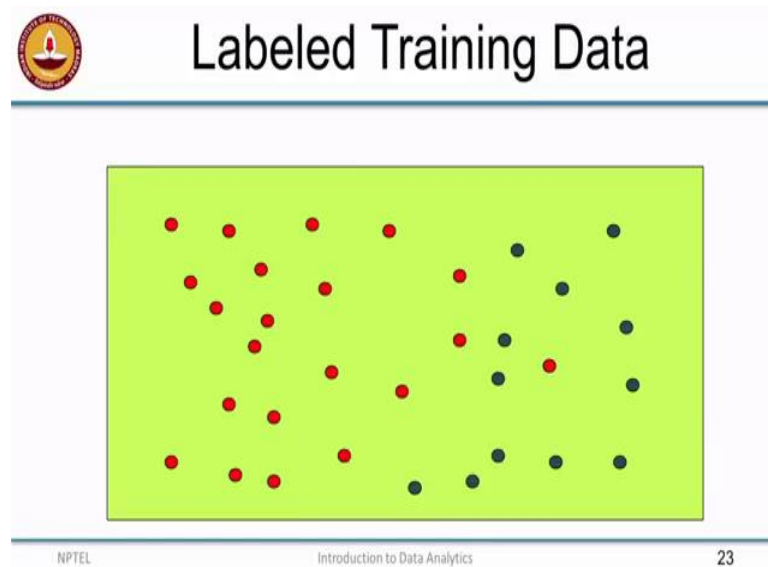


Introduction to Data Analytics
Prof. Nandan Sudarsanam and
Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 04
Lecture – 18
Unsupervised Learning

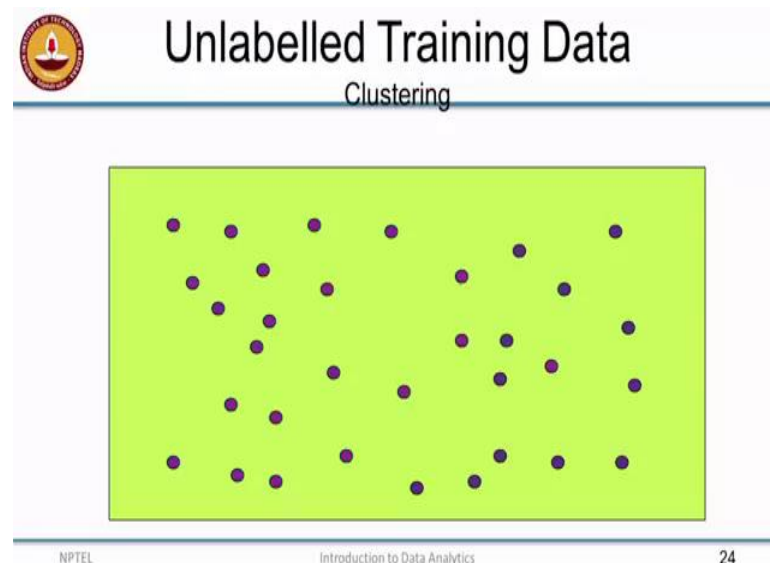
Let us move on to unsupervised learning.

(Refer Slide Time: 00:12)



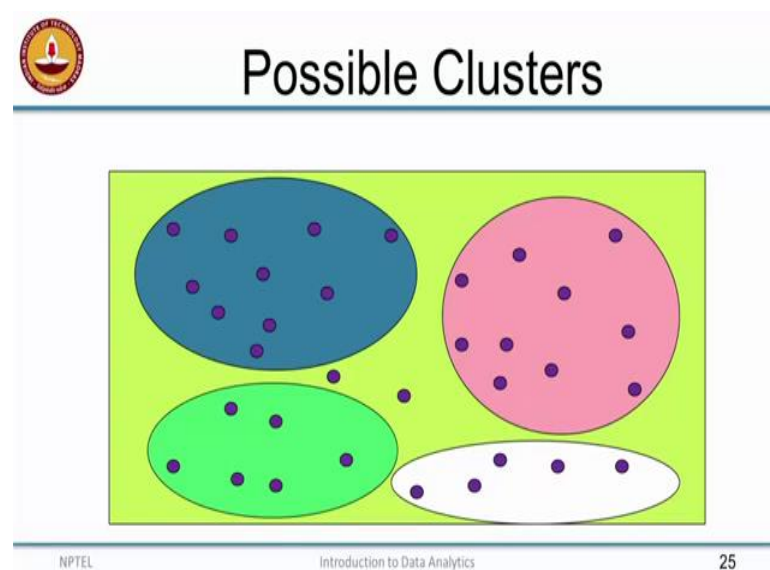
So, in unsupervised learning, so what is our experience going to look like. So, we have looked at label training data during classification.

(Refer Slide Time: 00:23)



In unsupervised learning, there is not going to be any labels, it is going to be completely unlabelled training data and, so the points of this going to look... The input data is going to look like points in n dimensional space. So, in this case it is a two dimensional space. So, now, whatever I interested in doing here, so in the task of clustering what I am interested in doing is to find groupings of similar data points in my input space.

(Refer Slide Time: 00:52)



So, for example, this could be possible clusters, so that could be one cluster. So, you can see that, there is some kind of a gap between the data points in that cluster and the others


and not a huge amount, but some amount and then this could be a another cluster and that could be another cluster and that could be another cluster. So, you could see that, a couple of things that we wanted to notice here. So, one is that, the clusters all seem to be ellipses, there are some kind of ovals in the input space and that is a choice that we have to make.

So, that again gives you one kind of an inductive bias, you know. So, so you have to make this kind of a language choice, even when you are doing clustering. And the second thing I want you to notice, there are the few data points that do not belong to any of these clusters. It could be because, they lie equidistance from both clusters or they lie far away from all of the clusters, but then, they do not belong to any of the clusters.

For example, look at this point, which seems to be far away from all of the clusters that we have identified so far. So, such data points which do not fall into any of the clusters are called outlier. So, what I would like you to note here is that, this particular outlier that I am pointing to actually lies in the middle of the input space. I mean, there are data points left of it, right of it, above it, below it, everywhere.

So, usually there is a conception that an outlier is something that is far away from the other inputs, it need not necessarily be the case. An outlier is something that does not fit into the current patterns that we have discovered in the input. It need not necessarily be something that is very far away from the input, so that something which you have to keep in mind.

(Refer Slide Time: 02:37)



Applications

- Customer Data
 - Discover classes of customers
- Image pixels
 - Discover regions
- Words
 - Synonyms
- Documents
 - Topics




Image Courtesy: <http://cs.brown.edu/~pff/segment/>

NPTEL


Introduction to Data Analytics

26

There are lot of different applications, so you could look at, you know again customer data you could discover classes of customers or in image processing, people try to discover regions in the image like shown in the figure that, which by doing clustering on the image pixels and it could take words, look at the occurrence of words and the context in which the words occur, try to cluster those context together and find synonyms or you could do even better cluster documents together and find topics or you could do the flip of it.

You know, you could cluster data together and try to find outliers. So, outlier mining, which is the very important task in several situations, where we have to find, say anomaly in a data, you know. So, you would like to build a secure system and you want to find a figure out if somebody is trying to track the system. So, any kind of anomalous behavior should be flagged. So, then you would not want data points that lie in clusters, but you want to find data points that lie outside clusters. So, this is called outlier mining. So, there are many different applications are for clustering and when we look at clustering in detail, we will see some more of these.

(Refer Slide Time: 03:54)



Association Rule Mining

- Mining frequent patterns and rules
- Association rules: conditional dependencies
- Two stages
 - Find frequent patterns
 - Derive associations ($A \Rightarrow B$) from frequent patterns
- Find patterns in
 - Sequences (time series data, fault analysis)
 - Transactions (market basket data)
 - Graphs (social network analysis)

NPTEL Introduction to Data Analytics 27

So, the other unsupervised learning task I want to talk about today is association rule mining. So, so the idea behind the association rule mining is that, I want to figure out what kind of entities are frequently, you know co occurring in my input and so, then I can say that there are some association between these. So, this typically goes in two stages, so the first stage I find frequent patterns. So, this is typically the stage, where you analyze the data closely and this is where, this is essentially the “analytics” part of it.


And in the second stage I want to derive associations of the form that, if A occurs, then B is likely to occur. So, A implies B, from these frequent patterns. So, this is essentially a two stage operation. First find the frequent patterns, once I have found the patterns, then try to find these kind of associations. So, more often than not, the challenging part here is finding the patterns, finding the frequent patterns. So, once you find the patterns, then deriving association is not too hard and as we will see, when we look at association rule mining in detail, but the performance measures that you are looking at, typically are associated with the association rules.

I mean, how useful is this particular association that I have discovered is and more than, another frequent pattern part of it. So, you could find patterns in sequences, you could find like time series data or a fault analysis, you could find patterns in graphs you know, where people typically use this in computational biology or social network analysis and

other domains or you could find patterns in transactions, which is essentially the first domain in which people introduce this association rule mining problem, you know.

So, association rule mining in some sense is an interesting problem, because that was the first problem to which the word data mining was properly applied to. You know, in some sense you could say that association rule mining kick started the whole world of data mining.

(Refer Slide Time: 06:26)



Mining Transactions

- Transaction is a collection of items bought together
 - A (sub)set of items is called an itemset
- Find frequent itemsets
- Itemset $A \Rightarrow$ itemset B , if both A and $A \cup B$ are frequent itemsets.

NPTEL

Introduction to Data Analytics

28


So, what is, what I am mean by transactions here. So, so transaction let say is a collection of items that are bought together. I do not know, you just go to super market and you could buy a set of items and then, so all of the items are go together in the basket that you bring to the check out would form a transaction. So, so this, the original form of this problem as post as a market basket analysis, so it is essentially what goes in to your basket when you do your shopping.

So, you look at the items that are there in the basket and then, try to figure out which all items that people buy often together. So, it could be, there need not be individual items it could be sets of items and then this community, these sets of item are called item sets, just as a terminology. So, I will be using item sets frequently when I talk about the association rules. So, the goal here is to find item sets that are frequent, once you have

the set of frequent item sets, then I can go ahead and form rules of the form that, item set A implies item set B, if both A and A union B are frequent. So, A is frequent.

So, I say I buy milk, I also buy bread, so if A is frequent and milk and bread, which is essentially if milk is frequent, so lot of people buy milk and lot of people buy milk and bread together. So, I could say that milk implies bread, so if you buy milk, you are likely to buy bread. So, this is the whole idea behind mining transaction data.

(Refer Slide Time: 08:13)



Applications

- Predicting co-occurrence
- Market Basket analysis
- Time series analysis!
 - Trigger Events


NPTELIntroduction to Data Analytics29

So, lot of applications here again, but just to highlight a few. It is, the first one is market basket analysis. People have actually tried to extend this to a point, where they try to look at the arrangement of items in a store based on what is frequently bought together and people could try to design promotions that take advantage of this kind of market basket analysis. And the second thing is looking at predicting co occurrence and where the things tend to occur together, this is the very generic application.

And in the context of graphs, so mining of frequently occurring sub graphs in the graph, allows us to have a lot of insight into a kind of, you know behavior you would seen biological data and social networks and so on and so forth. And you could also look at this in time series looking at frequent co occurrence of events in a time series can be looked at as identifying trigger events. So, if this event has happened when quite likely

another event is going to happen very soon. So, those kinds of a trigger event modeling, all of these can be done using frequent item sets or association rule mining.

(Refer Slide Time: 09:34)



Why ML?

- Data Mining is ML applied to
 - Large data sets
 - Scaling issues
 - Architecture aware algorithms
 - External memory algorithms
 - Data Selection
 - Feature Selection
 - Database design
 - Noisy data sets
 - Data Cleaning
 - Robustness
 - Missing Values
 - Real data sets
 - iid assumption not valid
 - Class imbalance
 - Stringent performance measures (medical data)
 - Resource constraints
 - Real-time
 - Limited computational power

NPTELIntroduction to Data Analytics30

Why we have been looking at machine learning in such detail. So, data analytics in particular data mining in my opinion is machine learning really, but applied to very, very large data sets, very noisy data sets and very real data sets. So, what do I mean? So, classically machine learning has been more concerned with getting accurate parameter estimation from small volumes of data and trying to do the best that you could do with little data.

But, now with very large volumes of data available, some of the focus is moving away from handling small data to things like, you know how I make sure that my algorithm will finish running in, you know in a few weeks. And, so looking at scaling issues, looking at things like data selection, do any to run my algorithm on the entire data, looking at feature selection is spoke about that a little bit earlier and looking at the actual design of your data base, so how are you going to represent the data and so on and so forth.

So, all of these issues now have come to the forefront. So, when you are working in these kinds of domains, you call it data mining and you then you go back and look at specific

algorithms and error measures and so on and so forth. Then, you call it machine learning, but then the lines are the kind of a, you know getting a very fuzzy between the two.

And then, the second difference error's mentioning is on noisy data sets. So, again I mention a little bit of this earlier, so you have to worry about cleaning the data, you have to worry about missing values and you have to make sure that your algorithm is robust, when something goes wrong that you did not expect earlier, you know. Something goes missing or some data gets corrupted, you should have some kind of a failsafe mechanism.

And the last thing is a little more technical which is that, real data typically does not satisfy some of the nice theoretical assumptions, many of the machine learning algorithms were operating under for several decades. So, typically the assumption is that the data is independent. So, one... So, we looked at the training data earlier, so their assumption was that each of those points in that age incomes space was sample independent of the other and then, there were sample from an identical distribution.

So, it is not like each data point came from a different population, they are all sample from the same population, the same kind of age income distribution and then, they were sample independent of one another. It is not going to be true, more often than not, because your friend goes to a shop and he likes a computer and buys something there, you are likely to go there as well, but it is not like the fact that the first guy bought a computer is not influencing, whether you are going to visit the shop or not.

So, these kind of independent assumptions are typically not valid in real data. So, you will have to think about that and then, so we looked at the classification problem, where we saw that we had, you know positive and negative classes like buys a computer, does not buy a computer and more or less equal, I mean that is about 60 40 split between buying and not buying computers. But, in reality it is not so.

For example, you take any medical domain, then fraction of people who are sick is very, very small thankfully, but still it is very, very small and therefore, it makes a machine learning problem very hard. So, if I tell that everyone who comes to the hospital, I mean or everyone I see in the population at large is healthy, I will probably be correct with, say

some 97 percent accuracy, which case it is not a bad predictor you know, but that is useless for us.

So, we have to worry about handling this kinds of class imbalance to what actually make sense, when the data is so imbalanced and the another thing is, in many real worlds scenarios like in medical domains and also in security domains, like you cannot be happy with the acceptable levels of performance, you know. They have to be near perfect for you to be able to deploy these things in practice. So, this causes a lot of trouble, you says no more the case being able to do the best effort and get away with it, you really have to be the best.

So, so such issues also are slightly different from what people worried about in a machine learning community. The last thing is, you have to operate under the resource constraints. You know, maybe want them things wrong on your hand, your handheld device not that there is a much of resource constraint any more, but it could still be and I have to work with limited computational power and you might not have the luxury of all the time in the world to produce an answer, you might have to produce answers in real time. So, how do you go about doing that?

So, there are lots of issues, so but at the core of data mining is machine learning. So, machine learning gives you the kind of algorithms and the data mining addresses all these issues on top of those algorithms. So, that is one of the reasons while we look at machine learning and we will continue to do so in future modules in greater details.

Thank you.