Transcriber Name: Saji Paul
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology, Kharagpur

Lecture - 49
Introduction

Hello guys, so we have come almost to the end of this lecture like this course basically, so this is the last module which we are going to discuss, this module is called classification. So, classification is something which all of us know it is something which we use in our day-to-day life in our day-to-day activities we get to see that we classify things we classify objects, even as a child also when initially when the style when the child just started and starts learning things.

So, basically at that time the child does not know what is if I talk of animals the child does not know whether it is a cat or a dog so maybe the mother teaches the child see this is a cat, this is a dog and so basically once he or she gets the when the child gets this knowledge that this particular animal we call it cat, this particular animal we call it dog and so next gradually when the child grows up the child can automatically whenever a child sees an animal which the child has already known the level of the animal.

The child can identify, this is a cat, this is a dog this is a cow whatever it is. So, desires is this classification it starts from the very beginning of her life. So, this classification we get to see in many aspects and coming to the technical aspect of it like when why we are going to learn classification, basically and in many of the disciplines I mean engineering disciplines many social science research many things in many domains medical domains we really need to classify some observations.

Classify objects; classify data we need to classify data based on some classified data into some basically relevant categories. So, now the thing is that so why do we classify data to some relevant character categories we can classify object to some relevant categories. So, we classify some observations different kind of some emails this may be various reasons maybe some for some medical diagnosis maybe some medical research from social science research and then maybe some business perspective for personal use.

And maybe for security reasons for different types of data we may have different security requirements like say mail we may need to classify a spam mail and a non-spam mail, so, this all this is what is called classification. So, we are going to learn definitely all the algorithms is not possible to learn in this course. So, we will be learning a couple of algorithms how we classify different object different observations different data.
(refer time: 03:17)

So, in this lecture first will be covering the introduction to classifications. Then there are two

broadly there are two different categories of classifications that is we call it supervised classification and unsupervised classification we will be discussing that in this lecture. And then finally I will give a formal statement of supervised classification technique. In this lecture I will restrict my description discussion to only supervised classification technique unsupervised classification techniques I will not be discussed.
(refer time: 03:51)

So, now coming to the; introduction of classifications. First now see there is a picture I am asking you to identify the objects. Now how you will identify these objects? Of course, as now you already know what are these objects so that this is an apple, this is an egg, this is an onion this is a potato all of us we know we can directly identify these objects. But the thing is that how did we identify this object.

This all these objects it has some characteristics it says some features or parameters or attributes whatever you say. It has some distinguished feature each type of classes when I call it a onion and when I call it a potato each has some distinguished characteristics that which we call it the features. Based on these features we try to classify the objects, this is an onion, this is an potato, this is an egg based on these features we do that.

So, now if we need to identify these objects since this and it is already we already know what is or when we call a thing as onion so when we see the peaches matches then this is an onion, that is how we identify the object.
(refer time: 05:01)

Now again there are some other way also where sometimes when we identify some data but some classify some data they may be it may be in and not I should not say this characteristics basically it is some sort of rule using some sort of rule sometimes we classify data. And rules means it is again specified some sort of features as well like here. If you see we are classifying the marks of the students we all we as two teachers we always do we give the grades to after exams we give the grades to the students how based on the marks.

When the marks are greater than equal to 90 we give a greater excellent great whatever it is. From 90 to 80 we give big grade so this is nothing this is again one sort of classification only here how we do the classification. It is directly by the rule base it is very much deterministic does mark above this is a great between this and this is B grade this and this is C grade it is a very deterministic rule.

So, are there based on the deterministic rule we specify these are the different classes of marks different classes of students grades. So, this is one way by which we can classify data. Once we have seen we can classify data based on some features different its distinguishing features, distinguishing attributes and another is that we can also classify this way by based on the some rule base.

If this how we are framing the rule basically again basing based on some features, features may

be this we can if we can consider this marks greater than 90 this is a feature. So, basic but there is nothing is just I have made a specific deterministic rule for that.
(refer time: 06:42)

So, now again some example if you see like from life science, I already mentioned in classification is very much to use in case of medical research medical diagnosis like predicting tumour cells as benign or malignant how do doctors how do Radiology does that. They realize we do MRI based on the MRI data this is what they basically do some sort of image processing. So, how do they do? So, there are different features.

So, they I did not they try to identify different features and they have the knowledge that these features meaning means these features implied this. So, based on the different features when an image is given so based on the different features they try to identify this is an benign tumour, this is a malignant tumour and then next coming to some example of security credit card transaction at the legitimate or fraudulent credit cards.

Sometimes what happened usually an average person and every person uses credit card every maybe once in a week or every two second day or every third day and even average person average middle class person takes out money. Say around every week if a person uses their uses maybe the card ATM card and then withdraws money say around 5000 or around 10,000 per week or every alternate day.

So, suddenly and this for the same card if the if it is founder it is withdrawing and lots of money in a very short interval so every half an hour it is at except hours suppose the person has withdrawn 1 lakh again after half an hour it is withdrawing another 2 lakh again after half an hour or one hour to withdrawing 3 lakhs very frequently it is we are drawing a huge chunk of money.

So, and this machine can understand that this is this may be in fraudulent transaction maybe some the credit card has got stolen and that is why the people the person who has who got hold of this character is using it this way. So, this is how this transaction can be classified, this is a legitimate transaction this is this may be a fraudulent transaction. Similarly, in case of it has many applications classification is something it has a varied application.

Like now again we can think of in the weather prediction like satellite what has the satellite collects data it collects the data of a humidity, wind speed, pressure, random and many other factors like it collects different data and then based on this data it can classify the next day it will it will fall in which category. It will be a cloudy day sunny day whatever it is. So, these are the different features they try to take the value of the different features based on different features it identifies to identify it.

Similarly, entertainment category new stories as finance, weather, entertainment, sports etcetera. You have might be you have noticed when you if you a user of YouTube If you use YouTube usually in YouTube a personal profile gets stored profile in the profile meaning the person's browsing history gets stored. And so, when a person if I load the YouTube and if some

other person load the YouTube they have informed the videos which I will get it will be different from what the other person will get it is.

Because on the things the staff I browse it basically tries to keep a history of that. And based on that it categories into my choices it has been categorized into a particular classes maybe I see lots of news stories. So, when I try to load YouTube, I will get to see a different types of news in my in the screen. Similarly, some other person maybe whenever he opens the YouTube, he usually watches movie so maybe we watch this new movie.

So, whenever he loads the YouTube whenever he opens the YouTube, he will get to see all the different new movies that has come that has been loaded. So, that is how it categorizes the persons based on the features of the person browsing history it categorizes the preferences of the person, so this here also we use classifications.
(refer time: 11:19)

So, now let me formulate we have seen lots of examples of classifications. So, let is formally define what is classification. So, classification is a form of data analysis to extract model, it is a form of data analysis to extract model describing important data classes. So, it is a form of data analysis to extract models will have will have to analyse the data and we will have to extract a model we will have to find a model which basically it is a black box a model.

What the model describe it will describe important data classes. So, given an input it will give the which data it will it belongs to which data class C 1 C 2 or C 3 which data class it belongs to which data class. So, that is the formal definition of classification. Essentially it involves dividing up objects whatever objects or observations or data whatever it is, so, that each is assigned to one of a number of mutually exclusive and exhaustive categories known as classes.

So, given an object it will be assigned to one of the mutually exhaustive and exclusive categories known as classes. Means when given an observation given an object it will be assigned to only one class it will not be assign to one object will not be assigned to more than one class and each object will be assigned to one of the classes at least it will assign to at one of the classes and one object which is assigned to one class it will never assign to some other classes.

So, it is mutually exhaustive and exclusive. The term usually exhaustive and exclusive simply means that each must be assigned to precisely one class that is exclusive each object; must be assigned to precisely one class. And that is never to more than one and never to no class at all never to no class at all, this exhaustive.
(refer time: 13:32)

So, now as I told before the classification consists of assigning a class level to the set of unclassified data and this classification broadly it can be classified into two different classifications. We can have broadly there are two different techniques for classification technique if you see broadly. So, one is called supervised classification and the other is called

unsupervised classification.

Now what is supervised classification? Supervised classification the data with possible class of each is known in advance. So, like as the example what I have given initially when I started with classification a child is taught this is this element and this is this animal this is cat this is dog. So, that is in supervised classifications. So, the child tries to put it in the child's in the brain however these are the features when these features are there and it is called a cat.

When these features are there and it is called a dog. So, this is the data features are known as well as its class level is known for a particular set of features this is the class level this then we call it a class supervised classification. Like your apple suppose it is round shaped it is red coloured and I mean what whatever may be the features whatever and so then this is called an apple. So, it is an again round shape it is yellow colour then maybe it is an orange.

So, it is based on this attribute is round shape red colour round shape yellow colour there are what these are attributes these are features and what is the label, label is apple orange these are labels. So, data with possible classes of each is known data attributes based on the attributes what is the class it is known. And then in such a case and then in next when we given data with an unknown level then it will be able to identify this belongs to this class.

This is called supervised classifications where the classes level are known in advance how many classes are there that are known in advance. The next is unsupervised classification. Unsupervised classification is something like class labelling of data is not known given a data set we first of all we do not know how many classes will be there. What will be the different class level? Nothing is known to us, the data are given.

Somehow when the data’s are given we try to group the data in such a way, the big data set is given we try to group the data in such a way the data belonging to a group are very much similar to each other compared to the of means all the object belonging to this group are quite similar to in each other in its features whatever may be in some of the features then compared to the object belonging to the other group.

So, like I can give you an example one unsupervised classification a very good example classic example actually it is used from long time back in data mining we basically use this customer segmentations like in what to say in a big retailed retail shop. So, daily data daily data is there or one monthly data we see monthly data used data many customers has come many customers has purchased many things.

So, this the used customer purchase data is with us so we do not know from this what sort of classes we can bring out and how many classes we can bring out nothing is that the data what is this data this customer purchase history. From this if we try to group data and group data and there can be different grouping now if as I told you it tries to form a group. So, that the characteristics of the object within a group are very much same because when we compared characteristics of the object belonging to some other group.

So, like here maybe one group may be people who are buying on electronic store, then there are some people who are only buying groceries there is some people who are only buying cloth material. So, this maybe we can somehow group this way this may be one can group or maybe some other grouping maybe people are buying when people are buying one x they usually they are buying y also so this can be one group.

So, again some other group members some customers they were quite active before now they have completely stopped. So, but on the contrary, there is some customer who is quite initially they were not there was no activity now they are very much active. So, these were the different groups there is different groups basically it is called clustering. So, unsupervised cluster and classification basically we call it clustering we form different clusters.

So, this is we will not be discussing an unsupervised classification in this like course basically we will restrict our discussion only to supervised classification. So, unsupervised classification the basic idea I have given I think you could understand that.
(refer time: 18:44)

So, now it is a good example of supervised classification here different emojis are there. So, it is different emojis according to its label of the emojis are also given. So, when we see the first emojis say this is this emojis indicates a good friend. So, when we get some this sort of icon similar kind so once we know this means a good friend so if I get that means this belongs to this class. So, this is supervised learning.
(refer time: 19:10)

Again, unsupervised classification you see here the different cats. So, now how will you all are cats only, so like as I told you in a retail shop there are all our data only purchase data all our purchase data how to classify this data how to think what can be the different classes of this data. Similarly, here all are cats only now how to classify this data, how to make different clusters of this data. Maybe we can have some maybe some cats which have say blue eyes.

So, this is a cat which a blue eyes this and cat we say blue eye so is there someone else no so maybe cats which are blue eyes we can have one cluster. So, then cats with a brown eye we can have another cluster so that way if we make the cluster based on the eyes only based on some other features maybe. So, that way but cluster label is not known how many classes will be there it is not known.

Then we just try to find out we just try to make the group with the characteristic that object within a group are very much similar and it is compared to the object belonging to the other group.
(refer time: 20:26)

So, supervised when you talk of supervised classification techniques. So, now any technique first we need to know what is the input of this check first any technique first what should be the goal of this thing why at all we are thinking of this technique. So, why at all we are thinking of

this technique that is the first thing we should know then only we will start preparing things, then that is goal that is known now then what will be the input to this technique.

And what is the input to this technique and then given this input and this is the goal this is the input this is the goal now what it will do to the input so that we get this goal. So, now for supervised classification technique, what is the input? Input is a collection of records we call we also call it a training set; it is nothing but a collection of records, collection of different observation of data. If I talk it in case of statistics it is this collection of different observation.

So, each record contains a set of attributes and among that one of the attribute is a class. So, it is a collection of observations and this observation each observation has some attributes and one of the attribute is the class label. And what is the task? Find a model for class attribute as a function of the values of the other attribute that means we have to find Y = f x I have to find these functions this x are the attributes for function of this attributes this x are the attribute.

For this attribute if these are the attributes then what is the class, basically I have to find out Y = f x that is my task for given attribute x what is the class. First initially for some so that I can understand before doing going to classified as things I should understand the stuff. These things belongs these things means it is belongs to this class. So, that for that is we have already modelled we have already trained it ourselves.

So, once we have trained ourselves like the child training so then we this is our task. What is the goal previously unseen records should be assigned as a class as accurately as possible. So, task is given a set of observations, this observation has also the class label. So, from this observation we have to find this function Y = f x for this class level what is this function somehow, we will have to find this function.

Once this function is known then what is the goal? Given any unseen record with a class level is not given just the attributes are given, we have to assigned a class as accurately as possible and how we will assign a class satisfying the property of mutually exclusive and exhaustive. A class has to be assigned to one observation has to be assigned to one class not more than an observation cannot be assigned to more than one class and each observation has to be assigned to one class.
(refer time: 23:35)

So, this is an example like see there are different we have total 15 observations. So, these are the different attributes we have total four attributes. So, for 15 observation we have 4 attributes, the first three I am calling it attribute 1 attribute 2 attribute 3 and a third one and the fourth one I am calling is a class. So, for these values this is the class is no let us consider that two classes is no and yes, maybe.

First, I have say 10 observations and the 10 observation all the attributes values and the classes are given using that I will try to understand if this is the features then this is the class. Once that is known then given it tests it given any attribute value, I should be applied to tell I

should be able to tell, this belongs to which class. So, this is induction learning model and this from this model I will apply the model and I will get a deduction for these classes.
(refer time: 24:40)

So, as now to define it formally the same thing given a database D it has some tuples what is this t 1 to t n different tuples and each tuple is nothing but a set of attribute each tuple consists of a set of attribute t 1 is a set of attribute A 1 to A n t 2 is a set of attribute another set of attributes A 1 to A n different attributes. So, given a database with a set of tuples and each tuple is defined by a set of attributes.

And a set of classes given a set of tuples and a set of classes C the classification problem is to define a mapping from D to C. But this is a classification problem classification problem is nothing just a mapping from the set of tuples to a set of classes where each t i is assigned to one class it is not that a t i will have no class assignment each t i has to be assigned to one class. So, note that the tuple t i belongs to is D is defined by a set of attributes.
(refer time: 25:54)

So, now the number of classification techniques I am talking supervise only. So, some supervised classification techniques are known which can be broadly classified into following categories. So, what are the different methods we know? One is statistical based method distance-based method decision tree-based method error-based matter. So, statistical based method basically we use statistical based method.

When we basically try to identify the level of a class when based on some known relevancy based on when we know this some classes with a known level some objects. We know some you know some object with no level and some other objects we try to put classified that object under some classes based on the relevancy of the known class level. So, here in this in statistical matter basically it is not very deterministic like.

If these attributes are there then definitely it will go to this class it is not that it is deterministically, we cannot say. Like for an apple we can deterministically say that if it is if the size is of this to this much means and this colour is this shape is this and top is this bottom is this then definitely it is an apple, like deterministically we can say. But in statistical base methods such deterministic rule does not apply.

Actually, did we cannot say if this is that then this it will belong to this class. So, in statistical based method we try to just find out a relevancy of what we have to predict and what already we know. From what we know we try to predict based on the relevancy of that so the that is the reason in statistical base method usually we use probabilistic terms. So, we will see and one of the most used statistical based method is called Bayesian classifier which we will be discussing in our in my next lecture.

Then next is the distance based method. So, and distance based method here also here how we try to classify the object we try to classify the objects based on the similarity index. Like if I

tell a cat and a dog, cat and a dog similarity index both has four legs one tail then two ears so the similarity index is quite similar does not it. So, as it and similarity index means both will belong to the same group.

But at the same time there are some features which will significantly differentiate between the dog and the cat. So, there it will the similar similarity index will be quite different for some other features which will differentiate between a cat and a dog. So, this similarity index we usually define in terms of the distance of one attribute distance between two attributes one attribute from one objects and one attribute from the another objects when we if we try to find out the distance.

The similarity index we try to find out the similarity index we usually consider the distance between distances of these two attributes. So, based on this distance based method one of the approaches K nearest neighbour which will be discussing in I think the next another after two three lectures. Then next is decision tree based method. Decision based method it is or it is very much like a rule base what I have told initially and the beginning of this lecture.

So, if this done this sort of that a rule base actually. So, here what we do is that from the attribute set we try to find out one attribute which we continuous when we go on continuously splitting those attributes. So, that it forms a sort of tree. Initially if there is a we will the for the best classification first is the good selection of a attribute which will act as a root node. So, first we will find out the attribute one attribute and then we will try to based on that attribute.

We will try to divide by using some rule we will try to branch it to different other nodes then again from that node we will further go to different other nodes. So, and gradually we will reach a leaf so that is basically the classes. So, distance that is the distance tree base method however in distance tree based method will not be discussing in this lecture. So, then the fourth is the error-based method, error based method I will be discussing support vector machine I will be it is a very important classifier I will be discussing the support vector machine in a couple of lectures.
(refer time: 30:47)

And so, coming to the conclusion in this lecture we learn about a basic introduction of the classification tasks the idea of supervised and unsupervised classifications also have put and have also put this taxonomy about some popular classification technique. In the next few lectures, we look into the working of some such classification matter.
(refer time: 31:07)

These are the references and thank you guys.