

Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture – 36
Clustering – V

Let us compare few clustering algorithms we studied.

(Refer Slide Time: 00:25)

The slide is titled "Summary of Clustering Algorithms" and lists the following:

- K-Means – fast, works only for data where mean can be defined, generates spherical clusters, robust to noise
- Single linkage – produces non-convex clusters, slow for large data sets, sensitive to noise
- Complete linkage – produces non-convex clusters, very sensitive to noise, very slow for large data sets
- DBSCAN – produces arbitrary shaped clusters – works only for low dimensional data

Handwritten diagrams include a circle labeled "Convex" and a crescent shape labeled "non convex".

The slide footer includes the IIT Kharagpur logo and "NPTEL ONLINE CERTIFICATION COURSES". A small video inset shows Prof. Pabitra Mitra.

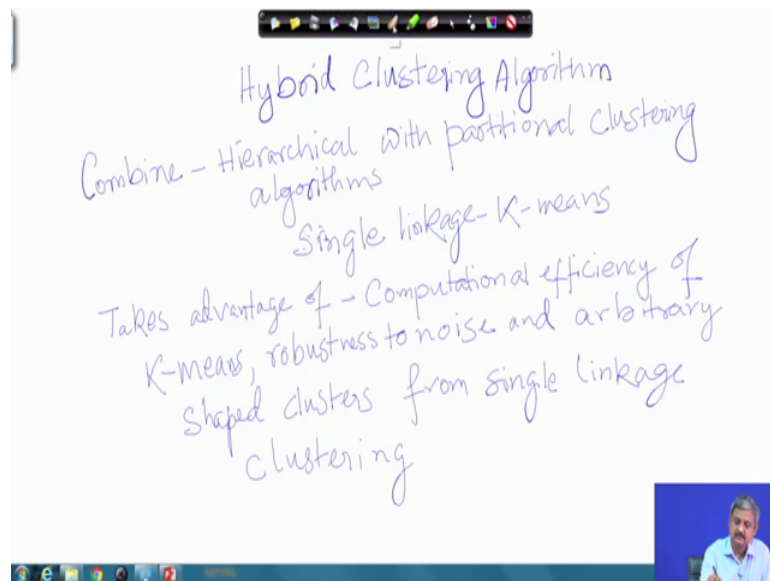
So, we studied three algorithms; the K – Means, the single linkage a hierarchical and the DBSCAN, density based. So, the advantage of K – Means is that is very fast, it is robust to noise, but it works only when the mean can be defined that is numerical attribute it does not work for nominal or ordinal attribute. So, it is difficult to define a mean. Also the disadvantage is it general spherical cluster.

So, if you the natural clustering is non convex, non spherical, by the way this is a spherical cluster would look like this. A convex cluster is anything which looks like this. So, this is convex and this is non convex. So, how do you know what is convex, what is non convex? You take two points inside a cluster draw a line the entire line will lie inside the cluster whereas, in non convex if you take two two points join them by line there will be some points on the line outside the cluster.

Ok. The single linkage provides non convex cluster if the natural cluster in it. It can also provide spherical cluster, but it is sensitive to noise and slow for large data whereas, the complete linkage is also non convex, but it is also sensitive to noise. Actually it is much slower than the single linkage this even more slower than single linkage, but it produces very good quality, elongated clusters if required.

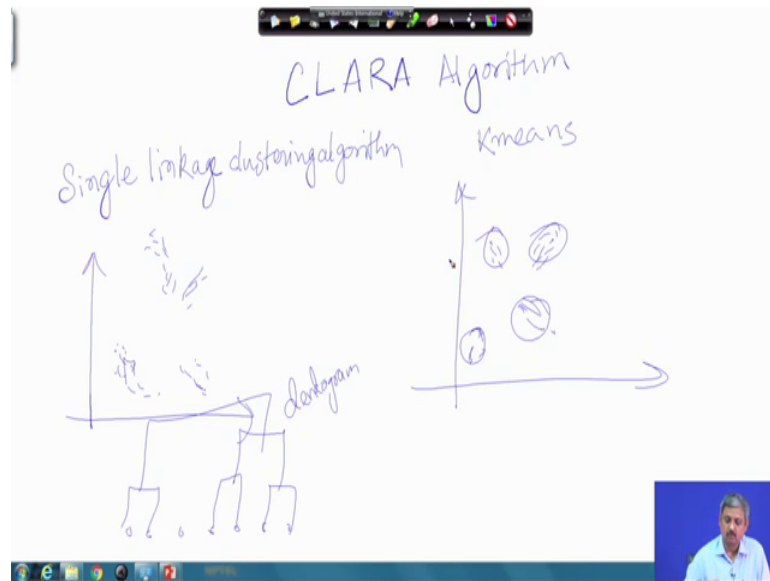
And, DBSCAN it provides density based clusters, you can get any shape, but because of the K nearest neighbour density estimation it will only work when the data dimension is low, less than 4. So, usually spatio temporal data there it works. There are many modifications to this algorithm for example, one modification to the K – means algorithm is the K-centroid or K-medoid, where other K-median also where instead of the mean being updated you take the centroid or the median getting updated.

(Refer Slide Time: 03:14)



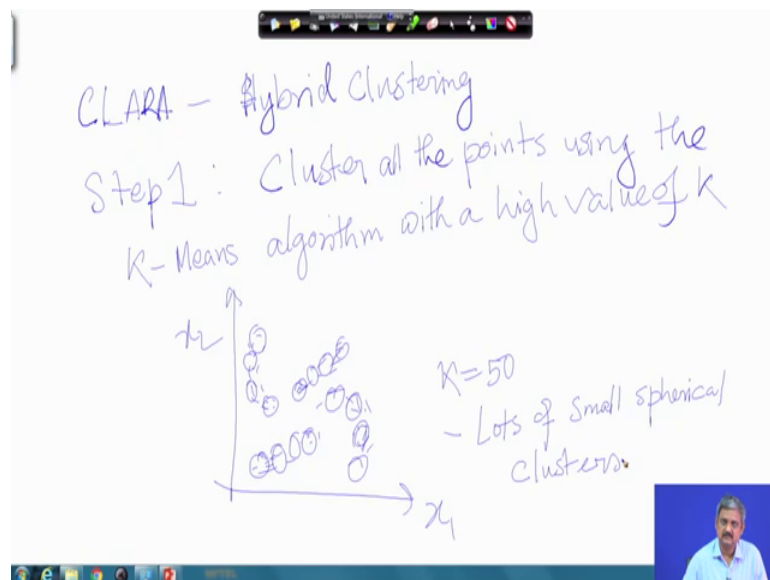
Ok. So, what people have done is that they algorithms say and clustering ok.

(Refer Slide Time: 05:44)



So, this algorithm is called the ok. So, if you remember what the does is it merges this point and creates a dendrogram, whereas, K – Means what it does is that it just forms a partition updating the mean, ok.

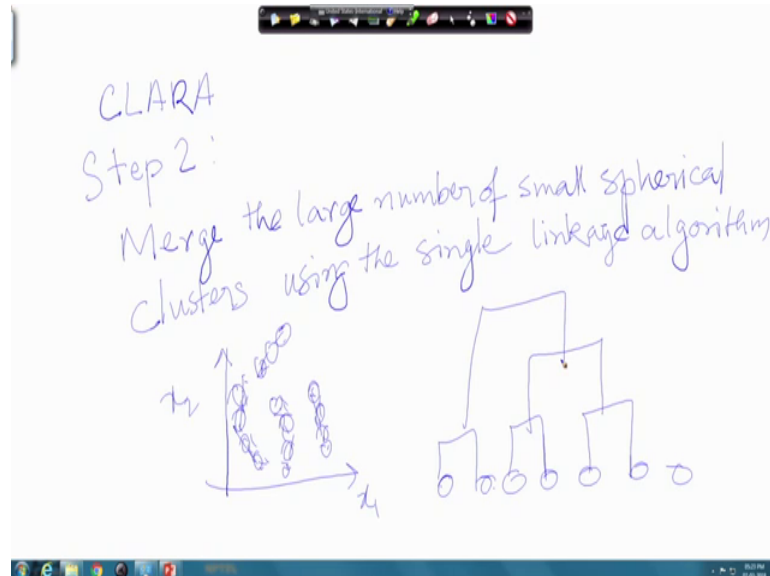
(Refer Slide Time: 07:20)



So, CLARA combines this two. How it combines? Choose a high value of K and run the K-means. So, these are really non convex clusters. So, you choose say K equal to 50 even the natural cluster is only 4. So, you choose K equal to 50 let lots of small small

cluster, clusters clear. So, run the high value of K with K – Means get small small circles to cover up the data oh, sorry merge the.

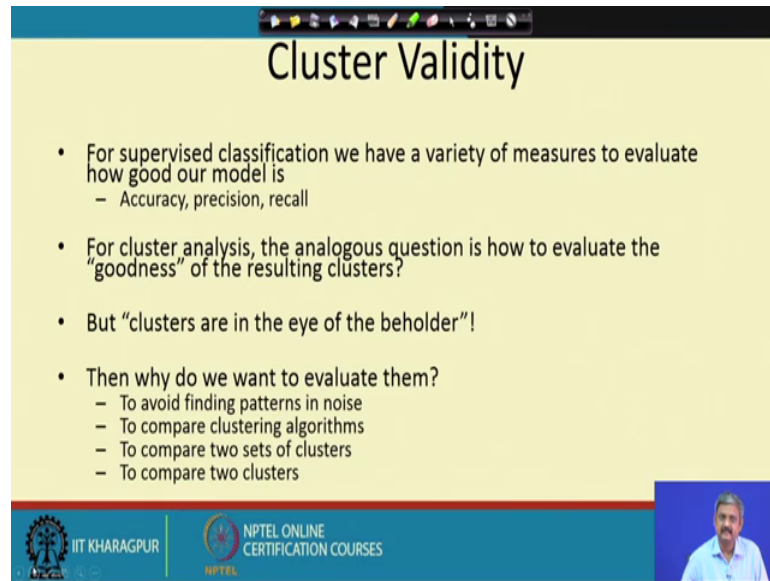
(Refer Slide Time: 09:33)



So, the points like this the first step you have what I do is I will merge them by, so, in single linkage interrupts single points swing leaps these small clusters will be leaps and maybe distance between them is distance between their centre of this point, the spheres merge.

So, here this since number of spheres are much less than the number of points, single linkage takes less time and also you get non convex cluster, ok. So, that is the idea. So, there are many other algorithms like that this is just an example. So, now, let us come to our next topic.

(Refer Slide Time: 11:34)



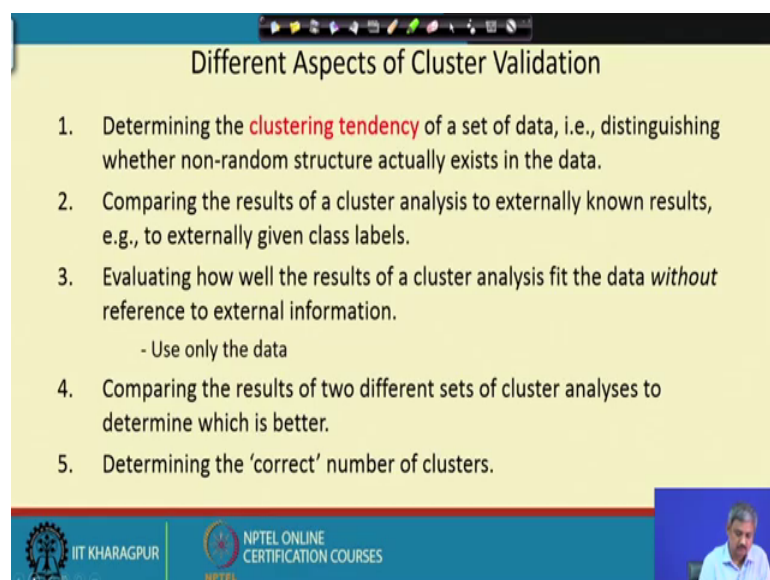
Cluster Validity

- For supervised classification we have a variety of measures to evaluate how good our model is
 - Accuracy, precision, recall
- For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters?
- But “clusters are in the eye of the beholder”!
- Then why do we want to evaluate them?
 - To avoid finding patterns in noise
 - To compare clustering algorithms
 - To compare two sets of clusters
 - To compare two clusters

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Evaluating clustering algorithms, how good are they? You have so much variety, which one do you choose? ok. So, for classification we have seen accuracy, precision, recall what do we do here, but the important thing is clustering depends on what is good cluster depends on application, but still we need to compare them because of this these this four reasons, ok.

(Refer Slide Time: 12:15)



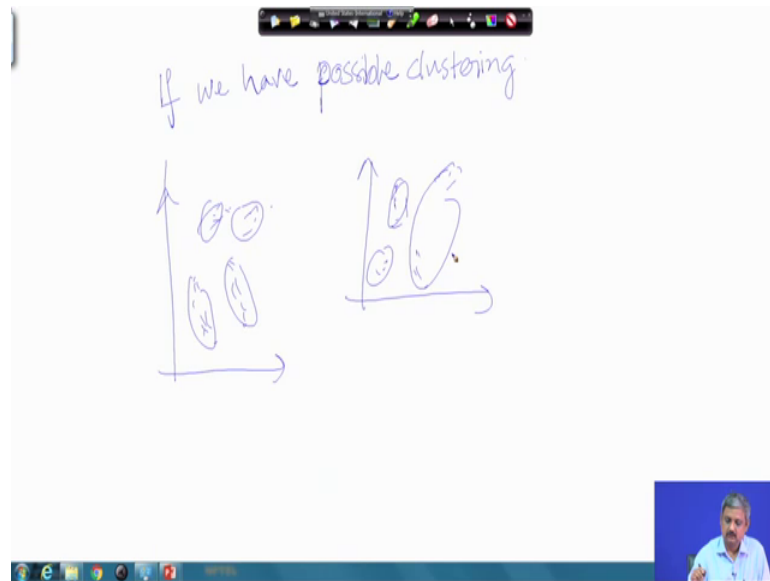
Different Aspects of Cluster Validation

1. Determining the **clustering tendency** of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.
 - Use only the data
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the ‘correct’ number of clusters.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So let us see you have to these are the regions, tendency. Can you compare with existing class levels?

(Refer Slide Time: 12:54)



Using only the data, how well can you evaluate? How does two clustering compare? So, this is one, this is another, ok. How good that, how does they compare?

(Refer Slide Time: 13:42)

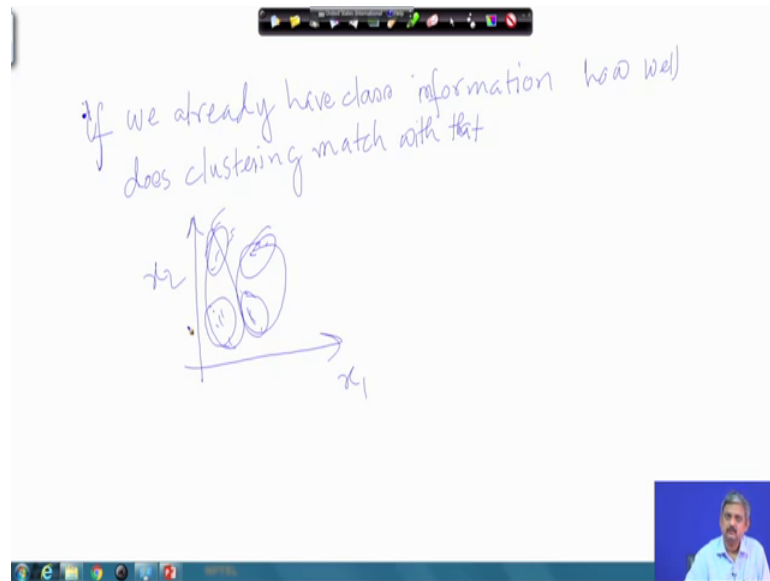
Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
 - **External Index:** Used to measure the extent to which cluster labels match externally supplied class labels.
 - Entropy
 - **Internal Index:** Used to measure the goodness of a clustering structure *without* respect to external information.
 - Sum of Squared Error (SSE)
 - **Relative Index:** Used to compare two different clusterings or clusters.
 - Often an external or internal index is used for this function, e.g., SSE or entropy

The slide includes logos for IIT KHARAGPUR and NPTEL ONLINE CERTIFICATION COURSES. A small video inset of a man is visible in the bottom right corner.

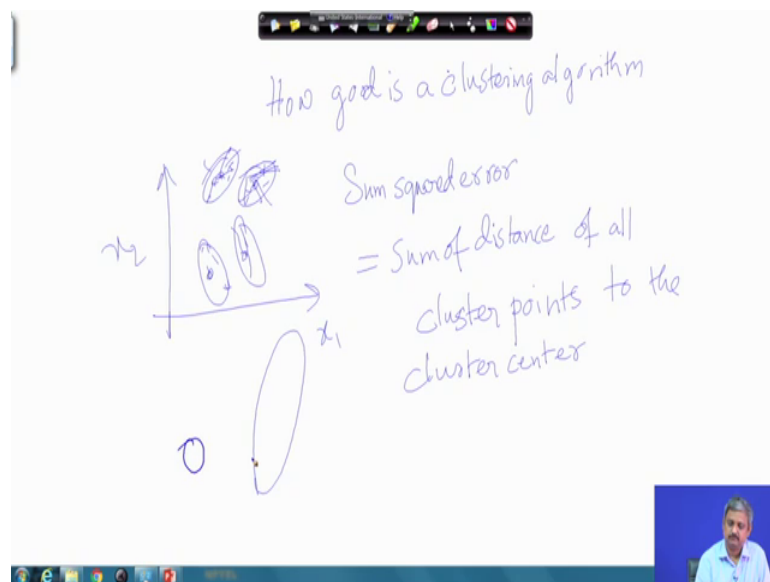
There are many measures called cluster indices index, external, internal, relative. External means there is already some ground truth some class level, how can we compare against them.

(Refer Slide Time: 14:10)



Meaning, suppose class level is this already known; given. How well if you cluster like this it matches with that? That is external.

(Refer Slide Time: 15:18)



Internal without knowing class level, how much can I say sum square error; so, in K – Means if some of this part these distances, ok.

So if the clustering is dense it will be low, it will be good. If error will be low, it is high, it is not good and that is a relative performance.

(Refer Slide Time: 16:43)

Scatter Coefficient

- Cluster evaluation index
- Ratio of average intra-cluster distances to intra-cluster distances (Sum Squared Error)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, the most common squared sum error is the scatter coefficient, which we discussed.

(Refer Slide Time: 17:07)

Avg. Intra cluster distance

Avg. Inter cluster distances

= Scatter

It is nothing, but the ratio of so, it is the; take any two points lying in the same cluster, this distances average of that, all these distances. Is the scatter, so, that is what we have defined.



(Refer Slide Time: 18:18)

Internal Measures: Cohesion and Separation

- **Cluster Cohesion:** Measures how closely related are objects in a cluster
 - Example: SSE
- **Cluster Separation:** Measure how distinct or well-separated a cluster is from other clusters
 - Example: Squared Error
 - Cohesion is measured by the within cluster sum of squares (SSE)

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$
 - Separation is measured by the between cluster sum of squares

$$BSS = \sum_i |C_i| (m - m_i)^2$$
 - Where $|C_i|$ is the size of cluster i





So, these two terms are as I have told you these distances, these are called cohesion and separation and scatter is ratio of cohesion to separation.

(Refer Slide Time: 18:54)

Internal Measures: Cohesion and Separation

- Example: SSE
 - $BSS + WSS = \text{constant}$



K=1 cluster:

$$WSS = (1 - 3)^2 + (2 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 10$$

$$BSS = 4 \times (3 - 3)^2 = 0$$



$$\text{Total} = 10 + 0 = 10$$

K=2 clusters:

$$WSS = (1 - 1.5)^2 + (2 - 1.5)^2 + (4 - 4.5)^2 + (5 - 4.5)^2 = 1$$

$$BSS = 2 \times (3 - 1.5)^2 + 2 \times (4.5 - 3)^2 = 9$$

$\text{Total} = 1 + 9 = 10$

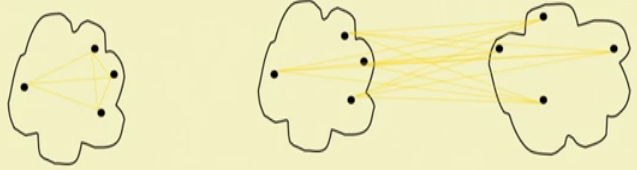



So, here is an example. So, this is within sum of error between sum of error. Here is an example of how we are getting these values, ok. So you, so, if you take one single cluster this is the mean this is the value, and if we take two clusters, red and green this is the value. So, you can use it to compare, ok.

(Refer Slide Time: 19:35)

Internal Measures: Cohesion and Separation

- A proximity graph based approach can also be used for cohesion and separation.
 - Cluster cohesion is the sum of the weight of all links within a cluster.
 - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion separation

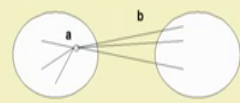
IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, as I have mentioned.

(Refer Slide Time: 19:42)

Internal Measures: Silhouette Coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, i
 - Calculate a = average distance of i to the points in its cluster
 - Calculate b = min (average distance of i to points in another cluster)
 - The silhouette coefficient for a point is then given by
$$s = 1 - a/b \text{ if } a < b, \text{ (or } s = b/a - 1 \text{ if } a \geq b, \text{ not the usual case)}$$
 - Typically between 0 and 1.
 - The closer to 1 the better.
- Can calculate the Average Silhouette width for a cluster or a clustering



IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

There is another important coefficient called the silhouette coefficient, ok. The silhouette coefficient, which is combination of these two so, you take this sorry. Similarly, you take this and if a is less than b silhouette is this, if b is less than s silhouette is this it is between 0 and 1; closer it is to 1, the better and you then calculate the average silhouette for a all the clusters. So, you can calculate the silhouette of this ok, how to get that, all right. So, this can be calculated.

(Refer Slide Time: 21:19)


External Measures of Cluster Validity: Entropy and Purity

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij} , the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j . Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = -\sum_{i=1}^L p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{j=1}^K \frac{m_j}{m} e_j$, where m_j is the size of cluster j , K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j , is given by $\max_i p_{ij}$ and the overall purity of a clustering by $\text{purity} = \sum_{j=1}^K \frac{m_j}{m} \text{purity}_j$.




There are some external measures also, which are entropy another which are defined here. This is an example for some data set entropy and purity, like that is (Refer Time: 21:33) we have discussed earlier you can consider that also.

(Refer Slide Time: 21:36)

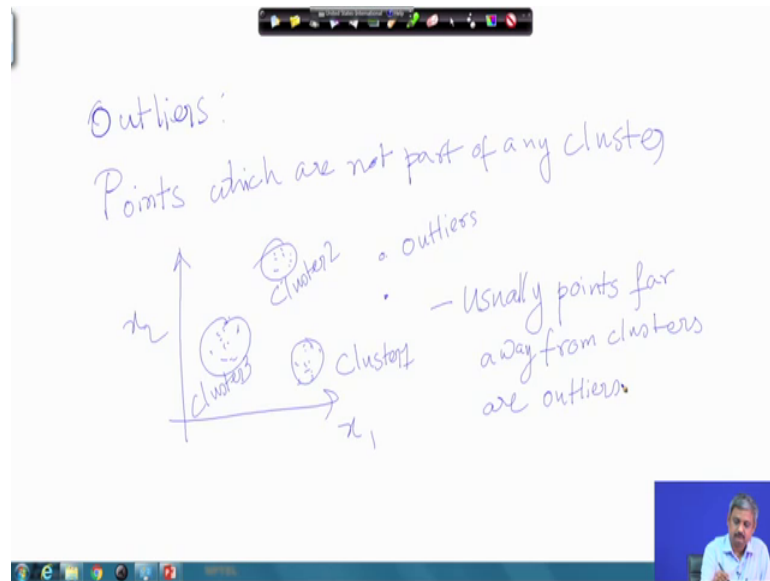
Outliers Detection

- Important in many applications like anomaly detection
- Outliers are points not belonging to any cluster
- Many outlier detection algorithms available



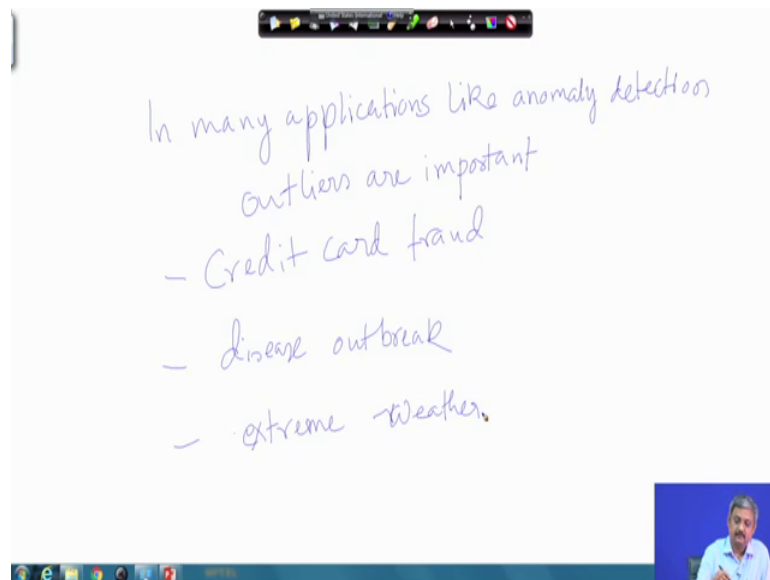
Another important related thing to outlier is into clustering his outlier detection this kind of complement of cluster, this is not a cluster, ok.

(Refer Slide Time: 22:12)



So, in many applications, actually outlier detection is more important than anything. So, ok, these are outliers.

(Refer Slide Time: 23:34)



Important things like this ok. So, that is what is outliers.

So, with this I complete my discussion on the clustering and outlier detection, an important class of unsupervised algorithm. In the next lectures we will look at other type of data mining tasks.

Thank you.