NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

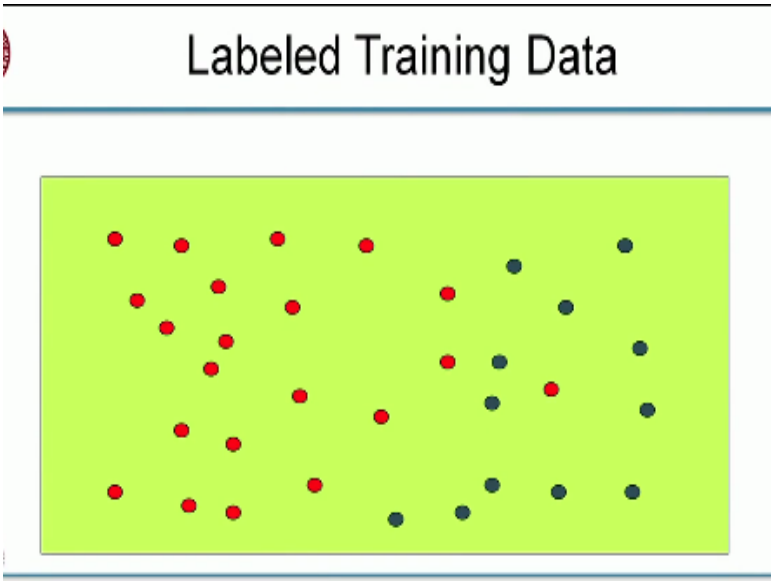
Lecture 3

Prof: Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

Unsupervised Learning

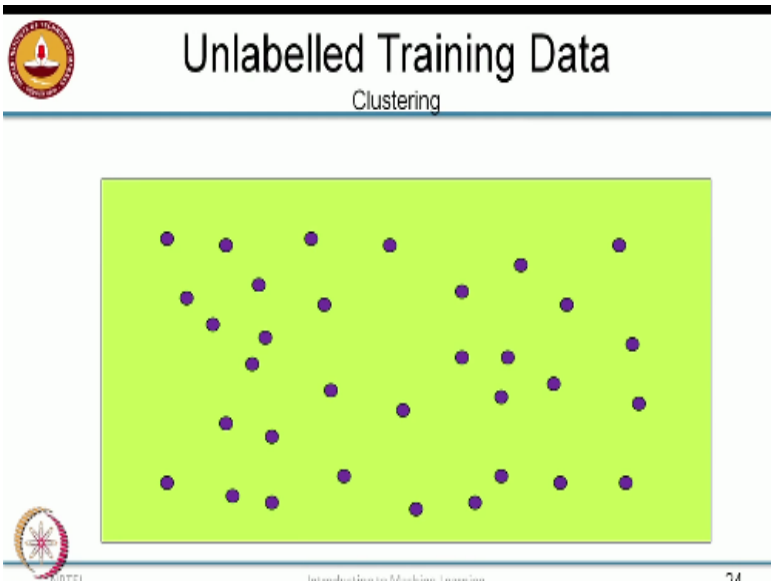
Hello and welcome to this module on introduction to unsupervised learning, right. So in supervised learning we looked at how you will handle training data that had labels on it.

(Refer Slide Time: 00:26)



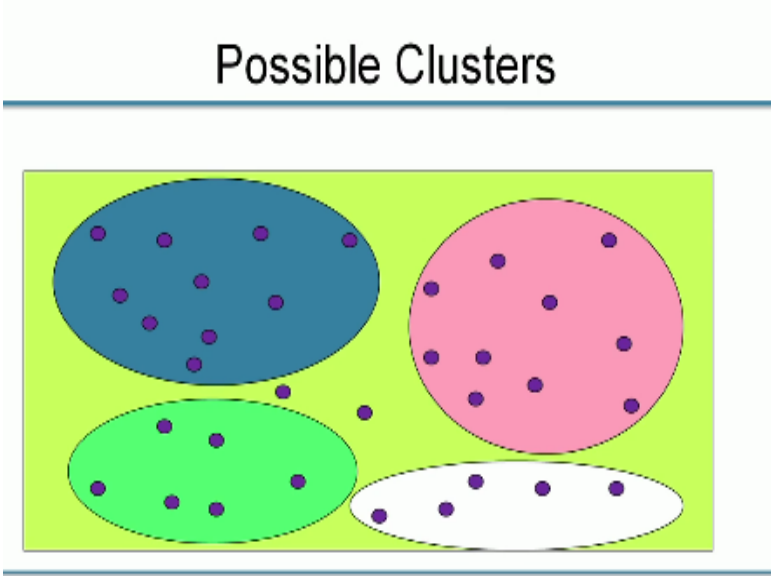
So this is this particular place this is a classification data set where red denotes one class and blue denotes the other class right.

(Refer Slide Time: 00:35)



And in unsupervised learning right so you basically have a lot of data that is given to you but they do not have any labels attached to them right so we look at first at the problem of clustering where your goal is to find groups of coherent or cohesive data points in this input space right so here is an example of possible clusters.

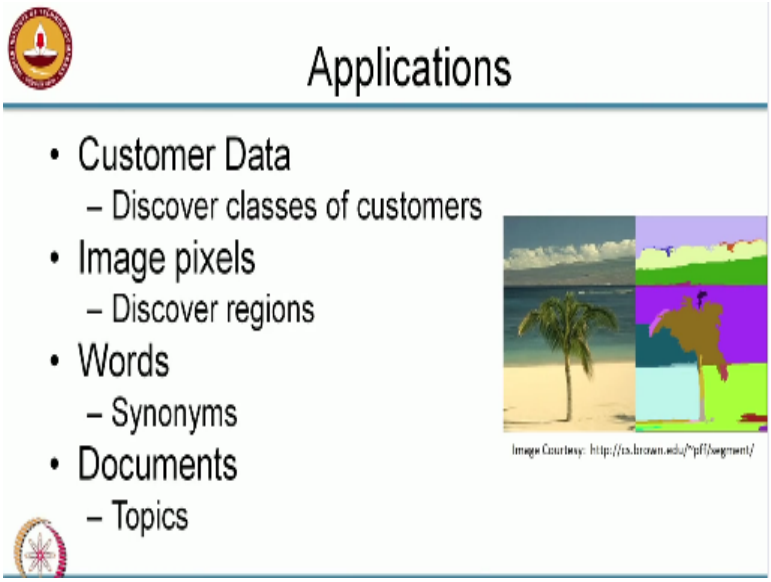
(Refer Slide Time: 00:57)



So those set of data points could form a cluster right and again now those set of data points could form a cluster and again those and those so there are like four clusters that we have identified in this in this setup so one thing to note here is that even in something like clustering so I need to have some form of a bias right so in this case the bias that I am having is in the shape of the cluster so I am assuming that the clusters are all ellipsoids right and therefore you know I have been drawing a specific shape curves for representing the clusters.

And also note that not all data points need to fall into clusters and there are a couple of points there that do not fall into any of the clusters this is primarily a artifact of me assuming that they are ellipsoids but still there are other points in the center is actually faraway from all the other points in the in the data set to be considered as what are known as outliers so when you do clustering so there are two things so one is you are interested in finding cohesive groups of points and the second is you are also interested in finding data points that do not conform to the patterns in the input and these are known as outliers all right.

(Refer Slide Time: 02:23)



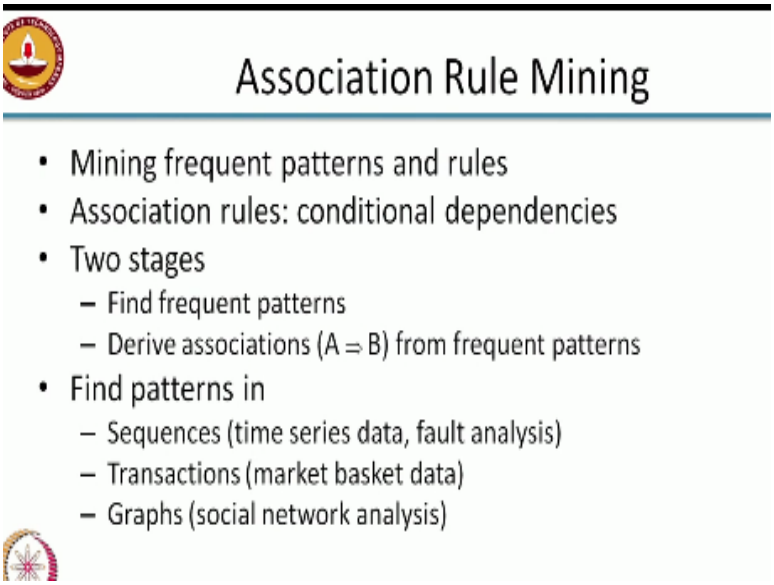
And that is as many mean different ways of an which you can accomplish clustering and we will look at a few in the course and the applications are numerous right so here are a few representative ones so one thing is to look at customer data right and try to discover the classes of customers you know there are so earlier we looked at in the supervised learning case we looked at is that a customer will buy a computer or will not buy a computer as opposed to that we could just take all the customer data that you have and try to just group them into different kinds of customers who come to your shop and then you could do some kind of targeted promotions and different classes of customers right.

And this need not necessarily come with labels you know I am not going to tell you that okay this customer is class 1 that customer is class 2 you are just going to find out which of the customers are more similar with each other all right. And as the second application which you have illustrated here is that I could do clustering on image pixels so that you could discover different regions in the image and then you could do some segmentation based on that different region so for example here it have a picture of a picture of a beach scene and then you are able to figure out the clouds and the sand and the sea and the tree from the image so that allows you to make more sense out of the image right.

Or you could do clustering on world usages right and you could discover synonyms and you could also do clustering on documents right and depending on which kind of documents are similar to each other and if I give you a collection of say 100,000 documents I might be able to

figure out what are the different topics that are discussed in this collection of documents and many ways in which you can use clustering rule mining.

(Refer Slide Time: 04:17)



And as I should give you a site about the usage of the word mining here so many of you might have heard of the term data mining and more often than not the purported data mining tasks are essentially machine learning problems right so it could be classification regression and so on so forth and the first problem that was essentially introduced as a mining problem and not as a learning problem was the one of mining frequent patterns and associations and that is one of the reasons I call this Association rule mining as opposed to Association rule learning just to keep the historic connection intact right, so in Association rule mining we are interested in finding frequent patterns that occur in the input data and then we are looking at conditional dependencies among these patterns right.

And so for example if A and B occur together often right then I could say something like if A happens then B will happen let us suppose that so you have customers that are coming to your shop and whenever customer A visits your shop custom B also tags along with him right, so the next time you find customary A somewhere in the shop so you can know that customer B is already there in the shop along with A.

Or with very high confidence you could say that B is also in the shop at some somewhere else maybe not with A but somewhere else in the shop all right, so these are the kinds of rules that we

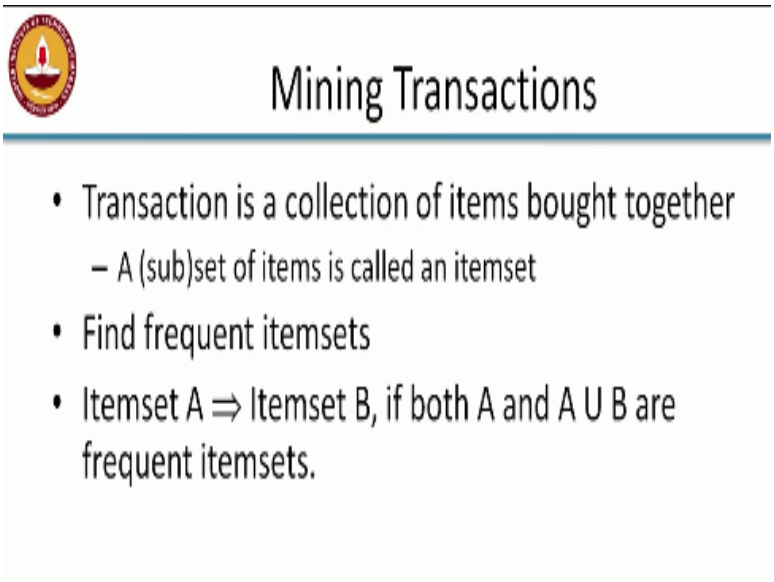
are looking at Association rules which are conditional dependencies if A has come then B is also there right and so the Association rule mining process usually goes in two stages so the first thing is we find all frequent patterns.

So A happens often so A is a customer that comes to measure the store often right and then I find that A and B are paths of customers that come to my store often so if I once I have that right A comes to my store often an A and B comes to my store often then I can derive associations from this kind this frequent patterns right and also you could do this in the variety of different settings you could find sequences in time series data right and where you could look at triggers for certain events.

Or you could look at fault analysis right by looking at a sequence of events that happened and you can figure out which event occurs more often with the fault right or you could look at transactions data which is the most popular example given here is what is called Market Basket data so you go to a shop and you buy a bunch of things together and you put them in your basket so what is there in your basket right so this forms the transaction so you buy say eggs, milk and bread and so all of this go together in your basket.

And then you can find out what are the frequently occurring patterns in this purchase data and then you can make rules out of those or you could look at finding patterns and graphs that is typically used in social network analysis so which kind of interactions among entities happen often right so that is a that is another question that is what we looking at right.

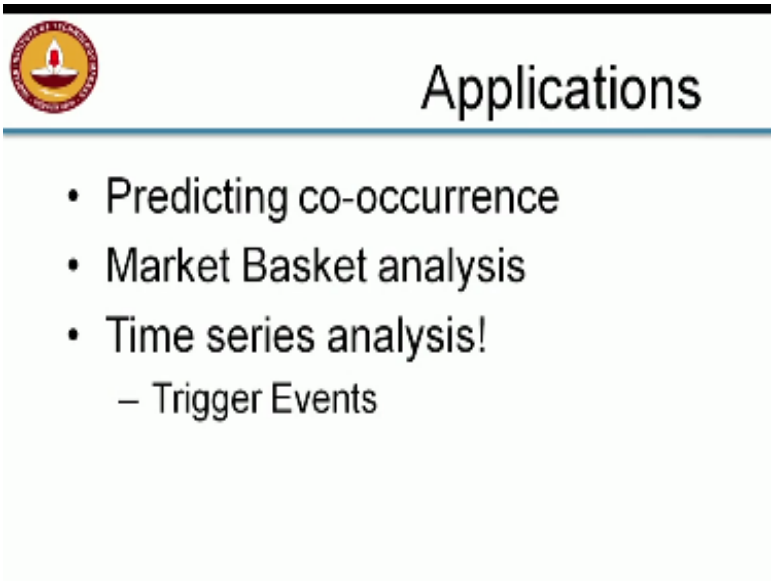
(Refer Slide Time: 07:31)



So the most popular thing here is mining transactions so the most popular application here is mining transactions and as I mentioned earlier transaction is a collection of items that are bought together right and so here is a little bit of terminology and it is a set or a subset of items is often called an item set in the Association rule mining community and so the first step that you have to do is find frequent item sets right.

And you can conclude that item set A if it is frequent implies item set B if both A and AUB or frequent item sets right so A and B are subset so AUB is another subset so if both A and AUB or frequent item sets then you can say that item set A implies item set B right and like I mentioned earlier so there are many applications here so you could think of predicting co-occurrence of events.

(Refer Slide Time: 08:31)



And Market Basket analysis and type series analysis like I mentioned earlier you could think of trigger events or false causes of False and so on so forth right so this brings us to the end of this module introducing unsupervised learning.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in