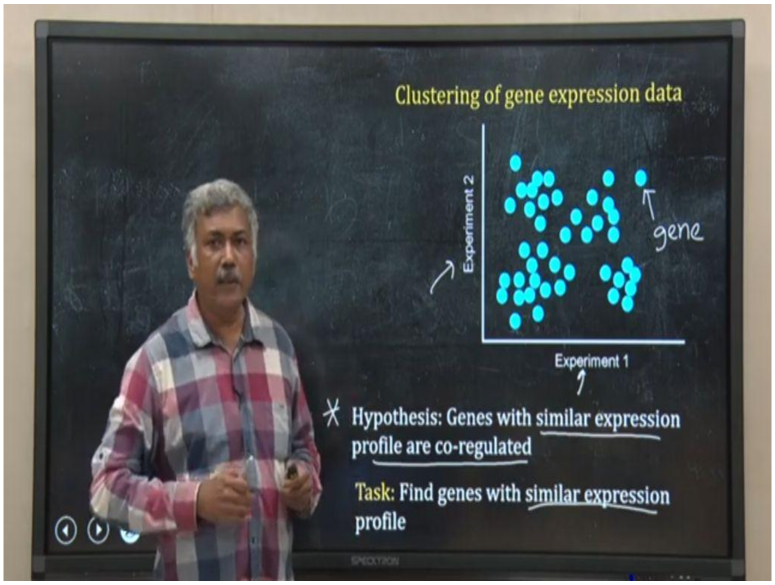
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences and Bioengineering
Mehta Family School of Data Science and Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture – 34
Clustering and Classification

Welcome back. Clustering of data and classification of data are two key component of machine learning. In this course, we will learn some algorithm of data clustering and data classification. In today's lecture, I will discuss the basic concept of clustering and classification of data; and I will try to explain what is the difference between these two, what is the difference between clustering and classification? So, let me start with clustering and let me start with the gene expression data.

(Refer Slide Time: 01:04)



Clustering of gene expression data

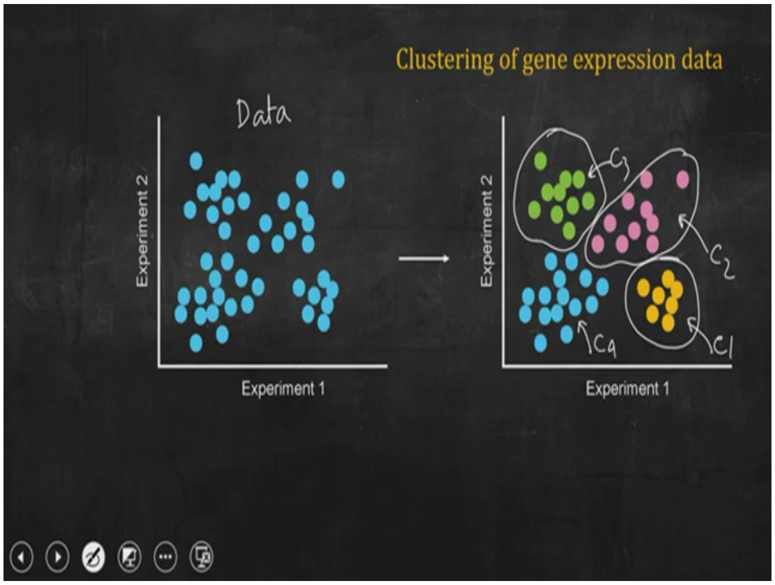
Hypothesis: Genes with similar expression profile are co-regulated

Task: Find genes with similar expression profile

Suppose, I am doing a gene expression experiment, and I have two experimental condition, 1 and 2. And each of these data points are actually one gene individual gene. You can imagine I am doing some high throughput experiment like RNA-seq or microarray. Now, usually when we do this type of experiment, we try to identify genes which are co-regulated. What do I mean by co-regulated, means genes which have control, which are controlled by same pathway, same transcription factors.

So, you expect them to behave in the same fashion in your experimental condition. So, the hypothesis for my work in this case is that genes with similar expression are actually co-regulated. So, my task would be that I have these hundreds of thousands of gene expression data for this experiment of condition; and I have to find out genes which has similar expression, similar expression profile. So, that is where the an algorithm of clustering will help us. So, what we will do?

(Refer Slide Time: 02:10)



Clustering of gene expression data

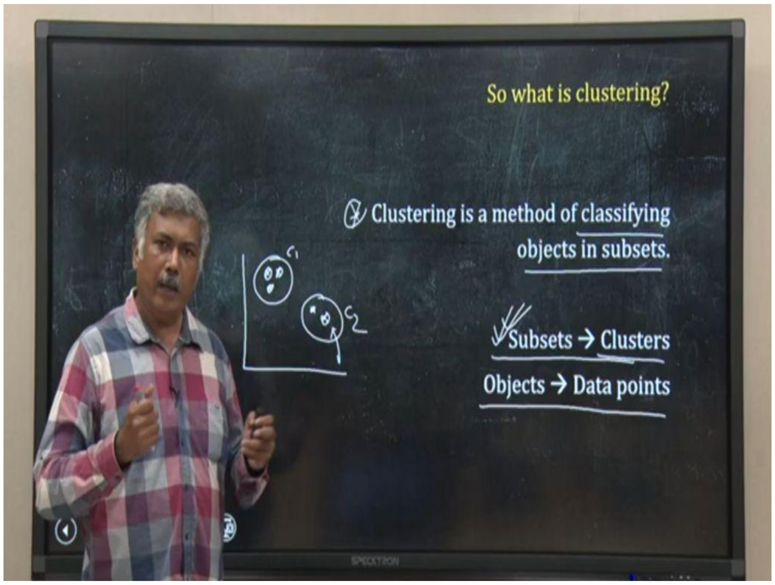
So, this is my original data. And you can see this is quite a heterogeneous data, different gene has different level of expression in these two different experimental condition. Now, out of this heterogeneous population, I want to segregate these genes into some closely packed or to some extent homogeneous sub-population or subset; and that is what the algorithm of clustering has done.

You can see it has clubbed all these genes which are now marked by green as one unit as one subset, because the algorithm has detected that these genes have closer similar gene expression level. Whereas, these are you can visually also see easily these genes the yellow one, they have similar or close gene expression level; that is why they are in another cluster.

So, maybe I can say this is cluster1, this pink one in another cluster, cluster2, this is cluster3, and the rest one is cluster4. So, in this particular case, I am showing a two dimensional data, I have two experimental conditions; but in your experiment it may be 10 dimensional, 10 different condition you may have. You may have samples coming from 20 different personal patients or it may be coming from different grade of the disease.

So, you may have n dimensional data and we can do this type of clustering on n dimensional data also. Although I will not be able to visualize in this two dimensional space, so I can do clustering that means segregation of these data points into subset even in n dimension. And many a time what we do and we will also learn that we can actually reduce this dimension from n to something lower two or three, so that that can help in visualization as well as in clustering. So, what we are doing in clustering?

(Refer Slide Time: 04:06)



So what is clustering?

Clustering is a method of classifying objects in subsets.

Subsets → Clusters

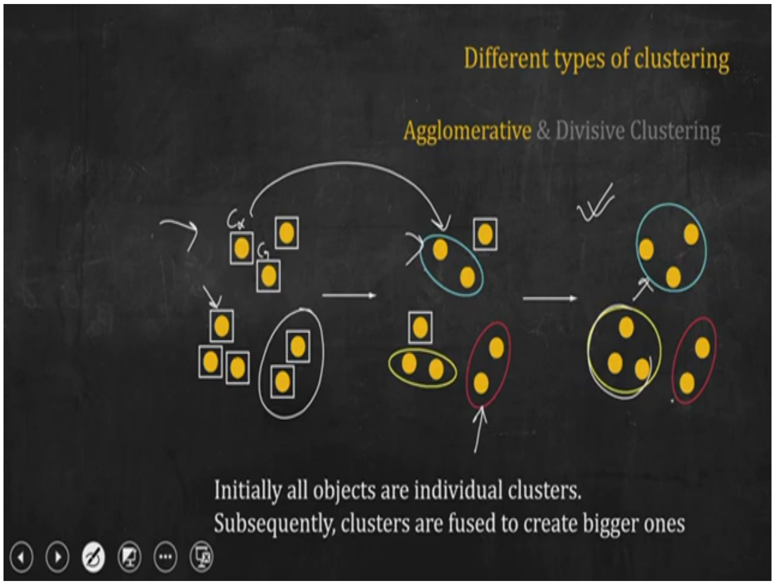
Objects → Data points

Let me write it down in words. So, clustering is a method of classifying objects in subsets. In this example objects are genes, and you are classifying them in subsets, those subsets are clusters. So, subsets are clusters and you are classifying or you are putting these objects, these genes in my example to these clusters to this subset. Based on what, based on some measure of closeness between them.

So, if I have this data, these are the genes; if these three belongs to this particular cluster C1 and I have another cluster C2. So, these three genes in cluster1, they are close to each other in by some definition, some by some measure. Whereas, in C2, these two genes are close to each other, whereas the genes in C1 and C2 are far away from each other; that is why they have been segregated in 2 different clusters. Now, what will be the measure of that closeness? We will have a separate lecture on different measures of distance that we use for clustering.

For the time being, let us accept that we have some measure of closeness between these data points, which we call objects; and based on them I distribute them in different subsets. So, what is happening in clustering, you are starting with a heterogeneous population, and you are now breaking them down into subsets of population or sub population, which are largely homogeneous. Now, clustering can be of multiple type; you may have heard about it also, agglomerative, divisive, flat, hierarchical; let me briefly go through what are those.

(Refer Slide Time: 06:01)



Different types of clustering

Agglomerative & Divisive Clustering

Initially all objects are individual clusters.
Subsequently, clusters are fused to create bigger ones

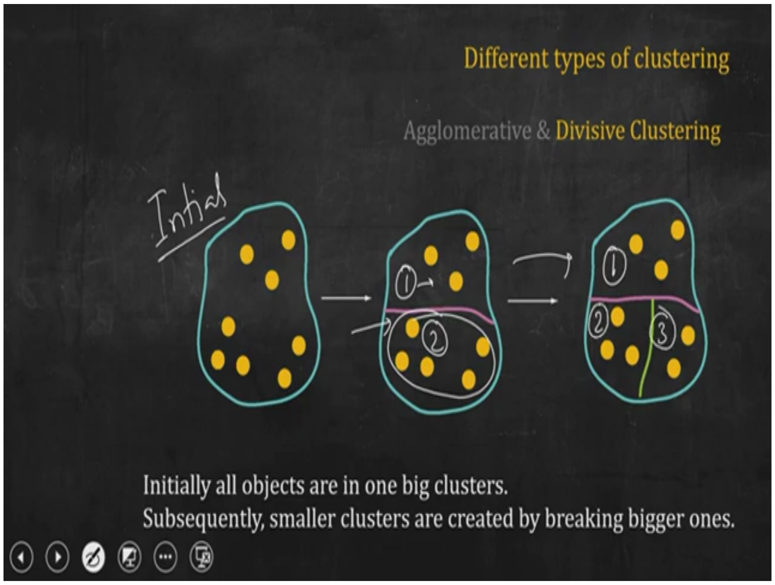
F First agglomerative and divisive cluster; that is what we will discuss and first I will discuss what is agglomerative clustering. You start agglomerative clustering considering that each object mean each data point is itself a cluster; so that is what I have shown here. These square thing represent a cluster and each of these cluster has only one object, so every object is a cluster.

So, if I have how many three plus three six plus two, eight data points; I have eight cluster right now at the beginning of my algorithm, when the algorithm is starting in the first step of the algorithm. Subsequently what I will do? I will find out which clusters in this data set, which clusters are close to each other. So, I can see these clusters suppose I mark it cx and cy.

This cx and cy are close to each other that is why I am now joining those two cluster to create a new cluster marked by this blue curve, blue ellipse. In this way, we have found that these two clusters are close to each other. So we have created a new cluster by fusing these two clusters together and a bigger cluster is formed. In this way, we keep on fusing clusters which are close to each other. And eventually I get this result where I have three cluster, this is one cluster, this is another cluster, this is the third cluster.

So, in agglomerative clustering, what you are doing you are starting with everything as a cluster; and you are then joining, fusing them together to create a bigger cluster. In divisive clustering, we will do just the opposite thing. What is that?

(Refer Slide Time: 07:43)



Different types of clustering

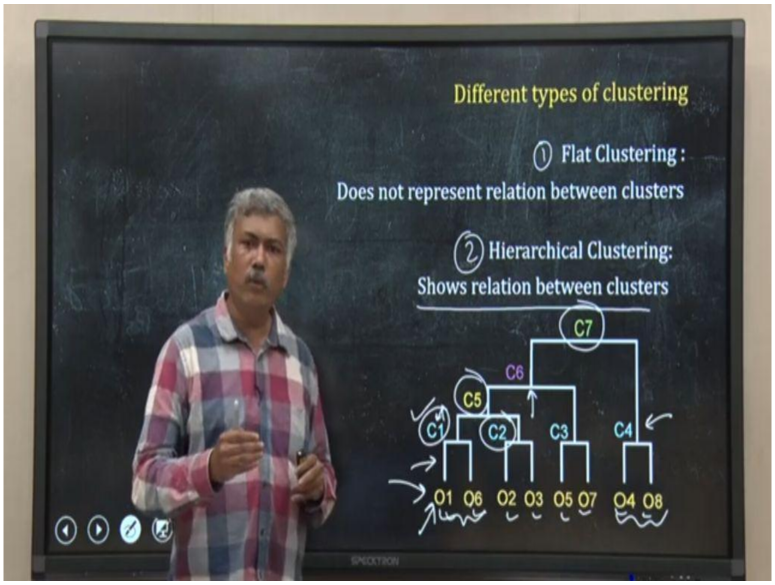
Agglomerative & Divisive Clustering

Initially all objects are in one big clusters.
Subsequently, smaller clusters are created by breaking bigger ones.

You initially consider this is the initial state. Initially, you consider that everything belongs to a big cluster, everything. All your genes are part of the experiment are part of that one single big cluster. Now using some rule, based on obviously closeness of this data point or objects, in this case genes, I try to break this bigger cluster into smaller one; that is what we have done here. We have broken them in two part 1 and 2.

The reasoning that these three things in one are closer to each other, then those object present in the second one. So, I am now starting with the bigger set, I am breaking down them in subset and then I continue; and finally, what I do I divide this whole thing, whole set into three subset. So, you can easily see that is why it is called divisive clustering. Another way of dividing clustering algorithm and methods is called a flat and hierarchical. What are those?

(Refer Slide Time: 08:53)



Different types of clustering

1. Flat Clustering : Does not represent relation between clusters
2. Hierarchical Clustering: Shows relation between clusters

So in flat clustering, let me start first. In flat clustering, what you will do? You will do clustering, whatever way you want either agglomerative or divisive, it does not matter. But, you do not want to show any relationship between these clusters. So, essentially flat, you just have the clusters given to you. On the other hand, for hierarchical clustering, you want to show the relationship between clusters. Now, what do I mean by relationship between cluster?

Let us see this diagram and that will explain what I mean by the relationship between different clusters. So, I have objects, for example in my gene expression experiment, these can be genes O1 to up to O8. So, you imagine this way that I am using some agglomerative clustering algorithm. So initially, each of these are individual clusters.

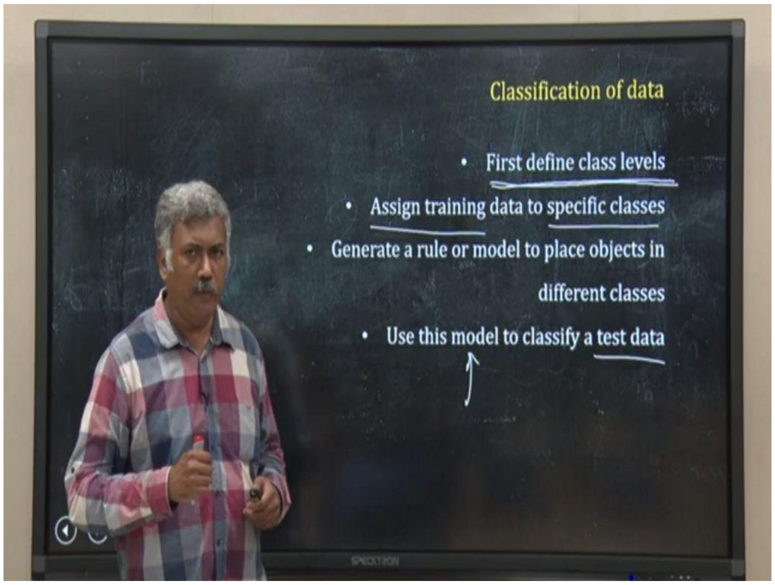
And then I found, calculated that which of these objects O1 to O8 are close to each other using some definition of distance or closeness; and I found that one object 1 and object 6 are close to each other. So, I fuse these two and create a cluster call C1; and I am recording this information that I have created C1 by fusing O1 and O6 by showing this type of tree like or dendogram like structure, that O1 and O6 gives rise to C1.

The same thing is done for C4, which has been created by fusing O4 and O8 and so on. Now, then in the next step, maybe I have actually created cluster by fusing C1 and C2, so I get a new bigger cluster C5. So, again the dendogram rises, you may have seen this in phylogenetic tree, the same dendogram. So, these are the leaves and these are the internal nodes. So each of these internal node is a cluster. So, I keep on fusing these clusters and eventually the whole data set is represented by these bigger clusters C7.

So in a way, when you represent your clustering data in this fashion, you not only know the clusters but you also know which cluster is close to which cluster, and how they are related to each other. So, in many gene expression studies thats why many time people try to draw this type; you must have seen in journal and book, in presentation you must seen this type of dendogram or tree to represent the cluster data.

Now, we have learned about two or three, four different types of clustering and we have understood what is essentially clustering. So, let me now move into what is classification of data, this is bit different.

(Refer Slide Time: 11:51)



Classification of data

- First define class levels
- Assign training data to specific classes
- Generate a rule or model to place objects in different classes
- Use this model to classify a test data

Now, in classification of data we are not just taking the whole data set or objects and then dividing them into some subsets. What we do? Apart from taking that data we also have some additional information. Based on that information what we will do? We will classify, we'll group this data which has with us into different classes. So, that means I need two things.

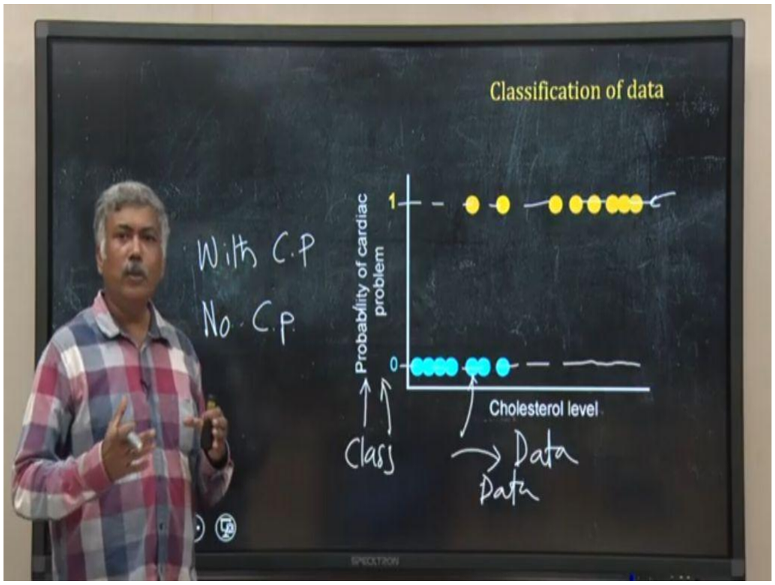
One: I need something called training data, for which I have the data. For example, gene expression or measurement of some blood parameters, I have those data. And at the same time, I have some other information; for example, whether those people who have given the blood or the sample for DNA analysis are diseased or not diseased.

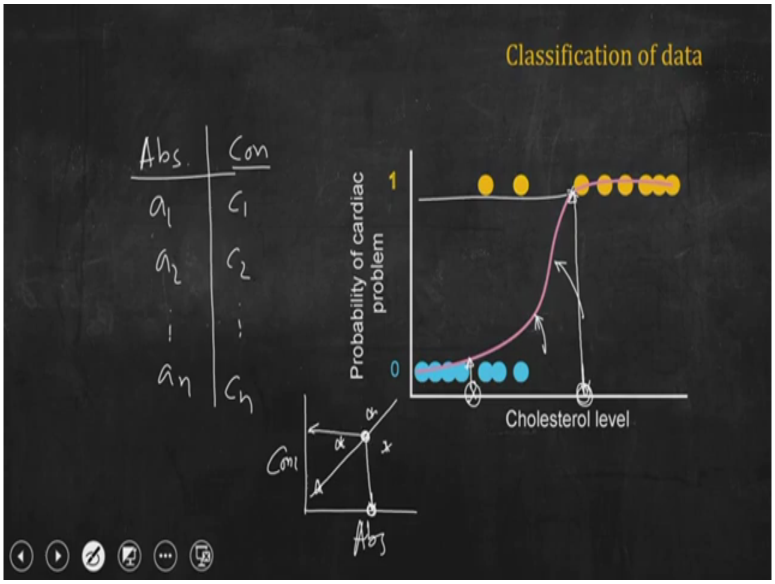
So, we define certain classes, certain groups first based on that additional information, and we call them class. And we call them these features which we use to define classes as class labels. Now, I have a training data set, I take that training data set and I will distribute this training data set into those classes; so, I assigned training data to specific classes.

Now, I try to create a model or essentially we try to extract some rule from this data and segregate it in different classes, so that we can explain or we can rather connect different classes and the data and the measurement. Eventually, when a new data comes that is a test data, we use this model that we have derived to predict in which class these new data test data should go.

It may look bit confusing, I will give an example; and it will be able to understand what we are doing here.

(Refer Slide Time: 13:45)





Classification of data

Abs.	Con
a_1	c_1
a_2	c_2
⋮	⋮
a_n	c_n

So, suppose I am measuring cholesterol level in some volunteers. I have suppose taken 20 volunteers and I have measured cholesterol level for each of them. So, now this is the data cholesterol level is the data or in a way you can call the feature. Now, apart from that I have additional information, what is that additional information? I know whether the person who is giving me the blood has cardiac problem, or he or she does not have cardiac problem.

So, I create two groups, two bins, two classes; one with problem with cardiac problem and no cardiac problems, these two group I create. So, I have collected the blood parameters of the cholesterol level for all these 20 people; and I have segregated them in two groups, with cardiac problem without cardiac problem. Those people those person who has cardiac problem, I have marked them by yellow; and those without the cardiac problem, I have marked them by blue.

And what I have done in the vertical axis? In the vertical axis I have kept probability of having a cardiac problem. Now, for those people who I know they have cardiac problem, that means their probability is one; that is why they are on this line. Whereas, those people I know they do not have the cardiac problem that means the probability of having a cardiac problem is zero. So what I have done here?

Unlike the clustering problem, in clustering problem I have only a data, maybe the cholesterol level, some other blood parameter or something like that. Apart from that, what I have here, I have additional information whether the person has the disease or not, two things, two levels.

And I have classified these people, these people, these 20 people whose blood cholesterol level I know I have measured and I know their disease state, we will call them training set.

So for this training set, I have classified them into two groups, two classes, having disease not having disease. Now, I will try to create a model to explain that the relation to understand the relation between what is in this axis? The probability of having disease and this data, the cholesterol level; mean this is the class and this is the data features.

So, I want to create a model to connect these two and what we will do actually? We will study this one. We will do logistic regression for this type of problem and the model will eventually give this type of sigmoidal curve. Now, this curve will be now used to classify people or test sample whom we have not studied earlier. Suppose, we have already done this I have got this pink curve, then suppose a new patient come and we found that that person has cholesterol level somewhere here.

Then I say your cholesterol level is here; that means your probability of having the cardiac problem is high, you are in the high risk group. Whereas, if someone comes with a cholesterol level here, I will say okay, that person falls into no cardiac problem. So, what I am doing now? I am using this line this model, which I have generated to predict in which class the test sample should go. So, now I am classifying the test sample in different groups.

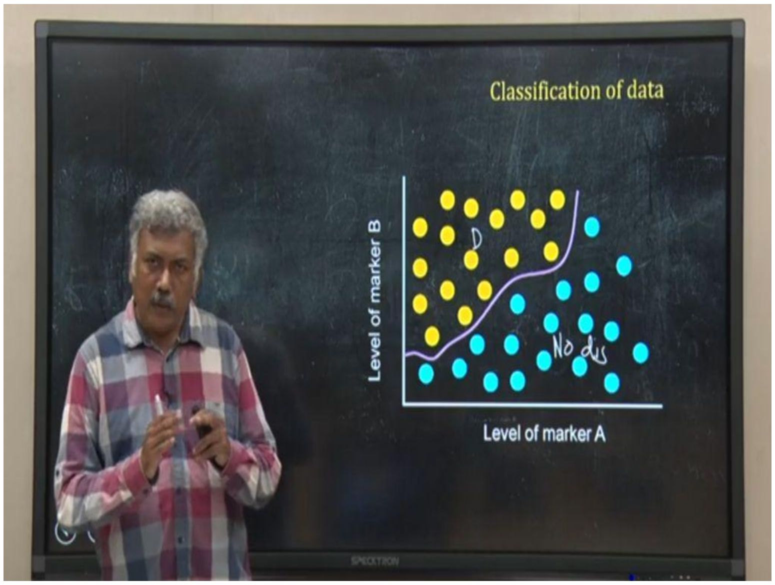
If you think over it, it is almost like doing regression, linear regression actually. Although this is not a linear one, but in conceptually it is almost similar, I will explain that. Suppose, you have different concentration of a protein sample and you have taken the absorbance of that; so you have absorbance versus concentration. You have a data, so you have a1, c1, a2, c2 like that an, cn. And then what you do you plot this data, is not it?

You plot absorbance versus concentration. So, your absorbance keeps on increasing as your concentration increases and you may have fitted line like this. Now, you take an unknown sample with unknown concentration test sample, so suppose that gives a absorbance here. So, from using this line straight line that you have got by regression, you predict what will be the concentration of that sample.

The same thing you are doing here in classification, you have the data points and you have additional information called label information. Based on that label information you are classifying the training data set. This is my training data set; these are my training data set in regression, my training data set in classes. In the example that I have shown here I have two classes disease or no disease, you can have more classes.

So, you are dividing, distributing the training data set in different classes; and then you are building a model to connect these classes and the data. And that is what we have done here using the pink line. And now you are using that new model the pink line in this case, to predict in which class a test sample should go, a test data should go. In this case, I have shown a sigmoidal line sigmoidal model, which you usually get from regression, logistic regression; but it is not to be like that for all cases, it can be a bit complicated also.

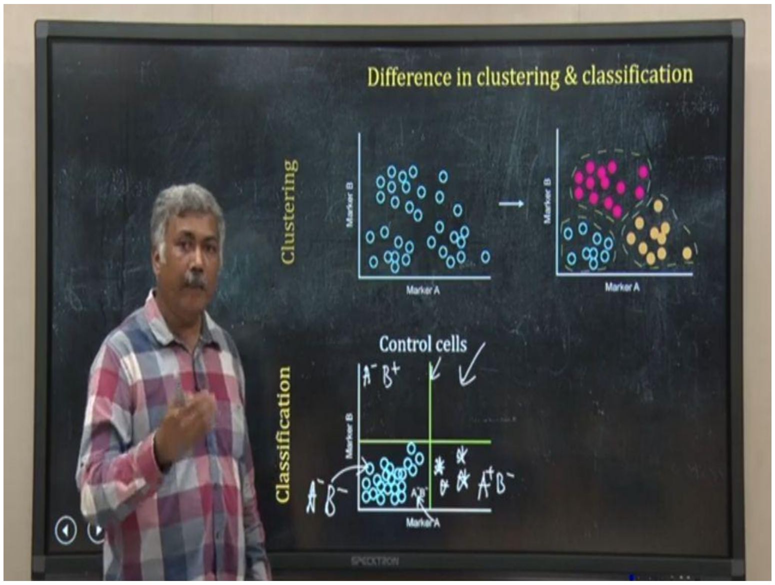
(Refer Slide Time: 20:00)


Classification of data

For example, suppose I have a label of a particular marker, gene expression suppose, and your two markers A and B. And based on that you want to classify a disease, you may have more markers there. So, you have a higher dimensional data and you want to class, you also know the disease state of this sample. So, you want to now create a model which will segregate these two and it will allow you to in future to classify new samples.

So, the model could be something like this, something which is dividing this is supposed disease, this is no disease. Let us take another example to clarify the difference between classification and cluster.

(Refer Slide Time: 20:49)

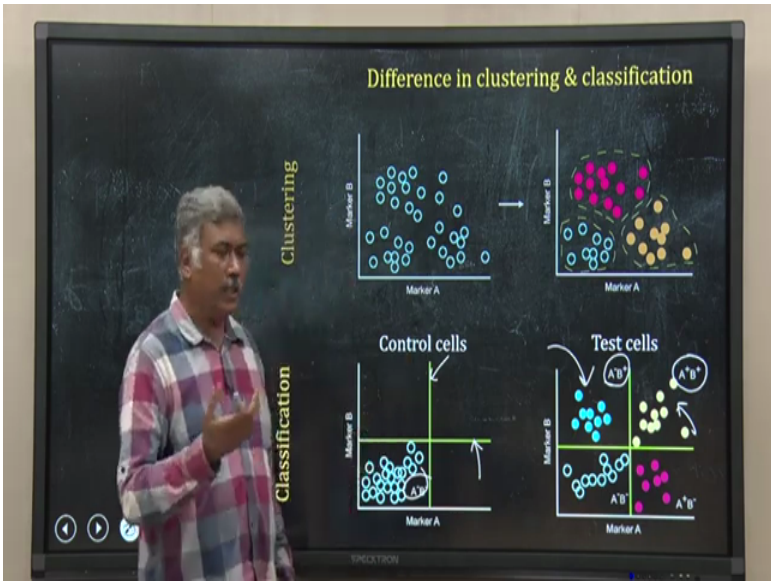


Difference in clustering & classification

Clustering

Classification

Control cells



Difference in clustering & classification

Clustering

Classification

Control cells

Test cells

I will take the example of flow cytometry. If you do not know what is flow cytometer is, let me explain what we do in that in brief. So, you have cells and in those cells, you have some expression of some molecules; and you have some probe which are fluorescently labeled. So, you mix those probe with the cell sample. And then what you do?

You pass this sample through the machine where a laser hit; and if the molecule you are chasing is there in that cell, in a particular one cell; then what will happen? A light fluorescent light will be emitted and your machine will capture and detect. So, for example, if I am doing chasing two molecules, marker A, marker B for example, CD-24 and CD-44, which are cell surface marker.

And I have used a antibody tagged with fluorophore for these two proteins separate, and maybe my data may look like this. Now remember, each of these points is an individual cell.

Now, I know in this population there is a heterogeneous cell population. Some may have one marker high, some may have two marker both the marker high, some may have all the both the marker low something like that. My data is simply for each data point I have only the value for the marker A and marker B.

Now, I use some clustering algorithm to find a pattern in this data set; this data set is a heterogeneous data set. Now, I want to break it into subsets which are largely homogeneous; that means their marker level expression are largely similar. So that is what I will do and I will do some clustering algorithm. And that clustering algorithm will tell me I have three type of cell here C1, C2, C3 because I have three clusters.

So, I have a heterogeneous population mixed together in my sample, and this has largely three types of subpopulations, based on marker A and marker B. Now, suppose I am doing the same experiment, but now I will do not clustering, but classification. How we will do the experiment? Say suppose initially what I have done? I have a sample for which I know marker a and marker B has have no expression. So, I can say that is A minus and B minus sample.

So, I will pass that sample through the machine and I will collect the data, so this is the data. Each of these cells here are A negative; that means they does not have A and B negative. Now, I put some quadrant line that means I divide this space in four quadrant. I have shown here just one experiment, control experiment; usually people will have multiple control experiment.

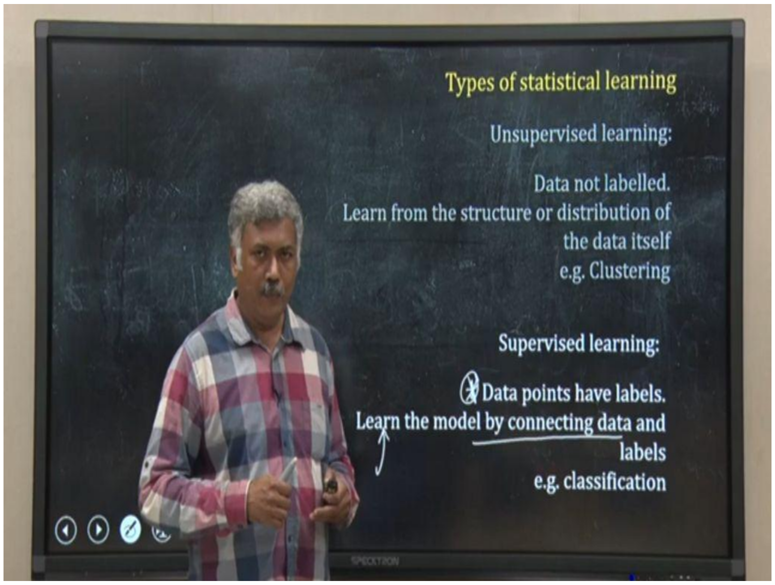
For example, they may have cells for which they know expression of marker A is very high, but the expression of B is very low. So that means those cells will go somewhere here; so this is A positive and B negative. You can have similar control for A negative and B positive; so these are your training data sets. These control cells A positive-B negative, B negative-A positive all these things are your and double negative earlier control samples, and they are a training data set.

Based on which you are dividing this whole space into four quadrants. In flow cytometry field, we will call them gates, we have different gates. Now, what I will do? I have a sample here, not control; treated cells or something, experimental sample, test sample. So, now that comes and I do the flow cytometry; these are the cells coming from my test sample. Some of them are here in the upper left quadrant; that means these cells are A negative B positive.

Whereas, those cells which are here in the in this quadrant are A positive and B positive. So, what I have done here, this is classification. I have some control sample that means training data set and based on that I have created a model. I have classified them in four groups; I tell what are the labels? Labels are A negative B negative, A positive B positive, these are the labels.

So, I have four labels, and I have distributed my control samples into those; and based on that I have created a model. What is the model? The position of this quadrant lines; and now you have a test sample, you run through that. And you see in which quadrant which cell is going and from that you can calculate what percentage of cells are AB positive, double positive, which are double negative, something like that. So, this is classification; hope the idea of clustering and classification is clear to you.

(Refer Slide Time: 25:55)



Types of statistical learning

Unsupervised learning:

Data not labelled.
Learn from the structure or distribution of the data itself
e.g. Clustering

Supervised learning:

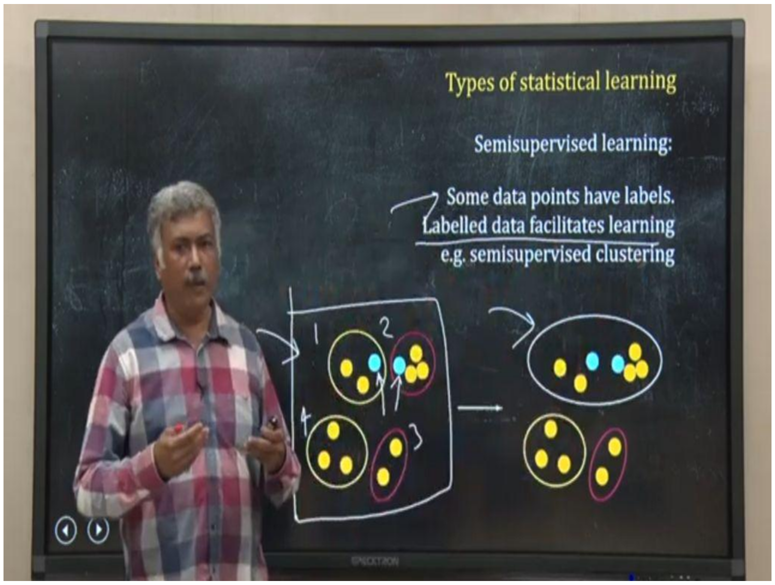
Data points have labels.
Learn the model by connecting data and labels
e.g. classification

As I said, clustering and classification of data are two main components of statistical learning. And this learning process can have many type. For example, one is called unsupervised learning; and that is what you do in case of clustering. What you are doing here? You do not have any additional label to data to say that which group does it belongs to.

No, you do not have, you have only the raw data. And you will look into your algorithm looks into that data to learn from the structure or hidden distribution of the data itself; and then segregate these data points into different groups, different clusters, so, it is unsupervised. Whereas, in case of supervised learning, which is usually the classification is a supervised learning technique.

What you have? You have data points with label; I know in which group this data point belongs to. That means I am giving a additional information, I am supervising the algorithm; and then the algorithm learns the model from that data. And then we use that model to predict or classify a test sample; so this is supervised. There can be in between also, we call it semisupervised.

(Refer Slide Time: 27:15)



Types of statistical learning

Semisupervised learning:

Some data points have labels.
Labelled data facilitates learning
e.g. semisupervised clustering

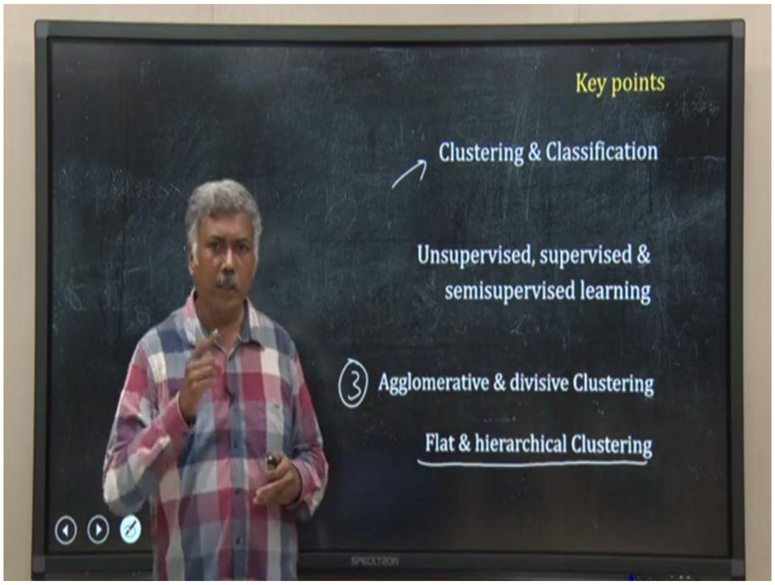
And let us that sometime helps in clustering; actually there are algorithm for semisupervised clustering. For example, let us take an example here. I am doing a clustering of gene expression and I have got this result. My algorithm says this is unsupervised; I have said nothing about label or anything just given the data. And it has found that there are four clusters in this data set 1, 2, 3, 4. Now, I as a biologist have some additional information.

For example, I know these two genes which are right now in different clusters and I marked them by blue color are actually controlled by the same transcription factor, or rather suppose same pathway, signaling pathway. That means I know that they are co-regulated, then my algorithm should put these two genes into the same cluster.

So, now I can give this information this additional biology information to my algorithm, to say that see it you do not break them apart. Then do not keep them in separate, separate cluster; put them into the same cluster and it will do it like this. So, this is semisupervised, you are giving some amount of information. What you are doing? Some data points have labels.

I have given the label that they are co-regulated, not for all, some of them I have given; just two I have given in this example, and then that labelled data facilitate learning. So, this is semisupervised learning.

(Refer Slide Time: 28:49)



Key points

Clustering & Classification

Unsupervised, supervised & semisupervised learning

3 Agglomerative & divisive Clustering

Flat & hierarchical Clustering

Let me jot down what we have learned in this lecture. We have learned about clustering and classification; we have learned what is the difference between clustering and classification. And then we also learned that there are supervised, unsupervised and semisupervised learning methods. Both clustering and classification are statistical learning or machine learning method.

They can be supervised, unsupervised, semisupervised in these three categories you can divide them. And also learned that clustering could be multiple type. For example, some algorithm will be agglomerative clustering, some algorithm will do divisive clustering. Whereas, some of them we call flat and whereas some of them are hierarchical, which give us the relationship between cluster in terms of a tree or dendogram. That is all for this lecture. See you in the next one till then happy learning.