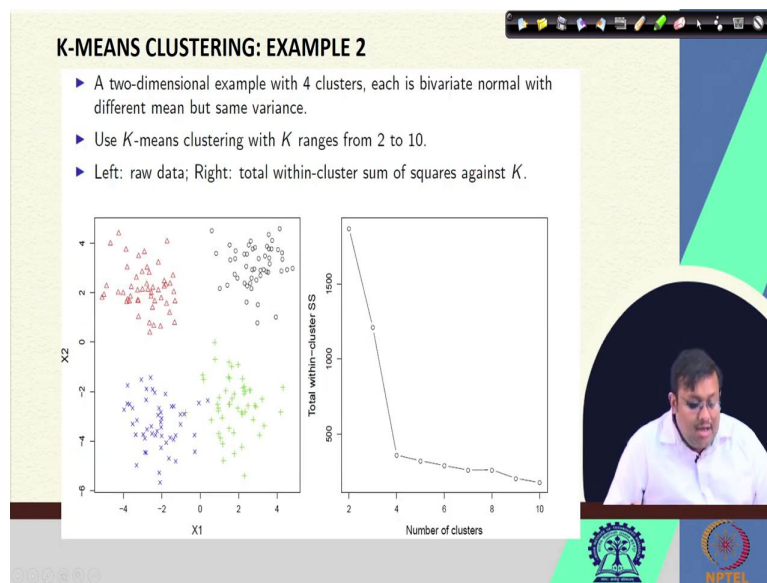
(Refer Slide Time: 25:43)



Now, here the example is given here a 2 dimensional example of examples with 2 clusters you can see here, so each is bivariate normal with difference mean, but the same variance. So, here we can do the K-means clustering with K equal to, 2 left is the raw data. And right is the K-means results, you can see that the samples are optimally clustered into 2 clusters, because in this fashion, we get the lowest value of WC of course.

Because here this is the class centroid, and here this is the class centered for this cluster, and we get the inter class variability lowest in this case, and ultimately, we get the WC values minimum in this type of clustering fashion.

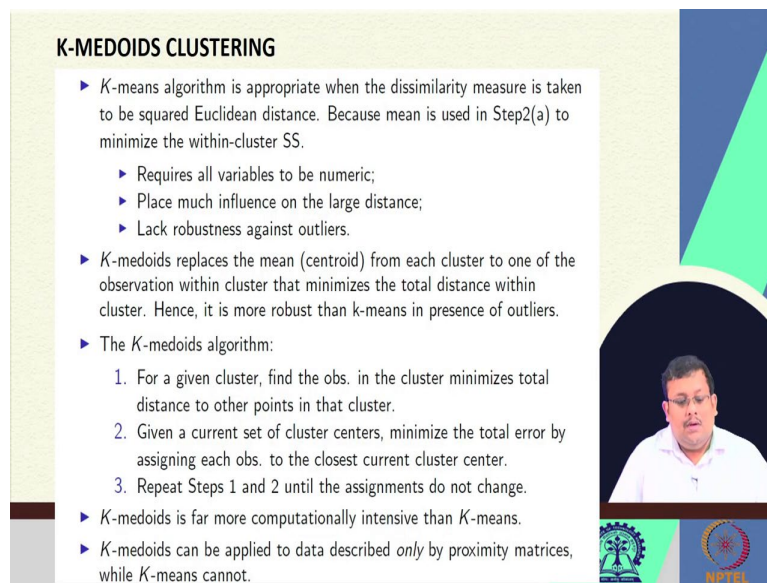
(Refer Slide Time: 26:34)



So, how to see the optimal number of clustering? So, a 2 dimensional example with 4 clusters is given each is bivariate normal with different mean but the same variance, so we use the K value as 4. So, we have generally vary the values from 2 to 10 to select the optimum number of clusters, how to select the optimum number of clusters? So, suppose we vary the optimum number of clusters of the cluster the number of cluster from 2 to 10 and then we see the total within class sum of square and we can plot them.

So, here on the X axis, we can see the number of cluster from 2 to 10 and here total within class SS is plot. So, we can see that in this point, there is an elbow. So, this will be the optimum number of clusters. So, here we can see the optimum number of clusters. So, whenever we see there is an elbow when the total within cluster sum of square will change with the along with the number of clusters. So, they are where we will see a kink or an elbow in this pattern that will be the optimum number of clusters remember at this point.

(Refer Slide Time: 28:00)



K-MEDOIDS CLUSTERING

- ▶ *K*-means algorithm is appropriate when the dissimilarity measure is taken to be squared Euclidean distance. Because mean is used in Step2(a) to minimize the within-cluster SS.
 - ▶ Requires all variables to be numeric;
 - ▶ Place much influence on the large distance;
 - ▶ Lack robustness against outliers.
- ▶ *K*-medoids replaces the mean (centroid) from each cluster to one of the observation within cluster that minimizes the total distance within cluster. Hence, it is more robust than *k*-means in presence of outliers.
- ▶ The *K*-medoids algorithm:
 1. For a given cluster, find the obs. in the cluster minimizes total distance to other points in that cluster.
 2. Given a current set of cluster centers, minimize the total error by assigning each obs. to the closest current cluster center.
 3. Repeat Steps 1 and 2 until the assignments do not change.
- ▶ *K*-medoids is far more computationally intensive than *K*-means.
- ▶ *K*-medoids can be applied to data described *only* by proximity matrices, while *K*-means cannot.

So, let us discuss another clustering method, which is different than *K*-means clustering algorithm we call it *K*-medoids clustering. Now, what is *K*-medoids clustering we know that in case of *K*-means algorithm, it is appropriate when the dissimilarity measure is taken to be squared Euclidean distance, because, mean is used in step 2a, we have already seen to minimize within the cluster some square.

So, here we know that we know we require for *K*-means algorithm we require all the variables to be a numeric. And it in case of *K*-means clustering it places much influence on the large distance and it lacks robustness against the outliers. So, there is a new method called *K*-medoids clustering, what is *K*-medoids clustering? So, *K*-medoids clustering replaces the mean or centroid from each of the cluster to one of the observation within the cluster that minimizes the total distance within the cluster.

So, instead of using the mean or centroid from each of the cluster now, we are considering one of the observation within that cluster that minimizes the total distance within the cluster. So, it is more robust than *K*-means clustering in presence of outliers, the *K*-means clustering is not very robust when there is an outlier, because we are taking the squared distance.

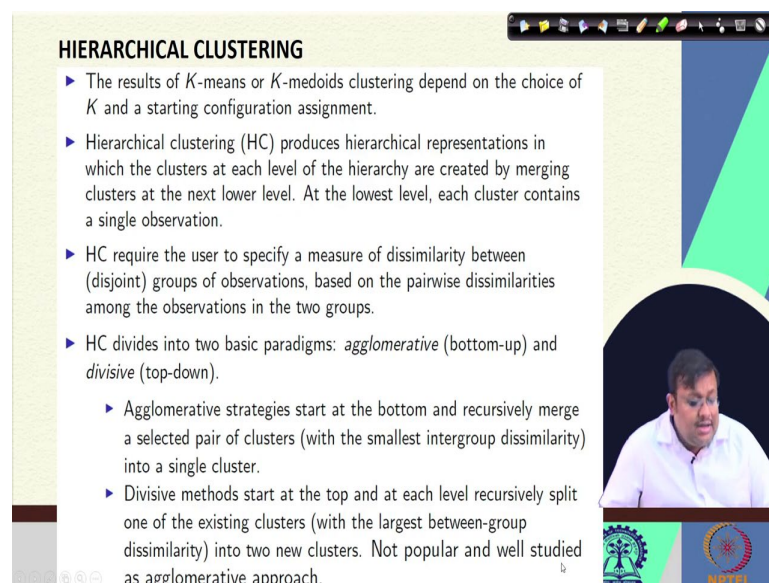
So, obviously, there will be much more influence of the large distance when it is when we are taking the square. So, in case of *K*-medoids clustering what happens? So, for a given cluster we find the observation in the cluster that minimizes the total distance to the other points in the cluster. And then, given a current set of cluster centers minimizes the total error by

assigning each observation to the closest current cluster center and then repeat the step 1 and 2 until the assignments do not change.

So, this is how we do that K-means clustering remember that that K-medoids is far more computationally intensive than K-means because in case of repeat the step one and two until the assignments do not change. So, this is how we do that K-means clustering remember that that Akemi diet is far more more computationally intensive than K-means because in case of when also in case of K-medoids, K-medoids can be applied to data described only by the proximity matrices while K-means cannot.

So, this is the difference between K-medoids and K-means. In case of K-means we are the all the computation is based on the centroid of each of the clusters. However, here in the K-medoids, we are considering to one of the observation within the cluster that minimizes the total distance within the cluster. So, this is the difference. And as a result of that, this K-medoids is much more robust to the outlier.

(Refer Slide Time: 30:54)



HIERARCHICAL CLUSTERING

- ▶ The results of K -means or K -medoids clustering depend on the choice of K and a starting configuration assignment.
- ▶ Hierarchical clustering (HC) produces hierarchical representations in which the clusters at each level of the hierarchy are created by merging clusters at the next lower level. At the lowest level, each cluster contains a single observation.
- ▶ HC requires the user to specify a measure of dissimilarity between (disjoint) groups of observations, based on the pairwise dissimilarities among the observations in the two groups.
- ▶ HC divides into two basic paradigms: *agglomerative* (bottom-up) and *divisive* (top-down).
 - ▶ Agglomerative strategies start at the bottom and recursively merge a selected pair of clusters (with the smallest intergroup dissimilarity) into a single cluster.
 - ▶ Divisive methods start at the top and at each level recursively split one of the existing clusters (with the largest between-group dissimilarity) into two new clusters. Not popular and well studied as agglomerative approach.

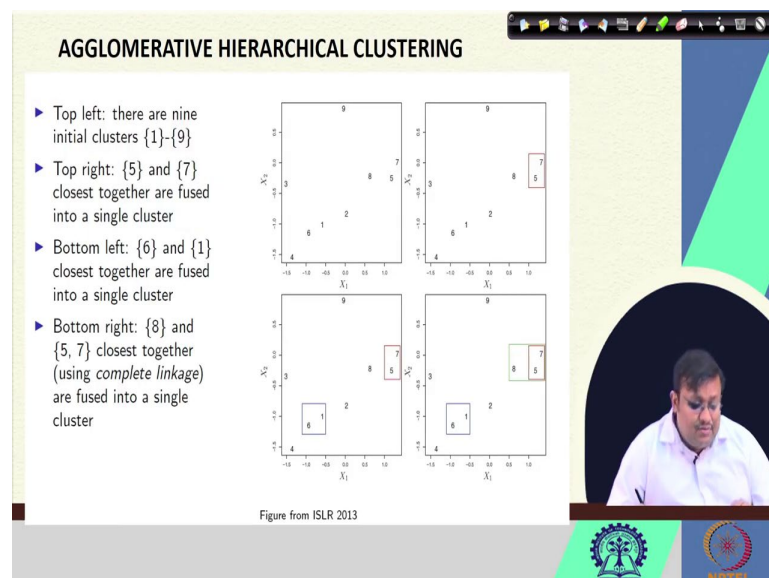
The slide also features a video inset of a man speaking, a navigation toolbar at the top, and logos for IIT Bombay and NPTEL at the bottom.

The last one the hierarchical clustering, so, the results of K-means and K-medoids clustering depends on the choice of the K and starting configuration arrangement assignment. However, hierarchical clustering produces hierarchical representation in which the cluster at each one level of the hierarchy are clustered are created by marching clusters at the next lower level at the lowest level each cluster contain only one observation.

So, this hierarchical clustering is used to specify the measure of dissimilarity between disjoint groups observation based on the pairwise dissimilarities among the observation in the 2 groups. And generally there are 2 types, one is agglomerative and other is divisive. Generally agglomerative strategy starts with the bottom and recursively.

Then keep on marching the classes together to a higher level and with the smallest intergroup dissimilarity into a single cluster. So, this is how this agglomerative hierarchical clustering is generally operated. Another one is a divisive method, but divisive method is not frequently used, divisive method is starts from the top and edit each level recursively split one of the existing clusters.

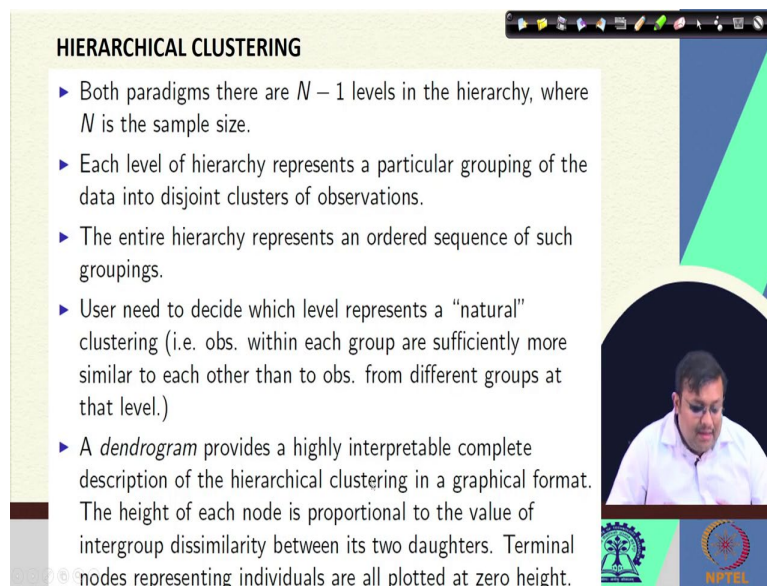
(Refer Slide Time: 32:15)



So, let us see one example, these are very good example here. So, in the top left cluster you can see there are 9 initial clusters. So, in the top right cluster, we can see the cluster 5 and 7 are closest together and are fused into a single cluster and the bottom left you can see that the 6 and 1 are closest together and fused into single cluster and bottom right you can see that 8, 5 and 7 closest together using the complete linkage and are fused into a single cluster.

So, in step by step, we keep on fusing this cluster together and that is why it is called agglomerative hierarchical clustering.

(Refer Slide Time: 32:58)



HIERARCHICAL CLUSTERING

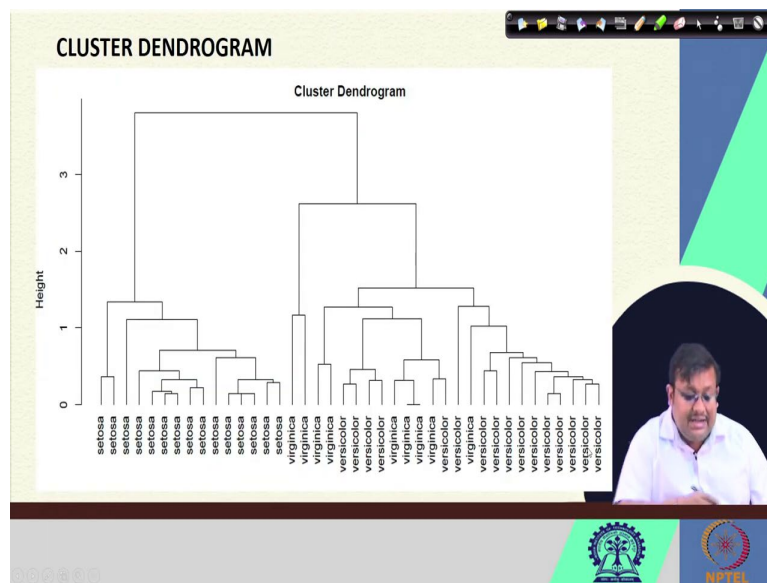
- ▶ Both paradigms there are $N - 1$ levels in the hierarchy, where N is the sample size.
- ▶ Each level of hierarchy represents a particular grouping of the data into disjoint clusters of observations.
- ▶ The entire hierarchy represents an ordered sequence of such groupings.
- ▶ User need to decide which level represents a “natural” clustering (i.e. obs. within each group are sufficiently more similar to each other than to obs. from different groups at that level.)
- ▶ A *dendrogram* provides a highly interpretable complete description of the hierarchical clustering in a graphical format. The height of each node is proportional to the value of intergroup dissimilarity between its two daughters. Terminal nodes representing individuals are all plotted at zero height.

So, both paradigms in both the paradigms in case of hierarchy clustering both agglomerative and divisive there are N minus 1 levels in the hierarchy where N is the sample size, each level of hierarchy represents a particular grouping of the data into disjoint clusters of observation. The entire hierarchy represent an ordered sequence of such groupings and use a need to decide which level represented natural clustering.

So, from this call the clustering you have to identify which level will show that natural clustering that is observation within each group are sufficiently more similar to each other than the observation from different group of each other at that level. So, in that way we produce a Dendrogram. So, it Dendrogram produce a highly interpretable complete description of the hierarchical clustering in a graphical format.

So, this height of the each node is proportional to the values of intergroup, dissimilarity between the 2 daughters and terminal nodes representing individuals are all plotted at the 0 height.

(Refer Slide Time: 34:12)



So, this is one Dendrogram using the highest data you can see here, so, this is how we can classify you can see there you remember that there are 3 different types of flyover like Setosa, virginica, and versicolor and how we can cluster them using agglomerative hierarchical clustering these and then this is called a Dendrogram.

So, you can see that how from the bottom level, there are, so, these 2 observation classes classified together and these 2 observations are classified together, these 2 observations are classified together and then these two sub clusters are close together here.

And then the clustering grows on goes on goes on and then merging, merging, merging and you can see all the Setosa grouped together. And then virginica and versicolor are also clustered separately. So this is called cluster Dendrogram. And then from this Dendrogram, we can have an idea of that the linkage between different groups of observations.

(Refer Slide Time: 35:10)

REFERENCES

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
https://commons.wikimedia.org/wiki/File:K-means_convergence.gif

Thank you

So, guys, thank you very much. These are the references which are used. And, I hope that you have got some good information in this lecture. If you have any queries, just please, feel free to email me, and I will be more than happy to answer your queries. And thank you let us meet in our next lecture for more discussion on clustering and classification. Thank you.