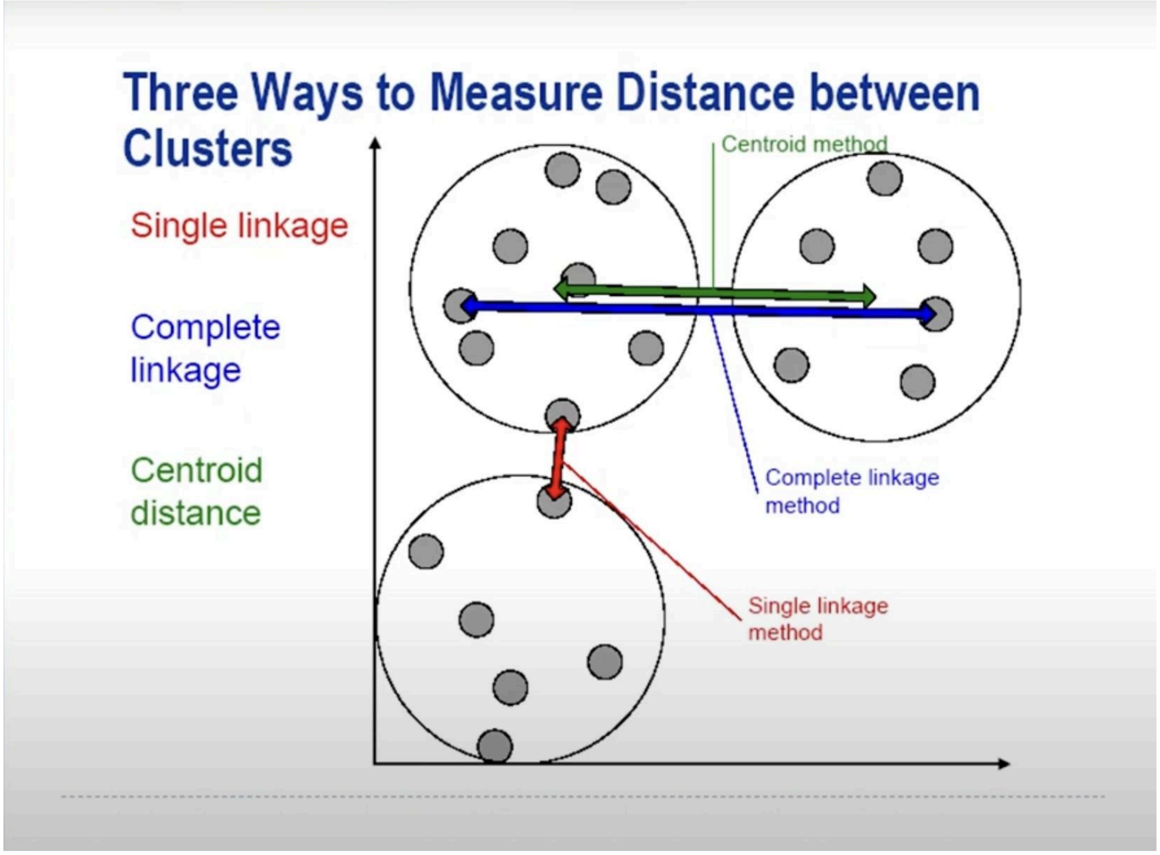
Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:09
Lecture:34

CLUSTERING TECHNIQUES Part 2



Three Ways to Measure Distance between Clusters

Single linkage

Complete linkage

Centroid distance

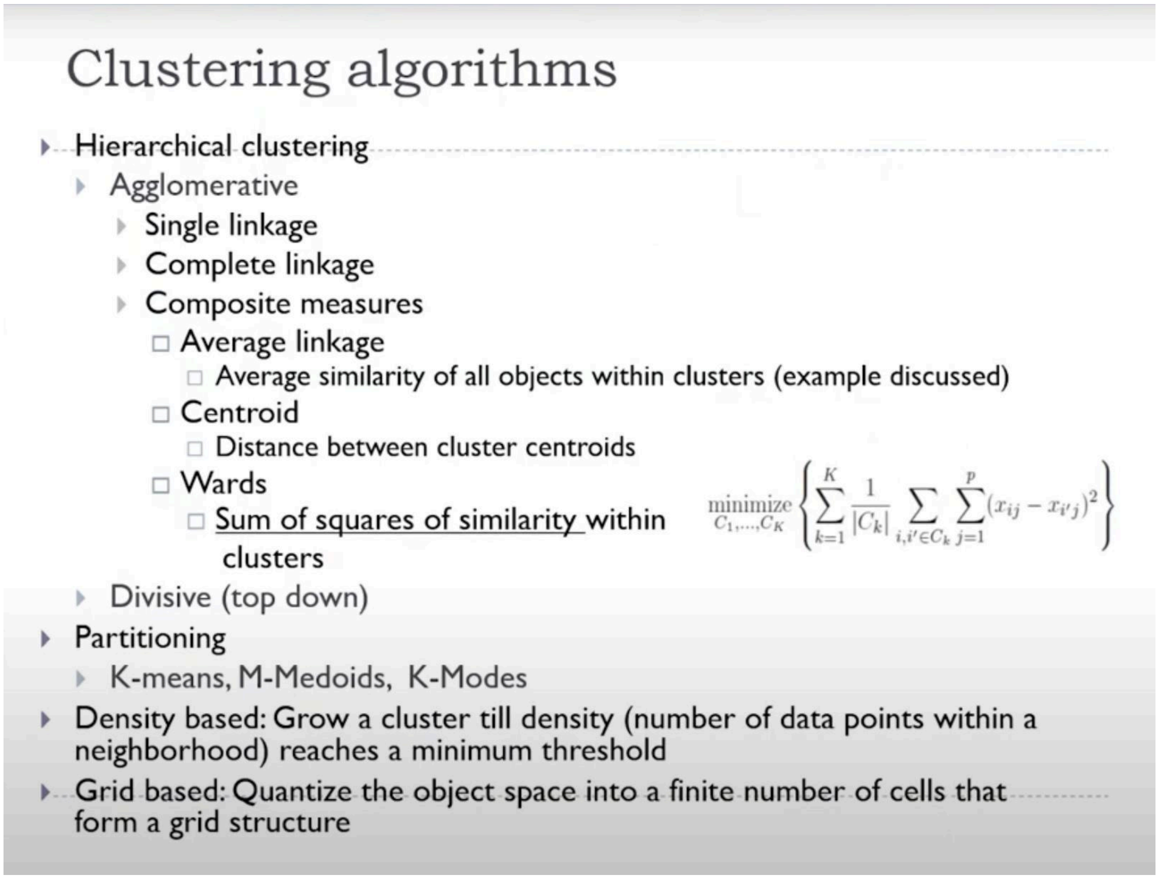
Now, yeah one concern for us here is, how do we have a measure of distance between clusters. So graphically, this diagram illustrates how the distance between clusters can be measured. There are three measures available, three common measures, single linkage, complete linkage and centroid distance. And looking at the colors, you understand what principles are used. Single linkage means the red line, the distance between two clusters is the distance between the nearest objects in the clusters. Each cluster has a given number of objects, the small circles, the gray circles are actually objects that belong to a particular cluster, the circles are the clusters.

And based on distances they have been formed into clusters. There are three clusters here and you can see in single linkage, it is a shortest distance or the distance between the nearest objects. And in complete linkage you take the farthest distance, you go in the pessimistic way, that is a maximum distance that exists between two clusters. And then there is also a centroid distance, a centroid distance which is the distance between centroids of the clusters. And what is a centroid? So, here it is a scenario of V1 versus V2, it is only two attributes or two variables, that is the simplest case.

So you can actually place the clusters in a two dimensional space, but you can imagine when the number of features go up it actually becomes a multi-dimensional space or a hyperplane. So that is, you have to imagine, you cannot actually easily represent that in a two dimensional space which I have, because clusters with multiple, more than two attributes would be in a hyperplane. But this is for illustration. So here you can see that suppose one object, this has a value x1, y1 or V1, V2 as we say, these are the coordinates and suppose this is x2 and y2. And let us assume that there are only two objects for simplicity.

What would be the centroid of this cluster with two objects? What would be the centroid value? Nothing but the centroid will be somewhere here which is having the coordinate (x1+x2)/ 2, (y1+y2)/ 2. That is the centroid, that is the centroid, the average value, the average point. So you average the attribute values and that is the centroid. So when you increase the number of attributes, it is the average of each attribute, each attribute of the object averaged is the centroid of that cluster. The term centroid will be often used in clustering and you should know what it means.

It is the average and as soon as you think of average or mean, you should also know that average would be influenced by the outliers. Suppose one object was here and another object was here and they got grouped into one. It gets influenced by the outliers. So that should be kept in mind when you actually use the term centroid. So centroid distance is a sort of compromise between the complete linkage and the single linkage, basically to measure the distance between clusters.



Clustering algorithms

- Hierarchical clustering
  - Agglomerative
    - Single linkage
    - Complete linkage
    - Composite measures
      - □ Average linkage
        - □ Average similarity of all objects within clusters (example discussed)
      - □ Centroid
        - □ Distance between cluster centroids
      - □ Wards
        - □ Sum of squares of similarity within clusters

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

  - Divisive (top down)
- Partitioning
  - K-means, M-Medoids, K-Modes
- Density based: Grow a cluster till density (number of data points within a neighborhood) reaches a minimum threshold
- Grid based: Quantize the object space into a finite number of cells that form a grid structure

Now in one shot or in one slide, if I am making an attempt here to give you a picture of different clustering algorithms. So that you can also compare, how they are similar or different. So it is not given this way in a textbook but this is one way to sort of understand different clustering algorithms and what principles are used in the clustering algorithm. So you can see that there are 1, 2, 3, 4 categories of clustering algorithms. This, by looking at different sources of information I have compiled this.

So, but the most widely used clustering techniques are the hierarchical clustering and the partitioning clustering. But hierarchical and partitioning are two categories. Within them, there are different types of clustering methods. So the clustering technique which we just discussed, as a simple example, that is a type of clustering technique which is called agglomerative. Agglomerative meaning you start with the shortest distance or you start from the bottom and build upwards, that is agglomerative.

There is also divisive clustering which is also a hierarchical clustering where you start from the top. You start with the longest distance and then come down by division, by the method of division, that is known as top down clustering not very popular, not very highly used. Agglomerative clustering has different sub techniques. Agglomerative is a subcategory and within them, within agglomerative there are different methods. One is agglomerative with single linkage, in the process of building clusters, you know that you have to also calculate distance between clusters.

So an agglomerative can use a single linkage or a complete linkage or composite measures. Centroid is one where you average and then find the distance between clusters, that is a centroid method I just explained to you. There is also average linkage which we just used in the agglomerative clustering method we used. So just finding the simple average. You know in the example, I illustrated it was average similarity of all objects within the clusters.

We can use centroid distances, once a cluster is formed, the centroid of that cluster to another object or another cluster can be used as a measure. Then it is agglomerative clustering with centroid as the measure of distance. And there is a very widely used technique within agglomerative clustering known as Wards method. The popularity of the clustering techniques depends on the usefulness of the solution. How good solutions you get.

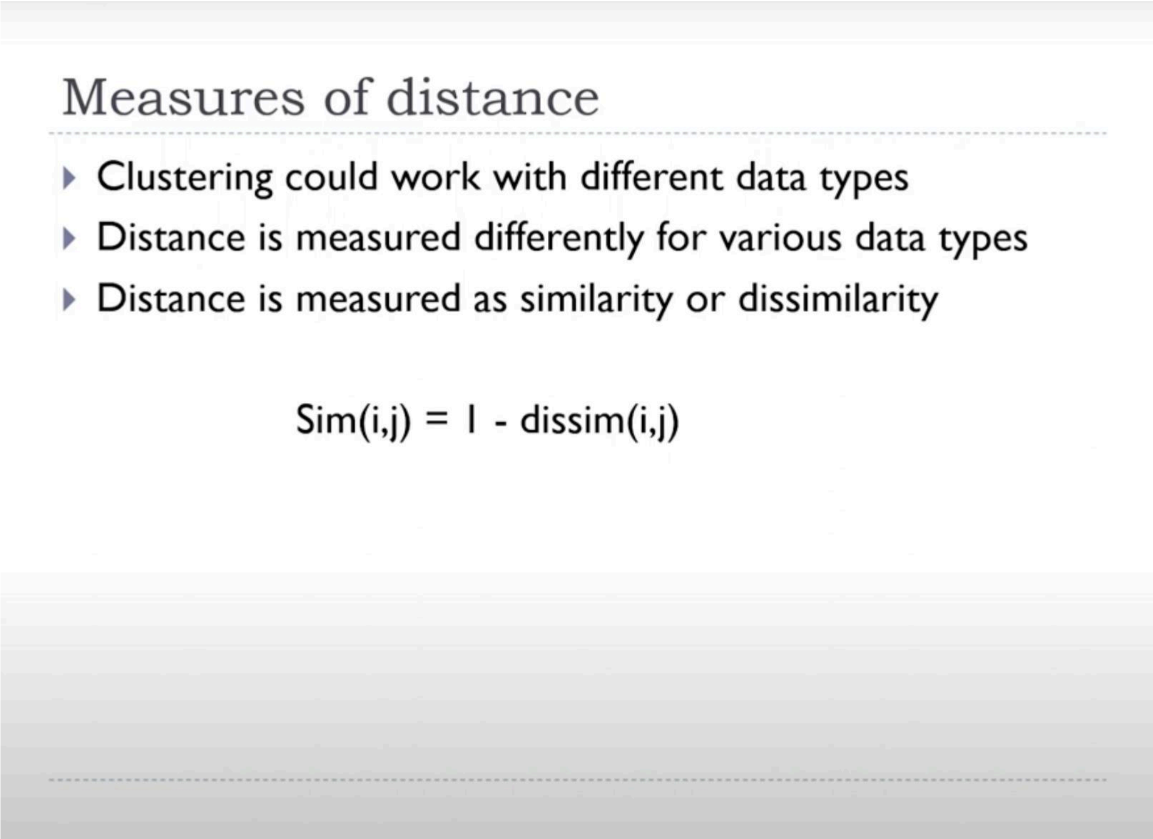
So in practice Wards method is very useful and usually yields useful solutions in the sense, usefulness is by the minimum error and also minimum, highest discrimination between clusters. So Watts method is highly used and in Wards method, you use sum of squares of similarity within clusters. It is not just the average but it is the squared sum of errors that is used. The squared sum of errors that is used to calculate the within cluster distance, you know the total within cluster distance.

Ultimately when you build cluster, you need to have a measure of total within cluster distance. And this can be estimated by different methods or this not estimated, this can be computed by different methods. So, and so you see that all the methods are listed here of which Wards method uses the squared sum and usually Wards method yields very useful solutions. And there are of course explanation for it, I am not entering there, I am just giving you a sort of picture or a broad view of different clustering techniques and how they differ and which are yielding good results. Wards method and moving to the next category, this is category 1 and this is category 2.

So the category 2 is also widely used, the partitioning technique and within partitioning technique there is K-means, M-medoids and K -Modes. These are three techniques available in the partitioning technique. I will explain the medoids and the modes towards the end, as to how they are different from K-means. But k-means is a widely used clustering technique. Wards method and K-means are widely used and in some problems, you first use Wards method and arrive at certain number of clusters and then subsequently use partitioning technique. So suppose you use Wards method, you have an initial solution and initial solution means, suppose you formed four clusters and each cluster has a centroid.

Each cluster will have a centroid. Now in partitioning technique, you start with centroids. You start with random seeds or random points in space and then you assign each object to those random points, based on the distance of each object to those random points. So in K-means, instead of starting with the random points, one could start with the centroids that is obtained from Wards method. This is a another improvement in method, suggested in literature.

There is a comparison of clustering techniques in one research paper. So I am just explaining one method that is followed in research for clustering. There are also density based and grid based techniques which are covered. I explained to some extent in your textbook but I am not actually using those in this session or in this class the other techniques that are available for clustering.



Measures of distance

- Clustering could work with different data types
- Distance is measured differently for various data types
- Distance is measured as similarity or dissimilarity

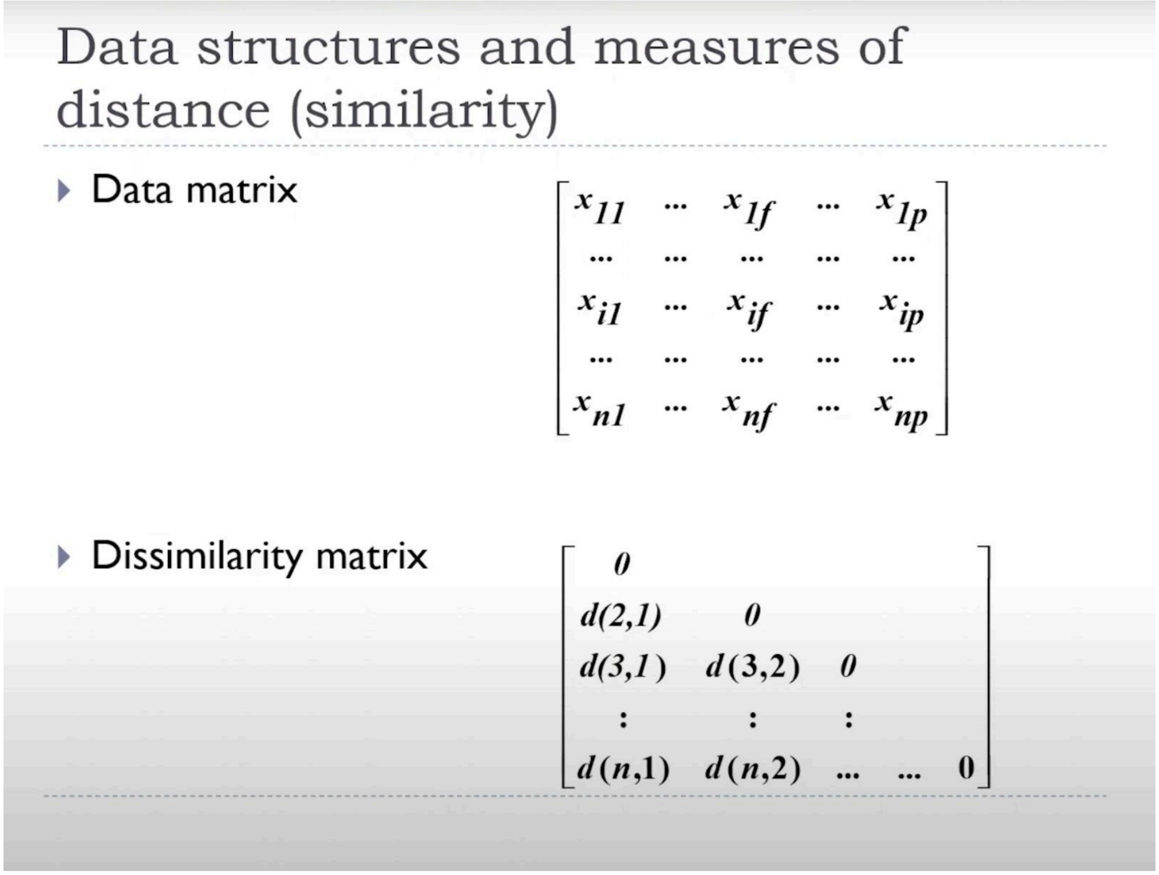
$$Sim(i,j) = 1 - dissim(i,j)$$

So broadly, hierarchical and partitioning. Now hierarchical, I actually give you a sense of how hierarchical clustering works from bottom up. Now before we step forward, now we have to use algorithms and obviously you know when you use a software like Python or R for clustering, you can choose the algorithm. There will be, choice of algorithm is up to the researcher or to the analyst. And our idea is to understand how different algorithms work.

What are the principles? So hierarchical is one type and then there is partitioning type of clustering. But irrespective of which method you use, there are certain basic measures that you require. One of those basic measures for any algorithm to work is a measure of distance between objects. Not distance between clusters. In distance between clusters, we saw there is single linkage, distance based on centroids etc. complete linkage. That is for distance between clusters. Now for cluster solutions to be formed, you need to calculate distance between objects. And we saw Euclidean distance as one distance, measure of distance between objects when the values are continuous. You can use it for integers, you can use it for rational numbers. But when it is categorical or when it is categorical variables or even ordinal data, you have a challenge in using Euclidean distance. Because you cannot, in categorical variables you may have numeric categories but they are not considered as numbers. They are only labels. Numbers used as labels. For example 0 and 1, yes is 1, no as 0. Does not mean that you can simply do averaging of those values. How many 1s are there, you know. Or what is the average number of 1s etc. We can have that but they are not, they are labels. They are not numbers to be used in algebra.

And therefore you have a challenge how to work with different data types. How to work with different data types. So let me give you an overview of working with different data types in clustering. So to start with, what is a measure of distance? Measure of distance is a value of the distance between two objects based on their attribute values. So suppose there is v1 and v2 here, v1 and v2 here.

So you try to find the pairwise differences and try to sort of sum them up, to have an average, to have an overall value of distance. That is what we are doing. So there is also this notion of distance and notion of similarity, as I said it is inverse. So there is similarity of object i and object j which is 1-dissimilarity between objects i and j or 1-distance between i and j. It is same as 1-distance, because distance is an inverse measure of similarity, i, j.



Data structures and measures of distance (similarity)

- Data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

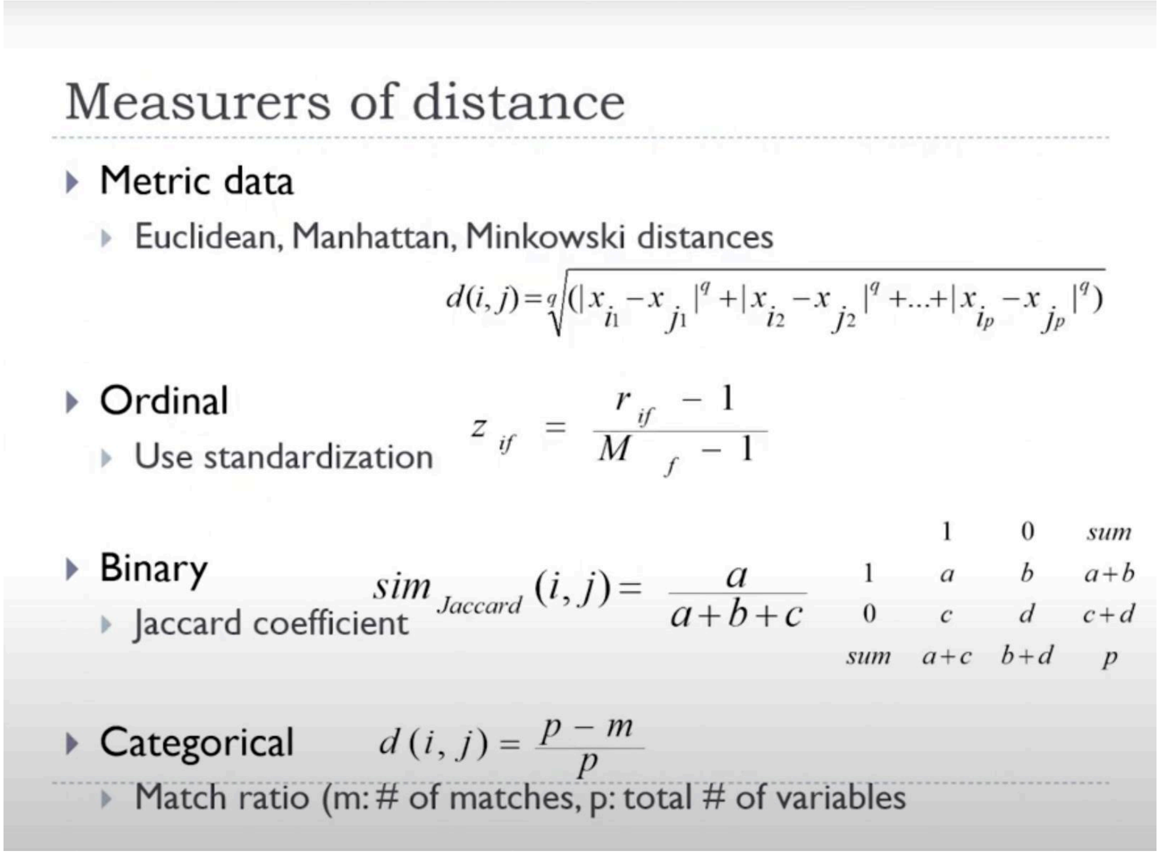
When you here notate i and j keep in mind i and j are objects. i and j are objects. And each object has multiple attributes. And how to measure and structure the distances? So you have a data matrix typically. A data matrix consists of rows and columns.

Each row, you know that each row is a record or a tuple and number of rows is n. Meaning what? This is data size. The size of the data is n. There are n number of records. There are n respondents or n objects or n records. n is equal to number of objects. Each record is an object. Assuming that each record has originated from some one, some object. It consists of the features of certain objects. So therefore what are the features or attributes? They are in the columns. A given object has how many features or how many attributes? p number of attributes.

Each individual or each object is characterized by p attributes. So therefore you can say it is a n×p matrix. x n p. It is an n × p matrix that would actually capture your whole data. So column stands for the attributes or the features and rows stand for the different objects or different records.

And that is how you represent a data set and interpret a data set which we use for clustering or a classification exercise. And now when you have to extract the pairwise distances between objects, for example the distance between object x1 and object x2, that is the distance between d(2,1). Distance of this object and this object, that is d(2,1). That you can calculate using some measure like Euclidean distance and that forms the first element of the diagonal matrix. And you know that in a diagonal matrix, you can actually capture all the pairs.

All the pairs can be captured in a diagonal matrix and therefore that becomes the dissimilarity matrix. And this is how you can actually represent the data set and the distances. And we have seen this already in our example. All right. Good.



Measurers of distance

- Metric data
  - Euclidean, Manhattan, Minkowski distances

$$d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+...+|x_{ip}-x_{jp}|^q)}$$

- Ordinal
  - Use standardization

$$z_{if} = \frac{r_{if}-1}{M_f-1}$$

- Binary
  - Jaccard coefficient

$$sim_{Jaccard}(i,j)=\frac{a}{a+b+c}$$

	1	0	*sum*
1	*a*	*b*	*a+b*
0	*c*	*d*	*c+d*
sum	*a+c*	*b+d*	*p*

- Categorical $d(i,j)=\frac{p-m}{p}$
  - Match ratio (m: # of matches, p: total # of variables

Now measures of distance. This is a whole chapter in Han and Kamber and again I am summarizing this in one slide for you to understand or to give an overview of measures of distance when the data types are different, when the data types are different. The easiest type of data to work with is, of course the continuous valued data or the metric data and each data type has a definition which is available in your textbook. I am not getting into those details. But here in metric data or continuous valued data you have measures of distance which are, which can be generically represented or measured using Minkowski distance.

Minkowski distance is a generic formula for calculating distance between objects. And what is the generic formula? It is, there are two objects i and j. i is a object, j is an object. And both i and j has, how many number of attributes? P number of attributes.

There are P number of attributes. So pairwise distance xi1 xj1. There is a first attribute distance. Second attribute distance. Third attribute distance.

What happened? And pth attribute distance or difference. These differences are calculated. Then you can, of course find the absolute value and if you are going to square it, you do not need the absolute value. But in Minkowski distance you can say you can actually have the qth degree. So Minkowski distance is generic formula. You can actually take the qth order or qth degree of the difference and then, of course when square root it, you can actually find the qth root of that sum of distances.

So you can see that Euclidean distance is a special case of Minkowski distance where q=2. When q=2, it becomes Euclidean distance. And when q=1 that is Manhattan distance. In Manhattan distance when q=1, it should be a absolute value of the differences.

Otherwise you have a problem. The pluses and minuses would finally lead to a value which is, which may not be the ideal value of the actual distance. And therefore you take the absolute value of the differences. And that is known as Manhattan distance. So both Manhattan distance and Euclidean distance are special cases of Minkowski distance. So in literature when you come across something called Minkowski distance, keep in mind there is a q value that you have to assign.

The q is the root. And the value of q will determine whether it is Euclidean, Manhattan or otherwise or higher degrees. And both Euclidean and Manhattan are commonly used. So this is about the data, that is continuous value. But suppose you have ordinal data.

Ordinal data is a special type of data. For example low, medium, high. You can actually put values 1, 2 and 3. They have an order. They have an order. The only difference is that the distance between 2 and 1 need not be the same as the distance between 3 and 2.

There is no fixed interval between these differences, this fixed interval between the numbers. And therefore when the interval is fixed, it becomes interval data. So that is the difference between ordinal data and interval data. But an approximation that is done to work with ordinal data in clustering is to use standardization, wherein you convert an ordinal data into a continuous value data.

And then you do, you use the same measures as in metric data. For example, suppose you have a scale of say 1 to 3. 1, 2 and 3 are your ordinal data points. Like low, medium, high, 1, 2, 3. So what you do is suppose some somebody responds with a value 2. That value 2 is converted or standardized using a ratio or it is actually found as a ratio of the span.

The span is the highest value which is maximum value, that is (3 -1) /(2-1) or upon (2 -1). (2 -1)/ (3- 1). That becomes a ratio which is substituted for the value 1. The value 1 will be standardized into a ratio. And once it is standardized, you use the metric data measures for finding the distance.

This is one approach. In minus rank, r i x n minus rank, total number of ranks possible minus the rank is there. Yeah, yeah, yeah, yeah. Total number of ranks. No, yeah, yeah, you can say total number of ranks.

That is suppose it is 5, then the denominator will be 4. 5- 1, 4. And r is the actual response that is obtained. If he is in the second rank. If he is in the second rank, then it is. And if suppose number of ranks are 3, then (2- 1) /(3- 1).

That is the value that corresponds to that rank of 2. Then you are actually converting that into a metric data. That is the effort here. This is one approach that is given in literature. And then there is for binary and categorical.

Binary is a special case of categorical. For binary, there is a common measure or a well-known measure known as Jaccard coefficient. Jacquard coefficient, you can read further in your textbook. It explains and also gives examples. You have a contingency table like, with two axis, vertical and horizontal to represent number of similarities. So when you compare two objects i and j and there could be, say a p number of, there are p number of variables.

P number of attributes. P number of attributes between i and j. And assuming all of them are binary, binary attributes. Then when you compare i and j there may be a number of similarities. When i=1, j is also 1.

The count could be a. When i is 1, j is 0 that is b. When i is 0, j is 1 that is c. When i is 0, j is 0, that is d. So each of them are counted. You know as you count the true positives and false positives and so on. And Jaccard coefficient is defined by a divided by, that is number of similarities in terms of 1s, divided by a+ b+ c. a+ b+c does not make it p. It is less than p. What it has not considered is d. Any intuition as to why Jaccard formula or

Jaccard coefficient or Jaccard index, you know it is called by different terms. Why Jaccard coefficient is indifferent to d which measures number of similarities with respect to 0. You can also add d.

There is another measure which is not Jaccard coefficient, where you also add d also. You can also have a measure a+, in terms of number of 0s. It could be (a+d)/(a+c+d). That is also a measure of similarity between i and j. You can, you have all these different measures but in Jaccard coefficient you do not actually consider d.

In certain context, you know Jaccard coefficient is useful in certain context. Not that that is only measure. Any intuition? 0s and 0s.

Both of us have 0s. It does not matter. I do not want to, I am indifferent about it. That is a good thing. That is a good, that is the best example to give. There can be other examples. Suppose it is in medical diagnosis.

And what matters is, in comparing two patients whether they have similar symptoms. But when you do not have sickness at all, that is the large number of people. Two people with those who do not have symptoms. There is no point in comparing.

We only want to compare those who have certain sickness. In which case, only the 1s matters. The 0 case is not important, not relevant. So therefore d is a count of irrelevant similarities which we do not want to consider. So in those context, Jaccard coefficient is useful. I am not saying that is only coefficient, but this is widely used.

The fourth case is of course, the pure categorical variables. Not just binary but categorical data. In categorical data, you know that there are p number of variables. And one measure for similarity between two objects, i and j with p number of categorical variables is p-m. m is the number of similarities, number of times i and j have similar values, same value. Suppose you have three categorical variables. One could be say age bracket, other could be income bracket, third could be, say geographical location. So m you count for number of times, there is a match between two objects. For example, you two are in the same age bracket, then m becomes 1.

You two are also in the same income bracket, m becomes 2. But you are from two different places, then that is not counted. So p is 3, then it becomes (3-2) / 3. That is your sort of similarity value for a categorical pair of, not pair, categorical set of variables. So this is another measure.

So you have a data set of n number of objects, n number of records or n number of rows.

And you have p number of variables, for each object has p number of attributes of variables. Now the challenge is, the final challenge could be that out of the p a few are metric data variables, others are ordinal and even a few others are binary like the gender or, gender need not be binary. But assume that there are some variables which are binary and some are categorical. Then you have a challenge because at the end, you need to have one measure of distance between objects.

You have to have one measure of distance between objects. So there are methods to sort of or you can actually have a algorithm which is explained in your textbook to combine each of these cases and to have a composite measure of distance. The algorithm should first detect what is the data type. First is detection of data type, then apply the corresponding formula and you can see that in each of this formula, the measure gets or the value gets converted, the attribute value gets converted to a continuous value data. And then you can apply one of the measures of the continuous valued data to find, for example that is the case with ordinal and categorical.

So, sorry, that is the case with ordinal. Metric and ordinal actually gets converted to metric. Binary and categorical actually gives a value of the distance already, i and j which is again a sort of measure, a standardized measure of similarity between the two objects with respect to the binary part of the attributes. Then you sum them up. So that is the way, a composite measure of distance is used in calculating distance between objects. Because we are all going to use data sets which consist of multiple data types. And therefore we need to be aware how this distance, a measure of distance works for objects with multiple data types.

This is an overview. This is covered in one chapter in your textbook. So I am trying to give you an overview. Excellent point. Once you get this, instead of taking the absolute values, you standardize it.

That is a normal procedure in clustering. We will standardize before we cluster. So that it does not get carried away by the absolute values. It is all distance. This is inverse of similarity.

What do you measure is inverse of similarity. Jaccard coefficient is a measure of similarity. That is true. Because it is counting number of items which are similar. So this is, you can see this is similarity. There also there is challenge when you actually add them. You may have to inverse it.

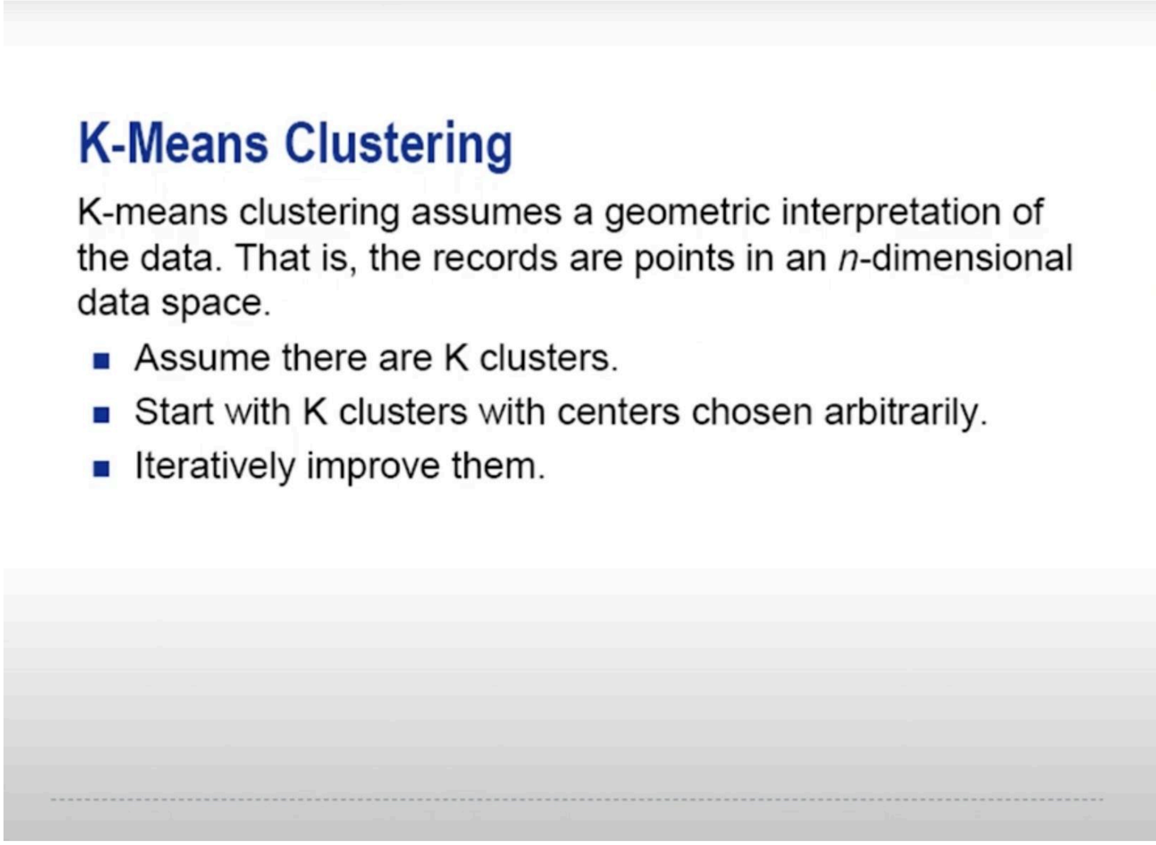
So the formula, the final formula can be fairly complex taking into consideration all this. Good point. This is similarity clearly mentioned but this is distance, this is distance which is dissimilarity. This is also similarity.

Actually when the value is close to 1, this is very similar. That is what we are saying. No? This is p-m. So this becomes dissimilarity. Is my statement correct or wrong? So when m=p, what happens? 0.

0. So, what is that case? p=m means both are very similar. So, it is again an inverse scale. So I am actually wrong, you are right. You can see the notations closely, which I did not observe closely. You see it is actually clearly mentioning it as distance.

It is a measure of distance. Whereas it is clearly mentioning it as similarity. Therefore this scale is inverse, this is inverse, this is inverse. This is also distance. Good that you noticed it.

This is distance, this is distance, this is distance and only this one is similarity. Jaccard coefficient is similarity. Good. Now let us take up the case of a clustering technique which is different from the hierarchical clustering. So I am taking you to a method which is here.



K-Means Clustering

K-means clustering assumes a geometric interpretation of the data. That is, the records are points in an *n*-dimensional data space.

- Assume there are K clusters.
- Start with K clusters with centers chosen arbitrarily.
- Iteratively improve them.

First we discussed the agglomerative with, yeah, with this case. This was the example that we discussed. But now we are going to discuss a very different method. That is what you have to keep in mind. This is not hierarchical clustering. This is partitioning, clustering by partitioning.