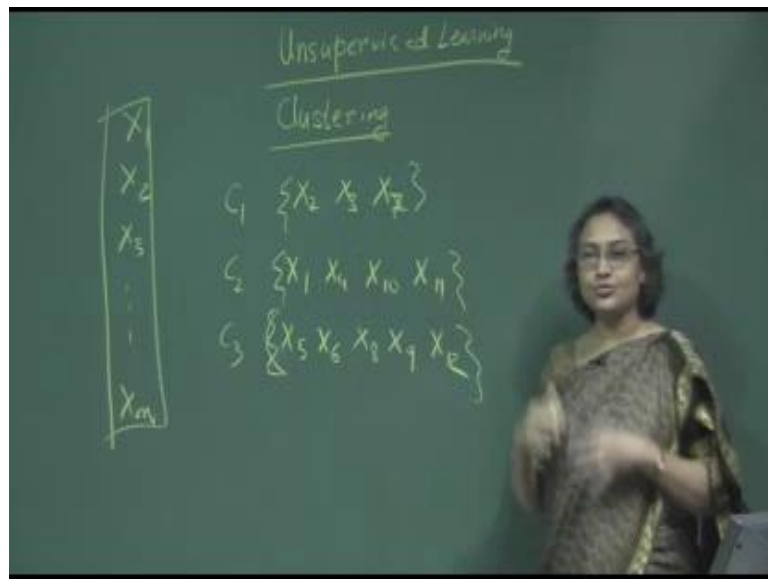


Introduction to Machine Learning
Prof. Sudeshna Sarkar
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Module – 9
Lecture – 37
Introduction to Clustering

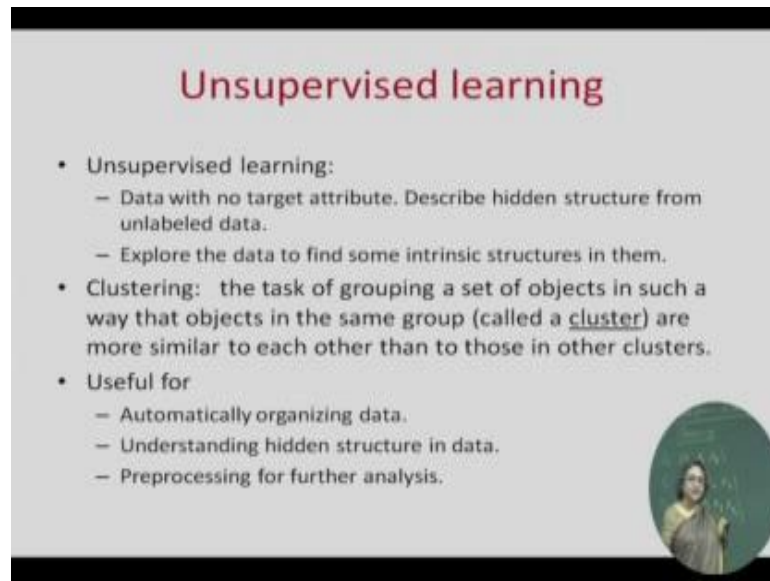
Good morning. Today, we will talk about Clustering. In the early classes so far in this course we have mainly talked about supervised learning. In supervised learning, we have some data which are labeled and we try to learn a function, we try to come up with a method to label unseen instances correctly.

(Refer Slide Time: 00:48)




Today, we will look at unsupervised learning and we will look at one specific type of unsupervised learning which is called clustering. We will introduce what clustering is and in the subsequent classes, we will give illustrations of different specific clustering methods.

(Refer Slide Time: 01:17)



Unsupervised learning

- Unsupervised learning:
 - Data with no target attribute. Describe hidden structure from unlabeled data.
 - Explore the data to find some intrinsic structures in them.
- Clustering: the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other clusters.
- Useful for
 - Automatically organizing data.
 - Understanding hidden structure in data.
 - Preprocessing for further analysis.



In unsupervised learning, as we have discussed in the first class of this course that we have data, but those data have no labels, so we have unlabeled data instances and in supervised learning we want to find hidden structure in the unlabeled data. We want to explore the data to find some structure in them, but there is no old standard or there is no label to say that this is what we get.

Now, why would we be interested in this when we have a large amount of data, unless we are able to group the data we will not be able to do studies, do proper exploration of the data. For example, let us take the example of the plant and animal kingdom. So, biologists have studied the different species of plants and animals and come up with a grouping of them and these groups can be called clusters. They have grouped animals into vertebrates and invertebrates, and then vertebrates as mammals, fish etcetera.

So, these labels were not given to the early scientists, but they have come up with these groupings based on species which share similar attributes and based on that this grouping has been done. Now, clustering is the most popular type of unsupervised learning. There is few other type of unsupervised learning which we will not talk about in this class. In clustering the task is given a set of data instances. So, here we do not have; earlier our sample contained $X_i Y_i$ values.

Here the sample will contain only X1, X2, X3, etcetera. So, only the data points are given and no output or no output label is given. So, given a set of these instances objects in clustering we want to group the objects into clusters. We can find the clusters C1, C2, C3. So, C1 will contain a number of these items suppose X2, X3, X7 go in C1; X1, X4, X10, X11 go in C2 and X5, X6, X8, X9, X12.

So, suppose these are 3 different clusters. We come up with a given this sample, we come up with groups and how do come up with groups. So, instances which belong to the same group with in some way similar to each other right. So, X2, X3, X7 should have some similarity among each other; X1, X4, X10, X11 should have some similarity with each other and X3 and X4, 2 elements which belong to different clusters will be in some way dissimilar with each other, based on this we can come up with groups.

Now, clustering is useful for many applications, for example, it can be used to automatically organize data. For example, the plants species or animal species or news documents or books. It can be used for understanding hidden structure in data and sometimes clustering is used preprocessing for further analysis of the data.

(Refer Slide Time: 05:23)



Now, let us look at this slide to look at an application of news clustering. Some of you

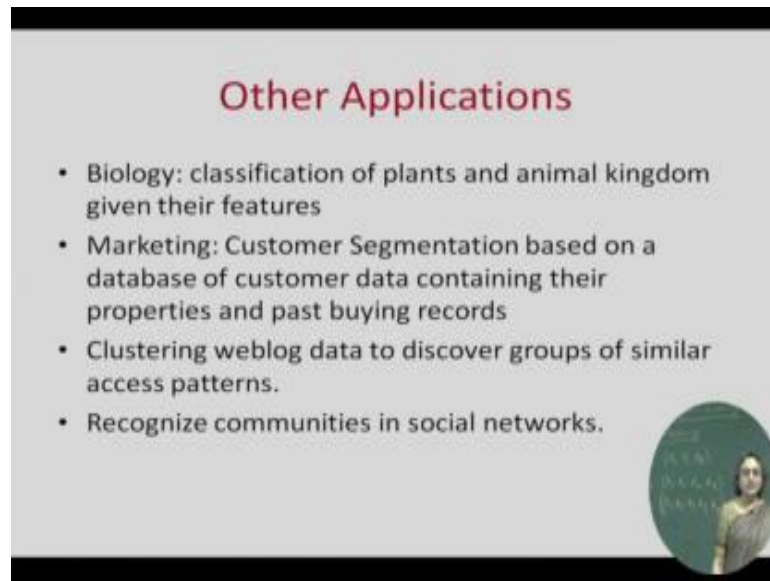
will have used Google news. In Google news, you will notice that the different new stories are grouped, for example, this is one new story about the Bombing of Istanbul airport on 29th June, 2016 and the new stories under that they are grouped together. So, this is unsupervised clustering because this particular new story was not given as a label to this, but the related news items were grouped together. So, Google news is an example of a system which does news clustering.

(Refer Slide Time: 06:12)




Then this is an example of clustering of gene expression data. So, these genes are clustered based on the gene expression.

(Refer Slide Time: 06:25)



Other Applications

- Biology: classification of plants and animal kingdom given their features
- Marketing: Customer Segmentation based on a database of customer data containing their properties and past buying records
- Clustering weblog data to discover groups of similar access patterns.
- Recognize communities in social networks.



There are other applications, for example, as I have talked about in biology the classification of plant and animal kingdom given their features. In marketing customer segmentation based on the database of customer data containing their properties and past buying records. These are very useful to marketing companies because based on the grouping of the customers they can decide what type of promotions to target to each customer.

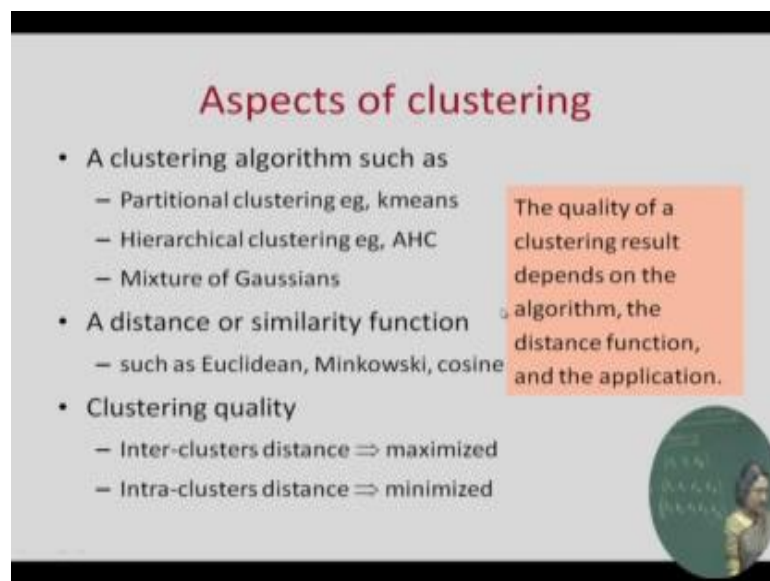
Third example is clustering of weblog data to discover groups of similar access patterns and the fourth example is to recognize communities in social networks based on their similarities.

(Refer Slide Time: 07:07)



Let us look at in order to talk about clustering algorithm. Let us look at some example data. So, these are the data points, only the data points are given, their labels are not given, but in this particular data you see visually that there are 4 natural clusters and what a clustering algorithm is required to do is take this and come up with these 4 clusters, ideally this is what the clustering algorithm will do.

(Refer Slide Time: 07:41)

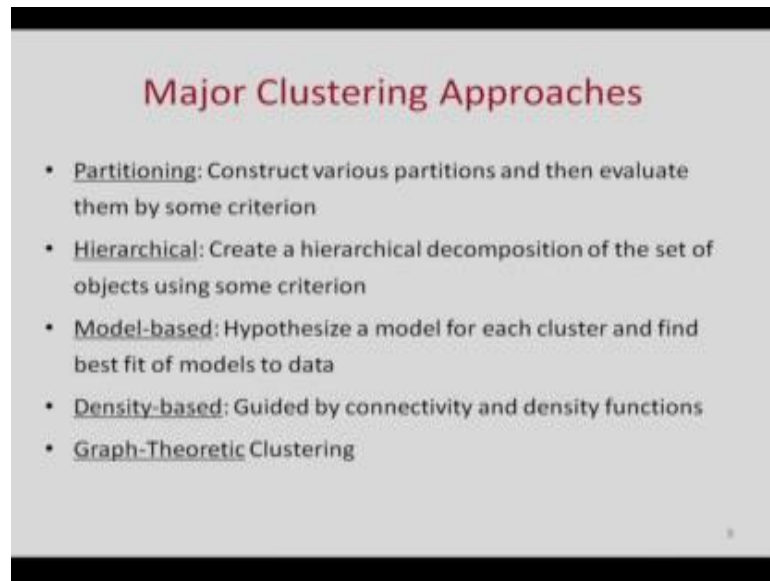


So, there are different aspects of clustering. First of all there is a clustering algorithm and there are different types of clustering algorithm. We will discuss a few of the clustering algorithms in the next few classes. There are algorithms which are partitional or divisional in nature which takes the data and divides them into a number of groups; kmeans is an example of a partitional clustering algorithm. Then there are hierarchical clustering algorithms which hierarchical divides the data into clusters hierarchical algorithms may be based on top-down method or bottom-up method, which is done in adapted hierarchical clustering.

Again, we will see in that class. There are other methods like model based methods, for example, mixture of Gaussian, density based methods like DB SCAN, etcetera. So, there is clustering algorithm. Secondly, there is a distance of similarity function which the clustering algorithm uses or tries to optimize. We told earlier that in a cluster the elements in cluster are similar to each other and the elements belonging to 2 different clusters are different from each other and to measure how similar or dissimilar 2 elements are we have to use metric or similarity or dissimilarity measure. Some possible measures are Euclidean distance, cosine distance, Pearson and correlation coefficient etcetera.

Thirdly one has to have a way of evaluating, how good the cluster is for that one can look at different methods, for example one can try to minimize the intra-cluster distance of elements which belongs to the same cluster and maximize the inter-cluster distance elements that belong to different clusters. The quality of a clustering result depends on the algorithm the distance function used and the application for which you are using it.

(Refer Slide Time: 09:50)



Major Clustering Approaches


- Partitioning: Construct various partitions and then evaluate them by some criterion
- Hierarchical: Create a hierarchical decomposition of the set of objects using some criterion
- Model-based: Hypothesize a model for each cluster and find best fit of models to data
- Density-based: Guided by connectivity and density functions
- Graph-Theoretic Clustering

So, these are some of the major clustering approaches partitioning a base method, which involves constructing various partitions; hierarchical methods which creates a hierarchical decomposition of the set of objects; model-based methods which hypothesize a model for each cluster and finds the best fit of models to data; density based clustering algorithm which are guided by connectivity and density functions; graph theoretic clustering based on the under construction of a graph and looking at some graph theoretic measures like and so on. These are some of the different clustering approaches, few of them we will talk about in this class.

(Refer Slide Time: 10:39)

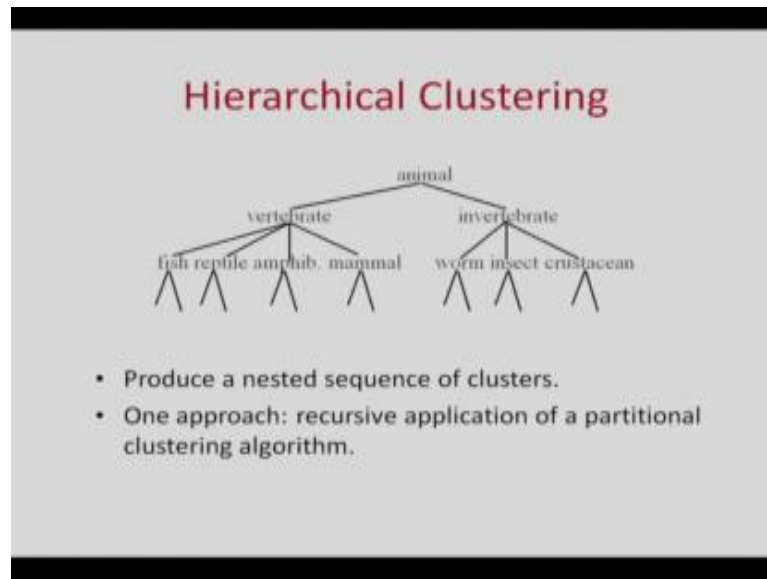
Partitioning Algorithms

- Partitioning method: Construct a partition of a database D of m objects into a set of k clusters
- Given a k , find a partition of k clusters that optimizes the chosen partitioning criterion
 - Global optimal: exhaustively enumerate all partitions
 - Heuristic method: k -means (MacQueen, 1967)



So, partitioning algorithms construct a partition of a given data sample. The partitioning algorithms will construct a partition of these into k clusters. So, k is given to the algorithm, it will find k clusters that optimize the chosen criteria. You may come up the algorithm may come up with a global optimum of the chosen criteria or use heuristic method and come up with a local optima. So, we will explore in detail the k means algorithm which comes up with a local optimum based on certain criteria.

(Refer Slide Time: 11:24)

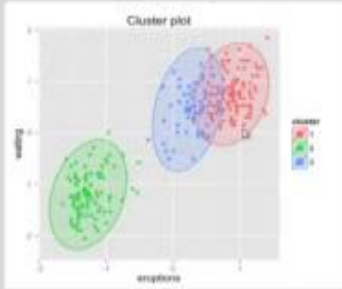


The second type of clustering algorithm is hierarchical clustering, for example, animals may be grouped into vertebrates and invertebrates; vertebrates may be broken into fish, reptiles, amphibians, mammals, etcetera. This is an example of using a tree or hierarchical clustering, we will look at some methods for hierarchical clustering which produces a nested sequence of clusters.


You can, for example, use a partitional algorithm and recursively apply it to get a hierarchical clustering or you may do a bottom-up clustering where you start with a large number of each cluster containing 1 item and repeatedly go on merging the clusters until you get 1 cluster and as a result you can get a nested tree. We will talk more about in the later class.

(Refer Slide Time: 12:18)

Model Based Clustering



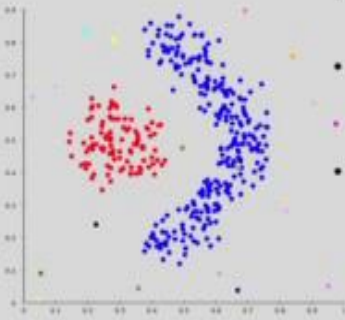
- A model is hypothesized
- e.g., Assume data is generated by a mixture of underlying probability distributions
- Fit the data to model




A third type of clustering is the model based cluster where given the data points you hypothesize a model, for example, you can think of each cluster being represented by a Gaussian distribution with a mean and a standard deviation and you try to fit the data to the model and, for example, you can come up with this three clusters.

(Refer Slide Time: 12:45)

Density based Clustering

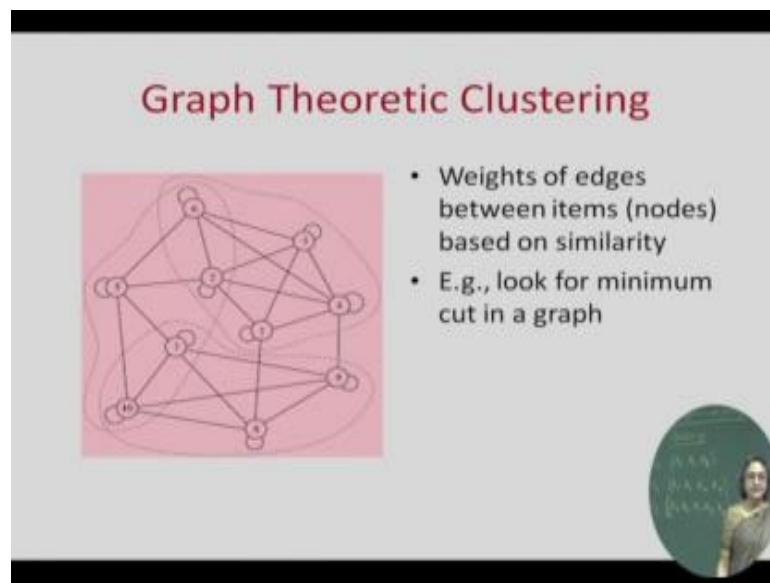


- Based on density connected points
- Locates regions of high density separated by regions of low density
- e.g., DBSCAN



A fourth type of algorithm is called density based clustering. We will not talk about it in this class if we do not have time, but it is based on the similarity, based on the density of a region. The density of a region is the number of instances in a region in feature space. It locates regions of high density and connects those points together. So, DB SCAN is a popular density based clustering algorithm.

(Refer Slide Time: 13:14)



Graph Theoretic Clustering

- Weights of edges between items (nodes) based on similarity
- E.g., look for minimum cut in a graph


The slide features a central diagram of a graph with 10 nodes and numerous edges, some of which are highlighted in red. To the right of the graph is a list of two bullet points. In the bottom right corner, there is a small circular inset image of a person standing in front of a chalkboard.

Then we have graph theoretic clustering algorithms which takes nodes to represent the different items and the weights of the edges is based on the similarity of the items, and based on this a graph is constructed and certain graph algorithms are used to find strongly connected components, for example, looking for minimum cut in a graph, again we will not talk about these type of algorithms in this class.

(Refer Slide Time: 13:46)

(Dis)similarity measures

- Correlation coefficients (scale-invariant)
- Mahalanobis distance
$$d(x_i, x_j) = \sqrt{(x_i - x_j)\Sigma^{-1}(x_i - x_j)}$$
- Pearson correlation
$$r(x_i, x_j) = \frac{\text{Cov}(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}}$$



The third aspect of a clustering algorithm is the metric that we use; a distance metric or a similarity metric. So, there are certain distance metrics for example, the Minkowski family of distance measures, where given 2 items X_i X_j .

(Refer Slide Time: 14:10)

Unsupervised Learning

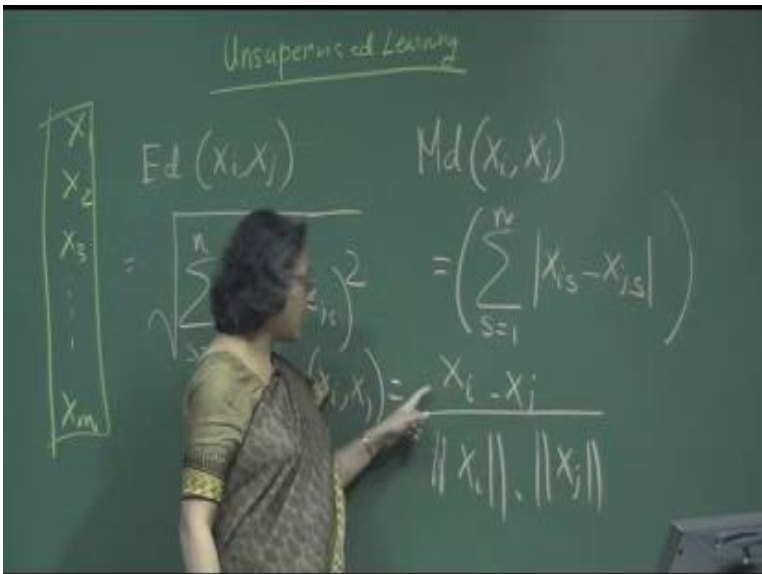
$E_d(x_i, x_j)$ $M_d(x_i, x_j)$

$= \sqrt[p]{\sum_{s=1}^n |x_{is} - x_{js}|^p}$ $= \left(\sum_{s=1}^n |x_{is} - x_{js}| \right)$

$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \vdots \\ x_{im} \end{bmatrix}$ $x_j = \begin{bmatrix} x_{j1} \\ x_{j2} \\ x_{j3} \\ \vdots \\ x_{jm} \end{bmatrix}$

$x_i - x_j = \begin{bmatrix} x_{i1} - x_{j1} \\ x_{i2} - x_{j2} \\ x_{i3} - x_{j3} \\ \vdots \\ x_{im} - x_{jm} \end{bmatrix}$

$\|x_i\|, \|x_j\|$



The Minkowski distance between d X_i and X_j is computed as it takes the summation

over the training examples. It takes the summation over, sorry the number of input attributes, let us say n is the input attribute X_{is} minus X_{js} to be power p the whole thing to the power $1/p$, this is the Minkowski metric. So, one popular now, if you said p equal to 2 you get the Euclidean distance, what you have is Euclidean distance of X_i, X_j is Minkowski distance, if p equal to 2 which gives you root over sigma s equal to 1 to n $X_{i1} - X_{j1} - X_{is} - X_{js}$, this is the Euclidean distance, or can you have the Manhattan distance between X_i and X_j as summation over s equal to 1 to n $X_{is} - X_{js}$, if you said p equal to 1 you get Manhattan distance which is defined like this; p equal to 2 you get Euclidean distance and you also use p equal to 3 4. So, this is 1 type of distance metric.

A second metric which is often used when you work with text data in the (Refer Time: 16:04) model is the cosine distance metric. So, given 2 objects; X_i and X_j are the vectors of these 2 objects you find the cosine between these 2 vectors and the cosine between these 2 vectors can be computed. So, the \cos of X_i, X_j can be computed as the dot product of the vectors $X_i \cdot X_j$ divided by root over the sum of this the coefficient of this. So, which we can write as $\|X_i\| \cdot \|X_j\|$, this is the normalization factor and this is the dot product. So, basically we are doing is that you are finding the cosine between the 2 vectors.

Then apart from such matrix there are correlation coefficients which are scale invariant and there are different such measures, for example, Mahalanobis distance, Pearson correlation coefficient, Spearman relation and so on. Mahalanobis distance $d_{X_i X_j}$ is taken as root over $(X_i - X_j)^T \Sigma^{-1} (X_i - X_j)$. So, this Σ is the co-variance matrix if they are independent then this is equal to the Euclidean distance and then we have the Pearson correlation coefficient which is given by co-variant of $X_i X_j$ divided by $\sigma_{X_i} \sigma_{X_j}$. So, σ_{X_i} stands for how far each of the items in a cluster is from the mean of the cluster and this is co-variant of $X_i X_j$. So, these are some of the similarity matrix that is used in the clustering algorithm.

(Refer Slide Time: 18:11)

Quality of Clustering

- Internal evaluation:
 - assign the best score to the algorithm that produces clusters with high similarity within a cluster and low similarity between clusters, e.g., Davies-Bouldin index

$$DB = \frac{1}{n} \sum_{i=1}^k \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)}$$

- External evaluation:
 - evaluated based on data such as known class labels and external benchmarks, eg, Rand Index, Jaccard Index, f-measure

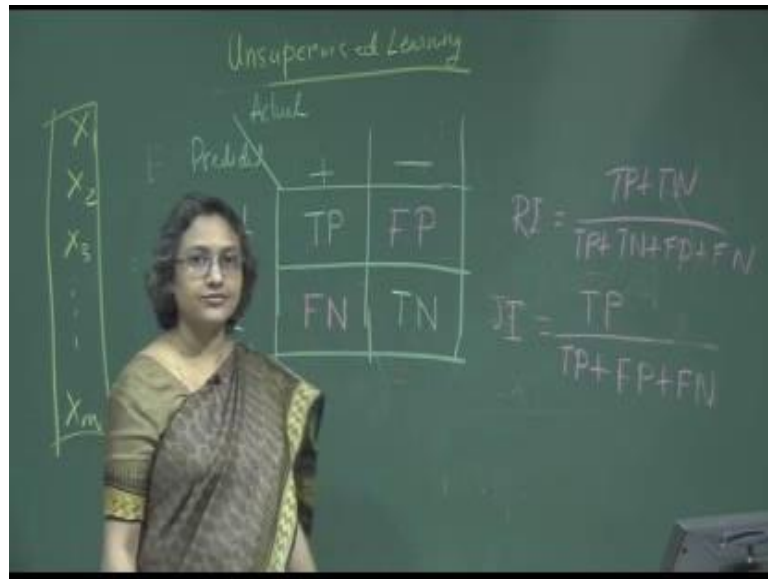
$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}$$

The next aspect which is of importance to us is how to measure the quality of the clustering algorithm. There are 2 types of measures of quality; it could be internal evaluation or external evaluation. In internal evaluation, you evaluate the cluster quality by the clusters and the data that you have whereas, the external evaluation you use additional data. For example, internal evaluation may try to look at how close the point in a cluster is with each other and how far they are from members to other clusters. There are many several such measures one of them is the Davies-Bouldin index.

Davies-Bouldin index is computed as $\frac{1}{n}$ summation over the number of clusters. You take items belonging to the same cluster which are j not equal to i . We look at $\sigma_i + \sigma_j$ divided by $d(c_i, c_j)$ this is the distance of c_i with c_j and this is the spread of σ_i and σ_j . So, this is one measure of the quality of algorithm. Then, if you have some label data, but you ignore the labels and come up with the clustering of the data you can compare the cluster that you have got with the labels that you have got and this gives you external evaluation, for example, there are different external evaluation matrix like f-measure, Jacquard index, Rand index etcetera.

Now, we have earlier talked about defined what we mean by true positive, true negative, false positive, false negative. To refresh your memory let me draw this table.

(Refer Slide Time: 20:31)



Suppose this is the actual label of the data and the actual label is plus or actual label is minus and this is predicted by an algorithm predicted plus predicted minus. Now, if you collect a set of items from which the actual label is plus and your algorithm has predicted plus they are true positive. If the actual label is minus and your algorithm has predicted minus it is called true negative, but if the actual label is plus and your algorithm has predicted minus, this is called false negative and if the actual label is minus your algorithm has predicted plus it is called false positive.

Now, the rand index, TP and TN. So, these are the 2 regions where your algorithms have worked correctly. Rand index is given by TP plus TN divided by the sum of all those, so the fraction of examples from which your algorithm has predicted the correct task. The Jacquard index is given by jacquard index of 2 sets AB is given by A intersection B divided by A union B. In this case, if you take the intersection of the predicted and the actual what you get it true positive. These are the ones which these are the intersection of those which both of the algorithm have predicted correctly divided by the union of those that any of the algorithms have predicted correctly. So, this is divided by TP plus FP plus FN. So, these are some possible matrix.

Then there is another common matrix called f-measure which is a harmonic mean of

precision algorithm. Precision is true positive divided by true positive plus false positive. Recall is true positive divided by true positive plus false negative and f-measure is the harmonic mean between them. So, these are some measures that are used for external evaluation.

With this I stop today's lecture, in the next class we will start with kmeans, which is an example of a partitional clustering algorithm. Then we will talk about one hierarchical clustering algorithm and we may talk about one model base clustering algorithm, which is mixture of Gaussian.

Thank you.