

**NPTEL**

**NPTEL ONLINE CERTIFICATION COURSE**

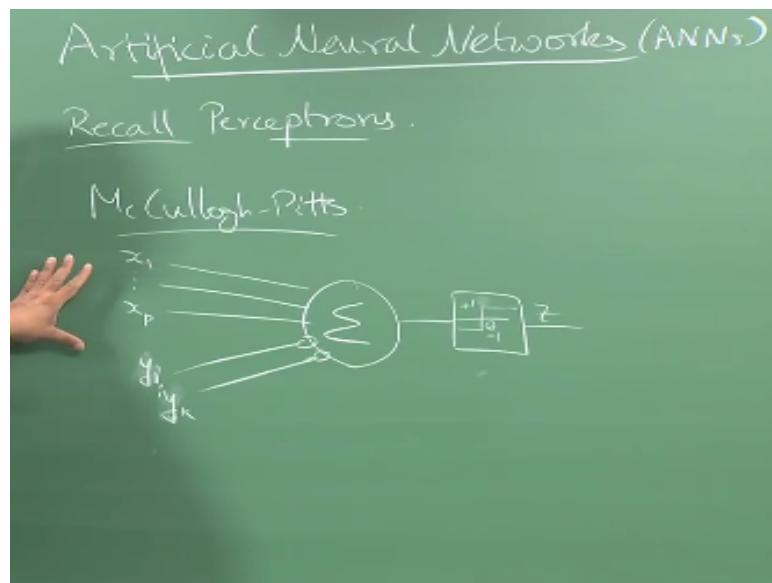
**Introduction to Machine Learning**

**Lecture 32**

**Prof. Balaraman Ravindran  
Computer Science and Engineering  
Indian institute of technology**

**Artificial Neural Networks I –  
Early Models**

(Refer Slide Time: 00:14)



Okay so we have been having discussions about neural networks off and on right, the last time we had a discussion about ANNs was went and we did perceptions right. So, basically so the whole class of solution methods which are lumped under ANNs are artificial neural networks were primarily inspired by trying to emulate the brain architecture right yeah, so then after a while the field split into two right one class of researchers who are looking at neural networks as just computing elements right trying to interpret it in terms of linear algebra right and partial analysis and other mathematical tools and then trying to understand what computing these artificial elements were doing.

And others who are still trying to make the neural models biologically relevant right, so now the communities have become fairly divergent right, so there are this set of people neuroscience people who are trying to build computational models of the brain right and then there are machine learning people who just use neural architectures without worrying anything about biological relevance right, so yeah we will take the latter approach right there will not be looking too much about the biological relevance of the neural networks I will just try to understand it in terms of computing.

So do not two people have ever been seen a neuron synapse did not write all those pictures so you do not have to really tax my drawing skills right so I will not draw anything on the board about the neurons right, so the whole idea behind all these biological neurons is that it is pulling in inputs from a variety of other neurons right and some computation is goes on within the neuron and there is a result of this it might actually fire right in which case it will cause it another input to be activated for yet another neuron right or a set of other neurons depending on who they are connected to.

Then these neurons are all connected in a very complex networks and even though it is a very simple computing element each neuron can be thought of as a very simple computing element the whole the fact that they are connected together in a large network allows them to do all kinds of cool things right, I mean if you really want to know what a neural network can accomplish they stop and think what you can do right, so one of the earliest, all of the earliest models of a neuron was the Mcculloch Pitts model right.

So it is a very simple model so it essentially it has a  $\Sigma$  unit right so it has a set of inputs right and then it has not set off inhibitory signals okay right, so it has a set of inputs right and a set of inhibitory signals if we what we are talking about artificial neural networks and I started explaining the Mcculloch Pitts model so you have a simple unit which has  $P$  inputs right and then it also has  $k$  inhibitory inputs right, so what does this neuron do is essentially adds up these  $k$  inputs right.

And if the if they exceed a certain threshold right some threshold  $\theta$  then it will output a 1 if it is below a threshold  $\theta$  it will output - 1 or a 0 depending on how you are encoding it right and the inputs are all considered either to be 1 or - 1 or 0 is depending on how we encode it right what

are these inhibitory inputs for if any one of these inhibitory inputs was one there was no output no I 0 okay yeah if any one of these inhibitory outputs is inputs is one then there will be the output will be -1 okay or 0 depending on how we were encoding it again right.

So that is the basic McCulloch Pitts model but then just to give you the historic perception right and then what happened people modified this way to propose the perceptron right.

(Refer Slide Time: 06:21)

$$P$$

$$E(\beta) = \frac{1}{2} \sum_{i=1}^N (y_i - \hat{f}(x_i))^2$$

$$\nabla E(\beta) = \sum_{i=1}^N (y_i - \hat{f}(x_i))(-x_i)$$

$$\beta \leftarrow \beta + \eta \sum_{i=1}^N (y_i - \hat{f}(x_i)) x_i$$


So if you think about what the perceptron does it is essentially saying that right so instead of doing  $x_1$  to  $x_p$  and adding it up I am going to multiply it by  $\beta_1$  to  $\beta_p$  so  $X_1 \beta_1 + X_2 \beta_2 + X_3 \beta_3$  and so on so forth now if this entire thing is greater than a threshold okay I will output a +1 right the entire thing is lesser than a threshold I will output -1 so what is the threshold meet or not right so another way of doing it is to actually add a add another input label that  $\beta_0$  I mean way is that as  $\beta_0$  and put a 1 and then and right okay.

So you could think of it that way right so usually it is written without the - sign right that essentially means  $\beta_0$  will be a negative quantity okay so this is essentially what our perceptron this you can see it is very closely related to the original McCulloch Pitts model except that there are no inhibitory inputs and they the actual inputs are weighted okay and we all know how to estimate the parameters of the perceptron right how did we get there we use the gradient descent rule right.

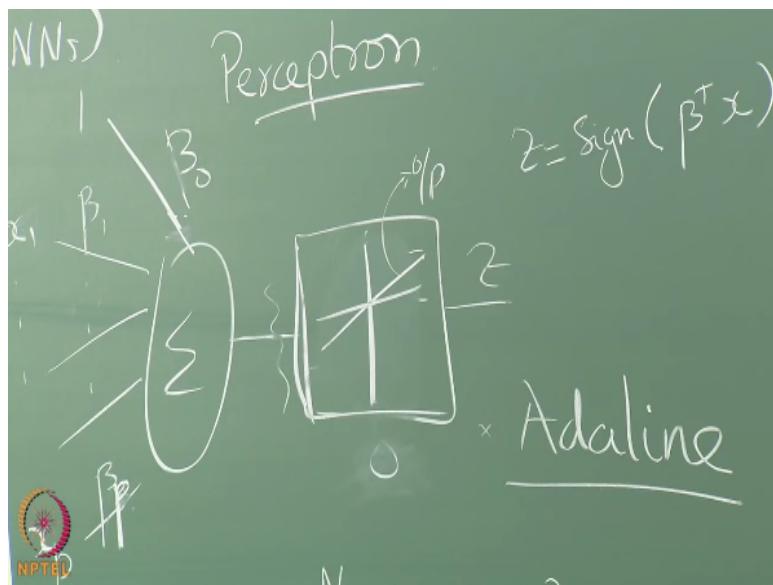
We started with that error function and then we took the derivative and quite look like the perceptron ruled as it does not quite look like the perceptron what it do the perceptron update so wherever there was a missed classification what did you do with this we just added those data points to added those data points times  $Y_i$  to so we did essentially read  $X_i Y_i$  to the input right we did not do anything like this right no right, so what we essentially done here is minimize the squared error all right.

So this is another way of training perceptron is it something wrong here what is wrong here yeah so with the with the perceptron training so we did something very different from whatever I have done so far so it is perceptron training algorithm we had a very different objective function that we were optimizing right, so what were we trying to do people remember that you have a quiz like two days away should have revised all of this by now yeah minimize the mean you might see what minimize the distance to the hyper plane of the misclassified points right.

And the way you can minimize it is we get a distance of zero total distance of zero is when all the points are correctly classified right, so what we had actual objective function we wrote down was we are trying to minimize the distance to the hyper plane of the misclassified points we remember we wrote it only over the misclassified points right so in this case what we are doing you are writing a plane with plain old squared error function right but this is actually the squared error there right.

If you think of it otherwise this is a non-linearity right this is a non-linearity the threshold is a non-linearity right I can just take the derivative of  $z$  with respect to  $x$  it does not make sense I cannot take the derivative of  $z$  with respect to  $x$  because that is a nonlinear function of  $x$  right.

(Refer Slide Time: 12:09)

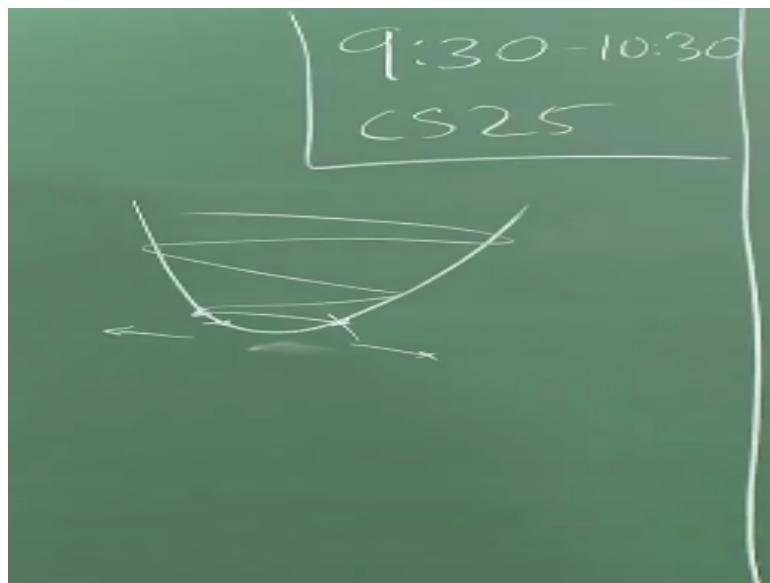


So  $z$  is what right and sign is a nonlinear functions non differentiable so I cannot really take the derivative. So in fact what I have written down here is for a slightly modified model of the neuron where the output is taken at this point okay the non-linearity is not there, so one way of thinking about this is that is the output like it is like a straight line right, so whatever comes as the input it will be produced as the output right, so fool yourself for a minute that this line has a slope of 1 okay.

So whatever comes as the input will go as the output right in that case I can take the derivative of the output with respect to the input does it make sense right, so this is not the perceptron by the way they are sometimes called the but what is either line stand for any guesses adaptive closed linear adaptive linear so it is called adaptive linear units so they are Adaline right so on the way you train Adaline is just you straight forward gradient descent right and you end up with this okay.

So people are familiar with gradient descent right I do not have to explain gradient descent to people everyone familiar with gradient descent right okay, so why am I using  $\eta$  here step size yeah why do I need a step size okay so why would what would produce oscillations okay great yeah.

(Refer Slide Time: 14:42)



So let us look at it in a 1 dk may suppose I have a function that I want to optimize right let us assume that I cannot measure the gradient properly and I am actually making estimates of the gradient right so I am somewhere here right and I find the gradient what direction is the gradient here that way hey that is the way I have to change  $x$  to go up right and what they have to do I have to move in the opposite direction.

So if I take a large step in the opposite direction what will happen is I will actually end up this same look I will end up somewhere here now again I will I the  $x$  the gradient is in that direction here again I take a large step I would end up here like I could keep going back and forth then I might not actually converge okay that is it make sense in fact it is even worse I could go back and forth and I might even diverge if I am taking very large steps right so that is why I have to take small steps when you are following the gradient right.

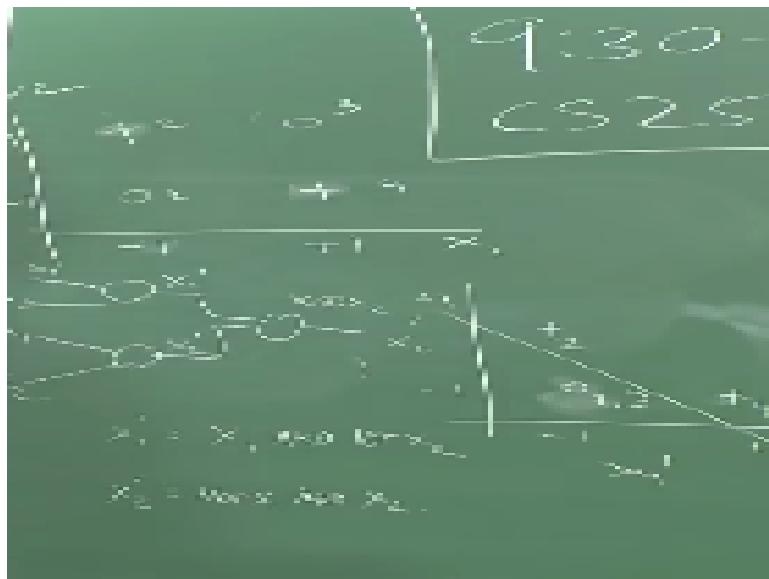
And this these kinds of methods are called stochastic gradient why no it is essentially the so if you think if you just think about it a little bit right where do these excise come from for you not from some underlying distribution you do not know right so what you are essentially doing is there is an error function which is the error function of  $\beta$  on the entire input space right what I am really trying to minimize is the error on the entire input space right but I cannot compute that gradient because I do not have the data distribution I do not have the the entire data set available to me right.

I only have a sample and the sample was chosen in a stochastic fashion, so if I am trying to minimize the error with respect to the entire data so what I am actually finding out here is some sample radiantly it whatever sample data said was given to me I am finding out the gradient with respect to that sample data right that is one part right the second thing is typically I end up doing this one data point at a time instead of doing it n data points right if you I spoke about this in perceptron, so when I am doing it one data point at a time that is essentially not even the correct computation of the gradient.

So even given the data that is available to me I am only doing it one data point at a time that means I am actually doing some incomplete computations of the gradients right and if I take really tiny steps in the direction I am hoping that on an average I will move in the right direction so if I had computed for all the data points and taken a step then well given the data that is the best direction I can move in what if I computed one point at a time then I have to really take tiny steps.

So that I am sure that on an average I am going in the right direction okay, so that is that is idea here, so remember this is useful for the rest of the class so what is the problem with the perceptron people remember let goodness in something as simple as Excel because it was not linearly separable right.

(Refer Slide Time: 18:48)



So it is not able to solve this is there some way you can think of solving this let us call this  $x_1$  and  $x_2$  should be -1 how do I move it to a different space change the basis function right I cannot rotating the axis would not be sufficient for me to still be not like I still not be linearly separable right I think I said that right so what do we do now too many  $x$ 's on the board yeah so I can define a new feature but what do what should I define will it one  $x$  times okay  $x_1 x - x_2$  and  $x_2$  into  $-x_1$  okay.

So now what will happen if I do the projection of this data okay, so we are at this point go -1 -1 -1 is it, so where would 1 go -1 -1 where would 2 go 2 is here, so there will be  $-1 x -1$  will be +1 okay then this will be +1  $x +1$  will be +1 is it okay does not sound very promising what about 4 is also +1 what about 3 which is +1 4 so where will 4 go is my question 4 when there 4 is what  $+1 x +1$  okay  $-1 x -1 +1 +1$  that is 4 goes there right now this is -1 that will be -1 is +1 yeah there you go what about that k it is -1 -1 - what are they the same see you okay you negate  $x_2$  you first negate  $x_1$  and then pick the product okay.

Is essentially the same no see the idea is that I want my  $x'_1 = x_1$  and  $0x_2$  they are not the same thank okay now tell me so we know that we can actually implement and using a perceptron right you can implement or not using a perceptron okay, so now do this so where will  $x'_1$  be for 1 to next one is -1  $0x_2$  so there will be -1 right  $x'_2$  will be -1 okay what about 2 -1 that is not the same thing okay does that make sense, so in this transformation where  $x'_1 x$  is  $x_1$  and not  $x_2$  right and  $x'_2$  is not  $x_1$  and  $x_2$  right where true is +1 false is -1 and if you do that then my projection will be 1 and 3 will get projected to -1 -1 and 2 and 4 will get projected to the same coordinates as they were earlier in this modified space okay.

Now is this linearly separable right, so now I can separate this right, so I can listen sleep construct a perceptron that does that so what have we done here is that  $xr$  is hard to solve using a single perceptron but I can hookup several of them and I can solve the problem right, so is that computable using a single perceptron yes or no that is computable using a semi perceptron that is computable using a single perceptron huh first one is not yeah so I am just giving people a chance to change their minds you need to for each of them one for the not do you need to do a not can I feed in the  $x_2$  to do that can you do or not okay so if I can do  $x_1$  and  $x_2$  is it sufficient in invert the weight on  $x_2$  to get  $x_1$  and not  $x_2$  you think an answer anyway so the point is essentially what I have done this I can actually build a perceptron that can compute  $x'_1$  I can build another

perceptron that can compute  $x_2'$  so take these fill it to another perceptron which will actually compute right.

So essentially we have somebody sink another so essentially we have found they have shown that I can actually put this perceptron in layers and I can solve what was originally thought to be a problem that cannot be solved by perceptron right if you remember I was telling you in the beginning neural networks are very powerful because they are all hooked up in a very complex network right.

Individually they need on saw a very simple computing elements individually the neuron was a simple computing element and therefore you could not find the answer to  $x^*$  but then by hooking them up in appropriate Cascades you can find the solution to  $x^*$  right does it make sense, so sorry for the initial confusion on the feature transformation but now it should be clear right yes okay great.

### **IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

## Introduction to Machine Learning

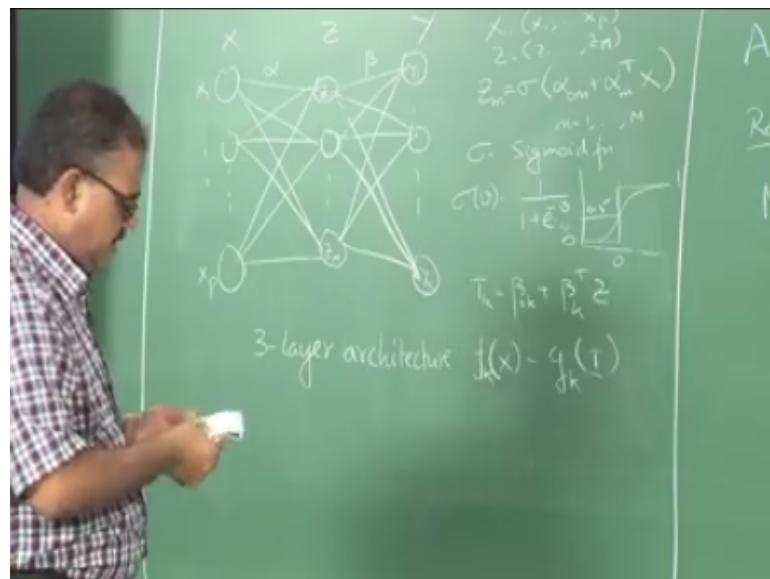
## Lecture 33

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

**Artificial Neural Networks II –  
 Backpropagation**

Now we move to artificial neural networks I suppose to looking at neurons.

(Refer Slide Time: 00:26)



So we are going to start hooking up neurons in multiple layers like this and the way we are going to train this is using gradient descent right, where we are going to Train this neural network is using gradient descent but what is the problem with using gradient descent for perceptrons the non-linearity right. So we had a threshold function which was not differentiable for the Adaline we got around it by getting rid of the threshold function altogether as I said we will use a linear output right.

So what is the problem in using linear outputs in multi layers multi-layer perceptrons so if you think about it let us call the first layer weights  $\alpha$  and the second layer weights I will call them  $\beta$  right, so if you think about it so what you are producing is some  $\alpha^T x$  is the output here if it did not have any non-linearity if I just use a linear neuron the output from here will be  $\alpha^T x$  right so  $Z$  will be  $\alpha^T x$  and what will go in here  $Z$  so  $Y$  will be that which is equal to just like having one layer of weights given by  $\alpha \beta$ .

There is no point in doing all this layering so if you only have linear neurons then I do not get the power of doing all the layering and I might as well have done a single layer of neurons, so I really need to do something nonlinear in the middle I need to have a threshold function for me to get the power of layering right so the threshold is actually needed but whatever we did that if I did kept it as linear neurons run into trouble great, so we need the threshold so how do we get around the fact that it is.

So I am going to say that okay now you get the drill  $Y_1$  to  $Y_k$ . Yeah, so sigma is a. So sigmoid is like a soft threshold right, so I can throw in a slope parameter here which I have not done that will give me different rates at which this sigmoid will assign so the actual threshold will be that way sigmoid will give me a soft way of doing the threshold the nice thing about the sigmoid is it is differentiable. There are many choices that you can have for the non-linearity a differentiable threshold function.

So the sigmoid is one of them so the thing with the sigmoid is that it will be between 0 and 1 so if you want it to be between -1 and +1 multiplied by something, is it? That will go from -1 and +1 right so you could have different choices, so for the time being I will stick with the regular the sigmoid that we are familiar with okay so each  $Z$  here, will be given will be given by that expression right so  $T$  is the quantity that you will see that goes as the input to the sigmoid in the output neuron  $Z$  the output from here.

From the middle are a hidden layer this one not one of this one  $0 m \alpha_0 m + \alpha m^T x$  can you zoom into that a little bit. At back, can you see it better now? People at the back who did not complain the fonts are small can you see it better now? So what are you doing on the laptop actually seen the video feed of this or something right so that is the output of the first layer okay and the output of the second layer is given by some other function  $G$  acting on this input  $T$  okay right.

So like to see if anyone notices something funny? Actually all can be different. Each one of these is a unit like this of that like that right, each one of those it is a one block this evil there are a good point okay yes there are three layers and this is sometimes called the standard three layer architecture okay but the first layer is really a fake layer this layer is really a fake layer it just takes  $x_1$  and gives  $x_1$  out on all the outputs okay, this one takes  $x_2$  in and gives  $x_2$  out on all the outputs okay this is this is really not a neuron okay so sometimes I like to call this as a two layer network because for me there are two layers of weights okay so it is a two layer Network right but in the literature for some reason this is called a three layer architecture.

So this is called the input layer. this is called the output layer okay one in the middle is called the hidden layer, so why is it called the hidden layer because I do not see the outputs of that layer directly okay so they are called the hidden layer so this is called the standard three layer architecture but there are other ways of doing it where actually you take outputs from the middle layer as well right and we can do and we can have inputs feeding into somewhere in the middle so you can have all kinds of craziness okay.

So the standard architecture is EC and we will stop with that okay I am not going to going to all the crazy neural network architectures are out there so you might want to take another course on ANNs specifically, if you want to I know more about thee all the crazy architectures outside so there is time permitting I will come back and do something very quickly at the end of the course not today but this is the standard architecture will stick with okay right so still people have not told me is there is something odd about this yeah.

Lastly it always has to be linear not necessarily but it did not also be sigmoid that is why I written it as decay, so the last layer could be sigmoid it could be linear right when we do you want the last layer to be linear when we want it to be linear is when you want to do regression for sure right when you want it to be a sigmoid is when you want to do classification and still people have not told me what is odd about this I am assuming people are thinking but I am waiting for the answer.

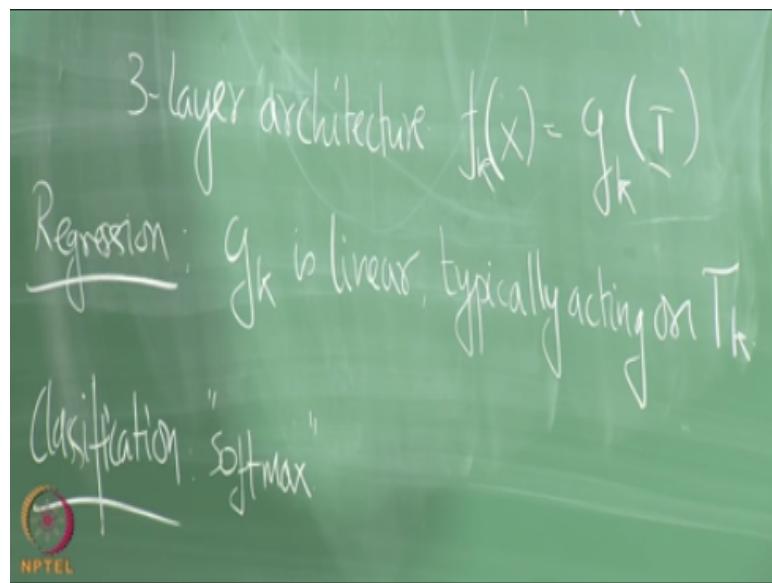
So that I can go that is why I am stalling yeah, so why did I write this directly as  $\alpha m + \alpha m^T x$   $\alpha_0 m + \alpha m^T x$  but here I split it up into  $T_K$  and  $F_K$  sorry good point, but why is it TY is it not  $T_K$  so if I am doing regression I might as well do it  $T_K$  right, so usually I because my regression my regression variables right if I am doing multiple output regression okay I am not talking about

multiple input regressions multiple output regression my output variables are typically taken to be independent right.

So I mean, so what is the what value I predict for one I will usually does not affect the prediction I make for the other right so I do not know if you read the book I did not I did not talk about multiple output regression before this but if you read the book they would have actually told you that you can do that regression independently right but if it is classification really they are not independent.

If I am going to be outputting class probabilities right they cannot be independent right if I am outputting class one probability is higher than class two probability necessarily has to be low okay so I had to say do some way of normalizing the outputs to produce probabilities right, that is why I am saying that this will operate on the entire  $T$ , I can produce a output probability vector so in case of classification you need to operate on the entirely so we will come back to that in a second right for classification. We will do a soft max like we did for logistic regression. Do you remember that so  $E$  power.

(Refer Slide Time: 15:16)



Yeah so I need the  $\sum$  over all the outputs right.

(Refer Slide Time: 15:23)

3-layer architecture  $f(x) = g_k(T)$

Regression:  $g_k$  is linear, typically acting on  $T_k$

Classification: softmax  $\frac{e^{T_k}}{\sum e^{T_i}} = g_k(T)$

I need a  $\sum$  over all the classes in the denominator and that is why my  $g_k$  operates on  $T$  so this will be the soft max thing, so  $e^{T_k}$  divided by  $\sum e^{T_i}$ . At this will be the  $g_k$  of  $T$  so what will happen if it is a single class I mean 2 class problem, it will reduce to a sigmoid it reduces sigmoid I can pick one class and have that output as the sigmoid and I can just say that okay, if this is greater than 0.5 then it is that class if it is lesser than 0.5 this other class correct so this is a 2 class problem that  $g_k$  will reduce to a sigmoid.

So if I am doing classification with only two classes I can straightaway keep a sigmoid as my output, output neuron and then I can solve it one how many neurons I need one right suppose in

solving a three class problem how many neurons I need as output 3 and you can always have the third one as dummy and then you can say I am going to do  $1 - \text{the sum of the rest}$  right but if you are going to do this right, so I can always say that okay I am going to have three outputs which will give me the probability of each of those three classes okay.

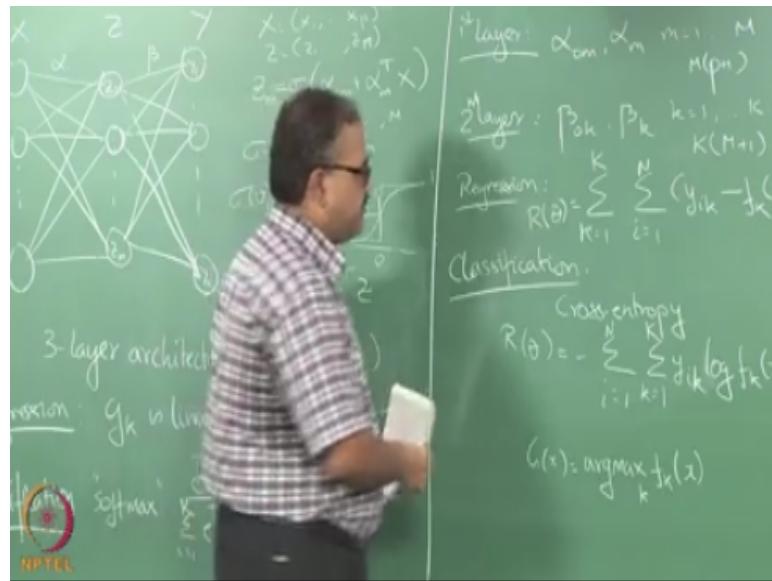
That is typically how it is done for two you just have one right but then for more than two classes you typically tend to have as many at the output is there any problem for doing that with doing that think about it I am not going to give you the answer right okay, so how do we fit the neural network parameters now and I have two layers the first layer  $m$  into  $P + 1$  parameters  $\alpha_m$  for right, so  $m$  for each of the hidden neurons right  $P$  for the  $\alpha_m$ s and  $+ 1$  for the  $\alpha_0$ s and in the second layer okay likewise, so I have that many parameters that they have to fit so I have to find all the  $\alpha$  and all the  $\beta$  so why do not you do more layers than this where did I stop with only three layers yes so empirically people observed that it is harder and harder to train why does it become harder and harder to train.

I will tell you in a minute okay but there is another reason for stopping with two layers right, so if you think of these as some kind of Boolean gates right if you think of the neural networks are some kind of Boolean gates it turns out that I can implement any Boolean function just using two layers of neurons except that the branching will become very large right but I can still implement any so as long as they do not give you any kind of gate with right I can have as many inputs coming in to a neuron.

As I want and I can implement any Boolean function in just two layers of neurons so why is that all of you know that right you can write midterm expansions right, so all of those things you know that and so you can essentially implement it in two layers of neurons and people thought oh, two layers is sufficient is a universal function approximator I can represent any function I want so let me not even think of what higher layers so that is one school of thought so people stop there but then there are others who are interested in going into more complex neural networks.

Because they did observe that when they got it to work okay adding more layers worked well so people kept at it and they made it work more robustly and so there like I was telling you that third wave of neural networks is all about having deep networks where you have more than two layers.

(Refer Slide Time: 21:08)



So regression what will be a loss function so my I am going to say define my regression loss for the parameters  $\theta$  what our  $\theta$  here all the  $\alpha$  and  $\beta$  okay, so I wanted a single notation for the parameters instead of saying  $\alpha$   $\beta$  everyday so I will say  $\theta$  is given by essentially this quiet loss that should look really familiar to you guys by now because they have been writing squared loss almost once every class if not often so what about classification what can we use for classification.

But 0/1 is incredibly hard loss function to optimize, right? So you could use squared error itself right so what is the rationale for using squared error same rationale that we use to linear regression right, so  $y_{ik}$  is an indicator variable that this one that gives you the probability of this particular data point being class  $K$  right and you are trying to fit the probability anyway that is what you are trying to fit and therefore for every data point you can take its probability of being class case 1 or 0.

And then you can try to do some squared error and try to make a prediction okay that is one way of thinking about it the other way is to use what is called the cross-entropy error or the deviance which is related to whatever we did in logistic regression, so here I will define my so we will all

put the problem the actual class table as the one that has the highest  $F_K$  of  $x$  like we did in the discriminant based classifiers except that  $F_K$  of  $x$  is no longer a simple discriminant function right.

And so this is essentially a error term okay that stands in for likelihood that we maximized earlier, so and to optimize this error term we will essentially train the neural network using maximum likelihood right so I am not going to go there so we will look at a more popular more mechanism for training neural networks which essentially looking at the gradient the squared error and do gradient descent okay.

### IIT Madras Production

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**Introduction to Machine Learning**

**Lecture 34**

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

**Artificial Neural Networks III-  
Backpropagation Continued**

(Refer Slide Time: 00:33)

$$\text{Let } Z_{mi} = \sigma(\alpha_{0m} + \alpha_m^T x_i)$$
$$Z_i = (Z_{1i}, \dots, Z_{Mi})$$
$$f_k(x_i) = g_k(\beta_{0k} + \beta_k^T Z_i)$$

So let  $Z_{mi}$  correspond to the output of the  $M_{th}$  unit in the hidden layer corresponding to the  $i_{th}$  input this is not the  $i_{th}$  component of the input corresponds to the vector  $x_i$  right so it is the  $i_{th}$  input in my training data of  $n$  elements okay and I am going to say that and  $Z_i$  corresponds to the, the entire activation of the hidden layer for the  $i_{th}$  input. it will find so far right.

So now we got rid of over what did we get rid of here? the  $T$  right so this is what I was saying in regression  $G_k$  is linear and typically  $d$  is acting on  $TK$  right and so this is acting only on  $TK$  so this whole thing is so because we are only talking about the regression, I got rid of that this will make our life a little simpler when we write the write the gradient so I am going to take the basic

I am going to use gradient descent right so I have squared error I am going to use gradient descent.

So I am going to take the derivative of the error with respect to the single output layer weight okay this is a weight that runs from some neuron  $M_{zm}$  right to some output  $K$  right so that is  $\beta$  okay just this one, one weight I am taking here right I am taking the derivative of  $R$  with respect to that one weight okay is the setting clear right so I am taking the derivative of  $R$  with respect to a single weight here.

Let us just designate that as  $\beta_{KM}$  so what will this be equal to yeah okay let us do it in a slightly simpler fashion so I am going to assume that each term inside is denoted by  $RA$  then I just do the summation over all  $i$  okay so that way I do not have to write the summation over all  $a$  everywhere so I am going to say this is  $\frac{\partial R}{\partial \beta}$  right if you remember the earlier that what we had the thing that I erased here there was just this right  $Y_i - f(X)$  was what we had earlier right.

But the input in this case is actually  $ZM$  right if you think about what is there on the other end of this weight right so the input that comes from here is actually  $ZM$  right so mathematically if you think about it just let them  $\beta_m$  okay right so that is what is happening so essentially that is what you are going to get so the  $i$  indicates that you are considering it only for the  $i^{th}$  input right this is clear so far we just then just taken a derivative right but exactly the same computation that we did earlier the only new thing here is they are the derivative of  $GK$  earlier.

We did not have that because we are assuming that  $GK$  was linear so  $GK$  is linear this will again vanish now comes the interesting part they will just disagree some the single input layer wait we will consider that so I am calling it  $\alpha_m$  how will I take the derivative of the error with respect to  $\alpha_m$  you look at the error  $\alpha$  does not appear directly at all it appears indirectly so what is the best way to do this name this using the chain rule.

So  $\alpha$  is going to affect the output of the hidden layer right and the output of the hidden layer is obviously going to affect the error right so I am going to take the output of the hidden layer right so I am going to chain it through the output of the hidden layers are going to take  $\frac{\partial E}{\partial Z_m}$  by  $\frac{\partial Z_m}{\partial \alpha_m}$  and  $\frac{\partial E}{\partial R_i}$  by  $\frac{\partial R_i}{\partial Z_m}$  right so one thing to note is that  $\alpha_m$  is going to affect the output only of  $ZM$  right it is going to affect only  $ZM$ .

(Refer Slide Time: 08:32)

$$\frac{\partial R_i(\theta)}{\partial \alpha_{ml}} = \frac{\partial R_i}{\partial z_{mi}} \cdot \frac{\partial z_{mi}}{\partial \alpha_{ml}}$$
$$\frac{\partial R_i}{\partial z_{mi}} = \sum_{k=1}^K \frac{\partial R_i}{\partial f_k} \cdot \frac{\partial f_k}{\partial z_{mi}}$$

So I just need to chain through ZM okay is it clear so then let us do each one of these in turn so this is rather easy so is that you have that already so what is  $\partial Z_M$  by  $\alpha_M$  what if they did be more consistent okay that makes sense right the derivative of  $\sum$  yeah can you zoom in so the derivative of  $\sum$  times  $x_{il}$  right so  $\sum$  prime of  $\alpha$  transpose  $X_i$  plus  $\alpha^0$  into  $X_{il}$  so that is essentially the, the derivative of  $Z_M$  with respect to  $\alpha_m$  it is straight forward differentiation if you are having trouble with it I do not know now is the tricky part so I am looking at  $\partial RA$  by  $\partial ZM$  right.

So what is  $ZM$  it is the output from here right but unfortunately this  $K$  goes to all the output neurons right so  $ZM$  can affect the output through all the output neurons okay so far there is been a single path that we have been considering but at this point we really have to consider all the paths of reaching the output from  $M$  okay.

(Refer Slide Time: 12:19)

$$\frac{\partial R_i}{\partial \alpha_{ml}} = -2 \sum_{k=1}^K (y_{dk} - f_k(x)) g'_k(\beta^T z_i) \left| \begin{array}{l} \beta_{km} \\ \sigma'(\alpha_m^T x_i) \cdot x_{il} \end{array} \right.$$

$$\delta_{ki} = -2(y_{dk} - f_k(x)) g'_k(\beta^T z_i)$$

$$\delta_{mi} = \sigma'(\alpha_m^T x_i) \sum_{k=1}^K \beta_{km} \delta_{ki}$$

$$\frac{\partial R_i}{\partial \beta_{km}} = \delta_{hi} z_{mi} \quad \frac{\partial R_i}{\partial \alpha_{ml}} = \delta_{mi} x_{il}$$

So what we really have to do is look at okay so, so ZM can affect RI through FK right so the derivative of FK with respect to Zm and RI anybody with respect to FK that is a chain rule again do this over all K because I can have multiple parts of reaching the output so what is  $\partial R / \partial K$  okay  $\partial R / \partial K$  it may which should be able to rattle it off just the derivative of GK so putting everything together.

I can write that is a big expression and I did nothing I just took this and wrote it here I took that and wrote it there okay I just took the product of the two terms so what we will do now is just to introduce certain simplifying notations let us think about it I have made my job a lot simpler so that is this term  $\Delta K$  which ever define so  $\partial R I$  with  $\partial \beta$  is essentially  $\Delta K$  into  $Z_{mi}$  right  $\partial R I$  with  $\partial \alpha_{ml}$  is essentially  $S_{mi}$  into  $X_i$  that is the  $\Delta$  part right.

And there you have a  $\beta$  and then you have your  $\Sigma$  prime so this all put together gives me mass  $S_{mi}$  so there nothing you just applied chain rule and done some manipulation to simplify this right if you go back and do it again okay you will find that it is very straight forward gradient computation but it took people a couple of decades to nearly a couple of decades to realize that they could do something as simple as this chain rule.

And apparently this technique which is very popularly known as back propagation so why is it called back propagation so when you take the input right and you compute the output that you are propagating the values forward through the network right but when you are updating the

gradients so if you think about it so what you are doing is first you are computing the  $\Delta$ 's right and then you are propagating the  $\Delta$ 's back through that weights  $\beta$ 's right.

So essentially what you are doing is  $\Delta$  times  $\beta$  it like when you are going forward you do  $x$  times  $\Delta$  and  $Z$  times  $\beta$  right so here likewise you are doing something like  $\Delta$  times  $\beta$  right so this is something like a back propagation of this  $\Delta$  term through the weights so as to update the first layer weights right so that is why it is called back propagation okay so the forward thing is whatever you do this, this is the forward pass okay and the equivalent backward passes are given by that right so the actual equations are right.

(Refer Slide Time: 18:47)

$$\begin{aligned}
 \beta_{km}^{(l+1)} &= \beta_{km}^{(l)} - \eta \sum_{i=1}^N s_{ki} z_{mi} & x_i &= (x_1, \dots, x_p) \\
 \alpha_{ml}^{(l+1)} &= \alpha_{ml}^{(l)} - \eta \sum_{i=1}^N s_{mi} x_{il} & z_m &= \sigma(\alpha_{0m} + \alpha_m^T x) \\
 \sigma'(v) &= \sigma(v)(1 - \sigma(v)) & \sigma(v) &= \frac{1}{1 + e^{-v}} \\
 \tanh(v) &= v - \sigma^2(v) & T_k &= \beta_{0k} + \beta_k^T z \\
 \text{for architecture } f_k(x) &= g_k(T_k) \\
 \text{Regression: } g_k &\text{ is linear, typically acting on } T_k \\
 \text{Classification: softmax } \sum_{i=1}^I e^{T_i} &= g_k(T)
 \end{aligned}$$

So we still left some things in there so I left a Gprime and a  $\sum$  prime and so on and so forth so if  $G$  is your linear function great right what about  $\sum$  prime  $\sum$ ,  $\sum$  is the sigmoid function then now

you can take that derivative of the  $\Sigma$  with respect to X and that is what you will get and if it is at tan H right instead of the sigmoid if I use the tan H function then my  $\Sigma$  prime will be  $1 - \Sigma^2 V$  you can work it out but sadly easy differentiation always people get thrown off by back propagation but it is really nothing but differentiation and a lot of algebra right just manipulating things around it is nothing more than that everyone knows the chain rules right that is it that is it, it is just a chain rule.

### **IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**NPTEL ONLINE CERTIFICATION COURSE**

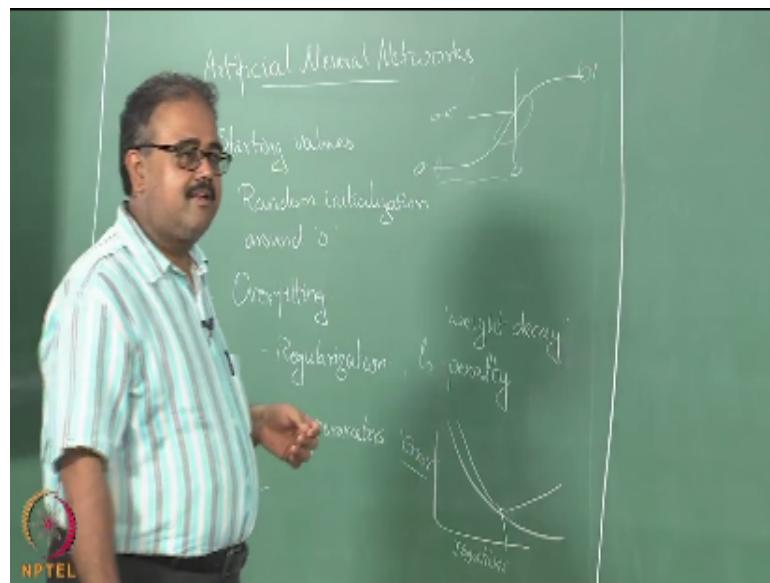
**Introduction to Machine Learning**

**Lecture 35**

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

**Artificial Neural Networks IV –  
Initialization, Training and Validation**

(Refer Slide Time: 00:21)



The first thing concerns the starting value of the weights right, so you have all this  $\alpha$  and  $\beta$  you have this whole set of parameters in a neural network right, so we talked about gradient descent but gradient descent starts at some point in the weight space right, so you need to have an initial guess for what your  $\alpha$  and your  $\beta$  should be right. So what should you do what are the good guess? set them all to 1 is it yeah, so setting all of them to the same value whether it is one or sometimes people say this 0 right setting all of them to the same value is usually not a good choice right.

More often than not you will end up in some weird part of your gradient space and you will find it hard to get out right it is very rare that you are going to come up with a actual solution where all the weights are the same right. So if you were going to start off with a solution where all the weights are same right it is going to be hard to specialize right, so typically what you do is you do use random initialization, but there is one more constraint let us be feeling really lonely one more constraint to the random initialization.

So what do you think that would be the constraint yeah of course you do not really want to have infinite weights yeah, so but then what should the min and max is, so you really want your weights to be rather small okay? So just think about what is the implication of having really small weights what is the implication of having really small weights, so you remember your sigmoid right, so if your weights are really small where do you expect the outputs to lie? You know your  $\alpha$  transpose  $x$  or your  $\beta$  transpose  $z$  will lie somewhere in this region because this is where 0 is right.

So this is 0.5 or 0 depending on whether you are using tan H or the sigmoid, so if you are using the sigmoid, so right around 0 you will have this 0.5, so if you look at this region this is almost a linear region correct. It is almost a linear region so I would really like to start off my network so that most of the outputs of the neurons are in the linear region. Why is that? Can you think of it and you have enough information to answer this question. The gradient will be larger right so the gradient will be larger if you are somewhere around here.

So even small changes in the input space or small changes in the weights will actually cause a large change in the output right, so if you end up going somewhere here or somewhere here right you can see that you are already saturated right, so you really have to drag yourself all the way here to see any change. So if you are somewhere in the middle you are more sensitive right to what the input right you are more sensitive to the weight changes right. So all of this helps you learn more rapidly, so this is one of the reasons you start off with random initializations around 0.

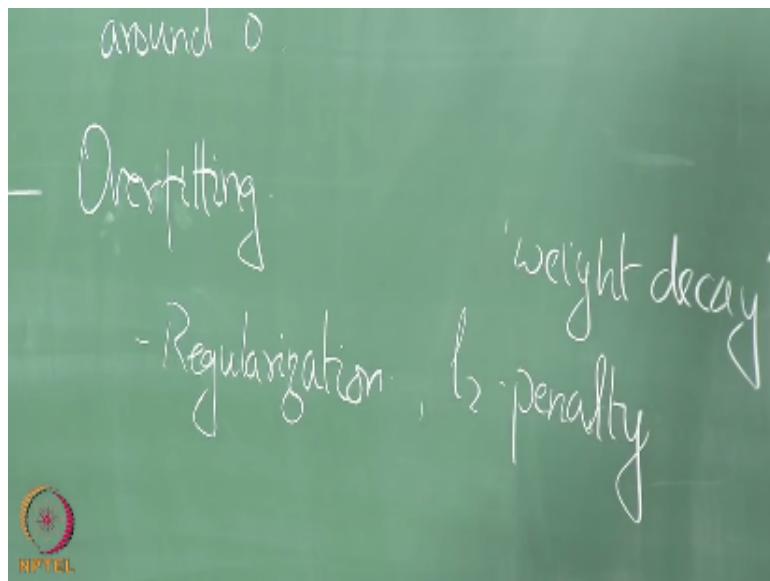
Of course you can make everything zero but then that will put you in a very weird part of the part of the search space right. So it is good to have some kind of randomness so that each wait can specialize to different things okay. It used to be the big bane of neural networks over fitting so why is that the case? You know all of you remember what over fitting is right, sorry yeah so I

mean people remember what over fitting is yes, the training examples right there essentially you are fitting the parameters very closely to the training examples.

So you are not able to do any kind of generalization to unseen examples right and the reason why neural networks do this over fitting is exactly because they have too many parameters you have so many weights here. So if you remember we actually counted the number of parameters in a neural network right you know that  $M \times 3 + 1 + \text{whatever } K \text{ time's } n / M + 1$  well that is a huge number of parameters. So it is very easy for you to over fit so you have to be careful about it, so there are two ways of avoiding over fitting.

So can you think of what are the two ways of avoiding over fitting one we already know regularization right, so one way of avoiding over fitting is regularization, so what you do here is you essentially add a quadratic penalty for the weights right so you do a norm  $\alpha$  squared + norm  $\beta$  squared and then you try to find the gradient with respect to that and then you try to minimize things, right it becomes a little bit more complex and if you add a squared error penalty right if you add a squared error penalty it is sometimes called weight decay right because it makes your weights go towards zero. So sometimes called weight decay, so can you add a  $l_1$  penalty

(Refer Slide Time: 07:16)



You could you could add anything I mean so it is it just makes it a little bit more complex but the question is does it induce sparsity right, they said another way of avoiding over fitting yeah you can tie parameters together to avoid over fitting that is good but in which wherever is new tied together? In fact one of the ways that deep learning actually has been made efficient despite doing this tie of parameters right but then the architectures that look at are very complex right so you could tie parameters together and try to reduce this okay.

So let me put that as a different kind of thing but it is actually a form of regularizing but it is yet another approach which is a purely empirical way of doing things which is to do what is called validation right I actually mentioned this in one of the earlier lectures right, so you train on a training set and then you have a validation set and then people remember I drew even a picture right, so as you are training so the error on your training set keeps going down right but the error on the test set or the validation set will initially go down, at some point it will start going up right maybe not that dramatically but go up nevertheless.

Like that right so the point here is where your right solution is okay, so I am putting it in quotes right solution okay, so what is the x-axis and what is the y-axis on this figure? Whatever is your measure of error it could be miss classification error or whatever right, so fiction you ideally want miss classification error if it is the regression it is a prediction error whatever and the x-axis is usually iterations right but you could also think of having a figure like this for complexity right.

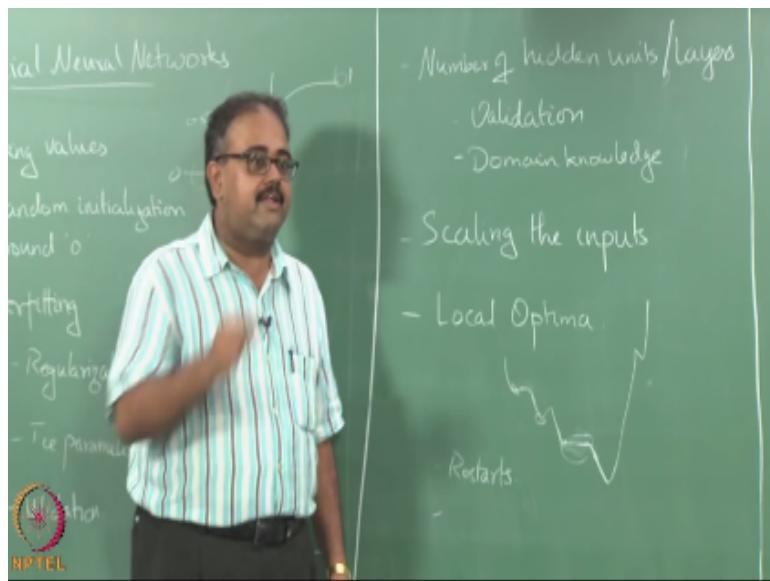
So you could say that I am going to keep adding more and more neurons right or more say I can keep expanding my  $M$  right I can keep adding more neurons in the hidden layer right.

I cannot change the new neurons in the input layer I cannot change in the output layer usually right because that is their depend on the problem that I am solving right when I can keep increasing the neurons in the hidden layer right but then these are more or less you know standard techniques for doing, I mean for avoiding over fitting not necessarily tailor to neural networks right. So you should remember that there was about a decade and a half when people worked a lot with neural networks in between right, they came up with many techniques for avoiding this kind of over fitting.

And they have explored many variations on parameter tying right and also many different kinds of regularizing so there are some really interestingly named algorithms for avoiding word fitting. so one that I particularly like is to be called optimal brain damage, essentially the idea was to remove weights from the network right so you train the neural network and then try to find out the sensitivity of the output with respect to certain weights okay. If I am changing this weight how much do the outputs change you know how much does the error change right.

So weights that have low sensitivity or right or yeah weights that exhibit low sensitivity on the error right well then remove and then you just retrained we keep doing this. So that way you are removing the number of parameters reducing the number of parameter heavily but we are not affecting the output too much. So like that there are many variations on it but these are the three things right that we have to think about.

(Refer Slide Time: 11:57)



Related to the over fitting question you see how do you figure out the number of hidden units and layers. Like the very expensive way of doing it is to do a similar validation kind of a setup right keep increasing the number of layers and then our number of neurons and check that out it is incredibly hard right and so people came up with automatic pruning techniques right, people came up with ways of growing your neural network. So they start off by having one neuron so something very similar to your forward feature selection or what is the thing a stage by stage wise selection right, so people remember stage wise feature selection what did he do?

At stepwise stage wise or something different see you later right so you could do the same thing with neural networks right, you start by training a single neuron all right and then what you do is once that neuron actually starts making some predictions right then you train another neuron that actually nullifies the prediction error right. Now you know a third neuron that adds up both of these and gives you the output right, so you could do something like that right so you do not have to make a decision as to how many neurons you are going to put in from the beginning right as you go along you just keep training more and more.

But the problem is such a network was that it will not look like your layered architecture right I will start off with 1 neuron okay that will give an output then I will add the other neuron then they will go so all of this gets the input directly, then I add another neuron right so the layer architecture is gone no wall right. So now how many layers this is our two or three you know so

this neuron seems to be at the third layer right but it is connected directly to a neuron which is certainly first layer because it is taking inputs from here.

But then well you do not have to be very dogmatic about having the standard three layered architecture if you remember when introduced the standard three layer architecture I said there are a lot of different deviations from this that people have proposed right and will not be looking at most of those in detail. So these kinds of networks where you are actually trying to minimize the residual at every point they are called cascade correlation networks, so there are. So the most statistically sound ways to just do validation, what is the other way? To do it is kind of a cheat it is slightly better what can you do you can take an educated guess using domain knowledge.

I said you have some information about how complex the system is and then you can use ideas from that you can then try to see okay one layer two layers three layers right, so a lot of deep learning network that nowadays happen essentially it is more empirically driven you try one layer okay see what is the best you can do right see thoughts the best in performance, as you can get and then you try to add another layer and see if we can improve that another layer another layer until you are happy that you are performing well right. Of course you just cannot train the network into the ground you have to always make sure that you are not over fitting it but you can still this right okay.

So I did not explicitly mention this while talking about the numerical training but you can kind of imagine, wherever that I am going to be using the data as it as a real valued vector right I would have to worry about scale, so if I have one variable that has a very large range another variable which has a very small range, at the variable with a large rate it is obviously going to dominate my gradient computation. If you remember the gradient has a  $X$  the  $x$  ml component to it right that is the input variable is part of the gradient. So if the variable some of the variables can have a very large range and some of the variables might have a very small range and the large range variables will dominate the computation.

This numerically by being large they are going to dominate the computation right, so we do not want that to happen whether they are actually needed or not just by being numerically large right they will dominate the computation, so we essentially make sure that all the variables will have the same range right. So we talked about this in couple of other scenarios also but in this case again it is important so this is something which people typically forget. When they are using

either neural networks or SVM's you try you take the raw data right and you just try to run it through a neural network or run it through an SVM and then produce a classifier right and quite often things do not look work that well right.

You might find some reported results that are much better than what you are getting by using is SVM, nine times out of ten okay the reason to fold with SVM is a two-fold with neural networks one thing you forgot to scale the input okay, the SVM's what happened? So you have those kernel functions we talked about right you forgot to tune the parameters of the kernel function you just took the kernel function as it is and you are trying to use it so that the performance will be bad, so you have to tune the parameters of the kernel function and you have to scale the inputs, if you do not scale the inputs sometimes the performance can be arbitrarily bad.

And this is a problem which SVM do not have and that is one of the reasons they became so much more popular than your networks in late 90s and early 2000s right, so the neural network error surface is fairly complex. So what do I mean by error surface? So what is it so error surface what will be the x axis the y axis the z axis whatever can you describe mathematically, what the error surface is? with respect to what aha on the what parameter is not the inputs okay, so the error surface is something which people have difficulty okay there is areas on I am making a fuss out of it.

The error surface is the function of the error with respect to the parameters okay, so as I change my  $\alpha$  and  $\beta$  how does the error change okay, so that is that is the error surface that we are talking about right and so how does the error surface how will it look like for the case of SVM and minimizing something quadratic they are right in terms of  $\beta$  right, so it is actually very nice quadratic thing, so it always has a single Optima right when the optimal hyper plane formulation the very nice thing about it is it has got one Optima and then if you run the optimizer on it you will always get that solution right.

The error surface for neural networks if you think about it it's got those stupid exponents in there right your sigmoid there is the derivative of, your sigmoid is in there so the error surface is going to look incredibly complicated right. So it is going to have lots of little valleys right the error surface is going to look something like this, ever even look something like this okay, so if I am doing gradient descent I might come here and get stuck, that looks like I mean whatever

direction I tried to go there is increasing. So I might say okay this is the good place to be so I might just stop there right.

I could get stuck here what about here huh very slow or not at all because the gradient is 0 I mean if you are in the middle of a point the gradient is zero because it is flat it is flat, I declared that to be flat okay and so the gradient is zero at that point right and there you go okay and so you might not just move right you are essentially drifting around there you and whatever happens you are not able to make any progress. So the error surface can become really complicated like this right and this is just on one dimension right, they say this is a no one it is a single neuron with one input that is what we have drawn.

So imagine this generalized to a very large dimensional space right  $mp + mk$  dimensions right, so the surface can be really complex and again the plethora of solutions were getting out of local optima right, so we are not going to get into most of those and let us tell you one very practical way of doing it essentially do restarts right. So you start off with some random initialization close to 0 right you do gradient descent until you do not change weights very much. Remember those weights right remember those weights and remember the performance. No reinitialize the network again close to zero random weights close to zero.

You say different random see please all right and then rerun the experiments right and again you will go off to some other optima remember those and keep doing it. There are other techniques which people use right, so they can make the, you know there is something cleverer gradient descent techniques right which all of you to get over these local optima not all of them but at least some of the shallow or local optima it allows you to get over easily. So for example this is a shallow local optima right with a little bit of effort I can actually get over and how do you provide their effort.

So think of it from a very dynamics perspective, so people have added something called momentum right, so if you have been moving in a particular direction I have been descending little bit a little bit a little bit okay do not stop just keep going in the direction for some more time right that is momentum right. So in this case these kinds of shallow things you can get over you know the gradient has become a slowed down significantly right this becomes 0 here but I will still be going in the same direction I went for a little while longer because I have momentum it is going to take me forward.

So that is going to get me out of this little valleys but if you are in a deep valley then still cannot get out right but so these kinds of tricks help right and then more recently with all the with deep learning, that one of the reasons that deep learning is become so popular most people do have very powerful gradient based techniques which allow you to navigate the error surface more efficiently a lot to avoid local optima, but allows you to navigate the surface more efficiently right good. So any questions so far, no why well I am going back to my zero yeah when I start restart a little again be small weights right.

I will be very far away from this is optima I have converged to after a lot of training, you maybe not see that, this is one thing which you should get your hands dirty then you will see what I mean. Even small changes in the starting weight configuration can lead you into very different Optima that so there is surface or so complex right and remember I am not just moving in one direction or the other I have a very large dimensional space in which I am moving right. So even though I am taking and I am constraining it to be around zero right, the volume that I can actually start in is very large because of the high dimensionality of the weight space right.

So and each random starting point can be very different because it is also possible that you start in the same location or start very close to the same location. I will end up with the same Optima but that is probability of that happening is very small especially with large networks. So if we do the D start will actually end up with somewhere else. Any other questions so we have till this point we have the exam rate just checking yeah, so that is the question nobody asks how many times do you restart right.

There when you have a budget you just say that okay I am going to restart this many times right and yeah you might have actually re absolute minima but you may still be doing research that is one of the reasons I told you to remember the weights right, it could very well be that the best weight best solution, you got could have been oh the first one and then all the fuel further restarts that you do could actually be leading to worse collisions right. So I am not guaranteeing that we do a restart we will get a better solution the restart just allows you to explore different local optima and pick the one that is best.

It all depends on your budget right I mean if it is it is as expensive as to train the network the first time around right and it is really expensive to train the network if you are doing deep neural

networks because the number of parameters are really large runs into several hundred thousand right and therefore doing a start is expensive, we do fewer of them and I also tried to come up with other gradient descent techniques, that allow you to avoid local optima. The whole idea behind simulated annealing is that you would want with some probability of ignoring the gradient right.

So what the whole algorithm here says at every point follow the opposite direction of the gradient right, you would descend the gradient direction you find which is the maximum ascent direction you descend it right. The whole idea behind simulated annealing is to say that no I allow you to ignore the gradient, you can move in another direction in fact you can move in the direction off the gradient also if you want right. So the gradient opposite direction of the gradient is a choice that is given to you can choose that or you need not choose them. As the number of iterations become larger and larger the probability of you choosing the direction of the grade up the gradient direction is going to become higher and higher.

This is essentially what we call the temperature parameter the temperature parameter is very high right if you can think of this particle that is going to be jumping all over the place right, so if the temperature is very high you can move in whatever direction you want you are not necessarily constrained to following the gradient direction and that is the temperature becomes lower and lower you, then you are constrained to follow the direction of the gradient .So the reason is called temperature is because it is actually used in modeling physical systems right, so the so if you look at the Boltzmann distribution which people typically use in this context that is actually a parameter called temperature.

Which behaves very much like this right, so one another way to think of what the simulator annealing will do to your error surface it is like it will pull your error surface is flat if the temperature is very high it is like your inner surface is flat it essentially means I do not have any gradient information. I can move any which way I want right whatever direction I move it looks the same then I can this move randomly and then what happens I slowly start regaining the shape of the inner surface.

So what will happen is first the deepest dink will form okay the shallower ones will still be farther away and the deepest thing will the first tip that will appear in my error surface and likewise as I keep cooling it cooling it cooling it will completely go back to the original inner

surface. That is one visual way of thinking about it I mean there are more formal ways of explaining why that visualization works but I do not want to get into simulated annealing today but that's another way of avoiding local optima okay done.

**IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

## NPTEL ONLINE CERTIFICATION COURSE

## Introduction to Machine Learning

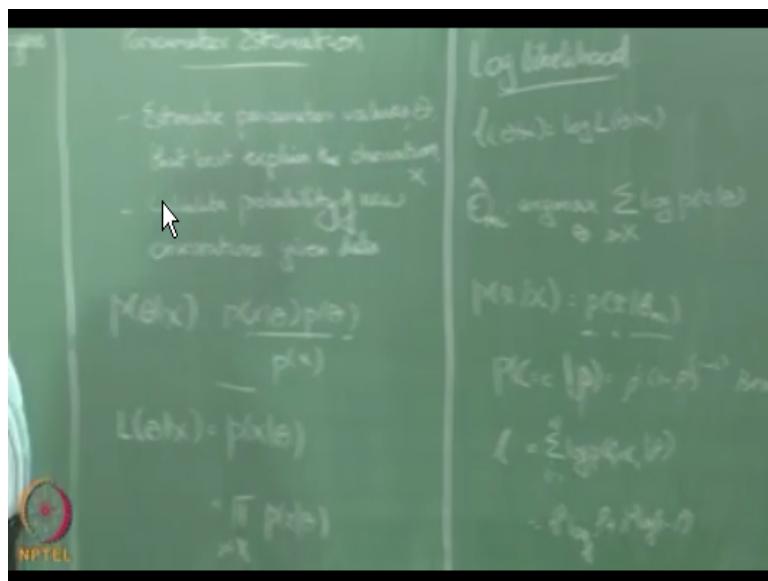
## Lecture 36

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

**Parameter Estimation I: The Maximum Likelihood Estimate**

More of a Bayesian approach to parameter estimation,

(Refer Slide Time: 00:19)



so I would say that there are two goals to us we want to estimate the parameter values right. That best explains some given data to us right. We already looked at this in the context of logistic regression right, so we assume some kind of a model that was generating the data and then we estimated the parameters of the model that somehow best explains the data right.

And we used the notion of a likelihood right we are just going to look at it again little in detail. And then go on and look at a couple of other ways of doing parameter estimation. The second thing we look second problem that we are interested, so essentially calculating the probability of

new observations given the old training data right. So I am going to assume that the parameters are given by  $\theta$  right.

The observation is given by  $X$ , so what I am interested in is per probability of  $\theta$  given  $X$  right. That is all the familiar Bays rule; so what is that what is that prior, likelihood right. So I only looked at it right, so I am going to like write the likelihood does likelihood of  $\theta$  given  $X$  right. So if I write it like that people get a little confused, so likelihood of the data right. I already mentioned this is likelihood of  $\theta$  not data.

Its likelihood of the parameters given the data even though we write it as probability of  $x$  given  $\theta$ . Is it clear? Why is it a function of  $\theta$  of  $X$ ? Because  $X$  is fixed in our context  $X$  is fixed right. I have given the observation  $X$  right I am interested in finding the parameter values  $\theta$ . So the way I am going to set this up for a given  $X$  okay, for different  $\theta$ . What is the probability of that  $X$ ? Let us say I can consider five different  $\theta$  for  $\theta_1$  what is the probability of  $x$  for  $\theta$  towards the probability of  $X$  for  $\theta_3$  what is the probability of  $X$  and so on so for.

So that gives me the likelihood of  $\theta$ , sometimes people say the likelihood of  $X$  with respect to  $\theta$ . In fact it is so widely used I do not know if it is even right to say it is incorrect anymore. Like pre-pone a meeting, but so why are we interested in the likelihood why are we interested in the likelihood? Yeah! So, what I am really interested in is to find that  $\theta$  that best describes the data right.

So essentially what I am interested in is finding the  $\theta$  that has the maximum probability here. That given the  $X$  which  $\theta$  has the maximum probability right, so the  $X$  is fixed right so this does not matter and if I really do not have any information about what  $\theta$  is the best  $\theta$  to start off with right. This is also irrelevant because it will be the same for all the  $\theta$  correct, so if I want to maximize  $P(\theta)$  given  $X$  all I need to do is maximize  $P(x)$  given  $\theta$ .

Because this is constant that will also be constant across all  $\theta$  all right, so it is enough if I look at likelihood right. So if I make an assumption that we make the assumption that my all my data samples are generated independently right as well right my likelihood as the product of the individual probabilities. Then we do not want product or probability so we typically end up using, people agree with that?

So suppose a new data point the  $X \tilde{}$  comes what is the probability of  $X \tilde{}$  with respect to  $X$ . Sorry, I mean given  $X$  sorry given that I have already been given some training data  $X$  okay. I am asking you what the probability of this new point  $X \tilde{}$  is. What will be the probability right. In fact is exactly what we did in the logistic regression case, we found the maximum likelihood parameters for  $\beta$  maximum.

Likelihood estimates for the parameters  $\beta$  and then we plug them back in and say okay. This is how you estimate the parameters right. So let us look at a simple example, it is considered a simple coin tossing experiment right, so there is a random variable  $C$  okay. Which has some outcome lowercase  $C$  right, so lowercase  $C$  if it is one in his heads, if it is 0 it is tails okay. What is the parameter that I have did not coin tossing experiments probability of coming up its okay. Let us change the symbol still looks like  $P$  but it is a  $\rho$  so the probability of whether you come up with heads or tails right.

Given the parameter  $\rho$  okay, so what is the probability that should also look familiar? We already saw that right in the context of class label being 1 and class payable being 0 right. And the probability of coming up class 1 versus probability of coming up class 0 right. It is like heads and tails. Now the probability of coming up heads, well  $\rho$  power  $C$  is 3 well. If its 1 then it is  $\rho$  if it is 0 which is tails since  $1 - \rho$  okay.

So this is the expression in simplified form, I always have written I would have to write it as  $\rho$  if  $C=1-\rho$  if  $C=0$  right. In stuff that I can write this using a selection function so what is this probability density called? Bernoulli yeah! Okay. The Bernoulli is so what is the likelihood when look like for each of the  $I^{\text{th}}$  toss you would have an outcome lowercase  $C_i$ . So what is the probability that the random variable can be lowercase  $C_i$  given the parameter  $\rho$  right.

And some this over all the right, so there are  $N$  one times head says occurred. Let us say and  $N_0$  times tales has occurred right. Then I can simplify the summation as  $N_1$  times  $\log \rho + N_0$  times  $\log 1-\rho$  right simple enough. Now take the derivative of the likelihood equated to 0, and tell me what  $\rho$  is right. So our common sense way of estimating probabilities from experiments is what toss coin  $N$  times find out number of times it turned up heads okay. Divided by  $N$  that gives you the probability of it coming up heads. And a turn out that is the maximum likelihood estimate assuming. That your coin is obeying a Bernoulli distribution.

**IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**NPTEL**  
**NPTEL ONLINE CERTIFICATION COURSE**  
**Introduction to Machine Learning**

**Lecture 37**

**Prof. Balaraman Ravibdran  
Computer Science and Engineering  
Indian Institute of Technology Madras**

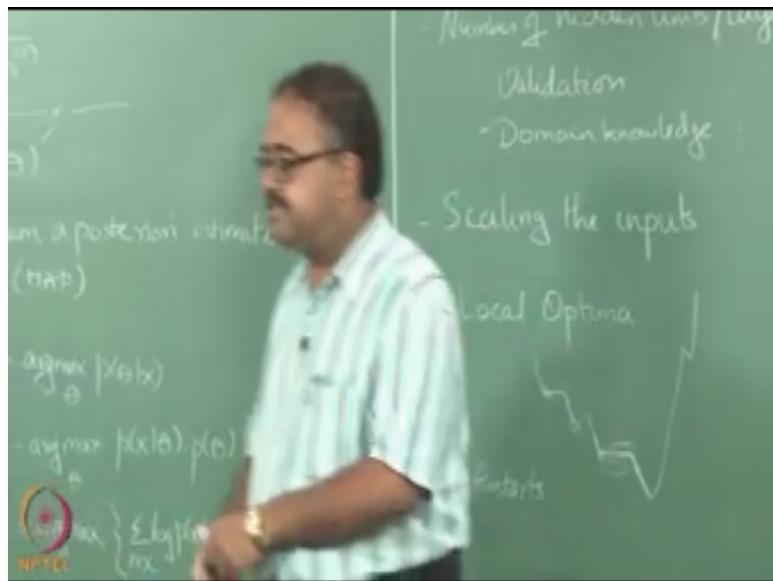
**Parameters Estimation II: Prior and the  
MAP estimate**

Ok so far we assume that the whole motivation for doing maximum likelihood was hey I did not know anything about the parameters .Before I started the experiment right before I gather the data I did not know anything about the parameters suppose I did know something about the parameters. So what could you know about the parameters just stick with the coin tossing experiment give me something from the coin toss case yeah well yeah I know the high probability it is fair.

If you know it is fair or not that is not a prior information that is insider trading okay so with the very high probability and think it is it is fair right so he hands mean coin we look at his face I mean obviously he is not going to cheat me right. So I will assume it is a fair coin to begin with right so what I can do is I can have a prior probability of it being fair being very high right now I can in fact think of having a Gaussian with the with a peak at 0.5 for  $\rho$  right. I can think of having probability of 0.5 for 0.5, for  $\rho$  being 0.5 there are two probabilities here all right do not get confused as a probability of the coin coming up heads and there is a probability of that being 0.5 right. So that is the prior probability we are talking about here and I am saying that I can think of it as Gaussian and our Gaussian is a good idea why not great probability is not only problem it can be even greater than one also mean either side is a problem right.

So what is a good distribution to do is useful you already seen that in your tutorials probability tutorials. Distribution that is  $\beta$  distributions limited between 0 and 1 in fact it seems to have been invented for putting priors on probabilities. Right in fact it was so you can think of that as your prior right so I have some information about right I want to use that in my optimization right.

(Refer Slide Time: 02:43)



So these are called alright so we looked at the maximum likelihood or ml here now we are going to look at maximum a posteriori or map right so this is a prior information right is this prior information about as  $\theta$  is the posterior information about  $\theta$  right. But are we actually computing the posterior here we do not know it right so we will have to see that anyway.

So what we are interested in is finding out  $\theta$  that gives me the maximum posterior right so if you think about it I do not have to actually compute the posterior to find out  $\theta$  that gives me the maximum posterior why because  $X$  is common I can ignore that I can only down arc max on the numerator I do not have to do the arc max on the denominator right so I do not actually have to compute the posterior.

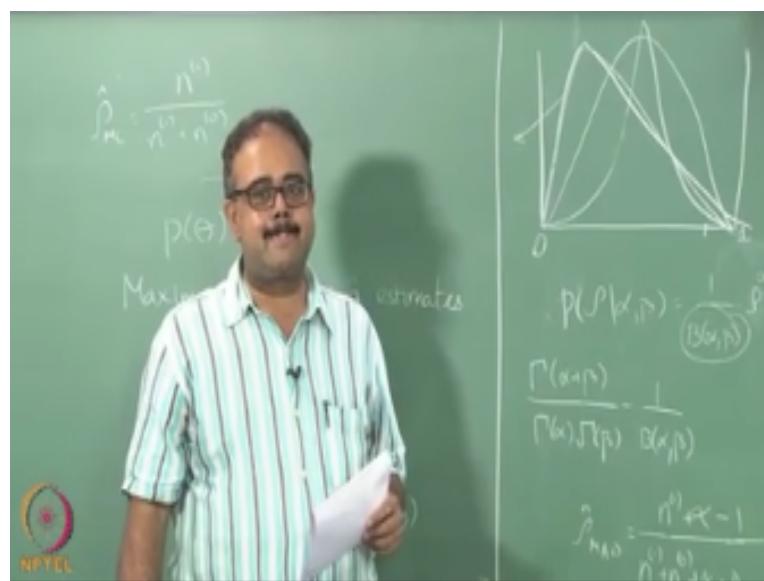
So as long as I have a convenient form of just dealing with the numerator I am happy right so i can just do the max over the numerator and of course I can take the logarithm because this is a nasty term because it has a product in it so I take the logarithm converted the summation and basically do the max of this.

Right so there are a couple of things which I want to point out here so the one is yeah if you have some prior information you can use that right. So if you believe in the honesty of the person you can use point five right but there are other cases where I do not really have a prior information

about how the true solution will be or what the true solution is right but I have some prior over what I want the true solution to be you have cases like that.

We could try to do this when we did ridge regression or when we did lasso right we wanted the parameters to be small the way I achieve that was by putting a quadratic penalty right instead of that what I can do is I can say that hey I have this prior which is very low probability to high values of the parameter right I can make this prior you know is all of you know but  $\beta$  distribution I can have all kinds of weird shapes to the  $\beta$  distribution.

(Refer Slide Time: 07:01)



Right let us see nice prior right the honest prior right so is a really honest prior and that is something like the  $\text{Beta}(1, 1)$  prior sorry wait so the probability of the higher value is  $\beta$  being high is small a low probability of  $\rho$  being high is small probability of  $\rho$  being small is high right so this is one way of thinking about enforcing regularizer right does it make sense so I can use the priors for enforcing my regularizer the second say no do not give me things

That have very high parameter value of interested in smaller parameter values and if you this is just about single parameters if you want to talk about multi dimensional case you can also say that hey I will give you a low probability half having a solution which has more than thirty percent of the parameters nonzero so what will that enforce for me sparsity right that will enforce sparsity so then the probability.

So if I need to have a hence equation suppose I really need to have my  $\rho$  here my row is actually there right but I start off with a prior that looks like this will I will I reach my current estimate for  $\rho$  somebody said it depends so I'm happy depends on the amount of data I have right so it depends on the amount of data I have so if you have an infirmity prior grade the amount of data that you actually need is actually low.

If you have prior is correct right if you put the maximum probability on the right solution the amount of data unit is low but if you put these the prior the maximum weight on the prior on the wrong solutions the amount of data you need is going to go up significant amount of data is mean is going to go up significantly so I said to I made two points about priors right so this remember what are the two points price can be used for regularization okay so wrong priors need more data to corrector and a completely bullheaded prior can never been corrected.

So what is this it is actually the  $\beta$  distribution where I have written the normalize in a slightly different format you are used to seeing the normalizer as okay so what is this call that is the  $\beta$  function okay this whole thing is the  $\beta$  distribution so you actually have three  $\beta$  is here so the  $\beta$  distribution okay and the  $\beta$  parameter lowercase  $\beta$  as a parameter and then you have a  $\beta$  function thing okay.

So the thing to note here is that you are and your  $\beta$  parameters right almost act like as if you have seen heads and tails right. So your  $\alpha$  is just increasing the count of your head s right and the  $\beta$  is increasing the count of your tails the make sense right  $\alpha$  is just increasing the count of heads. So if I had done the actually done the experiment I would have seen  $n_1$  heads right, but I am assuming I in addition.

I saw  $\alpha-1$  heads also right so if I am going to have a prior like this right can you imagine what would be the values of  $\alpha$  and  $\beta$  as I will be less  $\beta$  will be more because  $\alpha$  adds to the heads right so  $\rho$  would be higher if  $\alpha$  is larger row would be higher so if I am going to skew it like this so they should have a  $\beta$  larger than  $\alpha$  you can start reasoning about all of these things just if you understand what is happening so these things are sometimes called pseudo counts.

## IIT Madras Production

Funded by  
Department of Higher Education  
Ministry of Human Resource Development

Government of India

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**NPTEL**  
**NPTEL ONLINE CERTIFICATION COURSE**  
**Introduction to Machine Learning**

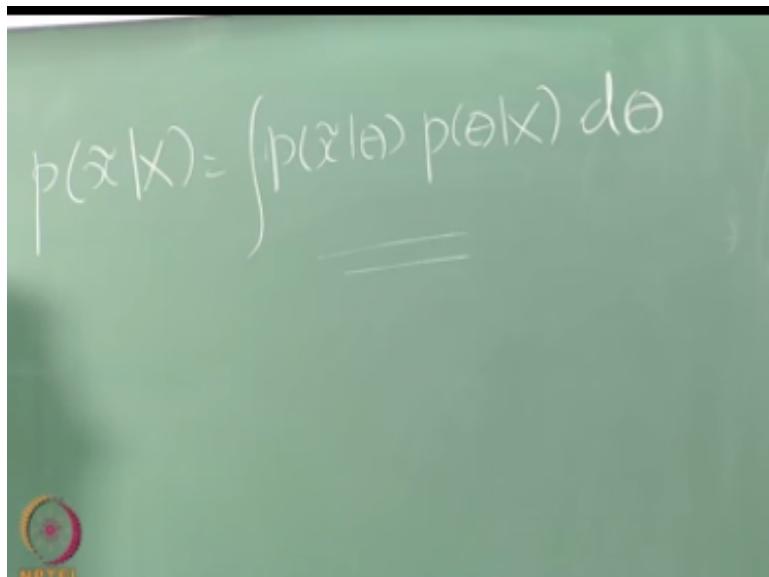
**Lecture 38**

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

**Parameter Estimation III**

So there is one thing that we are doing here see if you remember what was our two stated goals which I have erased of the board, what are the two stated goals that we had? One was to find some parameters that best explain the data, what is the second goal? Exactly right I am doing the best I am actually finding the best parameter like one single setting for the parameter that best explains the data that was given to me that is what we have been doing so far. But in terms of finding the best prediction for a new data point am I doing the right thing so far? is it the right way to do it right, so if you think about it right.

(Refer Slide Time: 01:01)


$$p(x|X) = \int p(x|\theta) p(\theta|X) d\theta$$

So probability of ( $\sim X / X$ ), it is a probability of  $\sim X$  given  $\theta$  times the probability of  $\theta$  given  $X$  summed over all  $\theta$  if  $\theta$  was a I mean first a discrete probability distribution but since we have

been considering Bernoulli and other things it is going to be a integral over all  $\theta$  right I am not talking about the outcome I am talking about the parameterization so the parameterization is a continuous parameter right,  $\theta$  is a continuous parameter.

So I cannot sum over  $\theta$  it is not like I am only considering 01 02 03 I am considering  $\theta$  in the interval 0 to 1, right? So is integral over  $\theta$ , right. So this is this is the actual outcome right but think about what happens in the case of the prior or any one of this case right I am only picking one  $\theta$  it could very well be that there is another  $\theta$  which also has a high probability of being correct.

But since I am picking only one  $\theta$  I am sticking with that, there could be two different  $\theta$  which I could have used them right in fact I should ideally be using all the  $\theta$  because for a certain parameter setting some  $X \sim$  might have a high value right, so even if that probability of that happening is very small and I should still be accounting for that in my prediction, that makes sense why this is a much better predictor than using ML or MAP. But why do not people use this then computationally hard, why?

(Refer Slide Time: 03:21)

$$\begin{aligned}
 p(\alpha | x) &= \int p(x | \theta) p(\theta | x) d\theta \\
 p(x) &= \int_{\Theta} p(x | \theta) p(\theta) d\theta \\
 p(\theta | x_1, x_2, \dots, x_N) &= \frac{\left( \prod_{i=1}^N p(x_i | \alpha, \beta) \right) p(\theta | \alpha, \beta)}{\int_{\Theta} \prod_{i=1}^N p(x_i | \alpha, \beta) p(\theta | \alpha, \beta) d\theta} \\
 &= \frac{\beta^{\alpha} (1-\beta)^{\beta} \cdot \prod_{i=1}^N (\alpha + \beta)^{-1}}{B(\alpha + \beta, N + \alpha + \beta)}
 \end{aligned}$$

So far I was actually trying to avoid computing probability of  $\theta$  given  $X$  right here I did that by assuming everything else was constant right and I just had to do the likelihood right here I said okay I am just doing a point estimate so I can ignore the denominator I can only do the numerator right but when I go here boom, I have to do the full computation right this essentially means I need to know  $p(x)$  right.

And that becomes hard but computing that is actually harder to actually multiplied over all the data points that you have right so it becomes a little tricky right and what is  $P(X)$  by the way, no yeah but what is  $P(X)$ ,  $p(x)$  is a probability of seeing the data right what does the probability of seeing that I do not know that I have only given you the data right I do not I do not know the distribution from which the data was drawn that is exactly what you are trying to do so the  $\theta$  gives you the distribution over which the data was drawn right, so what would be  $P(X)$  right so that is  $P(X)$ , how do you compute that? Good point, so whenever we talk about parameter estimation right so you need to have some parameterize form of a function for you to do the estimation of the parameters right.

So if you remember in the logistic regression it was not Bernoulli it was the logic function that we were trying to estimate the parameters for and I also told you in when we looked at LDA I told you we could make a lot of different assumptions about the parameters in the LDA we made an assumption what did we assume it was a Gaussian right, anything else? Covariance was the same right this is for LDA right.

And I at that point I told you could use mixture distributions as well and you could use whole bunch of other things I also said you could use nonparametric techniques right but I told you it is a misnomer it is a misleading name because nonparametric really means that you just keep adding parameters and things like this so there is a very flexible very powerful modeling paradigm so you could do parameter estimation for nonparametric methods also right.

Where you have to actually figure out how many parameters you need as well so then the distributions you consider become more and more complex now we are looking at very simple forms right but the distributions become more complex and it is the parameter estimation consequently becomes harder right, so infact most of machine learning research nowadays is essentially on parameter estimation for all kinds of different things like non parametric models how do you do the parameter estimation things like that lot of research is going into that. And a lot of powerful models have come out.

Let us go back to our Bernoulli case for a minute right I have Bernoulli and my prior is a  $\beta$  distribution right now I can try to do this so this is this will be what, this  $P(x)$  given  $\theta$  right that is  $P(\theta), p(x)$  given  $\theta$  is  $p(\theta)$  divided by; make that makes sense right this just the is  $p(x)$  right this is  $P(p/X)$  right so  $X$  is your  $C$ ,  $C$  is the set of experiments that we were talking about that right so  $P(p)$  given  $X$  is equal to  $P(x)$  given  $p \times p(p) / p(x)$ .

Next  $p(x)$  is given by integral over the entire row space which is 0 to 1 okay  $p(x)$  given  $p \times p(p) d p$  and it is just the base rule I have it now, right. one thing you notice here what is gone here or nice convenient logarithms are gone right but it does not matter too much why? We are not doing any maximization here we are not have to take the derivative or anything now right and this actually interested in computing this whole functional form again and I am not interested in taking derivatives and trying to maximize.

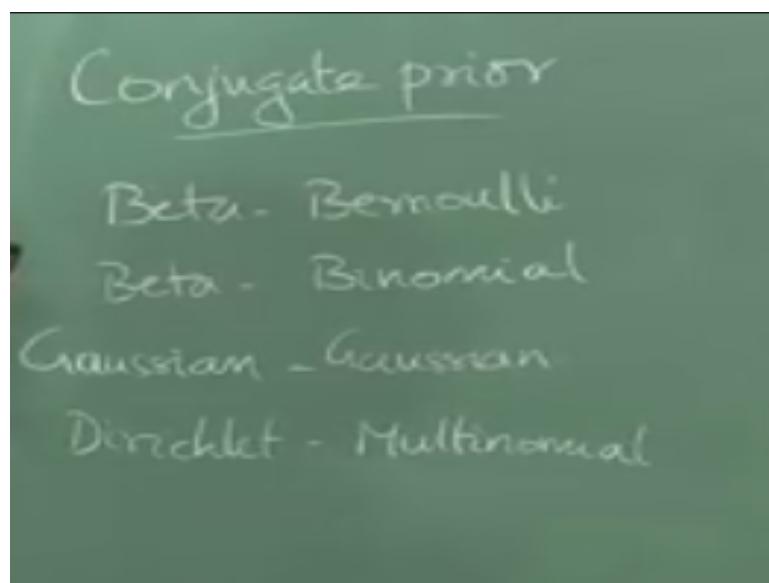
So it is okay if we look like if I do not have logarithms but it just makes the whole thing more of a nasty right, when it turns out this is pretty easy to compute well so I skipped a few steps in between but you can figure that out so I wrote out probability of  $p$  given  $\alpha \beta$  which is essentially this right and this product I can write like this as we did earlier that we have done both of this before so what I have left out here is a normalizing function.

There should be a  $1/\beta$ ,  $\beta$  function of  $\alpha$   $\beta$  right and then I have this integral also and it turns out that this whole thing including that normalizing factor is actually equivalent to the  $\beta$  function of  $n_0 + \alpha$  I mean  $n_1 + \alpha$  and  $N_0 + \beta$  okay. These are  $P$ 's these are  $p$ 's yeah if you remember the  $\beta$  function right it is  $p$  power  $\alpha - 1$ ,  $(1-p)^{\beta-1}$  and the normalizing factor is a  $\beta$  function of  $\alpha$   $\beta$  right, so it is  $p^{n_1 + (\alpha - 1)}$ ,  $(1-p)^{n_0 + \beta - 1}$  and then there is the  $\beta$  function of  $n_1 + \alpha$ ,  $n_0 + \beta$  this actually itself a  $\beta$  distribution right.

So it is exactly the same  $\beta$  distribution so we started off with the  $\beta$  distribution as the prior over the  $p$  right and then we did this computation and the posterior turned out to be a  $\beta$  distribution as well is very convenient right? Such distribution which allowed us to do this are known as conjugate pairs that are conjugate distributions so what are the two distributions are talking about here  $\beta$  and  $\beta$  and Bernoulli right.

So the data distribution was Bernoulli prior distribution was  $\beta$  right if that is the case then the posterior will also be  $\beta$  right people know the difference between Bernoulli and binomial, what is the difference between Bernoulli and binomial? Single trial is Bernoulli repeated trials is binomial right it turns out that  $\beta$  is also conjugate prior for binomial, right.

(Refer Slide Time: 15:12)



Any the famous conjugate pairs that you guys know? So both the data and the prior can because it so remember what we mean by the prior distribution right so the prior distribution is the distribution over the parameters of the data distribution when I say Gaussian-Gaussian that means that okay I am assuming my data is coming from a Gaussian and I am going to assume that the mean of the Gaussian is coming from another Gaussian right.

The probability of the mean is going to be given by another Gaussian so that is what I mean by a Gaussian-Gaussian prime like the like in the  $\beta$  Bernoulli prior I am assuming that the probability of heads is  $p$  and the prior distribution  $p$  is a  $\beta$  distribution, so when I say Gaussian-Gaussian I am assuming that the data is coming from a Gaussian distribution and the mean of the Gaussian is coming from another Gaussian distribution that is what Gaussian – Gaussian. There is also another very famous and so Derichlet so people about multinomial is what is multinomial?

It is a distribution that will describe multiple roles of a dye for example, binomial is when you have two outcomes multinomial is when you have multiple outcomes right, so the single experiment single trial version of multinomial is called not too many people know it and multinomial, binomial is called Bernoulli the single trial of a multinomial is called know the unimaginative name is called the discrete distribution okay.

So but so multiple trials is called the multinomial distribution and the prior the conjugate prior for it is an original a distribution which is nothing but the multi-level extension of the  $\beta$  distribution so it is like the  $\beta$  distribution when it is a multi-dimensional extension of  $\beta$  distribution okay and there are a bunch of others okay so but there are several that are known and so typically what you do is you look at your data right look at the data and figure out what distribution is a good distribution for the data right.

So for example coin tossing experiments we figured out that Bernoulli is good right, so die rolls right we will figure out that multinomial is a good distribution what about text, people typically use multinomial distribution so you can think of having a very large dimensional die on one word written on each side of it right so what is the next word you use roll the die that will tell you or the next word to use right so that is that is do not laugh I mean that is the seriously the model that people use for modeling text you know they use multinomial distribution so they assume that each word is a generated independent of the previous word sometimes when okay fine let another yeah.

So that each word is generated independent of the previous word and so you can model that as a multinomial distribution right, so they have actually have different names for it, it is sometimes called the Unigram model right also it is called the roughly a bag of words model right where the sequence do not matter and each word is generated in differently so many ways of describing the same idea right but at the end of it is nothing but using a multinomial as comical as it sounded that is what it means.

Having this huge die and rolling it every time I want to add a word to the document okay so that is multinomial right so once you have decided what is the distribution that you think is appropriate for modeling the data then you go and decide on what your prior should be right so sometimes the choice of the distribution for modeling the data is driven by not whether there is a conjugate prior is available for the distribution or not right so maybe there is a different distribution that is perfect for modeling data.

But because there is a very convenient conjugate prior for multinomial right people want to use multi no means so like that there are other instances where even though Gaussian is inappropriate for example people want to model discrete value data right the Gaussian cannot do discrete values right but then coming up with the distributions which have allowed discrete values and have nice conjugate priors is hard right so people just go with Gaussian sometimes you end up operate they use Gaussian.

Because it has a nice conjugate price so  $\gamma$  and  $\gamma$  are conjugate priors important, because an easy to do things in iterative fashion because once I run some data through the  $\beta$  Bernoulli pair, okay I am going to end up with a  $\beta$  distribution over the parameters again so if I get more data and I can just happily just go ahead and do it and if I every time I run through it I keep getting a different probability distribution there is no functional form for me to stick these things into and then it becomes very hard for me to do this in any tractable fashion then I cannot come up with some parameter update equations or anything like that so it becomes very hard so, so the conjugacy is very important.

**IIT Madras Production**

Funded by  
Department of Higher Education

Ministry of Human Resource Development  
Government of India

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

# NPTEL

## NPTEL ONLINE CERTIFICATION COURSE

### Introduction to Machine Learning

#### Lecture 39

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

#### Decision Trees – Introduction

So we are going to shift gears and we are going to look at a very, very popular supervised learning algorithm and also provide learning model I should say because there are many different algorithms for estimating this model or based on decision trees right , okay people at the back hear me fine okay. So decision trees have a very I mean very special place in all this machine learning stuff in that they are very widely used right and very poorly understood know in terms of I mean we talked about all this bias-variance tradeoff classification area we can show convergence we can show approximations and whole bunch of other things for all the linear classifiers linear regresses let me know lot about all the linear stuff right.

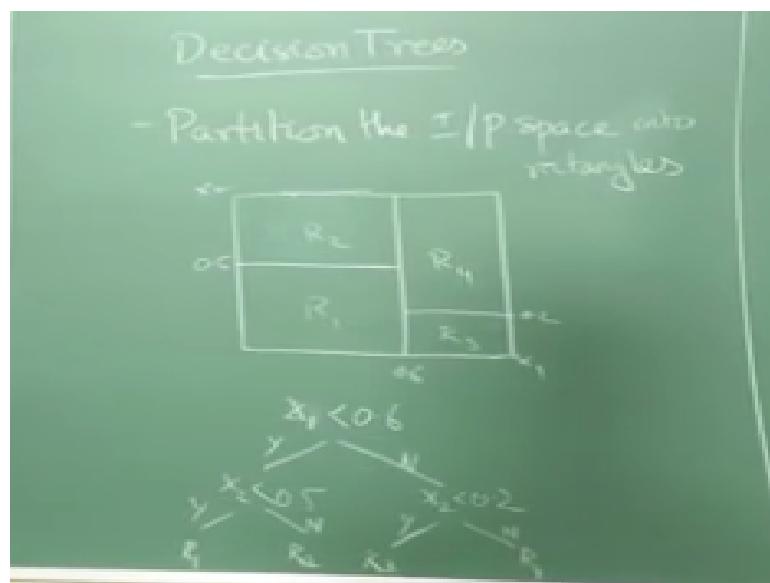
And some of the non linear stuff like SVM's and so on so forth again we have very strong theory we know about convergence and things like that right with decision trees I can tell you what the problem is and I can tell you what is the best known heuristic for solving the problem but I cannot even tell you how good the heuristic is right because there are isolated results under very special conditions people have some results on how good the heuristic is you know that is this best possible tree that you can learn right and how close will it get you to the best possible tree right.

So there are some very isolated results right but there is nothing that really something out there that we understand well right and it is incredible because it is such a simple idea it is a very simple classifier right. So it is more or less along the lines of how humans do their decision-making right, so when you are trying to decide whether something belongs to some category X or some category Y right how do you think you go about doing it right you do not do not create a hyper plane in your head right.

So typically what you end up doing is okay this is red? okay it is not red okay Is it round? oh yeah it looks round and round and okay it is blue maybe it is that right essentially what you are doing is querying some properties of the object right or some properties of the entity at hand right do I want the when do I think he is a studious boy or not right then I can ask all kinds of queries okay let us see show up for all the classes as you sit in the first row all the time is he smiling after quiz one.

So I can ask him in again build this thing so I can ask the series of queries right and then I can essentially am building some kind of a characterization of the object and then I say okay, so this is this person is class one okay this person is class 2 right. So on so forth that is the whole idea behind decision trees right you are essentially trying to if you think about what you are doing.

(Refer Slide Time: 03:23)



You are trying to partition the input space into certain regions okay right the feature one has value X feature two has value Y feature three has value Z and that gives me some region in stage space so what do I mean by that let us take a right. So the first question is a two dimensional data set now let us forget about all the all the relevance to real they real-life and things like this I have two variables X1 and X2 now I am going to ask the question first question there is a new data point is the x1 of this data point greater than 0.6 or lesser than 0.6 okay the first question I asked.

So what am I doing in some sense I am right so I am splitting this into two parts its greater than 0.6 it will be here which is lesser than 0.6 it will be here right next question I can ask is okay suppose  $X_1$  is greater than 0.6 okay then is  $x_2$  greater than 0.2 or not right then what do I do right, so this region is now  $x_1$  less than 0.6  $X_2$  greater than point I mean  $x_1$  greater than 0.6  $X_2$  less than 0.2 likewise this is  $x_2$  less than 0.2 and likewise here I can ask the question is what kind of question can I ask you come on let us just say something random and do not think too much about it you can make any conditions on  $X_1$  and then you did with  $X$  you can do anything right.

So typically they alternate but you can also ask a question of case  $X_1$  given that  $X_1$  is less than 0.6 is it less than 0.3 or not or you could ask is  $X_2$  greater than not give me some root 2 or something but 0.5 okay good right. So as soon as I write that thing that it becomes 0.5 so the regions that we can act arise each do class are in the rectangular but have some but lately on something i Function and lately a slant of the saying I can I come to that little later right.

So I am just trying to mimic the way we try to think of things right so normally what this is all of you are agreeing with me when I said okay I will think about attributes one at a time right so is the price okay is the TV screen of this as the right size then I am going to buy or not right so this is like that so I am just mimicking that process here okay then we will come to other things a little later right.

So I mean it will be really truly amazing if you are true class labels are going to lie like this right what is the problem what do you think is a likelihood that the class labels will actually be this kind of rectangular regions it could be high I mean depends on what process was used to generate the class label since the problem labels are generated by doing this kind of region splitting obviously they will you have to find the right regions but we are making some kind of an assumption right earlier when we made assumptions about linear right we are making some kind of an assumption about what the boundaries would be right.

Likewise we are making assumptions here that there will be rectangles right so if you do not want to make the assumption that there will be rectangles then it is not going to be little harder, so not only are these rectangles right there is something more special about these rectangles I mean they are all recursively generated right it is not like I can I cannot just take some arbitrary

set of rectangles and tile the space right they are recursively generated by first splitting it into two and then splitting each section into two and so on and so forth it turns out right.

So this kind of recursive splitting is what is most tractable to handle right and for most of our decision tree discussion we will stick with this right I will come back and address that issue little later right about having more complex boundaries but almost all decision tree algorithms try all the approximations that we do for decision trees use this kind of recursive splitting of regions right like one inside the other like that how will you describe this region it becomes harder right, see now each of these regions I can describe very easily right.

But if I start doing nested rectangles right it becomes a little tricky to describe the outer region but I can do something like this provided I am willing to accept that right now it is no longer a nested really no longer nested rectangle because I had to actually fragment the outer region that will give you the inner rectangle but the outer rectangle you have to actually exclude it right I have to I have to specify the outer rectangle and then say okay to remove the inner rectangle.

So it becomes a little harder to specific okay, yeah it becomes harder right I wanted to be easy I wanted to if I want to represent this as a tree but then the way you are going it make it harder and harder so the biggest advantage of decision trees is the interpretability, so for example this tree that I have this region segmentation that I have drawn I can represent it as I mean talking to you about trees right where it said tree right the segmentation I have made I can represent it as a tree.

So what I will do is I will first ask the question is  $x_1$  less than 0.6 rights then if we say yes I will go left then last equation is right. So I can very compactly represent this rectangle segmentation as a tree right you can see what is here I am asking the question is  $x_1$  less than 0.6 if it is true I go to the right and then again I ask the question is  $x_2$  less than 0.5 it is true I go to the right and I say that this r1.

Right so I am essentially here right otherwise I am here then go into the other branch which is  $x_1$  is greater than or equal to 0.6 on this side and if it is less than 0.2 mean r3.  $r$  is greater than 0.2 I mean r4 right so I can very compactly describe this segmentation as a tree right so what is nice about the streets is easily understandable that you show somebody okay so you are building your this you go out to become a data scientist or a data analyst or whatever okay.

Some manager who makes like 10 times what you do but who has ever heard of a hyper plane in his life right comes and asks you for a to build a classifier here is the data build a classifier and then you tell him okay, I built this classifier and this new customer you should label him as a buyer then he will ask you why right so at that point you just talk to him about optimal separating hyper planes and show him something okay well then the next day probably the manager is going to be running infinitely more than you right.

So what you should okay so what you should be doing is showing him a decision tree right because that people can understand right people even with an MBA can understand right so you can see what is my recommendation to you guys have to finish your B Tech or anything, do not do it MBA anyway so that is easily interpretable right you say oh that is oh yeah you should you should classify him as a buyer because well this is on this parameter he is so much on that parameter is less blah, blah and then there you go right the biggest advantage of decision trees is the interpretability always easy to explain the decision tree to people.

In fact so much so that at one point when neural networks were at their peak you know you know what is the biggest problem with neural networks the opposite of decision trees incomprehensibility right this is interpretable and neural networks are incomprehensible essentially you say okay it is a black box I do not even know what hyper plane it is learning right if I get if you think optimal separating hyper planes are hard I cannot even visualize what the neural network is learning right.

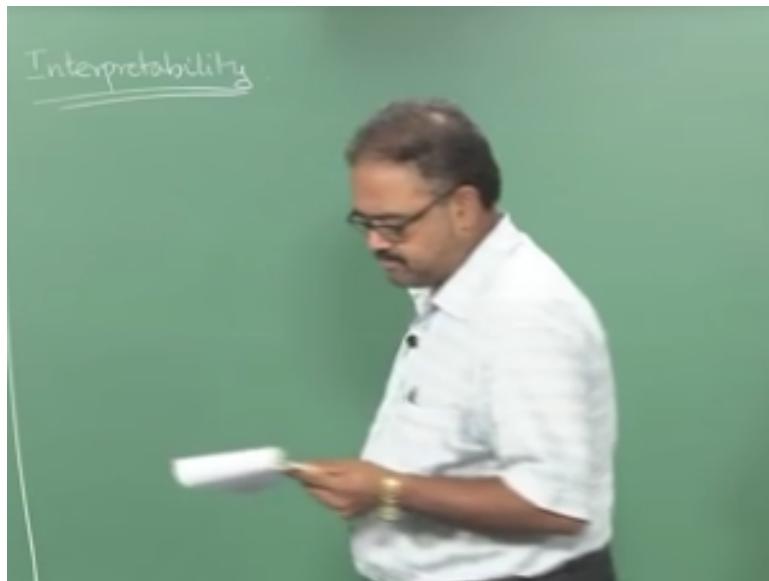
So you see here is a black box you throw in all your data at one end something will come out at the other end you just take it on faith right I so welcome to the Church of neural networks right, so that is essentially how neural networks are working so when the neural networks are at their peak there was this whole line of research where people took a neural network like that was trained on your training data etcetera, etcetera.

And then try to construct a decision tree that will give the same decisions as the neural network will give the same class labels as in your network so that you can actually understand what has what is happening it sounds weird right but remember that now I am no longer using whatever other heuristic I had for constructing the decision tree I am using a decision tree which makes the neural network so if the neural network learned some complex function of the data right I am trying to build a decision tree that mimics the complex function right.

So it is a different decision tree that I would come up with than the one I would have constructed if I had used any of my decision tree learning heuristics on the data from the beginning okay so that is value to doing this right, so people do see that right that is value doing this because your networks do something I cannot understand right but they seem to work now they give me a wonderful answer and I do not know what the answer means so can I use something for which I know what the meaning is and try to understand it in terms of that right.

So that is how useful decision trees are right even if you have a more complex learning mechanism at hand okay sometimes for interpretability sake so you can use decision trees right.

(Refer Slide Time: 15:28)



How expressive are decision trees. If you remember we have the discussion about neural networks I said if you have two layers of weights there is three layers of neurons and you can basically represent any Boolean function right as the branching factor might be very high but then you can represent any Boolean function so neural networks are universal approximators as in that sense is what about decision trees?

So you can it can be an Universal approximated as well right I can just keep dividing and subdividing this space okay just that my tree might become very, very large right as long as there is some kind of guarantee on the function yeah. Now I can define some variables here it is fine yeah in the eye I could do that I mean in fact he was pointing out in the beginning I could have as well draw another line here yeah okay.

Everywhere that the discussion was when we were discussing about this line versus that line yeah sure I can keep doing this if that is your question right now this becomes a much more complex tree right, so now I have splitting here once more right I will have another branch here will have another branch here and another branch somewhere there right so keeps becoming more and more complex I can but the point is conceptually you can represent anything right.

So it is powerful in the sense that it is a universal approximator right and it is just that the number of parameters can grow unbounded okay and that is another thing nice thing about it is nonparametric ever we talked about what nonparametric means right, so decision trees are actually nonparametric it can just keep growing right we can keep adding parameters as you go along.

### **IIT Madras production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**Introduction to Machine Learning**

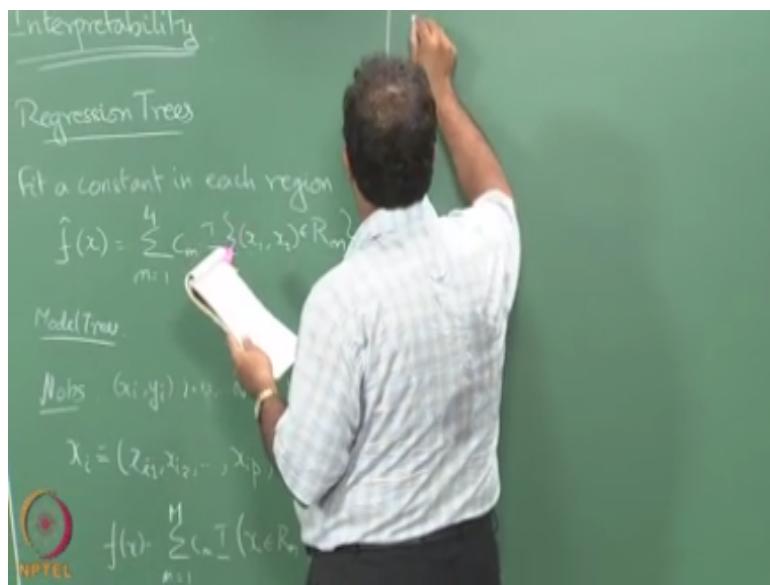
**Lecture 40**

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

**Regression Trees**

So as with the linear methods we will first start by looking at regression right, so we will see how to use decision trees for doing regression. So far I have just told you how to do the partitioning right.

(Refer Slide Time: 00:31)



Let us look at regression trees, so I split the region into four right, so I will first see if the data point that comes to me right, whether it lies in region 1 or region 2 or region 3 or region 4 and for each of those regions I am going to have a some constant that I will output right, so if you think about it the function that I will output from here right, so we will have one value in this region right one value in this region right, another value in this region another while in this region so it will be like a piece wise constant thing right.

So people understand that right, you are not going to test my 3D drawing skills right, so you can see that there will be one output for any point in this region one output for any point in this region, one output for any point in this region, one output for any point in this sense and in some sense it is similar to KNNs, because you are assuming that there is a piece wise constant assumption about the function that we are trying to model, right.

By the second parametric or nonparametric, parametric okay, what are the parameters, so is it parametric resonance is depends on N becomes larger what happens, so we can apply one of those ideas here instead of fitting a piecewise constant per region you can fit a linear function on that region right, I have done the splitting right, so I am going to have some training data right, so some of the training data points will be here, some will fall here some will fall here and some will fall here I can take all the points in R1 and fit a plane to that.

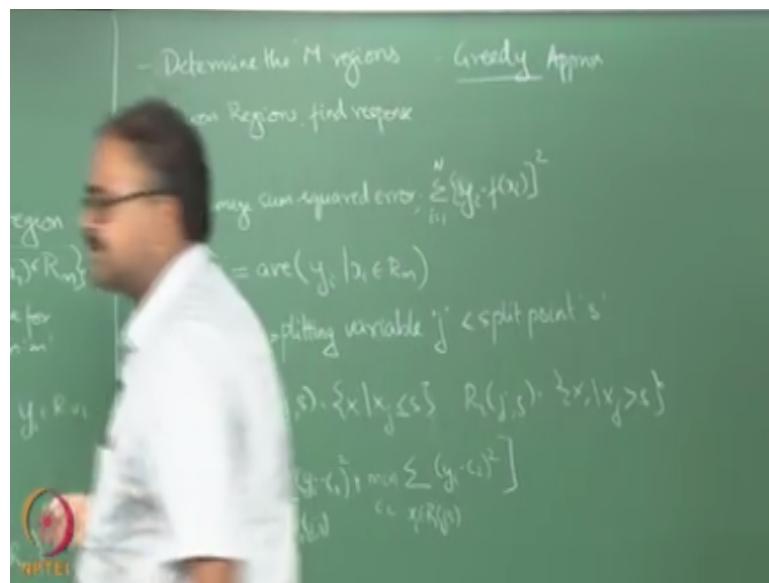
And I can take all the points in R2 and fit a plane to that likewise for R3 and R 4will that be better or worse than fitting a constant? Better. no actually it does not depend actually it is always better right, it is always better in the worst case you will fit a constant I mean if constant is going to be really better you will fit a constant because you are minimizing squared error and anyway end up doing that right.

So what is the problem with that little bit more work here that do more work and there is variance the significant variance we will come to the variance bit yeah, but the significant variance but such things are called model tree sometimes, model trees in fact you can do more complex stuff itself it does not have to be linear, a linear is easy right, I can do any kind of regression I want on this right, I can fit I can use a neural network if I want and to learn a curve only on the data points that lie in R1 right.

Not a good idea usually because I have already divided my entire training data into at least 1/4 if not smaller right, I mean some of the regions could have much smaller data some could have more right, but still I am cutting down on my training set that is available so it is going to be harder to train okay, so I am going to erase this stuff and try and generalize this or let me do it again. So I have n observations as you all know right, so this is our usual setting so where the input comes from Rp and the output comes from okay.

So the function that we are trying to learn is  $C_m$  and  $R_m$  right so sorry, the set of parameters that we have to estimate our  $C_m$  and  $R_m$  correct, so we need to know where are we splitting, how are we splitting the region and having found the regions right, what is the constant that we will fit within that region okay.

(Refer Slide Time: 07:07)



So there are the two questions determine the  $M$  regions and given the regions find the response right, but one thing to note is unlike KNN or anything the  $M$  is not given to you right, the  $M$  is something that you discover from the data and that is why I said it is nonparametric right, you can actually have more regions if the data requires it you can have lesser regions  $M$  is not given to a priori right, but sometimes as a regularizer you can decide to fix  $M$  as well you can say I do not want a tree that is more than four levels  $d$  as a complex  $d$  measure but again that is derived from the problem definition the model itself is not parametric.

So let us look at the second question first, because it is easier we actually have a proper answer to the second question right. Right, so this is essentially what we are trying to minimize right, and so we can try to do this region wise right, because the output that I produce for one region does not depend on the output I produce for another region so I can do this minimization region wise right, so I can pick yeah, every point would have its own box kind of that. Okay, yes. Could likely yeah, so we will come to that will again address this question later, right.

Yes, so assuming that there is some amount of so you will not okay, let us step back into this you are going to get training data at best what you would do is you would have regions set within a region only 1.6 right, so that is not saying that I am going to fit it tightly around that point right, so all of our for that might be only one point but still there will be some kind of regional segmentation that is happening on the input data right.

Second I have I am going to introduce some kind of regularization that prevents me from doing that okay, so that will not be recap that is exactly what he was asking so you could end up with that that is what I am saying and we have to find some way of regularizing it so that you do not do that right. Right so if you going to minimize this region wise but anyway right now I am talking about given a region right so that is easy so we will not be over fitting things so given the region right.

So I am going to find out what should be the output of the  $m$ th region right,  $C_m$  is the output of the  $m$ th region right, that is what we assumed, what should it be give me a simpler I am fitting a constant I am not fitting a straight line here, average of all the points okay, like average of all the points which lie in the region and take the  $y_i$  is corresponding to the points lying in the region and take the average okay, that is the best response that is that is easy okay that is done.

What is the harder part, finding the regions right, in fact it can be shown that finding the best possible  $R_m$  set of  $R_M$  right is actually NP-complete right, and is NP-complete now in the again NP exactly NP-complete right, so you can show that finding the best possible  $R_m$  is very, very hard right, so we have to come up with some kind of approximation so essentially we use a greedy approximation. No, they just tell you what XII I told you I told you what the training data is right, yeah this is all the training data is you get  $x_i, y_i$  your job is to find the regions and find the region I find the response.

Yeah, yeah find the best region you can given it a region you can tell me what the performance is right, but then finding what the best such segmentation is actually hard you have to search through the combinatorial really many segmentations. So the way we do it is following right, yeah, okay, now for a given  $M$  so I want to find the smallest  $M$  such that I get that performance smallest region, smallest region yeah, the smallest region said for which I get the performance that is really ideally what I am looking for right, smallest  $M$  sorry, given an  $M$  finding the region yeah, in general that is also hard but I want to find it for the smallest  $M$  as well.

Ideally you want to find it for the smallest, ideally like there is some data if you are right, like you would have to either specify the M and then you find the best or you see that when the best and find the smallest M for which you can. Ideally it should be find the best and then find the smallest M for which you can do the best right, but we end up doing compromise on that as well so what will what we will do is you just making me go do this all out of order by asking leading questions.

But what we are going to end up doing is we are going to say okay, here is a greedy algorithm right, greedy algorithm find the best that the greedy algorithm can do okay. Now find a smaller tree okay, that will achieve close to the greedy algorithms best performance again I will have to make a compromiser I cannot say that give me the smallest tree that will give me the same performance as the greedy algorithm.

Because if that is exists one in fact greedy algorithm would have found it right, along the way as it was growing right, and therefore we have to say that okay give me a smaller tree right, that is close in performance to what I get with the greedy algorithm right. See remember we already made an approximation by assuming we are doing recursive partitioning right, so you cannot get the best possible performance okay, that we have given up right by choosing a tree representation right.

So this is a lot of approximation that is why I said right, I mean there is no good understanding of how decision trees eventually work if you ask me two specific questions like okay, how good an approximation will this greedy algorithm converge to right, suppose my performance the best possible performance on this the Bayesian optimal error on this dataset is say 93% performance I can see 7% is optimal error okay, how close to optimal error will a decision tree algorithm get to no answer right.

While you can answer some of those questions for things like logistic regression and consider some splitting variable j so what I mean by splitting variable the splitting variable is the question that I asked here, okay this variable here in the question so  $x_1$  right or  $x_2$ , so in this case of splitting variable is  $x_1$  here the splitting variable is  $x_2$  okay, and what is the split point it is the number on the other side right so 0.6, so in this case the splitting variable was  $x_1$  and the split point was 0.6 okay.

Consider some splitting variable  $j$  and a split point  $s$ , okay so I am splitting my input data into two parts one where the  $j$ th variable is less than  $s$ , less than or equal to  $s$  the second part where the  $j$ th variable is greater than  $s$  okay, let us still to get to two parts so what we really want our  $j$  and  $s$  such that okay, I am seeking  $j$  and  $s$  is that if I fit the best value for the points that lie in  $R_1$  and if I fit the best value for the points that lie in  $R_2$  that is what the inner minimum minimization is right the sum of this is minimized, so sum of the squared errors over the two regions is minimized right ,so the  $j$  and  $s$  actually influence which data point goes to  $R_1$  which data point goes to  $R_2$  right.

So once I decide which points go to  $R_1$  and  $R_2$  I have a fixed optimization problem that I solve right, which one we already solved there. Yeah, I am just talking with the full data set I am at the root right then we can worry about the recursive splitting part right. So make sense for people so far yes, I want to find  $j$  and  $s$  such that this happens right, how do you solve this minimization problem.

So we can do this. No, this is not classification right, this is actually a regression I am solving right, so all these data points in our one I am going to output one value for all the data points in  $R_2$  I am going to output another value so you can think of saying there came grouping all  $R_1$  into one I am going to output one value and grouping  $R_2$  into one I am going to output one value find the right grouping such that I can output a value such that overall error is minimized okay.

So the first thing know it can be slightly better than that right, or worse I do not know I will tell you when depends a little better or not okay, the first thing you have to note here is I am going to do this for each and every  $j$  that I have okay, find the  $s$  and then I am going to pick the best  $j$  right I'm going to do this in turn for  $j=1, j=2, j=3$  from 1 to  $P$  I am going to do this, and then find the best  $s$  and that will give me a value for the objective function, right.

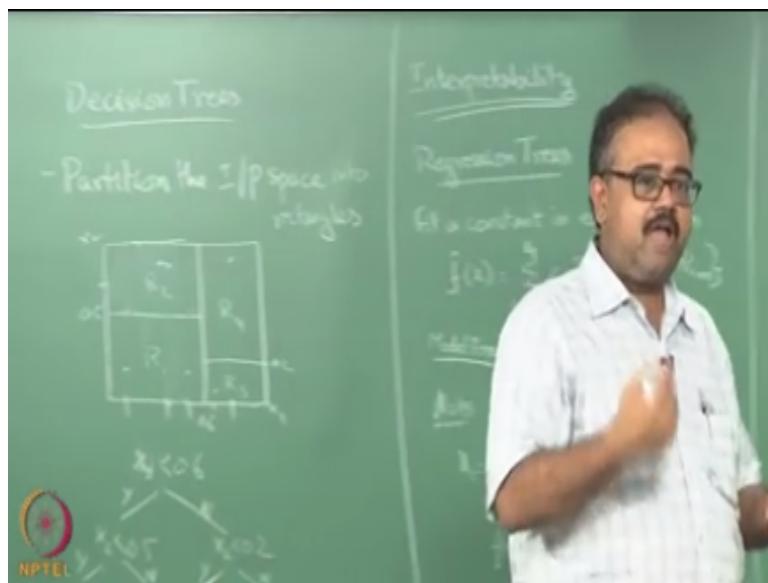
And I can use this to compare which  $j$  is better that I do not have to do this jointly okay, so that is the first thing you have to notice okay, so given a  $j$  right how will you find the right  $s$ , so once you have fixed a  $j$  you can think of it as just a just a line right, I have to find that  $s$  at some point so that I can split everything to one side to  $R_1$  everything to other side  $R_2$ , exactly so what are the steps I should choose for  $s$ , so he was talking about recursive doubling that is one way of doing it if you have no other clue right.

And then you have to come back and then you have to search through thee so people know about recursive doubling I start off by looking at 2 then 4 then 8 and I keep doing that at some point some sign will change right so I mean I will be fitting it one way then I will actually my error will start increasing again, so I stop and then now I will have a window of some power of 2 right, so I have to look between 8 and 16 and then I will do a search through that that is one way of doing it.

But there is a slightly better way we can do it any guesses something better than that imagine you are trying to do this not imagine so remember that you are trying to do this from data, I give you a training data set. Exactly, so order the training data along ascending order in that coordinate right, and then just keep hopping on that. Suppose I have data here, here, here and here somewhere there right.

So now if you think about  $x_2$ , let us say  $x_2$  is my, or  $x_1$  is my splitting criterion so there are only five different values of  $x_1$  that actually occur in my training data right, so that is  $x_1$  equal to this,  $x_1$  equal to this,  $x_1$  equal to this, this and this right.

(Refer Slide Time: 24:16)



It does not matter if I consider any other values for  $x_1$  because that is one of these five values will give me the same split, right suppose I consider this a splitting point does not matter I could have as well consider that as my splitting point you see that right. So I do not have to consider it

have to go smoothly along  $x_1$  I can just use any one of the data points that has come to me already right, so that is the easy way of searching for a splitting point in  $x_1$  essentially what we will do.

Right, we just start it one of the reasons I already used either less than or equal to here it is not easy, so how much work do we have to do for finding one splitting point  $n \times n$  log $n$  or because the sorting part is it okay, you do not have to sort here yeah, you do not want to sort here that is what yeah okay, you can get away you can just go through whatever already good yeah, it does not have to sort here. No, if you sort I mean the computation becomes a lot easier but for computing the complexity you do not have to sort, right you can leave it as it is you can it is  $NP \cdot N$  for each feature and you have  $P$  features right, so the amount of work you have to do is  $NP$ .

But if you sort then life is a little easier but you do not have to when you are writing the code you will know what I mean, but yeah an  $NP$  is the amount of work you have to do for one feature one level right, great so now what we do I have found the optimal  $j$  and optimal  $s$  have a the optimal and all our assumptions right, because I am actually doing an exhaustive search over  $j$  and  $s$  right, I am not doing any approximation here I am doing an exhaustive search so given the assumption that we are going to do something greedy and we are going to split on one variable at a time so we are finding the best possible variables great.

Now what do I do, I actually create the two sets  $R_1$  and  $R_2$  at having found the best possible splitting point and the splitting variable and the splitting value I find the two regions  $R_1$  and  $R_2$  and then I go into  $R_1$  I do the whole thing again, assuming that  $R_1$  is my entire data set likewise I go into  $R_2$  and do the whole thing again assuming  $R_2$  is my entire data set right. Thus, it makes sense to consider the  $j$  again the  $j$  that you split on say some  $j^*$  right does it make sense to consider  $j^*$  again yes, what does not make sense is to consider  $j^*$  along with the same  $s$ .

In fact does not make sense to consider  $j^*$  along with any  $s$  greater or lesser depending on which side you are right, so you can progressively you keep pruning your search based on, it but you do not have to worry about it because it is automatically taken care of because you are only looking at the values that are present in your data point, I am not written those things down right, you want me to write down the whole process all of you remember it right,  $j^*$  is the one that gives me the minimum in this that I am considering each feature in turn right.

So  $j^*$  is the feature that I finally choose to split on and  $s^*$  is the value that I finally choose to split that  $j^*$  on okay. there is another question somewhere okay, good okay great, so how far do we go so this is a question that we all had in our mind from the beginning right, if I do not put any restrictions on it I will keep going until I have one data point per region right, great yeah, so that is a really other people actually notice it could very well be that the number of features ends right.

How can you, yeah  $j$  can be repeated, but if  $j$  cannot be repeated can end up with something like this there is only one data point per leaf per region, where that is no more than one data point per region. Now you could do that but answer my question, we are allowing features to be repeated right, can you still end up with a point where you cannot grow the tree anymore but you have more than one data point per region.

If you keep getting your betting point as your border if your region you can traverse any further. Not, they are the same data point we can repeat it I never said your exercise have to be unique right, now it is not so it sounds like a trivial thing but no it is actually important I mean you should think about it right, so this is very important in cases where the  $y_i$ 's are different the  $x_i$ 's are same and the  $y_i$ 's are different there is no way you will get 100% correct maybe if you are assuming that it is a deterministic process truly underlying process is deterministic and it is corrupted by noises, but what if it is a stochastic process truly a stochastic process is generating all of these things for you right. Yeah, sure you can you should there is no question of you allowing it is life happens to you.

### IIT Madras Production

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

## Introduction to Machine Learning

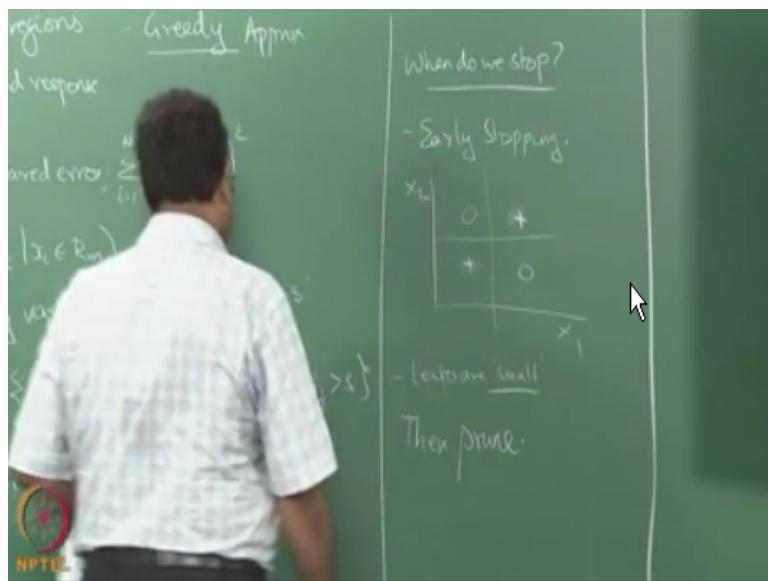
### Lecture 41

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

#### Stopping Criteria and pruning

Right anyway so the question is when do we stop.

(Refer Slide Time: 00:27)



So that is one technique called early stopping okay, where you say that. Hey, I am considering all of these regions right all of the split points but the amount of improvement I get in my error is very small and therefore I stop. I come to some point so I let us say I have like several regions here now so I consider  $R_1$  okay where I consider all possible ways where I can pick an  $X_1$  all possible ways where I can pick gets to and try to split and the error does not change by much.

I can stop I do not have to keep going into smaller and smaller regions even if there are many data points here I can stop right is it a good idea so let us go back to XOR right so this is not a

classification problem we are talking about regression so the excess have an output of + 5 this 0 is output of -5 right and you are trying to fit this. I can try to look at splitting it on  $x_1$  so I have  $x_1$  let me speak I have  $x_1$  and  $x_2$  right so if I were to try to split on  $x_1$  right what will I do right.

And the only thing I can do is this right so if i split anywhere else I will be just keeping a all the data in one side right and all the other mean other one will be empty right this the only meaningful breaking point right and what will be the prediction error here whatever is the half of it i will be predicting the average of this right so it is 0 and 5 then I will be fighting 2.5 so it is essentially  $2.5^2 \times 2$  right.

What would have been the prediction error if I had kept the entire region as one. Same thing the average--average sum square will be the same right so if I split on  $x_1$  I am not getting any improvement so let me not split on  $X_1$  okay what about  $X_2$  same we split on  $X_2$  I will not get any improvement let me not split to  $X_2$  that essentially means that I will just give you the average output for all the four data points.

But if I split on  $X$  well and then split on  $X_2$  okay now I can do really well right so early stopping is usually a bad idea because we will miss out such these kinds of interaction effects if we stop to early. Good point yeah! So in this case I mean one of them is a trivial split right this case is easy but in general yeah so you will have to take a call yeah! Case like your institutions day careful of them at one time but food is easy see yeah.

So just think about think about this optimization problem right if I m going to split on two at the same time this optimization problem becomes harder that instead of minimizing over  $J$  and  $S$  I allowed to minimize over  $j_1, j_2, S_1, S_2$  it becomes harder and harder sure you could think of other ways of optimizing it right so but the most common way of doing it this main reason people do not do two variables at a time is the interaction effect I made.

So if I start looking at two variable Saturday then I can start thinking a whole bunch of other things so why do I look at  $j$  greater and alone right I could think of other combinations hey  $J_1/x_{j1}/x_{j2}$  right  $X J X K$  I mean so then it just starts exploding so they said okay fine we will just do it this way and make sure that we grow a large tree very large tree so that we actually capture the interaction effects.

In fact you stop when the leaves are small I think of a tree right so the leaf of a tree is a region right so the region small I put it in quotes is not the extent of the region it is the number of data points in the region right so--so small would be like two or three or five or something of some really small number depending on how large your data set was you keep growing your tree. So if you view some of the standard tools right.

There will be an inbuilt parameter right which says how small the leaf should be right and you might have to go and pretty with it if you are going to use a decision tree function from either VICAR or MATLAB or something right they have an inbuilt parameter that says how small is the leaf and they stop right so for VICAR it is 2 now you might want to change it to five or something I am not sure what is the limit in MATLAB but you might want to change that.

So that is that is a parameter that you have to fix right and it matters, it actually matters. like he pointed out so if you if you set it too small then you might miss things like this right and then what you do? You build a very, very big tree right this is what I was telling you. You try to use your greedy algorithm and get the best possible tree that you can write so a tree with very very small leaves this kind of the best tree that you can build right right.

Once you get there now you are going to ask the question okay what is the smaller tree that I can get that performs almost as well as the big tree that I have right so there are two ways of doing it the first one she called reduced error pruning case is rather simple so I each leaf then the smallest the largest leaf is smaller than the threshold effect okay so every leaf should be less than that size so basically I mean I can stop each branch independently.

So whenever a leaf reaches size 2 I do not split it anymore so I keep doing but other branches can continue growing so the tree does not have to be of uniform hydrate at some part some sub tree might be shorten some sub tree might be longer right if you remember that picture here so I kept drawing lines in only one region so that means that path alone would have been a much deeper sub tree and others would have been much shallower that is fine.

so reduced pruning is something very simple and so I have a training day training set I built the tree fully on the training set and then I have a validation set may we talked about validation set long time back I have a validation set now what I do is I start greedily or not greedily it is very

safely pruning away my internal nodes right so what I can do when I erase the only tree I had onboard here right.

So what I do is I have this prediction that I am making right I will replace right an internal node with a leaf it does sorry exactly I am just joining the region is together now I see the performance of this with respect to the validation set is I had the original performance on the whole tree right now I look at the performance with respect to the validation set right it could go down right it could go up depending on how the validation set is right.

When it because the tree was constructed only on the training set when you do the pruning the error might actually go up I mean the error might go down sorry right for on the validation set if the error improves our state is the same I will keep this right but if the error mug becomes much worse right I will put it back try otherwise so I as I use the Yi's for making a prediction right and then I keep doing this in turn.

Yeah that could cost could cause more variation agree provided mean usually when you have a large enough validation set so you can actually trust it right and then you try this again right and then if does not work keep going yeah so it is like to have this region but in stuff that I just treat this as one reason now once I you collapse the region I again do average on this whole region and you start at the output right.

See once I collapse this question is each one of this could have been outputting a different value right what will you do with the combined node right so I will take all the data points in the combined node take the average output and we use that as the new output for this right I could take the average of these two but why is that not a good idea the number of data points could be different right so it is it not be truly the average of the outputs right.

So if there are having the same number of data points then I can take the average of these outputs and use it otherwise they should okay so I keep doing this suppose I was able to prune right and now I have pruned this and I have ruined this as well then I can go back and try to prune that also right I can replace this whole thing with this and see how the performance is on the validation set right.

No reduced pruning is only on one thing you see the problem we do cross validation is I will end up with five different trees after the pruning now the question is how do I combine the 5trees

right yeah so exactly so see that is this a very same pruning works it is only one validation say does not use cross-validation right so in that for that reason it is not that popular anymore I am just introducing reduce air pruning because is easy way to think about pruning right.

But like issue is pointing out first of all the variance will be very high depending on what you pick for the validation set right you will end up with a very different tree right so just like when already decision suffer from very high variance and the reduced pruning will actually make the variance worse but this is conceptually easy way of thinking about pruning and if I introduce a more complex pruning method right.

Then a little harder right yeah sorry as long as you are improving sure I will come to that I have a whole class planned on all those model selection methods right since he knew about cross validation he asked me the question I answered but I will come back to that right a whole --whole lecture planned on the model selection okay so cross validation is something guys should never forget.

Once you learn the other kind of pruning which we are all familiar with is called cost complex tree pruning right where you have your error function right and you also have your share in the name also have a cost for the complexity okay like you had here  $\beta^2$  in your ridge regression and things like that right and norm  $\beta$ . so you already know about this kind of cost complexity measures right so we looked at that in ridge regression we looked at that in lasso and things like that.

And here what we essentially do is we grow the full tree right and then what we do is for every possible non terminal node that you can collapse right you collapse that non terminal node so it should mean that the entire sub tree underneath it you consider as a single region and replace it with the average prediction for the single region like that you can collapse each of them on terminals and create many many different trees right.

So each of this is a sub tree of the original tree right so what do you do is once you created such a collapse tree you look at the average prediction error of the tree it essentially look at the prediction error for each data point divided by the number of data points you get the average prediction error and add a complexity term right the prediction error plus some size of the tree right

So what is the complexity that we are really if so  $T$  is at three and  $\alpha$  as a parameter will come to that so what is the complexity measure you think is good for a tree number of leaves right number of leaves number of regions you are split into so that is a measure that we use so when I say size of a tree it is the number of regions that the tree splits the input space into so  $\alpha$  is a parameter that controls how small a tree I want.

Large  $\alpha$  means small trees small  $\alpha$  means large trees so now I essentially find my  $T$  okay which is a sub tree of the original tree right so it is not any arbitrary  $T$  tree okay I have original  $T$  tree that I have grew that I grew with this procedure right with this procedure I grow a tree and then I stop when the leaves are small and then what I do is I try and collapse each of the internal nodes of the tree and you can do this in a slightly better fashion right.

You can try to collapse from the lowest level on up and then stop at some point and things like that but you should remember that it could very well be that may be collapsing one sub tree alone might not give you much of an improvement right but if I collapse everything above it right it might give me an improvement why so maybe collapsing this alone does not give me an improvement but collapsing here when might give me an improvement.

No see the point is the error reduction might be small right but then I might not have gotten rid of enough nodes and if I get rid of this whole thing that I get rid of a lot of regions the complexity of my tree comes down significantly so even if I am making a slightly higher prediction error I might be willing to accept that because I have reduced the size by such a significant amount right so that is one of the reasons you consider all possible things.

May be pruning lower down might not simplify the tree enough for you to accept error reduction but if you go higher up the tree you might actually get the same error reduction right I mean whatever by reduction error worsening right but you must have you met a reduced entry by a much larger amount therefore you are willing to accept that right so it is actually no a great idea to just go bottom up all right so --so the small things to remember right.

So that's something that you pick since the magic word has been introduced so you pick  $\alpha$  who ask the  $\alpha$  question okay since the magic word has been introduced you pick  $\alpha$  by cross-validation I will tell you what cross-validation everybody okay so all of you understand what validation is

right so cross-validation essentially is kind of a multiple rounds of validation and instead of just using a single validation set.

You in fact try to use all parts of your data as validation in a very systematic fashion okay we will talk about this more detail later but just to give you a rough idea and so this is clear so what we are doing here yeah no, look at all possible collapsing right so basically what i mean by collapsing remove an internal node the entire sub tree structure underneath it whatever regions it was covering you consider that as a single region and you replace it with that.

So I can choose any internal node to collapse full sub tree and yeah concept is an expensive process so i said if it is expensive you can come up with other mechanisms of ordering it right but the best way to do it sorry nope nope decision trees there is nothing that is optimal I mean right I mean everything is hard right so any questions on any other questions on cost complextive pruning.

Likely yes but we do not know until you actually fit it you wouldn't know for example in the XOR case what we call it over fitting or not so you would not know right until until you fit the data you do not know whether you are or fitting or not so you have to grow the whole tree and if you are over fitting then when you prune you will actually end up removing it I mean the error will obviously on the training data the error will obviously be lower when you over fit right.

And that is why you need the complexity criteria right so when a prune if I am do not lose too much in terms of accuracy then I am happy to so any other question so so essentially what you do with the  $\alpha$  as you pick a good choice of  $\alpha$  right and then try to do the pruning on five different validation sets then pick another choice of  $\alpha$  right on the same five validation sets you pick another different  $\alpha$  right on the same five validation sets.

And then pick an  $\alpha$  that gives you the best it depends on how you are normalizing the prediction error and as well as what is the expected size of the tree you are going to see right if the prediction error lies between 0 and 1and the tree sizes are order of, order of 10,000 right you would really want your  $\alpha$  range to be small right where the tree is also of the order of say 5 or 10 nodes then the  $\alpha$  is could be larger.

### IIT Madras Production

Funded by

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**Introduction to Machine Learning**

**Lecture 42**

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

**Decision Trees for Classification-  
Loss Functions**

So look at the probability of a data point in M region, belonging to class K which is  $P_{MK}$  right not talking politics but so you estimate that by there is no counting the number of data points of class k and region m and dividing it by the total number of data points is fairly safe further this is how I do the prediction right so what about how do I grow a tree to do classification it is exactly the same as this except that I do not use square error right can I use square error why not exactly.

(Refer Slide Time: 0:33)

The image shows handwritten mathematical notes on a green background. At the top, the text "Reduced Error Pruning" is written. Below it, "Cost Complexity Pruning" is written. A formula is shown:  $\hat{C}(T) = \text{Prediction error} + \alpha |T|$ . Underneath, the word "Classification" is written. A formula for classification probability is given:  $\hat{P}_{nk} = \frac{1}{|R_n|} \sum_{x \in R_n} I\{y_i=k\}$ .

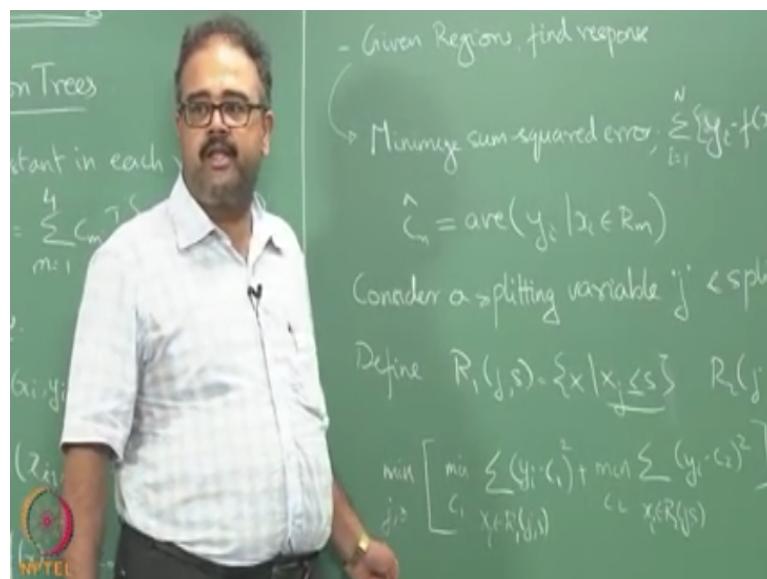
So it depends on how I encode it right I mean earlier in the linear regression you are kind of faking it by encoding it as indicator variables or whatever right so here I am going to have actual outputs I can still look at the distance of the prediction vector to the indicator variable vector

right and try to do that right I can still do that but there are better ways of doing it right so the first thing I can use is the miss classification error.

So denote by  $K(m)$  the class label that I am going to assign to the entire region  $M$  just like we did the  $\hat{C}(m)$  as the response that I am going to assign for the entire region  $M$  so  $k$  of  $M$  is the rest for the class label I am going to assign for the entire region  $M$  okay that just say arg max of this okay. so now the Miss classification error is I am going to count all the data points in  $R_M$  which do not have  $K(m)$  as their label right that is a Miss classification right is all those data points the label I will be outputting  $K(m)$  right for all the data points in  $R_M$ .

I will be outputting  $K(m)$  has the label so all the data points in  $R_M$  which do not really have  $k$  of  $M$  has their label or misclassified right and divided by the total number of data points that gives me the average miss classification error is there some way to simplify this  $1 - P_{m \in K(m)}$  okay so because the fraction of data points that will be correctly classified or  $P_{m \in K(m)}$  right.

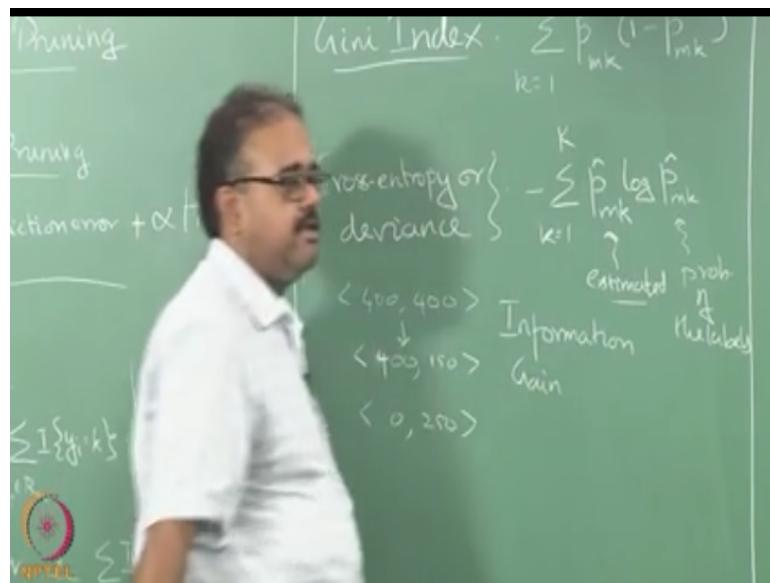
(Refer Slide Time: 04:35)



Because the two label is  $K(m)$  I am outputting  $K(m)$  there will be correctly classified right so the fraction that you be misclassified is  $1 - P_{m \in K(m)}$  okay so that is a Miss classification error so how do I use this essentially I plug it in here right find the split point and splitting variables such that the Miss classification error in each of the regions is minimized at the sum of the Miss classification error in each of the regions many ways.

So remember that this has a very specific solution for minimizing so you do not really have to do this minimization is a fixed process right as soon as you find the region you just take the most abundant class in that region and set that as the class label for the entire region okay right is it clear I will do the how you used to miss classification error rate.

(Refer Slide Time: 06:18)



So the next thing we would like to look at this so one of the downsides of not being able to say anything very theoretically formal about decision trees is that it leads to dogmas so there are two camps of people who are very sure that this is the right way to do decision trees right they, they just keep fighting each other right and there are two very, very popular measures for doing classification using decision trees okay.

So the first one is called the Gini index ok so the Gini index was actually originally proposed by economists to look at disparity of wealth right so let us look at the wealth distribution in a population okay so are there more rich people than poor people or their lot more poor people than rich people I mean how does a disparity of the distribution of wealth okay that is essentially introduced that so in that in some sense you can roughly see that right.

So are there more class one data points than class two data points or anything else suppose I have K classes okay in this particular region are there more lot more class one data points than 22 k if I am able to split my regions like that then I am doing something good right because I can output

the class level as one and I will have less error correct so if I am able to split region says that the class distribution is actually skewed within that region.

Then I am doing something good and if the class distribution is uniform within that region then I am doing something bad that because that is not a good region because whatever class table I output I am going to have a lot of error but if the class distribution is skewed in favor of one class over the other then I can output that class in fact the ideal leaf would be so skewed.

There is only one class present click so the skewness measure is what I have to look for and the more skewed the data is the better so the Gini index is actually more popularly given by this form so I do this for each region so this is for a single region I do this for all regions so the other popular measure is cross entropy or deviance but it is more popularly known by the name I will give it to you in a minute right.

And this is given by this expression this looks familiar to you guys Shannon's entropy kind of thing races cross entropy where is the cross part you have  $\hat{P}_{mk}$  and  $P_{mk}$ . there so why do they call it cross entropy okay it turns out that that they see the true output label distribution that you have right from the data that is given to you right and this is what you do for estimating this is the estimated label distribution.

And since you are using an unbiased estimator for the probabilities you end up actually estimating the true probabilities so that is why it is called cross, cross entropy this, this is supposed to be the that is the estimated okay and since you are anyway just counting the number of labels of each class and then dividing it and doing this so it is essentially end up with the same thing okay.

So the first one is the output label distribution this is the estimated one and so if you end up with the same thing right so another way of thinking about it is if you look at the prevalence of the labels in the data and I give you 100 data points right essentially if I am going to randomly pick a data point and look at the label right so this is the probability of seeing label k correct so going back to your ideas of Shannon's entropy so if I have if I have a sequence of 100 things I have k possible symbols that can occur right.

And this gives me the number of bits I need to encode these k symbols given the relative frequency of those symbols right if I had not done the splitting right if I had not split into M

regions right if I had kept the data as a whole I would have required some number of bits to encode the output level that make sense suppose let us look at it this way so I have my data so there are 400 data points of each class.

I will require some amount of bits to encode this right half, half the entropy is I mean the probability is half and half I will need some amount of thing to encode this suppose I split it up so that I get I get two regions one gives me 400,150 other gives me 0 and 250 that is how many bits do I need to encode the output variable here none right always the big improvement.

I do not need any bits for encoding the variable here and here I will need some but that certainly be less than this because we know half of the worst case right so in, in terms of the number of bits that I need for specifying the label I have some improvement when I do the split when I go from 400, 400 when you go from there and I get these two splits the number of bits I need has come down right.

So I have gained some information by doing this split right so how much information have gained? sorry right so the original entropy minus this quantity gives me the amount of information I have gained right so sometimes this is also known as the information gain criteria because of that right so either you, you minimize the cross center of PR you maximize the information gain let us information gain is essentially some constant minus this so that is information again.

Therefore you maximize the information gain or minimize entropy so again the process is very simple you for every feature J you try to find that split point S such that this or this is optimized right one of these three things but the most popular or actually the Gini index and the cross entropy so one thing I want to point out so when you are splitting this into two things right and then I have to find out the overall cross-entropy are devious right.

So what I need to do is so the entropy of this will be weighted by 250 the entropy of this will be weighted by 550/800 right both of these cases so I will have to have some kind of weighted combination of the code or the Gini index whatever it is I have to have a weighted combination of the Gini index of the individual partitions or the deviance of the individual partitions so I have to be careful about that just do not add the M up okay.

You have to use the weighted combination so for this it is fine because it is per region yeah so again you have to be we have to make sure you are combining it appropriately right yeah there is only one output will come right only one symbol is present there is only one symbol present you do not need any bits to encode it because that is only symbol this present class one will not happen so, so 400, 400 means class one there are 400 data points class 2 there are 400 data points 0 to 50 means class 1 there are zero data points class 2 there are 250 data points.

The symbols I am talking about are the classes right here there will be no occurrence of class 0 and only class 2 will occur okay so one again one other caveat we are using this for classification you are doing cause complexity pruning right almost always you are supposed to use the Miss classification error because that is eventually what you are trying to optimize so you grow that tree with whatever error measure you want but when you prove the tree use the Miss classification error.

Because at the end of the day I am going to evaluate you based on the Miss classification error not on the Gini index or information gain or anything and these are in some sense they are relative measures we are good for comparing one feature against the other right but the final performance measure is only miss classification error right so use that when you are doing the protein ok so I will stop here.

### IIT Madras Production

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**NPTEL**  
**NPTEL ONLINE CERTIFICATION COURSE**  
**Introduction to Machine Learning**

**Lecture 43**

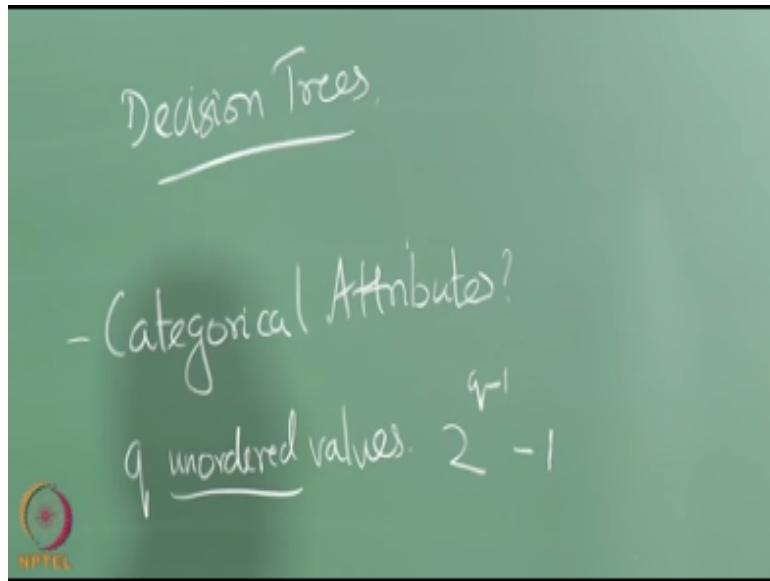
**Prof. Balaraman Ravindran  
Computer Science and Engineering  
Indian institute of technology**

**Decision Tree – Categorical Attributes**

We will continue looking at decision trees so like I said there is a little bit more about trees that I wanted to look at and then we will actually do a, an example today of how to construct tree, so starting from data set I will actually constructed decision tree right okay, so we already looked at a couple of issues with regard to addition trees what are the what are the things we looked at a well how will you we need to talk about that yet.

So when we look at how will you pick a cell a splitting attribute right what is the splitting value in the splitting attribute, so how large you should go a tree right and how do you prune it okay these are the issues that we looked at right whether several other questions that we could ask right, so when we talk about splitting attributes and split points inherently we are assuming that our attributes are continuously right. So that we can talk about a split point right so what happens if I have categorical attributes.

(Refer Slide Time: 01:24)



So no categorical attributes things that take some discrete values right, so it could be things like color red, blue and green, right or it could be things what you normally would believe our continuous variables like age right but for a variety of reasons they have been recorded as discrete values young, middle-aged old right, so most surveys and things like to if you look at it when you answer these things you know they do not ask you for an exact age they ask you are you lesser than 25 or in between 25 and 34 or greater than 35 or things like that right.

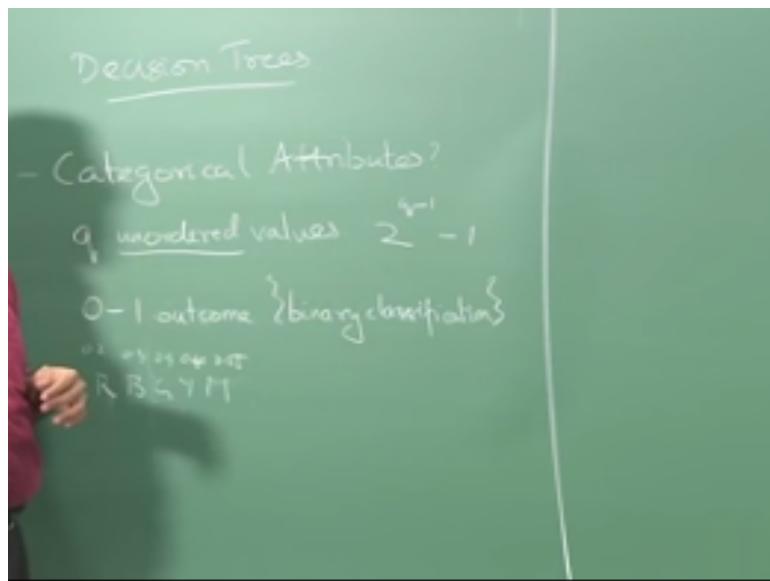
So somehow it is discretized either by for reasons of anonymization or for convenience or whatever you end up discretizing the values right, so you quite a lot of circumstances in fact especially if you are involving with the involved with medical domain right or with as kind of marketing kind of a domain right you will end up having discrete attributes right, so in which case what is the meaning of a split point, yeah I put this binary yes right suppose it is color right, so we think about it if I have q right so that is a key part here think of q unordered values right age itself has some kind of ordering in it.

So I can still fake it you know I might have only 3 different age entries but I can think of it young or middle-aged and old right young and middle-aged or old right it does typically usually does not make sense to look at young and old and middle-aged you know because it is I mean I can say you can fake it you can think of it as a ordered attribute and then you can do this but suppose it is unordered right.

So essentially what you would really have to do is think of splitting the values into two subsets, so like I said let us say take color as an example right I might have to put okay red, blue and yellow into one group and green and I do not know give me a few more color names that is about how I how much I know magenta, purple okay so all of this into another group right and so like that right I have to figure out some way of splitting it into two okay, so how many possible splits are there like that.

I have  $q$  values to  $2^q$  power? good yeah, right so that many possible combinations is not really not going to be feasible for me to go over all of them in order to pick the split point right, so that is exactly what we were doing right if you remember algorithm from the last class you are actually going over all possible split points and we said there are only finitely many such possible split points because we only have to look at  $n$  of them right but now you have to look at even though I have only  $n$  values I mean  $n$  data points for training and I potentially have to look at  $2^{q-1}$  split points right.

(Refer Slide Time: 05:37)



So that is not going to be feasible so there are 2 ways of handling this actually there are 3 ways of handling this the first one is if you have a we have a 0 - 1 outcome basically that is what people would call a binary classification problem if you have a binary classification problem you can do one clever trick what is it that you can do any ideas no I do not want to explore the number of attributes right I am making it some something very restricted here right I am looking at binary classification problems.

So says something that you can think of that you can do here, so what exactly are you looking at when you are trying to find a split point what you are trying to do is trying to make sure that your prediction right when I do it on one half versus other half is more accurate and the prediction I did on the data as a whole before the split right so that is exactly what you are looking at from the split point trying to find a split point such that it is more accurate than the other right.

So what you can do essentially here is you can pick one of the classes let us say you pick class one right let us say I have 5 predictors or I am sorry not yeah I have 5 predictors, so predictors or this the unordered values right, so let us say I have 5 values for a particular thing let us say colors right red, blue, green, yellow, magenta right as I say I have 5 colors, now what I will do is I will take red okay I look at all the data points that have color red okay I will see what fraction of them or class one right.

Then I will take all data points that have color blue I will see what fraction of them class one likewise for the other 3 colors it makes sense so far I look at each color figure out what fraction of

that color that data points having that color or of class one now I will arrange them in some order ascending order let us say of this probability when I say fraction it means what fraction that is a probability that a data point having color red will be class one right from the training data I shall ascend arrange it in ascending order of this probability.

Then I will just treat it like any other ordered variable and then I will split right does that make sense, so why does this help us think about it a little bit right suppose I suppose I have put it in some order right let us say that. So red has 0.2 % of the data having class 1 right let us say suppose something like this no need not be why should they be I am just looking at each fraction of the data points with color red that were class 1 more than one color no this one attribute that says color of the data point okay whatever is it the one attribute that says color of the data point right.

And that way that attribute can take 5 values red, blue, green, yellow or magenta right suppose it is taken color red I look at what fraction of those data points that have color red or of class one right suppose I find that there are 10 data points that I have color red and two of them are of class 1 then it is 0.2, so obviously so these numbers do not have to sum to 1, right because they are only for that right, now we can tell me what is a good place to split this okay before somebody asked me a question about what if it is exactly 0.5 okay.

There you go oh 0.2, 0.3, 0.4, 0.45 and 0.55 see how we go about doing the lot of one thing good point yeah so yeah, so you know how to do this come on pick up pick a thing and tell me what you know the Gini index and you know the you know Gini index or information gain or something like the right all of you know that all miss classification error let us use miss classification error as the splitting criterion, so for me to find an optimal split I do not really have to consider R and Y going to one part right B, G and M going to the other side right.

So it will either be here or here or here or here or here I mean that is a really bad attribute to pick if it is here right, but it will only be left to right all right, so these are the only subsets I need to consider I do not have to consider all of the other subsets right it does not make sense you can intuitively see this here right, so since this fraction of the class one keeps going up right, so either you break here or here or here or here, so that is a heuristic for this right in fact with a little bit of thing you can show that for two classes right.

You will get the same optimal split right by using this method as you would get by exhaustively searching through all the splits I am seeing too many puzzle looks we did this decision trees day before yesterday if remember decision trees okay, so split points if I have categorical attribute split points are going to be like subsets of the values the attributes can take right, so a split points would be okay do I consider our red and green to one side blue, yellow, magenta to the other side that is good potentially a combination right.

So in this case I am saying you do not have to worry about all possible subsets all you need to do is after you have done an arrangement like this okay wherever you choose to split depending on the criterion you are using so wherever you choose to split right so everything to one side will form one subset ever thing the other side will form another subset and these are the only subsets that you need to consider while you are trying to find the optimal space you do not have to consider all the  $2^{q-1}$  subsets okay fine.

### **IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

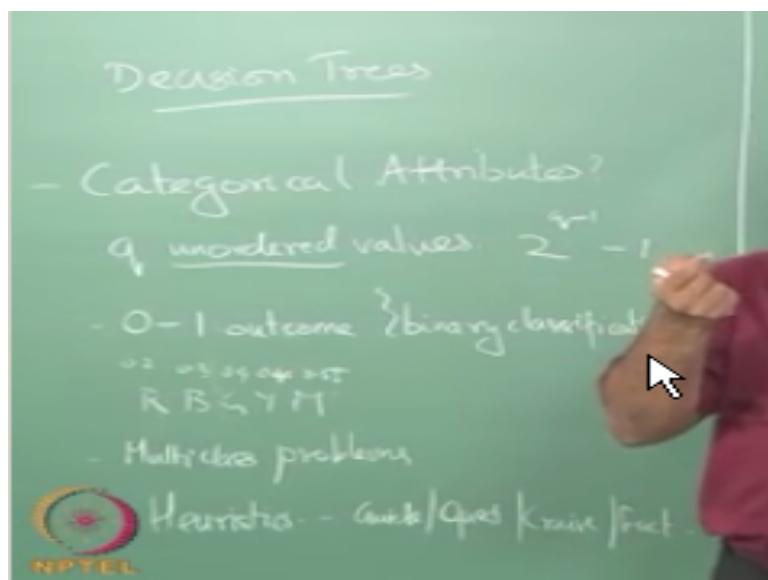
Lecture 44

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

**Decision Trees - Multiway Splits**

So what about multi class problems many, many classes not just 0 and 1 and I have like five classes for echo design labels from right what about multi class problem for multi-class problems this kind of simplification is actually not possible right so we have to end up using some heuristic or the other right so typically what people do is they end up doing some kind of very rough clustering on the values that attribute can take right and then try to define split points based on that and yeah so I am not going to go into the details of the heuristics I mean if, if at all you are going to use this and you will probably be using a packet but will be good to read up some of these.

(Refer Slide Time: 01:17)



If you are interested as you can imagine as soon as you enter heuristic territory right everyone can have their own favorite heuristic so there are many that have been proposed in the literature is guide quest Clues fact the one says annoying things in many of this machine learning data mining literature as people sometimes go out of their way to come up with pronounceable acronyms.

So people have clustering algorithms called chameleon imagine how much work they must have gone to produce chameleon as an acronym and so in fact I think in fact what you essentially end up doing is you use some kind of indicator variables right for each of these right this is something I think you suggested that right this is an indicator variable for each of these dimensions and then they try to do some kind of dimensionality reduction on that I try to pick a discriminating direction right.

And then project on to that and then use that dimension for splitting suppose I want to spit on color right I will not do it on color so I will create 5 variables okay which is essentially one variable or color is red one variable for color is blue one variable for color is yellow and one will for color is magenta but I will not use those as Boolean variables right and I will try to find some kind of a projection from this 5 five dimensional space on to a single dimension and then flick that single dimension as a continuous dimension and try to do my projection on it essentially ends up doing some kind of clustering instead bring some kind of clustering on that one dimension.

You talk about clustering little later but you but you know what Clustering is I already told you what the problem is in the very first class okay the other approach to doing this is to is to do multi very multi-way splits, so what do I mean by that if okay if I decide to split on color or I have to evaluate color in so splitting it into two groups I will split it into 5 groups in our case in our example because they are 5 values color can take I will split it into five groups right so in my decision tree instead of always looking like this will suddenly start looking like that.

So what are the problem with multi way split so why do not we use multi-way splits all the time too much computation in what way not each of the class right why are we determining the split point for each of the class talking about an attribute that describes the data right this is that some confusion people are having here when it when I talk about categorical attributes I am talking

about attributes of the data other than the class label the class label will always be categorical right if it is continuous then it becomes a regression problem right.

But then the values that are describing the data itself you normally assume that  $X$  comes from RP right I was telling you that that need not be the case right if suppose I am filling out a survey form you in stop filling in a rage or something will going to say less than 25 or something right so in such cases how will you test on that variable right how will I split on that variable that is a question we are asking so in now instead of saying that you see less than 25 and between 25and 35 will go left and greater than 35 and greater than 45 will go right.

Instead of saying that and say okay this will be less than 25 this will be between 25 and 35 this will be between 35 and 45 will be greater than 45 or something then splitting it all the ways in one go really does not it is a little bit more computation because when you are computing the score of each attribute that you have to do some additional work but it is not too much okay what is bad is it no yeah but moving you always remove the whole sub tree right yeah so interpretability becomes a casualty right.

So because if you are going to have multi-way splits becomes harder to interpret so the tree becomes very sprawling all right so remember as one of the biggest advantage of decision trees is that they are easily interpretable now if I am going to say okay there is a ten way split and then you have to go down the 10 way split and go down further then it becomes harder to interpret right.

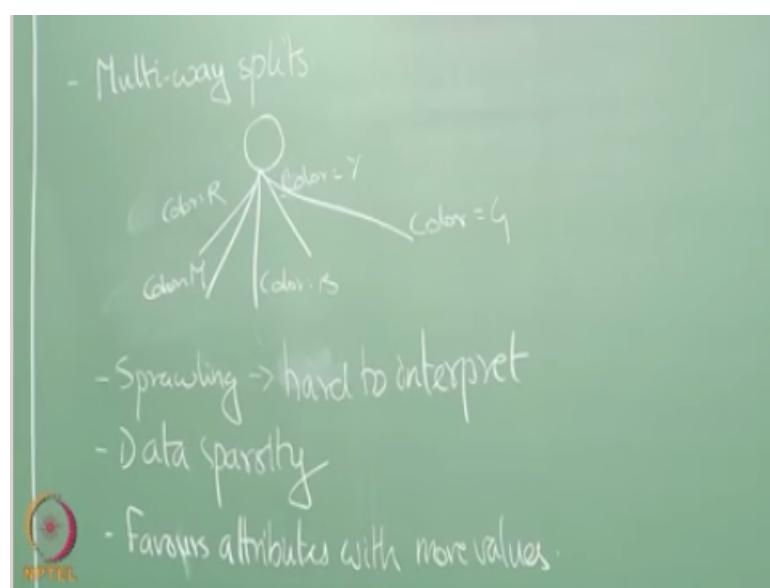
So like if she was saying you might lose insights right that is essentially saying that some amount of interpretability is lost but there is another problem with having sprawling trees yeah, so variance is more but the related problem to that right is the fact that if you do this multi-way splits the amount of data that is available might come down drastically right, so each path might have suppose let us say Magenta is an Rare color right so here that has nothing to do at this 0.55 okay magenta might be a rare color right it might be just only 10 people in my million customer database ever have magenta color shirts okay.

Right but then 55% of them might be positive I do not know see that that has nothing to do with it right so how predictive it is of the positive class is nothing to do with the size of the population there right but the problem is I will only have 10 people here on which to make further decisions

right so if I am going to do this multi-way splits I run into data scarcity problems very quickly does it make sense I mean I know I really cannot ask you questions and exams or things like that with all of these things more like practical guidelines for you to when you actually start using these algorithms.

What are the things you should be watching out for right it should we are using decision trees you should make sure that you are not running out of data points very quickly if you run the some branch in your tree becomes sparse quickly right then it becomes harder for you to trust the trick okay and this is related to the variance question because we are making decisions based on very small number of data points then naturally the variance is going to be high here decision branch that is what I say.

(Refer Slide Time: 09:19)



So if you want me to actually fill in some things here there are no two choices when you pick and that is the whole point I am eliminating the whole question of splitting again picking a split point right so at the color attribute I will say a color not is sorry color equal to R you go that way like that, so this is essentially how your tree will look up, so this how it is going to look like yeah exactly so this is though so see you remember you compute the quote-unquote the utility of splitting on a particular variable right.

So you pick a split variable and then you find optimal split point in that split variable and then you look at what is the least quality whatever you can achieve rate we look at squared error we looked at entropy in whole bunch of other things, so you essentially look at that so here instead of looking at the best possible split point once you pick an attribute you split on all the values attribute can take and then compute the measure whether it is squared error or entropy or whatever it is you can compute the measure.

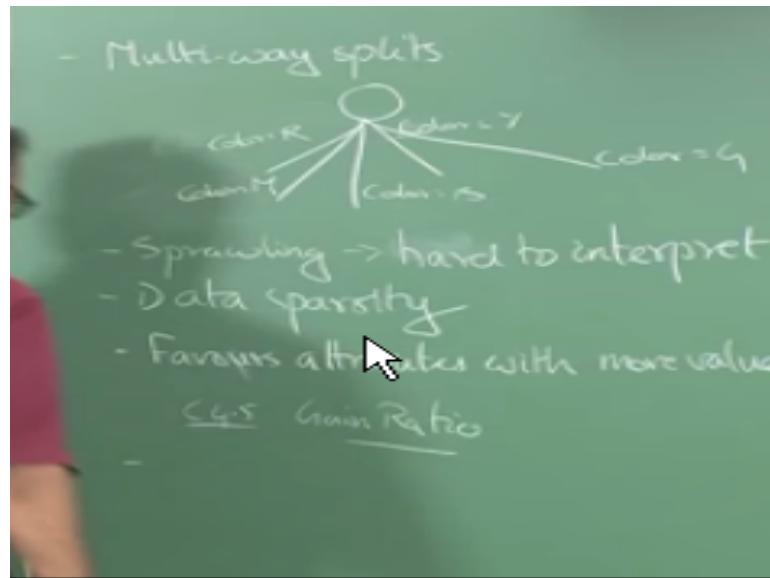
So all the measures we talked about where in contingent on the split being binary right essentially you just looked at  $R_1, R_2$  for simplicity sake but I could have had  $R_1, R_2, R_3, R_4, R_5$  and I was computing the expression if you do not and if you cannot relate to what I am saying this flip back if you actually were taking notes you would know what I mean right, so we have any way looking at the squared error measure right I wrote down as  $R_1$  and  $R_2$  where  $R_1$  was it is lesser than  $J$  and where  $R_2$  is greater than  $J$  or something or  $S$  whatever those split point.

But here we do not there is no choice of what is an optimal split point here once you pick an attribute you split on all the values it can take so it becomes sprawling so that leads to so if we are doing this multi-way splits it is green natural please favor attributes with more values, so let us say I have color and then I have something else like a tan color has 5 values and ages like 15 different bins I have split age into right so when I split on color I will split into 5-way branch when the split on age I will split into 15 way branch right of course there can be exceptions but I would more likely to find pure leaves when I split into 15 then when I split into 5 right.

I split it into 15 ways and more likely to find leaves that are pure and if I split into 5 ways I am less likely to find leave set of pure right so just pure in the sense they have the same class right so this kind of multivariate tends to favor attributes with more values right, so that is not necessarily the best way of doing the splits because you might not be generalizing properly later

right, so for this people use all kinds of tricks, so they are very popular decision tree algorithm called C 4.5 which uses something called gain ratio.

(Refer Slide Time: 13: 37)



So people recall information gain as you spoke about in the last class it is related to entropy right the information gain thing so information gain tells you how much less information you need right by splitting on a particular attribute for encoding the class labels right, so what again ratio says is hey forget about the fact that I have this way this variable suppose I split the data into 10 ways randomly how much information would again vs. splitting it into 10 ways based on this attribute you see the defense I take the data split it into just randomly split it into 10 groups right or I take the data and split it into 10 groups based on this attribute okay.

So that ratio is what I will use so if I can just figure it out arbitrarily split the data into 10 groups under out of still gain the same information as splitting on this attribute then I do not want to split on this attribute it is no better than random right and heaven forbid ration is less than 1 I really do not want this right, so the ratio should be higher than, so that is what I will be looking for so I can instead of using information gain I will use gain ratio in likewise you can order this for any of the attribute as any of the measures that you use that you can always adjust it for random splits. So that is essentially what we end up doing right.

So you gain anything special about expressive power for the tree we are doing multi-way splits as a tree become more expressive in the sense that can it represent more functions than you could with binary splits know whatever I do you do it is multi-way splits I can do a recursive binary splits and I can achieve that not adding to the explicitly it just avoids the question of picking a split point right that is not a trivial thing okay.

You have to come up with all kinds of heuristic to split bit points no pick split points but still if you can okay the recommendation is to avoid multi-way splits and stick with binary splits but in some cases just easier to do this especially if the number of ways in which you will split is small enough if you know if you are not going to split it into 20 different things or 50 different things right you can still do multi way splits like 5 or 6 should be fine.

### **IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

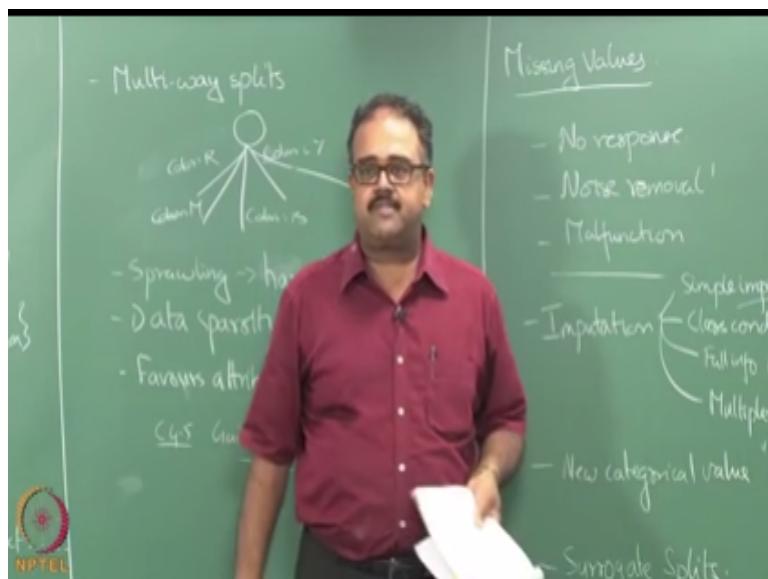
**Introduction to Machine Learning**

**Lecture 45**

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

**Decisions trees – Missing Values,  
Imputation, Surrogate Splits**

(Refer Slide Time: 00:15)



Missing values right, so what I mean by missing values? Some classes that are a different thing we come we will talk about class will much later. Suppose I have again let us go back take a real scenario right, you are filling in some survey questionnaire and then I am going to take all the data from you and then I am going to build a decision tree that allows me to predict whether you will buy a machine in my or buy a new computer from my shop or a new TV from a shop or something right and then the next somebody comes in and I am supposed just look at you and say okay he says he is going to buy a computer he is not right.

So I should be able to classify people like that but then when I fill in the survey I would just not answer some questions. So i will have a data point right so i am assuming that  $x$  is down from RP right or whatever the space that I am drawing  $x_i$  from but for each  $x_i$  assume I already know the values  $x_i$  to  $x_{ip}$  and so far we have never talked about the case where some of these might be unknown, some of this might be missing they might be missing for a variety of reasons right. So one could be right there could be a no response in a surveyor something right the other could be could be due to noise removal.

What do I mean by that? I look at my patient record data I find that somebody has a temperature of 223, so obviously some noise there right do I make it 22 or 23 when or 22.3 right nothing seemed straight right, I will tell you what scale it is in but still the still does not seem right so what do you do just remove it okay. So let us just assume that the nurse did not record the temperature of this patient right, so you remove noise from your data you might lose some attributes hi everything else about the patient I just do not know whether he was running a fever or not when he came into my clinic right.

So likewise you could just not have recorded it right that is equivalent to no response, so that guy messed up might have come with alike a bleeding right hand with it is just hanging off a wrist or something and you are not going to say hey first get this temperature and put it in there I do not want any missing values right, so this thing says it might not just get recorded you know so those kinds of things are there is an equal until no response, all right so anything else yeah exactly that is what is a malfunction right.

So it just that you might be recording sense of data from somewhere and the sensor just turns off for a while it may be it over heated or something went wrong and just for just a while you do not see any of this data being recorded, so the variety of reasons why you could have missing values in your data in fact if you work with real data right more often than not you will have significant missing values. In fact when I work with some data I have had cases where people have given me data where some attributes were missing in more than 80% of the data point's right.

So what you do in that cases remove attribute itself okay, you mean I shall not worry about the attribute because you are not going to be able to use it in any practical setting right so we just removed a trigger itself, but in other cases if it is missing only in like 5% or 10% of the data points you do not want to throw away that data point like throwing away 20% of the data point is

still a big thing right and yes and you do not want to remove the attribute also because it is available in 80% of the data points and you do not want to throw the attribute away you do not want to throw the data points away right.

There are two things you can throw the column of a can throw the row way right if it is missing in more than 80% you can directly throw the column away, but it is somewhere in the middle right summer small numbers then you do not know what to do throw the column or throw the attribute throw the rope, do not both exactly. So there are lots of different ways of handling this missing, very the statisticians have studied these new ones right so they have come up with many techniques for handling missing values and why am i bringing it up while we are talking about decision trees and not other classifiers.

Because there are some techniques which is peculiar tradition trees which are not available for other classifier. I am going to talk about all of these right so that I mean in general also you could use some of these techniques the first one all right so we will give it a fancy name called imputation. That imputation is essentially filling in a value for the missing attribute right, so how do you fill in the value for the missing attribute, you the mean the simplest thing is to do the mean you could do a regression on the attribute right.

So you could regress on the attribute in fact what is the best way of doing regression on the attribute? You should do it in a class conditioned fashion, use the class also because you are talking about the training data here right, and so use the class also as in part of your regression or part of your averaging. So what you can essentially do is okay to take all the data points that are of class 1 and use those to predict the value of the missing attribute for in that set of data points right suppose I have like a hundred thousand data points and let us say thousand or of class 1 and of which 4 of them are missing some attribute 3.

Let us take those thousand data points and I can i will do a regression and predict for those 4 data points i will use the remaining 9996 data points as my training data and fit the curve and now I can predict what that one missing point is, so why is this kind of conditioning on the class useful? You use that feature class, so if there is any kind of variation right the correlation between that feature in the class right this will help me preserve it right. If I am going to do this across the entire 100,000 data sets I lose the correlation I will lose the effect.

At least for these attributes it will get polluted right but this way I will be able to retain it, do not lose anything you do not do anything by doing it this way if that is correlation you actually preserve it if there is no correlation you do not lose, anything sorry? Exactly was asking just saying what if there is no correlation do not lose anything by doing it this way right. So this is imputation the different ways of doing imputation you can use the mean you can use the class condition mean right you can use regression for doing the imputation and there is something anymore complicated technique.

That something called multiple imputation on using the regression for doing the imputation is also called full information imputation, I told you only the statisticians have been at it for awhile right, so they have all kinds of the full information imputation is it because you are using all the known attributes, for predicting the unknown attribute right we are doing the mean you are only using that attribute in the same attribute in other data points and multiple imputation is a little weird thing. So what you do is you use all the data that you have right and setup a probability distribution over the missing attribute values right.

Like I said I have 996 data points in which that attribute is not missing right I will use that and figure out okay for if for red what is the probability for blue what is a probability for green what is a probability for discrete see there for continuous values i have to pick some distribution, let us say I pick precaution and else okay I will find what is a mean and the variance of the Gaussian that will predict the missing attribute value okay. Now what I do is I draw samples from this distribution and use those samples to fill in the missing values.

I will get one data set again other set of samples and fill in the missing values i will get another data so that is called multiple imputation. So i can create multiple copies of the data point by repeatedly sampling from this distribution and in some cases this has much better variance with much lower variance than using some of the other method. So even though this entails significantly more computation okay, so imputation is one that is another handle this I just introduced a new value for the variable right and I will call it missing.

Why would this be useful? exactly so there might be some kind of systematic reason for which the data goes missing and if I instead of trying to somehow guess what the value should be if I actually pay attention to the fact that it went missing right that would be useful, so I did not see who said that okay yeah, so infact it is actually a very practical practically useful thing it because

quite often the reason it goes missing is that is a specific reason for it, and you can in fact the fact that it is missing might be predictive of and you know.

So how likely is my patient to recover the temperature reading is missing, so those kinds of things, so use something called surrogate splits, so what is a surrogate split? okay so surrogate splits actually a slightly different function, it works for imputation and this can be used during training itself right but the circuit splits thing we typically use during testing you can also use it during training is suppose. The basic idea is this for every attribute right that I have I will try to pick another attribute okay, that tends to split the data in the same way right.

Suppose let us say again let us take the same example I have 100,000 data points I split on attribute say 3 can I get two groups right this has says 70,000 data point this as another 30,000 data points. I split on some another attribute let us say 4 okay again I get two groups one has 68,000 data points other has 32,000 data points and not only that it turns out that the intersection of the 70,000 and 68,000 is something like 65000, on the intersection of the other two is something like 25,000.

So essentially three and four give me more or less the same splits, we are finding correlation we are not really reducing it here, so we are finding correlation what we do is if we have selected attribute 32 split on our tree and then we suddenly find that attribute 3 is missing in the data point we just split on an attribute 4 and behave as if we split on attribute 3 and go on. So this is what it means a surrogate right it is like putting proxy right, so I have attribute for can put proxy for attribute 3 and then he just continued working with your tree right.

So that is essentially what circuit splits up and it does is it exactly finds, that right it actually looks at correlation between the attributes and tries to exploit that okay, so as you can see that imputation right and adding this new categorical values could work with any kind of classifier that you are working with right. As long as you have a way of handling categorical attributes it is just one more value that you are handling, while the surrogate split something very specific to trees right likewise we are going to look at fragment which is also something very specific to trees.

So this is a little subtle so what I am going to do is the following right, so I come to a point I am going to make some query, I am going to make this  $x3 < 5$  that is a query that I am going to make

it is a variable  $x_3 < 5$ , that is a query I have to make at this point in the tree right and what do I find my data point does not have  $x_3$ . That is for categorical drive talking about categorical attributes, so it will be like okay I am going to RM V YG missing like that that or if I am going to do it in two subsets I will put missing into one of the subsets right.

But suppose this is there  $x_3 < 5$  so what do I do  $x_3$  is missing right what I do is I look at that all the data points for which  $x_3$  was not missing okay, I will see what fraction went down this way let us say oh 0.6 went here 0.4 went here, so what is 0.6 all the data points that did not have  $x_3$  missing 60% of those we are  $< 0.5\%$  of those had  $x_3 > 0.5$ , so now what I do is I am looking at one data point right one data point that came here I am going to split it into two right, so it is going two point six of the data point is going to travel down the left and 0.4 of the data point is going to travel down the right.

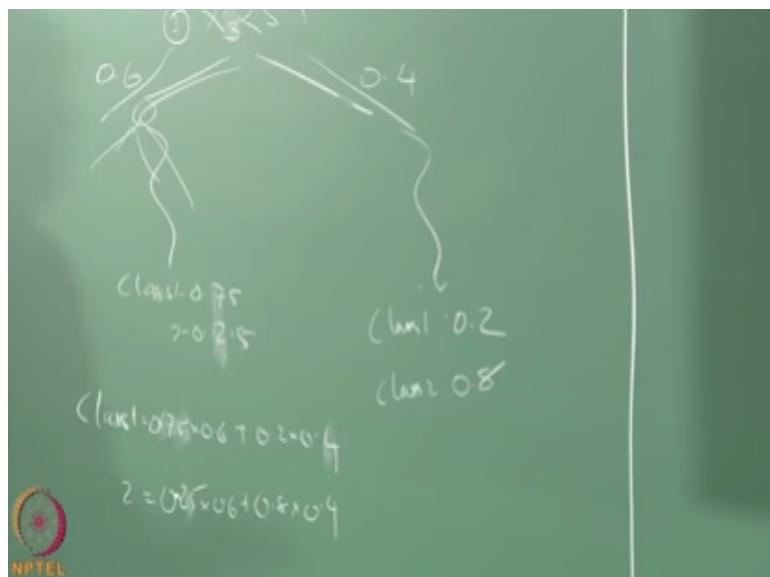
So what I do is it is essentially I am actually letting the data point travel all the way reach a leaf right and the leaf is going to make some prediction, so some probability it is class one some probability risk class to some probability is class 3 right. So the 0.6 part will make one prediction the 0.4 paths will make one prediction, I make a weighted combination of the two predictions and I output that as my final is it is seems like quantum mechanics name. So let us say I go down I finally reach here and I say its class 1 with probability I do not know 0.6 and class 2 with probability 0.4 and this one winds down somewhere.

And I say this is class 1 with probability 0.2 and class 2 with probability 0.8 right, so overall think that I will report is the probability of class 1 is okay, that makes sense no this is maybe this is a bad choice, do that make sense I am using it only once right. So what is the meaning of saying 0.6 of the data point goes down this side is essentially I will go all the way down and I will say that finally I will use the 0.6. I am not using the 0.6 anywhere here and these telling orders the semantics of saying 0.6 goes down this way.

So the reason we are carrying this weight along is at some point further down the line if I have another missing attribute and I decide to split it I have not been splitting one I will be splitting only 0.6 right, so this can get weird right, so I can have a data point which has multiple missing attributes traveling down more than two paths it will reach multiple leaves and then I eventually combine all the leaves so this is called when I am calling this fragmenting method right.

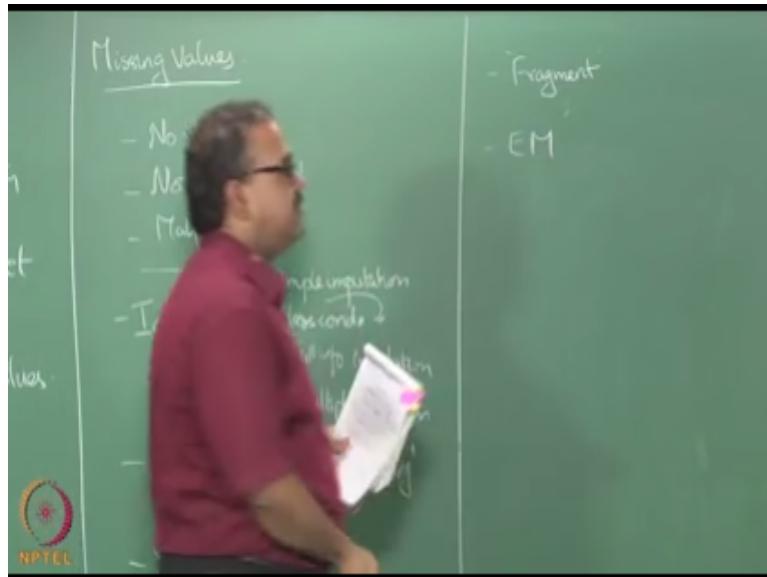
Again this is pretty unique to trees so if you think about it this is somewhat similar to doing multiple imputations. Of course the whole idea is to use training data to make a prediction on this data point right the whole the subject is predicated on using the behavior of other data points to predict output.

(Refer Slide Time: 20:27)



Of new data right, so it should not matter right.

(Refer Slide Time: 20:34)



So the last way of handling missing values is something called am expectation maximization right and it is going to keep cropping up all over the place as we go along but we will do it we will actually formally do with deal with expectation maximization much later. So just be aware that when we look at EM this is one of the applications of EM okay and link to missing values I am not going to get into this is a pretty involved thing and in fact if you think you have been having difficulty with any of the concepts we have covered so far in the class you will not seen anything yet right so am is the one thing which everybody struggles with when they look at it first time so we will come to that later.

### IIT Madras Production

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

## Introduction to Machine Learning

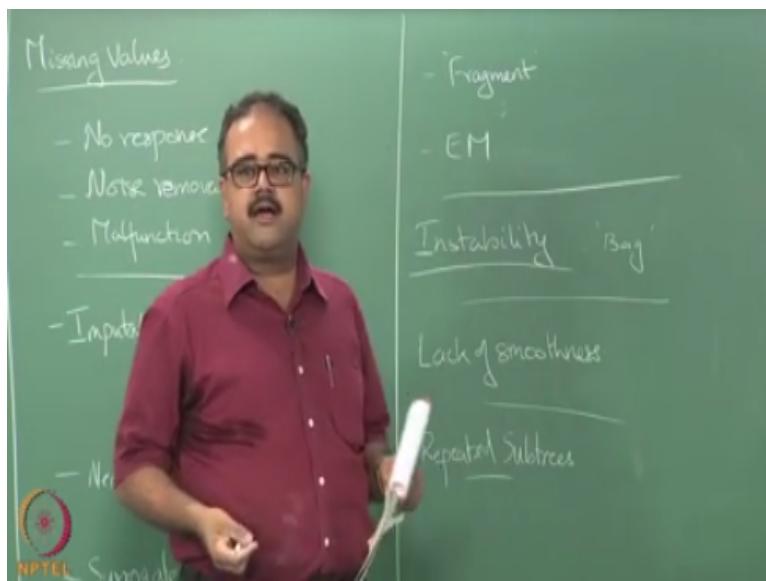
### Lecture 46

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

#### **Decisions Trees – Instability, Smoothness, Repeated Subtrees**

So the next thing would not talk about is instability.

(Refer Slide Time: 00:19)



So Decision trees are pretty unstable so what we mean by not stable all of you know what we may not, stable small changes in the training data will could cause potentially large changes in the decision tree, so what could happen things that you split at the root might go somewhere down because of some changes in the data right and if the data you start out with small where is the sample size is small then the variation is going to be very high right. So I said any way of getting around it some of it some of some regularization helps to some extent the solve the pruning and stuff helps to some extent.

But still not a lot right because the variance is really high these things is really unstable but still trees are very useful, so what we will do is we look at a very specific technique so that that is what we are going to do one of the ways of doing it yeah so there is a very specific technique called bagging right, so that not the next not the next four classes from now okay I will do bagging in more detail right but the basic idea is that to minimize variance is not just with trees you could do it with any unstable classifier.

So what you do is instead of training it on the data that is given to you train on slightly different versions of the data right maybe you can just take say 70% of the data randomly choose 70% of the data and then train a tree randomly choose another seventy percent train another tree keep doing this and then somehow combine the class labels predicted by all the copies of the trees.

That you have trained so this allows you to have slightly more stable classifier what is the problem with this yes anything else yes anyone else and I think I heard somebody say that you lose the biggest advantage of decision trees, which is simple comprehensibility so instead of having 13 or 15 trees or 100 trees now again you have the problem of keeping your job and your manager ask to explain what happened right so now I have 100 trees and somehow they make this magical prediction and you do not know.

So that is a problem so smoothness in your prediction is something that you are looking for a condition trees are not going to give, you that right there will always be this jagged jumping around especially regression trees right, so when you are talking about making some predictions if you are using a piecewise constant fit right for every region you are going to have some amount of jumping around so if you are looking for a smooth function for doing your prediction this is not going to work so you have to do some kind of post-processing after you build the tree in order to smooth the predictions right.

And there is nothing you can do this nature of the beast so the trees are so much convenient in other ways but smoothness is a problem, so if you are looking for a smooth fit right for your prediction that is not going to happen and problem of having repeated sub trees, so what do I mean by that multi way splits it does not have to be multi waste receiver in binary splits you can get into this problem think of XOR right how will XOR look like right I will split on  $x_1$  if  $x_1$  is 0 I will go down the left branch  $x_1$  is 1 I will go down the right branch and then what do I do in the left branch.

I will test on the test on  $x_2$  if  $x_2$  is 0 I will go down one branch  $x_2$  is one I will go down the other branch and likewise I will test on  $x_2$  on the other side with 0 I will go down one branch I will go is 1 I will go to the otherwise so we can think of it these two sub trees are kind of similar right, so it could very well be that I split on one attribute but everything underneath it could be similar the tree structure is very similar but I cannot collapse it because I end up with different conclusions right.

So if  $x_1$ ,  $x_1$  was 0 and  $x_2$  are 0 I would be outputting 0 right but  $x_2$  us 1 and  $x_1$  was 0 I would be outputting one so the outcomes are different so I cannot really club the two trees but then the test that I do are exactly the same right, so this shin trees are prone to having this kind of repeated sub trees right so you could have the same test set of tests that are actually implemented in many different points in the tree so it just makes it three more complex.

(Refer Slide Time: 05:51)

$$\sum_{k=1}^K \hat{P}_{mk} (1 - \hat{P}_{mk}) = \sum_{k=1}^K \hat{P}_{mk} \hat{P}_{mk} L_{kk}$$

But there might be other ways of reordering things, so that you get with the simply the XOR is a bad case right so if you reorder x are you still get with this we still end up with the same kind of repeated structure but there might be other cases where you might have just done the splitting in the normal way but end up with too much repeated structures but if you had we flip the ordering of some variables even though it is not the best variable to pick at some point but you might end up with a more compact tree but finding that is finding that is very hard finding that ordering is very hard as I so you have to just live with it just pointing out some of the caveats.

So far we have assumed that we are dealing with the 0/1 loss function so what is the 0/1 loss function for classification right yeah as good as a mile I mean I do not care there is no ordering in my class labels if I miss if I do not predict it correctly I penalize you with one if I predicted correctly it 0 but there might be cases where some miss classifications are more acceptable for you than others right so what do you do in such cases they so you are going to have some kind of a loss value right.

So I am going to have some kind of some LKK` which is essentially the prowl loss that I will suffer by classifying the data into k` when it is actually class k correct so I am going to have this so how do I accommodate that in the decision tree setup, how do I account how to accommodate that in the SVM setup optimal hyperplanes by the way if I am missing one thing we never actually talked about how you use SVM for multiple classes I will ask me the question what is the margin maximum margin.

What is a margin mean and you have multiple classes that is a topic for another day I will come back to that but yeah, so as she is not immediately clear right so how we do that so suppose you have neural networks like we all know about neural networks right, all of you are familiar with back prop by now I suppose right so how will you accommodate this kind of thing in back prop I think of ways of doing it, so it turns out there is no easy way of doing any of these but what you can do is at least in decision trees.

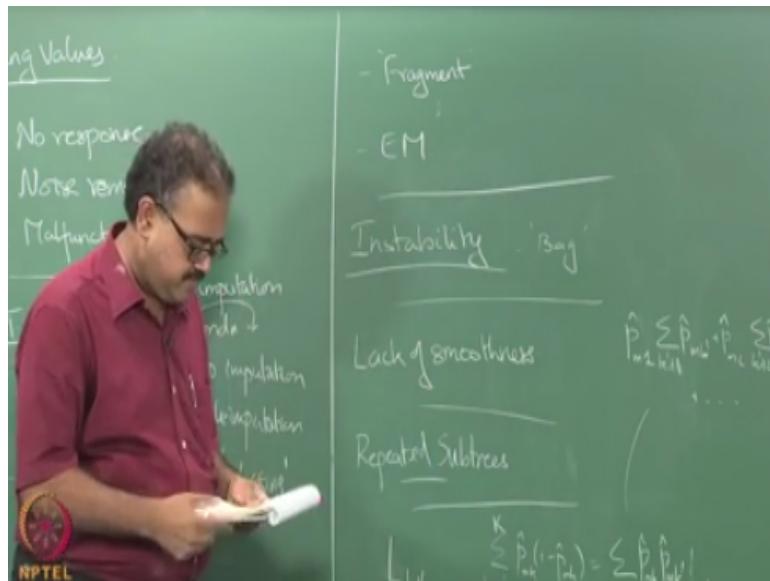
Try to incorporate this when you are computing your GINI index or your what information gain whatever right, so whenever you are looking when you are doing that so you can figure out okay what is the probability that I will miss classify this so what is the GINI index expression that we have right, so this is what we had right so this is essentially this was the probability that a data

point in region m will be in class K right times the probability that data point in region M will not be in class K right.

So as another way I can write this which is essentially right so probability that the point is k + k and probability that is K' so essentially from here to here what I need to do is take out all the terms where with  $\hat{P}_{mk}$  and some out some of our the remaining so which will be  $1 - \hat{P}_{mk}$  right, so that is essentially what I did so for each k, I take  $\hat{P}_{mk}$  out from here and sum over the remaining things and I will get this expression okay so this is some way of saying let okay the original probability is k okay.

And the estimated probability is K' this is the someway of looking at it so here what I can do is I can add my  $L_{KK'}$ , so you have to actually work this out for all of the measures that you are going to work with so if you are going to have a neural network mean squared error criterion you are minimizing or cross the deviance at your cross entropy or minimizing whatever is error function you are minimizing lot to figure out what is the appropriate way to use this class information this class specific loss information okay.

(Refer Slide Time: 11:40)



There was a not equal to the right yeah, okay without this yes they are equal without this year equal so essentially what is what I am doing here is so for every k right I am writing one term like this so he this I can simplify like this right, so that is  $\hat{P}_{m1} + \hat{P}_{m2} + \dots + \hat{P}_{mK} = 1$  right like that I can do that so I will get case terms and this summation is

essentially  $1 - \hat{P}_{m1}$  hat and this summation is  $\hat{1-P_m}$  hat to write like that so that is essentially what I get here right so like that you have to work it out for everything so if we have a different loss function okay.

**IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

# NPTEL

## NPTEL ONLINE CERTIFICATION COURSE

### Introduction to Machine Learning

#### Decision Trees Tutorial

Hello and welcome to this tutorial on decision trees in the preceding lectures we have looked at some of the theory behind decision trees, in this tutorial we will get some hands-on experience actually building trees using some of the concepts we have learned for building decision tree models with real data we will of course resort to packages such as VEGA, however in this tutorial we will be building trees for a very small data set in order to understand the process involved in building decision trees.

(Refer Slide Time: 00:51)

Data Set				
age	income	student	credit_rating	buys_computer (target)
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

This is the dataset we will use for this exercise as you can see there are four different attributes with a binary valued target hat is bias computer note that the attributes age and income can take three different values whereas student and credit rating are binary valued.

(Refer Slide Time: 01:14)

## Decision Trees - Options

- Multiway splits vs Binary splits
- Impurity measure:
  - Cross entropy:  $-\sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk}$
  - Gini index:  $\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk})$



In the theory lectures we have already seen the different options available to us when building binary trees for example the first thing we have to consider is the type of tree which we want to build we can either have binary trees or multi way trees depending upon the branching factor at each node another option available to us is the impurity measure used in this tutorial we will be looking at two different impurity measures which are cross and entropy and the Gini index. Another option which we do not consider here is the pruning technique used.

(Refer Slide Time: 02:01)

## Multiway Split using Cross-entropy

Consider the attribute 'age'



To start with let us try and build a tree using multi-way splits and cross entropy as the impurity measure, the first thing that we have to do is to identify the root node this is done by considering each attribute in turn calculating the cross entropy value for that attribute and identifying the attribute which uses the lowest value let us start by considering the attribute age.

(Refer Slide Time: 02:32)

age	income	student	credit_rating	buys_computer (target)
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	no	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no



5 points with age = youth; 2 with buys\_computer = yes & 3 with buys\_computer = no

From the table we observe that the attribute age can take on three distinct values which are youth, middle-aged and senior, going back to the formula for cross entropy. We see that for each node we need the proportion of class k observations in that node in case of a two class problem such as the one we are considering we can use the simpler expression  $- p \times \log p - (1-p) \times (\log 1 - p)$  where p is the proportion of observations for the positive class. Since, we need to calculate the cross entropy for an attribute with three distinct values we will have three components.

Let us first consider the value youth this is highlighted in the table we observe that out of the 14 different data points five observations have aged equals to youth among them two are belonging to the positive class and three belong to the negative class.

(Refer Slide Time: 03:51)

$$\begin{aligned}
 \text{cross\_entropy}_{\text{student}}(D) &:= \\
 &(7/14)(-3/7 \log_2 3/7 - 4/7 \log_2 4/7) \\
 &+ (7/14)(-6/7 \log_2 6/7 - 1/7 \log_2 1/7) \\
 &= 0.7885
 \end{aligned}$$

$$\begin{aligned}
 \text{cross\_entropy}_{\text{credit\_rating}}(D) &:= \\
 &(8/14)(-6/8 \log_2 6/8 - 2/8 \log_2 2/8) \\
 &+ (6/14)(-3/6 \log_2 3/6 - 3/6 \log_2 3/6) \\
 &= 0.8922
 \end{aligned}$$



Using this information we have  $-2 / 5 * \log (2 / 5)$  that is the proportion of observation belonging to the positive class and  $-3 / 5 * (\log 3 / 5)$  for the negative class this expression is multiplied by the ratio 5 : 14 which indicates which is a weight on the which is a normalizing factor since five out of the 14 data points had aged equals to youth continuing with this manner we take up the next value that is age equals two middle-aged and observe that among the 14 there are 4 points where age equal to middle-aged and for all of them buys computer equals two years that is they all belong to the positive class.

This gives us the second component as you can see we do not necessarily need to calculate this but we have put it there just for your reference the final component comes when we consider age is equals to senior again there are 5 points with age is equals to senior of them we observe that three belong to the positive class and to belong 2 the negative class putting it all together we get a value of cross entropy. that all logarithms used here are using the base 2.

Make sure that you are able to follow the calculations especially how we were able to write down each of the components of the cross entropy expression using the same process we now consider the attribute income and find the cross entropy width next we find the cross entropy for the attribute student and a cross and trouble for the credit rating note that here we have only two components because both of these are binary valued.

(Refer Slide Time: 06:07)

$\text{cross\_entropy}_{\text{age}}(D) = 0.6935$   
 $\text{cross\_entropy}_{\text{income}}(D) = 0.9111$   
 $\text{cross\_entropy}_{\text{student}}(D) = 0.7885$   
 $\text{cross\_entropy}_{\text{credit\_rating}}(D) = 0.8922$

From the above calculations, we select 'age' to be the root node of the decision tree



Finally we compare each of the cross entropy values and in this case observed that the attribute age gives the lowest cross entropy value and hence is the optimal attribute to use as the root of the our decision tree.

(Refer Slide Time: 06:24)

## Partial Decision Tree



Thus we obtain the partial decision tree with age as the root attribute and three branches corresponding to the three distinct values that the attribute age can take note that the middle-aged that is the branch where age equals two middle-aged has been labeled with yes indicating that this is a leaf node where any observation following along this branch will be labeled yes this is because if we go back to the table we observe that when age equals to middle-aged buys computer equals two yes.

Thus along this branch of the tree there is no need to further grow the tree since from the training data given to us we can directly conclude that if we observe age to be middle-aged then we can label the class and the observation as positive that is the person will buy a computer. Now we have created this partial decision tree so how do we proceed? Essentially it is a recursive process we started at the root node we were able to find the root node to be the attribute age. Now along each of the remaining branches where we have not found the note to be a leaf node we have to repeat the same process. So let us first look at the branch is equals to youth we have already considered the attribute age, so there are three attributes left to us using a process similar to what we have just seen we try to identify the best attribute to use at this position.

(Refer Slide Time: 08:29)

---

`cross_entropy_income(age=youth):`

\*



So we consider the cross entropy of income where age equals to youth, now we are not now we will not be considering the entire data set but will consider the restricted data set where age equals to youth.

(Refer Slide Time: 08:42)

age	income	student	credit_rating	buys_computer (target)
youth	high	no	fair	no
youth	high	no	excellent	no
youth	medium	no	fair	no
youth	low	yes	fair	yes
youth	medium	yes	excellent	yes

This is illustrated in this table where we have crossed out all observations where age is not youth so essentially we repeat the same entire process with this restricted data set note that the attribute age has already been considered so we are left with the remain three attributes and these are the values that are to be considered.

(Refer Slide Time: 09:11)

$$\begin{aligned}
 \text{cross\_entropy}_{\text{income}}(\text{age}=\text{youth}): \\
 & (1/5)(-1/1\log_2 1/1 - 0/1\log_2 0/1) \\
 & + (2/5)(-1/2\log_2 1/2 - 1/2\log_2 1/2) \\
 & + (2/5)(-0/2\log_2 0/2 - 2/2\log_2 2/2) \\
 & = 0.4
 \end{aligned}$$

$$\begin{aligned}
 \text{cross\_entropy}_{\text{student}}(\text{age}=\text{youth}): \\
 & (3/5)(-0/3\log_2 0/3 - 3/3\log_2 3/3) \\
 & + (2/5)(-2/2\log_2 2/2 - 0/2\log_2 0/2) \\
 & = 0
 \end{aligned}$$


---

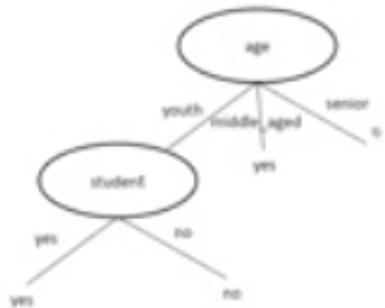


Thus we have cross entropy of income when age equals to youth you can go back and verify that these are the values you will obtain next we have cross entropy of student when they is equal to youth here we observe that the cross entropy is actually 0 going back to the table we see that in a equals to youth and student is no bias computer is no and when student is yes buy computers yes so this leads us to a pure leaf we can go ahead and calculate the cross entropy for credit rating as well when age is equals to youth but since we will not get a value less than 0.

We can stop the process here and get the partial tree where we have selected the attribute student with the least value of cross entropy.

(Refer Slide Time: 10:07)

## Final Decision Tree



As you can see we have labeled the branches yes and no because these are leaf nodes which are pure there is no mixture, now as you can see we have this branch this branch in this branch are all leaf node so last the remaining at branch to consider is when age is equals to senior again we look at the table where we discard all observations where it is not senior and follow the same calculations.

(Refer Slide Time: 10:45)

$$\begin{aligned}\text{cross\_entropy}_{\text{income}}(\text{age=senior}): \\ (2/5)\{-1/2\log_2 1/2 - 1/2\log_2 1/2\} \\ + (3/5)\{-2/3\log_2 2/3 - 1/3\log_2 1/3\}\end{aligned}$$

$$= 0.9510$$

$$\begin{aligned}\text{cross\_entropy}_{\text{student}}(\text{age=senior}): \\ (2/5)\{-1/2\log_2 1/2 - 1/2\log_2 1/2\} \\ + (3/5)\{-2/3\log_2 2/3 - 1/3\log_2 1/3\}\end{aligned}$$

$$= 0.9510$$

②

We get cross entropy and we look at income when h is equal to senior and cross entropy of student when age equals to senior.

(Refer Slide Time: 10:55)

```

cross_entropy<sub>credit_rating</sub>(age=senior):
    (3/5)(-3/3log<sub>2</sub>3/3 - 0/3log<sub>2</sub>0/3)
    + (2/5)(-0/2log<sub>2</sub>0/2 - 2/2log<sub>2</sub>2/2)

    = 0

```



And cross entropy of credit rating when is equal to senior, here again we find a cross entropy value of 0 which is the minimum and if we go back to the table wherever credit rating is fair we have bias computer equals 2. Years here as well and whenever credit rating is excellent you have by some high school to no. So this allows us to create a decision tree.

(Refer Slide Time: 11:21)

## Final Decision Tree



Where each of the leaf node is a pure nod in case we did not get a cross entropy value of zero let us say we have a different value cross entropy here we would again continue the process and the last situation is when we have exhausted all attributes available to us and we still do not have a pure leaf what do we do then essentially let us say when we follow this branch there were five points of which three were positive and two were negative then this would have been labeled yes.

Because the majority of the data points have a positive Cubs have a positive belong to the positive class fortunately for us in this example we have obtained all leaf nodes as pure but this will always this will not always be the case.

(Refer Slide Time: 12:24)

## Multiway Split using Gini index

Same process with the Gini index impurity measure

$$\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk})$$



Next we will look at decision trees using multivariate and the Gini index, essentially the process is the same except that we replace across entropy measure with the Gini index impurity measure this will be left as an exercise.

(Refer Slide Time: 12:41)

## Binary Split using Gini index

For binary splits, for the same attribute, we have to compute impurity multiple times for the different subsets of the attributes value

As mentioned previously, for a binary outcome, to reduce the number of partitions to be considered, we order the values according to the proportion belonging to the positive class.



Now the other type of tree that can be built is a binary tree for this exercise we will look at using the Gini index impurity measure recall that when we were creating multi-way trees at and to select an attribute for a node we have to consider each attribute only once, however in the case of binary trees since for each attribute there may be different subsets to consider that is where to split.

We may have to look at attributes multiple times also as was mentioned in the theory lectures in case of a binary outcome we can reduce the number of partitions that have to be considered by ordering the values according the proportion belonging to the positive class since our data set has binary valued outcome we will see how this process works.

(Refer Slide Time: 13:41)

Consider attribute 'age'



Again we start by considering the attribute age.

(Refer Slide Time: 13:44)

age	income	student	credit_rating	buys_computer (target)
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

Positive class proportions: youth: 2/5; middle\_aged: 1; senior: 3/5

As we can see from the table H can take three distinct values, now we look at the positive class proportions for each of the values we see that when age is equals to youth two out of five times two out of five observations belong to the positive class age is equal to middle-aged each of the four observations belong to the positive class and when a is equals to senior three out of five observations belong to the positive class.

(Refer Slide Time: 14:21)

Ordering for attribute 'age':

youth = 2/5; senior = 3/5; middle\_aged = 1

Possible split points:

{youth}, {senior, middle\_aged}

{youth, senior}, {middle\_aged}

Thus, for attribute 'age' we need to estimate  
the impurity measure for both the above splits



Thus we have an out ordering for the attribute H youth senior and middle aged, what this essentially means is that we have two possible spit points youth and youth along one branch and senior and middle age on the other or youth and senior along one branch and middle-aged along the other note that the attribute age actually has a notion of order that is youth middle-aged and senior.

If we want to retain that notion of order then we would only consider the split points where youth is along one branch and the rest the other two are along the other branch or youth and middle-aged is along the one branch and the seniors along the other we have considered the attitude age unordered here to illustrate how you would go about ordering values for the rest of this exercise we will use the specific ordering for the attribute age, now that we have identified the possible split points.

(Refer Slide Time: 15:39)

Consider attribute 'age'

Gini<sub>age ∈ {youth}</sub>(D):

$$\begin{aligned} & 5/14(2 * 2/5 * 3/5) + 9/14(2 * 7/9 * 2/9) \\ & = 0.6508 \\ & \text{(Note: for 2 class scenario, gini index} = 2p(1-p)) \end{aligned}$$

Gini<sub>age ∈ {youth, senior}</sub>(D):

$$\begin{aligned} & 10/14(2 * 5/10 * 5/10) + 4/14(2 * 4/4 * 0/4) \\ & = 0.3571 \end{aligned}$$



We estimate the impurity measure here using Gini index for both possibilities note that since we are in a two class scenario we use the simplified formula of the Gini index equals to  $2 P x (1 - P)$  where P is the proportion of observations belonging to the positive class. Go back to the table and verify that these are the values obtained we see that when we calculate Gini index where we are considering the split where youth is in one branch and the remaining two are on the other branch we get a value of 0.6508.

And where we consider the alternate split where youth and senior belong in one branch and middle-aged it belongs to the other we get a lower value of 0.3571 thus among the among these two this is the split that will be preferred. Now this calculation is just for the attribute age we need to repeat the same process for the remaining three attributes.

(Refer Slide Time: 16:48)

Consider attribute 'income'

Ordering:

$$\text{high} = 2/4; \text{medium} = 4/6; \text{low} = 3/4$$

$\text{Gini}_{\text{income} \in [\text{high}]}(\mathcal{D})$ :

$$4/14(2 * 2/4 * 2/4) + 10/14(2 * 7/10 * 3/10) \\ = 0.4428$$

$\text{Gini}_{\text{income} \in [\text{high, medium}]}(\mathcal{D})$ :

$$10/14(2 * 6/10 * 4/10) + 4/14(2 * 3/4 * 1/4) \\ = 0.45$$



Thus we have the attribute income, going back to the table we will see that this is the ordering for the three values that the attitude income can take and the corresponding Gini index values.

(Refer Slide Time: 17:04)

Consider attribute 'student'  
Binary attribute, hence single ordering

$$\begin{aligned}\text{Gini}_{\text{student}}(D): \\ 7/14(2 * 3/7 * 4/7) + 7/14(2 * 6/7 * 1/7) \\ = 0.3673\end{aligned}$$

Consider attribute 'credit\_rating'

$$\begin{aligned}\text{Gini}_{\text{credit\_rating}}(D): \\ 8/14(2 * 6/8 * 2/8) + 6/14(2 * 3/6 * 3/6) \\ = 0.4286\end{aligned}$$



\*

And next we have the calculations for the attribute student and credit rating here there is only a single possible ordering because both of these are binary valued attributes.

(Refer Slide Time: 17:17)

$$\begin{aligned} \text{Gini}_{\text{age} \in \{\text{youth}\}}(D) &= 0.6508 \\ \text{Gini}_{\text{age} \in \{\text{youth, senior}\}}(D) &= 0.3571 \\ \text{Gini}_{\text{income} \in \{\text{high}\}}(D) &= 0.4428 \\ \text{Gini}_{\text{income} \in \{\text{high, medium}\}}(D) &= 0.45 \\ \text{Gini}_{\text{student}}(D) &= 0.3673 \\ \text{Gini}_{\text{credit\_rating}}(D) &= 0.4286 \end{aligned}$$

From the above, root attribute = 'age' with split: {youth, senior} & {middle\_aged}

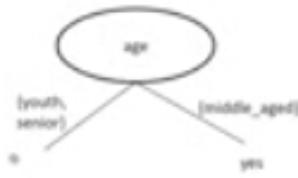


---

Finally we compare all the results and we observe that the attribute age where the split is with youth and senior along one side and middle later on the other is the optimal gives the optimal value and thus we select this particular attribute with this particular split point as the root.

(Refer Slide Time: 17:43)

## Partial Decision Tree



Thus we create this partial T three entry again from our previous exercise we know that when age is equals to middle-aged all our observations are positive, so we do not need to grow the tree beyond this thus our not now we focus on the branch where equals to youth or senior is.

(Refer Slide Time: 18:05)

age	income	student	credit_rating	buys_computer (target)
youth	high	no	fair	no
youth	high	no	excellent	no
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
senior	medium	no	excellent	no



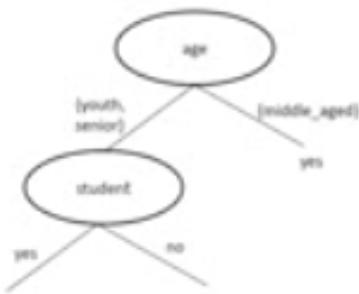
Ordering (income): high = 0/2; medium = 3/5; low = 2/3

So again from previous experience we know that to create and to grow the tree along this branch we need to consider only observations where ages youth or senior thus we have disregarded all observations where you age equals to middle-aged. Now we repeat the same process we have already consumed the attribute age we have three remaining attributes of these income has three distinct values so we need to identify the optimal split point that is we need to first consider the ordering shown here for student and credit rating both are binary values.

So there is only going to be in one straight point so if you repeat the calculations for this subset of the data set.

(Refer Slide Time: 19:02)

## Partial Decision Tree



We will identify that the next node should use the attribute student however this is not the end since we do not obtain pure nodes here and this process has to be continued again we will leave this as an exercise, hopefully this tutorial would have clarified some of the concepts that we came across in the theory lectures and helped you in understanding how decision trees are created of course for real-world data as well as the programming assignments that will be released we will be using a tool such as vacca where you have lot more options for example pruning which we were unable to cover in this short tutorial for any doubts regarding any of the concepts covered here please use the forums.

**IIT Madras production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

# NPTEL

## NPTEL ONLINE CERTIFICATION COURSE

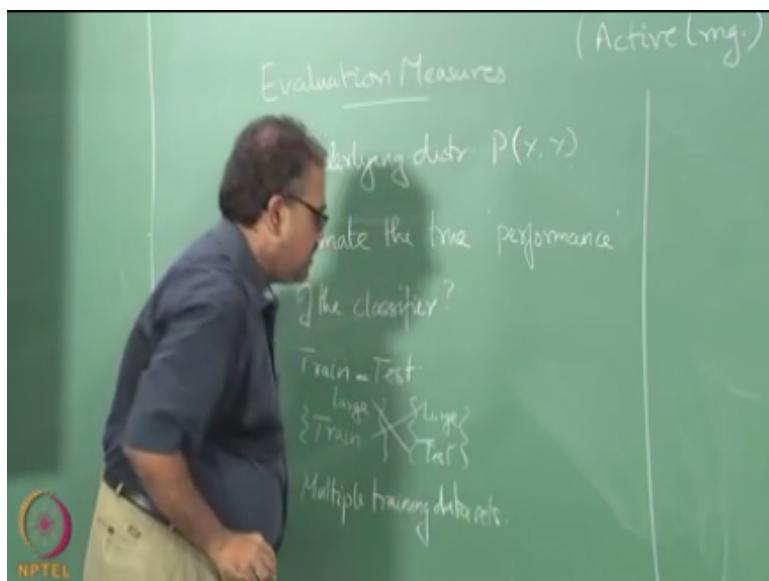
### Introduction to Machine Learning Lecture 48

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

#### Evaluation and Evaluation Measures I

Okay, so that leads us into some things to talk about.

(Refer Slide Time: 00:26)



I am going to talk about evaluation measures for a bit today so there is maybe like a little bit of hodgepodge class okay, I am going to talk about a few things which do not really fit into a bigger theme right, so far as evaluation is concerned we have spoken about some very standard think so far right, so in classification we talk about come on give me something yeah ,what would be the evaluation measure in classification, miss classification error then cross entropy then Gini index okay.

Gini index is not a evaluation measure right, Gini index is more of a parameter selection mechanism right, I can after and finish the classification I can actually compute some kind of a deviance right, and I can say okay this method giving me better deviance then that method and

things like that or the same thing for 0 1 error I can use squared error right squared error right, squared error of your thing to a target variable and anything else can I use, penalties in what sense, okay.

So there are a couple of things that we had to be careful about here, so there is something which I optimize right, to get to what I want right and there is something which I use for evaluating what I finally get if you limit yourselves to the classify supervised learning scenario right we are not even started unsupervised learning limit yourself to supervised learning scenario more often than not what we really want to evaluate ourselves with is the performance on the entire data distribution right, so I have some distribution of data right, I do not know the distribution apriori it derive go back to the very first class we started talking about something serious I mean not the one with the pictures right, the one with the Greek in it, right.

So in that class we talked about that being an underlying data distribution right, and that we did not know about this data distribution the only way we know anything about this data distribution is through training data points, right is through the samples that are given to us right, so there is an underlying so we had this distribution so what I am really interested in is finding out when you give me a classifier how well it is performing with respect to this distribution, right.

So in that sense how well it is performing I am not really interested in figuring out the square the ridge regression loss or anything like that right so I am using that to come up with a single classifier, but at the end of the day when I am looking at how well this classifier is performing with respect to the underlying distribution I have certain measures, so one of them is the 01 loss, so I do not care how we arrived with the classifier right I just want to look at the 01 loss and then I can do that right, that 01 loss gives me the miss classification error, right.

So ideally that is the evaluation measure that you should be using okay, so sometimes what people do because they are optimizing a different objective function they choose slightly different evaluation measures that may can make their method look better right, so squared error could be one evaluation measure, if you are doing classification 01 losses the measure that you should be looking at if you are doing regression while little tricky but square error is the most widely accepted measure for looking at regression.

But then you can look at other things also like deviance and other things you can use it for classification right, but having said that how do I estimate it is easier for me to write classifier or I will pick one I am going to pick classifier but some of what I talk about now works for regression as well right, and how do I estimate the true performance of the classifier, so what do I mean by that, I give you a sample data right I give you some sample drawn from  $P(x,y)$  right.

So based on the training that I that is all the information I have right, and I can use some of the data for training. I see find the parameters so how do I find out how good these parameters are or how good is my there are two questions to ask right, so the first question is how good are my parameters that I have found okay, the second question to ask is how good is the method that I use for finding the parameters is if you give me a slightly different data will I perform better or worse right, how will I perform right.

So I need to know something about the technique right, suppose I am proposing a new technique and I want you to go use it on your data later on right, but I should convince you that you can use it on whatever data you have right, so that means I have to convince you that my technique is good for finding the parameters this is the two things here for a given set of parameters I have to figure out how good they are right, and I also need to tell you how good my overall mechanism for finding these parameters are, right.

So for a given set of parameters how do you find out how good they are, on the training data I said good enough on the testing data, cross-validated okay, so if you okay, so we will get to this thing right, so one thing if we spoke about earlier was to split the data into train and test right, so if I estimate parameters on the training data again write it on the test data that will give you some performance, okay is that a good estimate of the true performance of the classifier.

Why not, Test data might not be independent of the training data okay, the training data may be biased I met a wharf the model then you are doomed right, no, that is actually a very, very valid point in fact that is something which you will face in real life right, but the assumption mostly we make in theory is that the training data that is given to you is a sufficiently representative data of the true distribution okay, that is not the case then you are doomed anyway, right.

So you assume that it is the, so in real life that happens what do you do, in real life that happens okay mean you cannot avoid it, you cannot just say okay, I am assuming it is a properly

representative sample of the underlying distribution then what do you do in that case, come on he is telling that you are not sample the entire range of  $P(x,y)$  so what is exactly so figure out where you are deficient right, so sometimes the most obvious thing is what you have to do, we have to sample more right.

But then do not blindly sample I mean of course you blindly sample you may actually return the same samples from whatever region you already have right, so what you should do is you should be more careful in how we do the sampling so you can use this is where you try to understand what the data is all about right, so you try to understand how the data is distributed the data that is given to you is distributed and figure out if there are parts of the input space which you believe we are important but are not covered in the data right, go and try to sample from that region, right.

So there are the different names for this okay, so one popular thing that people call this why this call, is called active learning, because I actively I am asking you for samples so I am not passively learning from the samples that were given to me okay, the learning algorithm comes back and says hey I want to know more about this part of the state space give me some samples from there right, I want to know more about this part of the input space give me so this is called active learning methods.

So your question this valid point of this discussion is yeah, one train one test is usually not a good idea right, so what do you do there are two things which you can do we can try to get multiple training sets from their data right, you can try to get multiple training sets from the data and try to so nobody is asking the obvious why are you getting multiple training sets from the data why not one large training set, okay why not pull everything together and then create one large training set.

Yeah, close yeah, so I will have to spend a whole week on weak classifiers right so we will come to that right, so it is a see that is one very, very amazing property that will look at later which probably the next class and later means pretty soon, on how you can take a lot of not so good classifiers right, which are just better than random of course it has to be better than random right, it cannot be worse than random so classifies that are just better than random and give you an accuracy of 51% okay.

So in the two class problem that is just better than random okay, I can take classifiers at giving accuracy 51% and they can produce arbitrarily powerful classifiers okay, it is an amazing, amazing insight that came about a couple of may decade and a half ago now maybe do not I am old more than two decades ago right, and it is one the girdle price and things like that is it an amazingly wonderful inside and we will talk about that right.

So that completely revolutionized machine learning once right, people then started saying so I really did not, do not have to build this super optimized classifier I can build a lot of this almost moronic classifiers but the operational word is almost right, I have a lot of them right, I have a lot of them and I will be able to do really well.

In fact in many, many applications that we have worked on right in real life where I have worked on with real data, I find it very hard to beat these kinds of classifiers you can think of whatever optimal classifier you want to come up with right, but beating these kind of groups of weak classifiers is actually very hard in practice alright so we will talk about that but no, that is not what I meant here I still have a point to make, right.

So yeah, so even if you do one large training set, to do one large one train one large test set right, you can get away with it provided largest large enough right, provided large is something that is dense in your input space right, if the large is so large that you essentially plaster your entire input space right, so any point in the input space if a pick there will be one point very close by in the training set that is what we mean by dense I mean there is an actual more mathematical characterization of dense.

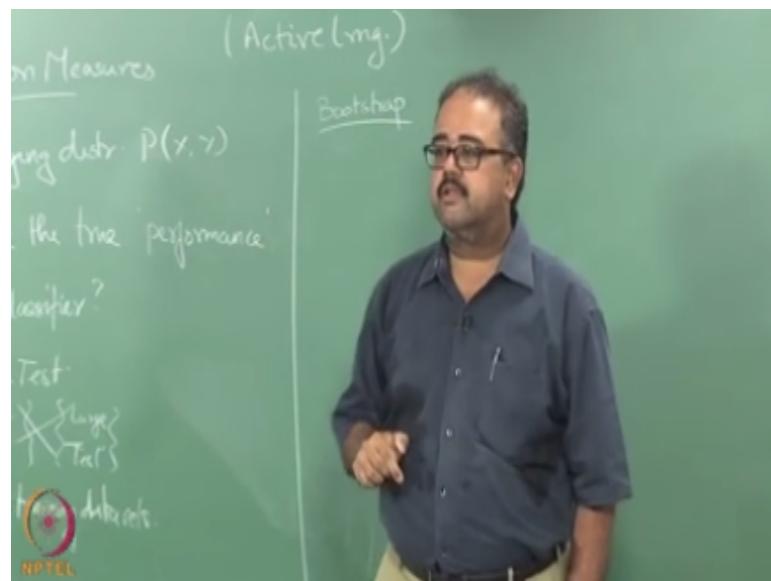
But if my training data is really dense in the input space then it is fine right, then I can get away with just doing one sample like one very large sample, but usually what is going to happen is you are not going to get such a large sample, so you are going to get a much smaller sample than that and therefore if you just use one sample and try to make an estimate okay, then the variance people remember what is the variance of the estimate, we talked about this again no, no that is unstable I am talking about variance yeah sorry, on data of similar the size we train a lot of models on data of similar size the parameter estimation I am going to make will be varying a lot right.

So it turns out that instead of doing one sample and then trying to train this if I take many, many samples right and then find the parameters on these samples individually right, and then take an average of those okay, turns out I can show that the variance will be lower in that case than what is the variance you are talking about the variance in the parameter that we are estimating, okay and what is the parameter we are talking about estimating here so here is the point where I am going to confuse all of you.

But the parameter I am talking about estimating here is the error is the miss classification error right, so I have the classifier right, what I am trying to measure is a miss classification error and that is what we have the whole discussion was all about I am trying to estimate the miss classification error right, so what I do is I start off with many samples of data right and on each sample I train a classifier separately again then I look at how the performance is on the test data and then take an average of all these performances and then I can tell you okay, if you give me a new data new set of data I expect to make this much error on the test data.

I am trying to figure out what the performance of the algorithm would be on a unseen train later I remember I was telling you I want to know how good my algorithm is right, so this way I can estimate the performance of the algorithm on the unseen data right. so there are many ways in which you can generate this, the many ways in as you can melt this multiple training data sets right, so many of these or I have strong roots in statistics and were typically designed in errors where the amount of data available to you was small, right. The amount of data available was small and we are trying to see how you can fake multiple data sets with a small amount of data okay, so the first technique is known as bootstrap okay.

(Refer Slide Time: 17:33)



So bootstrap is actually a very powerful statistical technique it is used in a variety of different places we will come back to another use of bootstrap a little later.

### IIT Madras Production

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**Introduction to Machine Learning**

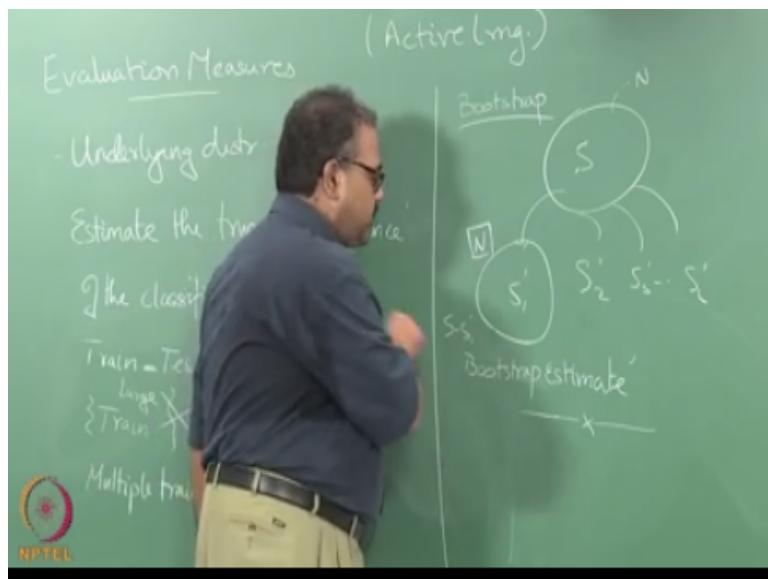
**Lecture 49**

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

**Evaluation and Evaluation Measures II:  
Bootstrapping and Cross Validation**

Right anyway so I will just talk about one news of bootstrap here so the idea behind bootstrap is very simple right so I have a large sample right I have a large sample of data so what I am going to do is I am going to sample from this data with replacement.

(Refer Slide Time: 00:43)



Let us assume that the set  $S$  has  $N$  elements in it so what I will do is I will sample from it another set  $S'$  that also has  $N$  elements then sound like a big thing right. I mean if I wrote sample with the if a sample without replacement I will essentially be duplicating it but I am sampling with replacement so what is the idea behind doing this no so if you remember I said that the

assumption that we are making is that exactly the assumption that we are making is that the data is truly representative of the underlying distribution right.

In which case given the data right the best approximation I can construct to the underlying distribution is the discrete distribution right defined on this data in one sense if I do not make any other assumptions all I can do is I can construct a discrete distribution on the data. It is a probability of sampling this point  $x_1$  is equal to the number of times  $x_1$  appeared in my state set  $S$  divided by the size of  $s$  right.

So how do I simulate this distribution sample from  $S$  with replacement is it make sense to people right I am going to assume that yes in some sense the set  $s$  is representative of the underlying data distribution and I am going to simulate the underlying gator distribution by using the discrete distribution form by  $S$ . So what do I mean with the discrete, discrete distribution I do not know may be the underlying  $p$  is actually Gaussian or whatever.

But my  $s$  has only  $N$  elements so only these end points will have some nonzero probability of occurring so that is what I mean by the discrete distribution so I can just construct the discrete distribution from is and I will sample from that that will give me a  $S'$  right so I will call its  $S'$  like that I can do that multiple times to get right I can go up to  $SL$  prime right I can do that you can create many many many such samples okay.

So now what I do to get a bootstrap estimate of the classification error so this kind of a sampling to produce this  $L$  sub subsets I have done is called bootstrap sample okay. So wonderful once I would derive such samples I can find out the bootstrap estimate of the quantity that I want right so in this case error so what will I do I will try on  $s_1$  prime right and what will I test on I will try non  $S_1$  prime and I will test on  $s - s_1$  prime right.

Because I am sampling with replacement so some of the data points will get left out right so whatever gets left out I will sample I will test on that so likewise I will try another classifier on  $s_2$  prime right using the same method if I am using back drop for training I will use back prop and train on  $s_1$  prime okay and test on  $s - s_1$  prime likewise I will try on  $s_2$  prime and test on  $s - s_2$  prime.

Like way so I will get how many estimates for the error  $L$  estimates I will take an average of that that gives me the bootstrap estimate for their right and you can show that the bootstrap estimate

will have a lower variance than just the error estimate on just using s and just randomly splitting into test and train it will have a lower variance right so what I go again--again I want to be clear what do I mean by lower variance here.

If I give you another training data point of size N right and then you do the same thing you do you do to estimates one just train on the original set that is given to you and test on the test set once or blue this bootstrap estimate likewise I give you another data point net the data set off size n another data set of size and so on so forth so now you have two estimates for each of these okay the second estimate will be more consistent than the first estimate.

That is what I mean by lower variance okay so that is a bootstrap estimate okay so this is sometimes I forget the exact number so its estimate of the error then but it is some parameter that i am estimating about the whole process that's why I said right so this is one way of estimating the performance right or thing right I mean you can estimate anything you want on s1 prime s 2 primes 3 prime s 4 prime you can do whatever right.

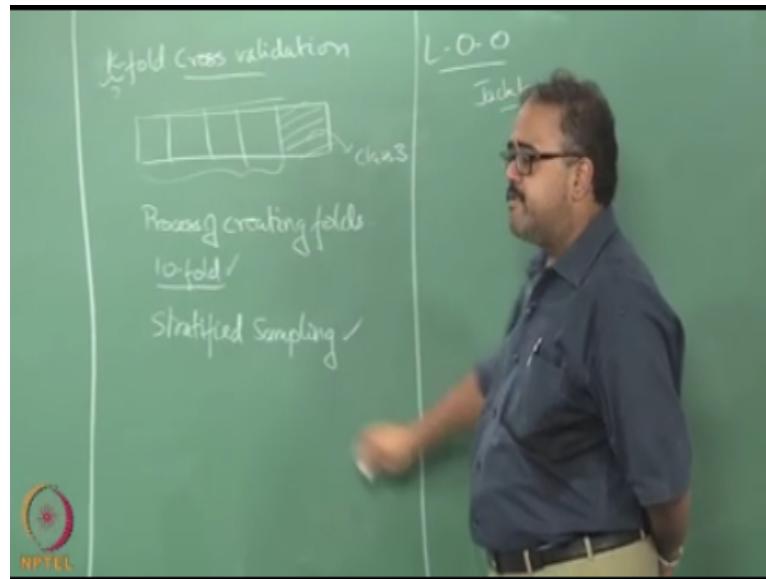
You can estimate the variance of the data on s1 prime just the variance office one prime not all here for anything that you can do this for each of those yell things and then you can have a bootstrap estimate of the variance right so you can do whatever--whatever statistics you want you can measure on these L sets right and then you can call that the bootstrap estimate okay so good stuff is a very general technique it is not just for error measurement.

Yeah so if you have been in--in a proper statistics course by now should have gotten into the variance reduction properties of bootstrap and we could have shown some interesting results right but I'm just going to tell you that I variance Goes Down and you can see intuitively it goes down and I will leave it at that okay so roughly about sixty--sixty nine percent of the data okay on an average rate sixty-nine percent of the data will be in s1 Prime.

And the remainder or sixty-three percent of the later sorry 63 will be in s1Prime and remainder will be in the test data right. So this is also sometimes called that ever get 0.632 bootstrap or something like that because sampling with replacement leaves a certain fraction of the data in the in the sample and leaves another fraction of the data in the test set okay. So it is sometimes that fraction also denotes what bootstrap estimated this right.

Remember okay so this is one way of doing it and this works fine provided I had a large enough sample to begin with right so it had large enough sample to begin with so suppose your sample is smaller suppose your sample was smaller you do something called cross-validation K fold cross-validation.

(Refer Slide Time: 08:25)



So what you do is you take your sample okay you divided into multiple bins and it divided into K bins what I do I train on some  $K-1$  bins and I test on the last bin right the clips I use the first  $k-1$  bins as my training data and one bin has the test data next what I do and servicing the last bin I use the second last bin as the test data and everything else has the training data right suppose I break this into k bins I will have k different estimates right.

I will take an average of those so which will give you a better estimate bootstrap or cross validation depends yeah okay that is not that is not the end of it we are done we want it depends on the size of the data I mean if you have a sufficiently large sample that your bootstrap assumption is true right so you might get a better estimate with bootstrap right one of the nice things about cross validation is that every data point is in the test set at least once.

Is it correct exactly once every data point is in the test set exactly once so in some sense whatever number that you are reporting is essentially the average over here performance on the entire sample that has been given to you right entire sample that has been given to you at some point you are using this so there are a few caveats that you have to worry about the first one is

what is K right and the second one is the actual process of creating folds so what should be k 5 to 10 yeah okay.

Yeah so those are the numbers typically used five or 10 okay those are the numbers typically used and so depending on the number of folds you have okay you have stronger variance reduction properties the more the number of folds the more the variance reduction property provided the folds are large enough that if you have K data points there is no point in creating k folds but people do okay.

That is a very special kind of cross-validation I will come to that later and I'm just going to leave you with the short form of it okay so we will come to that so which is exactly creating k folds if you have k data points it is called leave-one-out leave -one-out cross-validation okay leave one out that is what L over--over stands for right that essentially means that you will train on N- or K-1 data points in tests on one data point so in some cases.

No--no there are actually in there are in cases where this still gives you good estimates okay and earlier version of this was the one of the first do not ask me why it is called jackknife but one of the earliest is this kind of variance reduction technique used for a parameter estimation it was called jackknife a jackknife is very similar to leave one out okay. So going back here so I would recommend that do not split it into so many faults.

That you have very little data point left in each fold right and so the typical number do not do not go more than 10 folds right if you manage to get 10 folds out of your data right then you should just be happy right that gives you good enough variance reduction so typically people in a report empirical results do not expect you to do more than 10 folds right. But then if your data size is small right people end up doing only five folds right.

There are extraordinary cases even when your data size is not small when you have to do fewer folds right. So let us think for a minute suppose I have am solving classification problem right so I have created these folds right and this is entirely of class3 okay and there is no class3 in this data right. Since you think this is odd this can happen quite frequently if you are dumb about how we split to your data so the data has come to you sorted by class level.

I will give you the class well sorted by class level and you do the fivefold splitting by serial number right so what will happen is you will have all your class go into one fold right the other 4 folds will not have any data point of class3 now if you try to test on this what will happen yeah

right because you do not even know that exists class3 okay a training you didn't even know there was class3 so we are going to get 100 percent error right.

How do I avoid that shuffle the data in fact I do better than that I I do what I call what I call stratified sampling so I am not going to make much progress today so we will have to stop with the cross-correlation I have so much more that you need to talk about will do the next class so stratified sampling so stratified sampling essentially says that when you create the folds try to make sure that the class distribution that you had in the original data is maintained right.

Suppose I have five five data points of class 1 and 10 data points of class 2 right and I am splitting it into five folds right I should make sure that there are two data points of class two and one data point of class1 in every fold right so the class distribution the 2: 1 ratio is maintained in every fold. So this is called stratified sampling so this is something that you have to do so the recommendation is due 10 fold then you do stratified sampling.

And now can you answer my question why even though you have a lot of data you might be forced to have smaller number of folds class imbalance right so i might have very few data points of one class what if i have only 10 data points of one class right if I do tenfold sampling then forced at the cross-validation right I will essentially put one data point of that class into each fold that might not be sufficient for me to get a good good enough estimator.

And I might want to do a smaller number of force five may be 3 of course the other things which I could do but this is some case where you might want to you work with a fewer number of folds then what your data would suppose right so if you want to have a more formal description of cross-validation, bootstrap and leave one out and all of that you are encouraged to read has t write the elements of statistical learning book has very very nice discussion on all of these things right so right now I will stop here right.

### IIT Madras Production

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

# **NPTEL**

## **NPTEL ONLINE CERTIFICATION COURSE**

### **Introduction to Machine Learning**

#### **Lecture 50**

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

#### **2 Class Evaluation Measures**

Okay, good. So today we will look at some measures that are typically used in classification. And so, I will primarily focus on two class problems okay. The main reason is that more often than not the classification problems will encounter is for two classes, you know want to identify it as belonging to a class or not belonging to a class right so multiclass classification is little rarer to combine right.

And frankly two class classification has a lot more richer set of measures that people have proposed and usually the ones that we use for multiclass classification or extensions of these measures to multiclass okay.

(Refer Slide Time: 01:02)

		True Class		
		Predicted Class	+	-
True Class	+	True Pos. (TP)	False Neg. (FN)	
	-	False Pos. (FP)	True Neg. (TN)	
		$\frac{TP + TN}{N} = \text{Misclassification}$		
		$\frac{TP + TN}{N} = \text{Acc.}$		

So the first thing I want to introduce you to is something called the confusion matrix okay. So is nothing to do with the understanding of the course material or anything so far okay, it is something completely different right. So I say it is a 2 class problem, let us say the classes are 0 and 1 right. So I am going to say that, so I am going to form this matrix so the true classes are on the rows, the predicted classes are on the columns right.

I mean it does not matter if you do the transpose of this as long as you remember the meaning of the numbers that going here okay. Let us, I am going to change this slightly, so I am going to call class instead of calling them 0 and 1 as we have done so far okay. So I am going to call them positive and negative. So it makes it little easier for me, so what is the positive class, typically the class of interest to us right.

So what we will denote as positive class is the class of interest to us, what we denote as negative class is the class that we do not want right. So when I say that in some problem that the positive class is that person is suffering from dengue, it does not mean that dengue is something positive okay. It just means it is a class of interest to us, so just remember this okay. So when I say positive class it is a class of interest okay.

And more often than not your positive class in the population will be small right, hopefully right, I mean not too many of you have dengue right. So the positive class will be small and the negative class will be large right. So and we have to worry about getting the, what does it mean, of course yes. I mean that is my interest, so I am a doctor, I need patients to pay me right. So obviously the people were sick or more interesting to me then the people who are healthy right.

So that is the class of interest and the negative class is the other class, I am not so much interested in right. So things that go in here right, so with the true class is positive right and I am also predicting it to be positive okay, these things are called true positive okay, otherwise known as we will denote them as TP right, there are true positives okay. And what about these guys, they are true negatives right.

They are actually true class is negative and the predicted class is negative there are true negatives right. So that is why I said, you just had to remember which the true positive, which is the true negative and other things here. So whether you write it this way or whether you write the transpose it does not matter right, as long as you flip these three things, these two things around so what about these guys, false right.

So what about this, some of you know about true positive false for everything right, everyone has been telling me all of this, so where have we encountered this before, nowhere so glaringly obvious okay good, great.

So now what is the most common classification thing that we know about this classification error right so accuracy right so -1 the miss classification error is known as accuracy right what will be the miss classification error here where n is the total number of data points, right this is what miss classification so what about accuracy.

Yeah so n is the total number of radar points okay it is not the total number of negative points in case somebody is worried about that but I need a symbol for total number of negative points also I will introduce something later right this is known as accuracy right as you can see is 1- miss classification error right and what are the other things that we know of that are popular right so you can take a lot of different ratios of these numbers right and come up with different evaluation measures right.

(Refer Slide Time: 06:39)

$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$   
 $\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$   
 $\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FN}}$   
 $\text{Sensitivity} = \text{Recall}$



So I will talk about a few popular once so anyone know what prison is true positive so what does it mean so the classifier is going to tell you so many data points are of the positive class right the classifier is telling you so many data points are of the positive class how many of them are really positive right so the denominator is also those points which the classifier is telling are positive, right so true positive + the false positive are all the points that the classifier tells you are positive and the true positive are the once that are truly positive right this ratio give me the what is called the precision right.

So precision can be defined per class if you will if you want to rights so here I am doing this only for a two class thing that is why we are talking about positives but suppose I had k class problem I can treat it like a 2 class problem and I can give you the precision for any class, so suppose class I am interested in some class k okay I just keep that as the positive class and everything else as a negative class right now I can create a true positive true negatives, so I can actually talk about precision for class k precision class I and so and so forth.

It can for a multi class problem I can talk about precision right but typically this defined for 2 class problem, so there is a complementary measure for precision know as recall right so what is recall all is negative, so what is recall essentially there are so many positive points in the data right there are positive points in the data of these positive points what fraction is the classifier telling you are positive.

Right we get that all the positive points in the data which is true positive + false negative right how what fraction of it is the classifier telling you are positive so that is recall right so one way of thinking about it is so precision recall actually originate from information retrieval they are not originally for the classification domain they are originally proposed as a measures for evaluating information retrieval, so what do you mean by information retrieval so there is some repository of documents right.

Then you type in a search query right and then I give you back results corresponding to that search query right suppose there are 10 results that I give to you right of this 10 how many truly relevant to the query, what is that precisions suppose there are 50 documents in repository that are true relevant to the query how many of them appear in that 10 that is recall right so that is why it is called recall so how many of them to by actually recall from the repository so I have this huge repository of documents, how many of the truly relevant documents to by recall from the repository so that is why is call recall and procession you can see this is actually miss normal for people who are used to measurements right, what is precision in measurements sorry, how there is no not enough closeness in precision right.

No, elec guys come on how many elect guys are still in the class, 2,3,5,6,7,8 okay yeah, so what is precision sorry, how we elects valid precision if we go to measurements right, precession is essentially how many digits you are going to actually measure it to okay, that is precision you did not be accurate okay, the accuracy tells you how correct you are right, we can be less precise and very accurate and it can be very precise and very inaccurate I mean I can give you absolutely random number to 10 decimal places so I will be very precised, right.

But I cannot make any guarantees about the accuracy, but precession here is nothing to do with that okay, so precession here is essentially all to do with correctness, right so if I giving you 10 answers how many of them are correct. So why is this is a good measure, when this is a good measure already I told you an example, information retrieval right, so what characterizes information retrieval if you think about it, why cannot I use accuracy in information retrieval as a measure, why do I need something new, right.

Why cannot I use the following measure okay, I type in a query I give you back 10 results okay that means of the remaining documents okay, so I have rejected all of them right, so these 10 documents I have classified as being relevant to your query the remaining documents I have

classified as not relevant to your query, so among those I have classified as a relevant how many are truly relevant, among those I classified as not relevant, how many are truly not relevant and I can do an accuracy, right.

I can do miss classification error, so is that not a good measure for information retrieval. Yeah, so there will be like I said 50 documents at extra relevant to your query but you might have a 10 million document copy though if you Google you have several 100 billion documents as your copyrights right, and of which 10 or 15 might be relevant to your query right. So if I just use accuracy as a evaluation measure right, I need to be really, really precise in the measurement sense to make out any difference between two algorithms, because they mostly they will be correct.

And because of large fraction of the data I am going to say is irrelevant and I will be correct right, suppose I have two million documents and I am returning 10 to you right, and there are only 50 things that are relevant so basically I am right mostly right, a few things I miss here and there but I am correctly most of the time, because I have said large fraction of the irrelevant documents are truly irrelevant, right so that way I am good right, so that is not a good measure.

So when there is extreme class imbalance right, accuracy is not a good measure right, only 1% of your data is of positive class then if I say everything is negative class I am 99% accurate, but my precession will be what, you can define it to be 0 but in mathematically it will be undefined because I said everything is negative class I have no true positive, no false positive okay, so but you can define it to be 0 if you want and recall will also be 0 in that case, okay.

But quite often so what you will find is that if you try to increase you precision right, your recall will fall, when you try to increase your recall your precession will fall right, why is that so if you want to pull in more of the positive class right, if you try to pull in more of the positive if you want to predict suppose there are 40 documents that are relevant and I want to predict all 40 of them as a being relevant so the easiest way to ensure that is predict everything as being relevant, right so may recall will be very high.

Right which will be 100% right but the position will be too low depends on what is my universe of documents precision will be too low so and if you want to have very high precision what we

have to do predict only sure documents select no documents that go back and define zero base zero as one instead of defining it has zero right.

So we will define zero and zero minute ago and define it has one right because if undefined you can do whatever you want with it right so select no documents is saying as 100% precise obviously because you cannot point out the mistake I have made in giving documents back to you so we will recall and we will suffer because recall will be zero right.

So there is always precision and recall right so we have to figure out where you want to pitch your algorithm right so typically people draw what are known has PR curves right precision and recall curves and how do you think this PR curves look like, like this, like this yeah like that here it really want to be here right.

You do not have high precision and high recall but then you can compare algorithm so you compares again this PR curves right for example again let us go on I will tell you little bit more as we get long right so there is another measure which is especially popular in medical literature call specificity so what is specificity something different from what we are seeing so far it is what does it mean yeah so what does it mean what does its schematic of this.

So schematics of this if I say that something is if I have a high sensitivity sorry if I have high specificity it means that if I say something is negative then it is for truly it is really negative right so why it is good thing to have exactly right so this is very useful in medical test so I run the test and I said that you do not have malaria right or well must be topical you do not have dengue right.

Then you really should not have dengue it should not be say okay he said he does not have dengue but then really has as a 50% of chance he actually has dengue even though the test says that you do not have dengue okay that is the bad thing to have right so in such cases specificity is very important so if you are building a classifier right that predicts whether a person is suffering from a particular disease or not.

Then it should have a high specificity okay the flips side of specificity is right is a terminology that comes from medical domain right so sensitivity is true positive by true positive plus false positive and specificity is too negative by true negative by false negative okay it looks like at the other things that you leave out here okay this is specificity this is sensitivity.

The two measures that you just like a precision and recall and information retrieval in medical literature you have sensitivity and specificity okay sensitivity is just like precision well specificity is the opposite of that right so sensitivity says okay how likely you have to sensitivity is recall I think sensitivity is recall sorry yeah sensitivity is recall so it just says how likely are you to diagnose the disease right.

If there are so many patients with this particular disease how likely I am to find a patient with the disease right so what fraction of the patient is fully discover if there are the disease so that is the sensitivity and specificity is essentially if at all you do not have disease how likely is that you do not have disease okay right.

So for regression we already looked at a classify measure so more or less is the same thing right so basically you look at squared error right so all the interesting things are with the classification right for regression we look at a squared error or you can do an absolute error also if you want to evaluate the how good your regression fit is you can do absolute error we can do squared error whatever if we can use so one more thing which I want to talk to you about.

### **IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved



**Introduction to Machine Learning**

**Lecture 51**

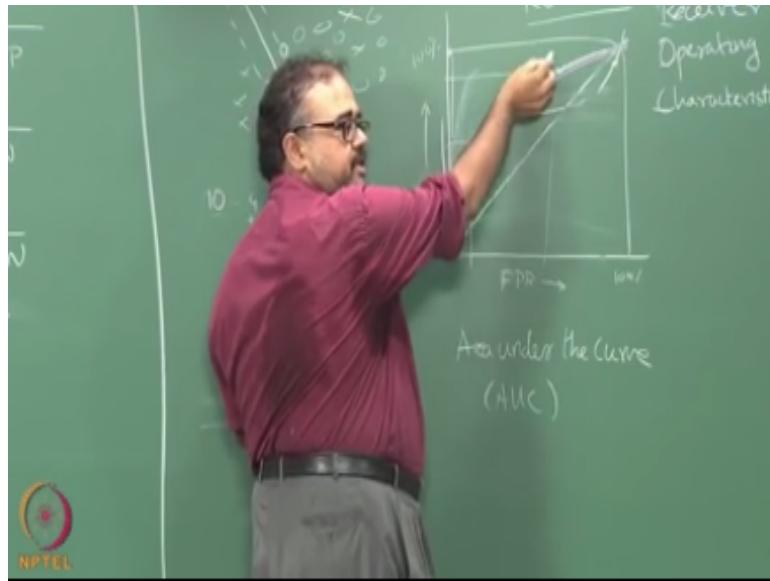
**Prof. Balaraman Ravindran  
Computer Science and Engineering  
Indian Institute of Technology Madras**

**The ROC Curve**

So when I build the classifier typically right, so what is going to happen if I'm whether I am building this hyper plane business or whether I am using a discriminant function right, there is an additional thing which I can use to tune how am I going to finally assign the labels right, so let us say I have a two class problem so then you can say that I learn a discriminant  $\Delta$  and typically what we do if  $\Delta$  is less than 0 we assign it to one class,  $\Delta$  is greater than 0 we assign it to another class, correct.

Yes ,so what if I tell you that no, no if  $\Delta$  is less than some  $\theta$  you assign it to one class if it's greater than  $\theta$  assign it to another class, right so what will this let us do let me move things around a little bit right, so essentially what will happen is I can by moving by changing this  $\theta$  I can take some points which are originally classified as positive and make them negative right, so one way or the other rights i just keep sliding  $\theta$  around, right, so essentially.

(Refer Slide Time: 01:38)



Right, so that is a line that marks where they are equal to 0 right, when I say that is greater than some  $\theta$  so that essentially means I could either have a line this side right or I could have a line that side right, and as you keep moving this what happens when you move here so this will become 0 that will become class O right this is already class 4 now this will additionally become class O at likewise when I move it that side now this will become class x right, so I can actually change where I pitch my line right and I can get different performance right.

So what is important here is okay, I have figured out that right but then I can move this that way or that way is little bit I can change me what is it whatever I want to do right, so I can increase the precision of my classifier by just saying okay I have learnt the classifier right I have learnt whatever it is i have learnt the hyper plane but instead of looking at 0 right, I will just move it a little bit that same right when looking at the point where the probability is 0 I will move it a little bit that side or that this side so that I can change my classification that I give right.

So given that I can do something like this right, so how do I know i have got a good classification right, how to put it I put it in another way right, so I am asking you to give me a classifier right, but I am not going to tell you right what is the precision I need or what is a recall I need from this classifier right, so you give me a classifier and then i am going to figure out what is the  $\theta$  I need to set so that i get a good precision or a good recall or a good the classification error whatever it is.

Next I want to be able to tune this and figure out where I am going to settle down right, so one way of summarizing all of this is something called the ROC curve right, so the x-axis is a true positive rate the y-axis is the false positive rate so what I mean by this so at any point I am going to look at how many true positives could I possibly get right, now that is my denominator how many of them have actually obtained that is my numerator, okay let us take a simple example.

Suppose I have, I say I have 10 data points okay I am going to make it very simple let us say I have 10 data points right and 4 are positive 6 are negative okay, 4 are positive 6 are negative right and I have a classifier okay, so of these 4 points it gives me 3 here and 1 here right, yeah, so I got it right okay, fine so it manages to tell me that 4 of the negative points are actually negative right, and 3 of the positive points are actually positive right and one of the positive points it classifies as negative and 2 of the negative points it classifies as positive right.

So the true positive rate for this is essentially  $3/4$  right, the false positive rate for this is false positive  $2/6$  there are totally 6 things which I can say as false positives right of which only 2 I get as false positive so I am not 2 bad right so that is the thing, so typically what you what you would want is maybe I flip this thing around right, is that TPR on this FPR and always get confused with this right, so when I make when I say something right I need to go up yeah, so that is TPR and this is FPR right, yeah so I ideally want a curve that goes like this okay.

So I should get a 100% true positive rate before I start saying anything is false positive right, does it make sense so here is 100% and here is 100% so at this point I am classifying everything as positive at this point I am classing classifying everything is positive right, so i will have a 100% true positive rate because i got everything is positive and I 100% false positive rate so everything that I could say is false, falsely positive I am saying is falsely positive.

So at this point I am classifying everything as positive right, but then what I would really like to see is as I when should my false positive rate start going up after I have achieved 100% in the true positive axis right, after I have classified every positive point as positive then if I still ask you to move your classifier more that side then I should start getting the negative points as positive right, so you should really go up all the way here before you start moving this way right.

And what about that guy this is essentially random behavior right, I mean so for every true positive guy I get a equal fraction of false guys as positive so essentially I am this flipping coins

and telling you whether it is positive or negative right, so I am just flipping coins and telling you that this essentially gives me that line right, so the further up you are if the probability of you saying positive is higher.

I am tossing a coin like you give me a data set it will give me a data point I will toss a coin and will tell you whether it is positive or negative okay, if the probability of it coming a positive is slow then it will be somewhere here the probability of it coming a positive is high then I will be somewhere here right, makes sense right so that is this line this is bad right, you want to be above that line you never want to be below that line right this is essentially random that you want to be above that line.

So typical curves you find will be something like this okay, so obviously you will not get that right ideally would like to get some curves like this right, make sense questions so far so the steeper this rise is the better it is for you right, so sometimes what happens you get curves that rise very steeply and then do that then a good or a bad ROC curve. Yeah, nobody said depends yeah, of course we all know the right answer is depends.

Depends on how yeah, what type of performance I am looking at right, so when will this be good right, so when I want to achieve a middling true positive rate right, so this is about middling rates about 50% true positive see that means of all the people with dangue I want to capture at least 50% of them okay, without putting too many people on quarantine right, so this is a good classifier.

But then if I see if you really want to get 90% right then this becomes unacceptably high false positive rates, well this might not be that bad right want to get 90% so this might not be that bad a false positive rate right, but then at this point this classifier is better so if i want for 50% false positive rate this classifier is better for 90% false positive rate this classifier is better, so there is no sure fire way of saying this is better that is better without knowing what you want from it just because I drew a curve that went below the random line really does not mean that this classifier is uniformly bad, right.

So if you want to show true dominance you have to show that one curve is above the other throughout right, this white line truly dominates the pink line that then now in such cases I can say white is better than pink right, but not in these cases in this case this curve is actually better

for some points some operating points it is not better for some other operating points, so nobody asked me what ROC stood for.

Yeah, receiver operating characteristics I think I wrote curve next to it ROC curve okay, so this essentially was used in olden days when people are talking about radars and things like that so false detection true detection versus false detection right, and then you choose your operating point right based on how much, how sensitive you wanted it to be do you want it to capture everything that came your way in which case it shows a different operating point for your detector right, so that is why these curves came about and you can use it for the same purpose in your machine learning evaluation right, you can use it to figure out which point in the space of parameters that you want your classified operatives okay, makes sense everyone get what ROC curves are all about okay.

So the thing with the ROC curves is unfortunately and people do not really use it this way when they have when they run experiments right, so what do they end up doing they do not want to look at the curve right they do not want to look at this curve and try to sit down and do an analysis and write papers because they want to run 100 of experiments they want to generate 100 such curves and they wanted an automatic way of comparing the curves,

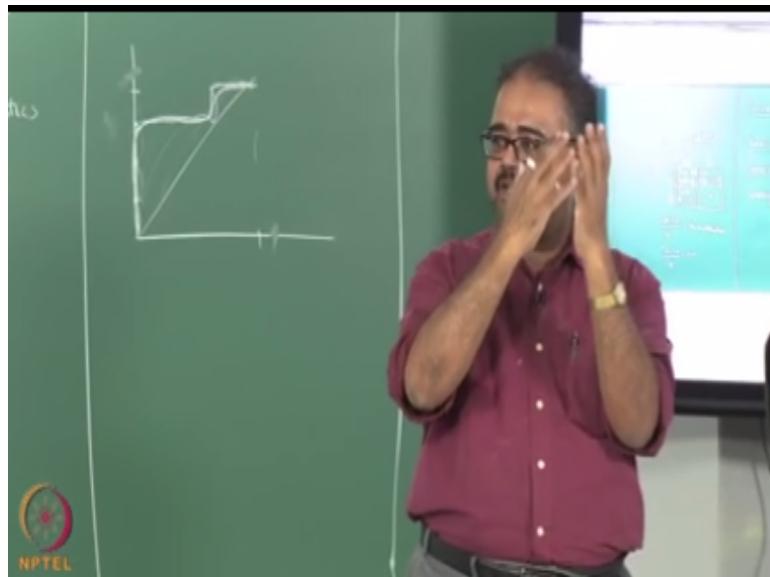
So what they did was they use this measure call, so area under the curve the uses measure called area under the curve or a AUC right, when they typically mean the area under the curve they do not mention it but they mean the ROC curve right, so the area under the curve is essentially this, so the assumption is the larger the area under the curve the better your classifier is because the ideal classifier is the one that goes like that right, so it gets all the positives before it gets a negative so the area under the curve is one for ideal classifier right.

And the area under the curve is 0.5 for random so if you are somewhere between 1 and 0.5 you are better than random so higher the value the better it is okay, all of this is fine provided all your curves are of similar shapes right, but you could get funky curves like this so what do you think went wrong here so the data is something like this right, so there are lots of 0 here and soon so forth and then are a few more x is that and then a few more 0 there right, so if I want to get those x as correct I have to get lot of the 0 as x's before I get those x is correct so that is essentially what is happening here.

I am getting more and more of the 0 both as  $x$  is here and then suddenly I get those last two pieces last two  $x$ 's and then I get them right, so that is essentially what this means so that your negative class okay, is lying between two bunches of positive data points so you are getting all the initial set of positive data points and then before you get the remaining positive data points you are getting a very, very large fraction of the negative data points and then you are getting the positive data points right.

So this could be an indication that your encoding feature encoding is wrong right or you need to go to a different dimension so you need to do an expansion of the dimension so that you can get all the positives before you get the negative ones, but people unfortunately do not actually look at the ROC in fact there is all this code bases for generating AUC directly you just done your experiments feed the data feed the classifier to your AUC generating package and then out comes the string of numbers nobody even plots the ROC and looks at it anymore right, so you can actually get insights about what is happening by looking at ROC okay.

(Refer Slide Time: 17:11)



So how would you actually go about plotting the ROC okay, so here is a very simple way of thinking about it right ,so I am going to take all these data points right and arrange them in descending order of their likelihood of being positive it could be anyway I choose to do the right so if I am going to slide this thing around okay, the farther I am from the hyper plane right, so the

less likely I will be further I am on that side from the hyper plane the less likely I'm going to be positive right so I will start off from here and then I will slowly increases or if I am doing a neural network I can look at the probability of the classification right.

I can look at what is the probability this data point is going to be positive right, so like that so whatever is it I am going to arrange it in descending order let us say that I do not need to do the particular class so let us say that I am re arranging it like this right, so this is the true classes I have arranged it in this order okay, now I choose a threshold above which I will classify this as positive right, so let us say choose a threshold here right, so what do I get no,  $1/4^{\text{th}}$  right, so I will just go up by  $1/4^{\text{th}}$  here right.

Next what do i do next I move it down one more data point then what we get up I move up by another one for if I see another plus I move up with another  $1/4^{\text{th}}$  likewise next down I will move up to another  $1/4^{\text{th}}$  okay, I am assuming all of this is  $1/4^{\text{th}}$  so that means that will be 1 okay, now what do I see say negative point what do I do, I do not go down I go right I go right by  $1/6$ , and I go right then another negative point I go right another negative point I go right, right and then then right, right ,yeah.

So my ROC curve is actually like that because I have only ten points it does not look like a curve anymore right this is all the estimates I can do but this is my curve right, that make sense right and this is my random in fact I should cross the random curve at some point because  $3/4^{\text{th}}$  what is that  $2/3$  okay that is fine I will still be about random okay that is fine right, so that is my, that is how I draw the ROC curve right, fairly simple right.

So you arrange it in descending order of it being positive right, so whenever you see a positive data point as you go down that list you keep going up whenever you see a negative data point you keep going right the step you go up is one by the number of positive data points step you go right this one by the number of negative data points. Once you do this now you can compute this in the end of this curve fairly easily. So the probability with which my classifier thinks this point is positive right, or whatever measure weight whatever measure I am using so it could be the distance from the hyper plane whatever is the measure the closer it is to the hyper plane the less likely it is to be positive right.

I mean the further it is from a hyper plane the more likely it is positive, so whatever is the thing I arrange it according to the criteria that I am using right or if using a discriminant function the larger the value for the positive things then the higher up the ranking it will be right, so like that and I just do this in descending order of what is the probability i think it will be positive or what is the likelihood i think it will be positive right.

So this is roc curve, any questions on ROC yeah that is why i said you start of from the leftmost end and then you keep ranking them according to whatever you start of from the one end of it and then you keep ranking them according to that when there are other ways of actually once you train a classifier when you do the hyper plane thingy there other ways are figuring out what the probability should be right, so and from that you can you do not have to actually shift the hyper plane around so you figure out what the probability of the classification will be and then arrange them in the descending order of that.

Because you can always say that okay, get my SVM to run give me the output give me the distance from the hyper plane for all the data points I will convert that to a probability and then I will use the probability for making the prediction if it is at least point three probable that it should be positive and classified as a positive class and in decision trees again it is easy to do probabilities so how will you do probability in decision trees, look at the leaves the data point lies in and look at the fraction of data points belonging to one class right, then you can do probability in decision trees.

So likewise SVM is the only thing that is tricky but you can there is a way of converting the distance from hyper planes into probabilities, anything else I want to say here so I am not sure if I am going to get to this later so let me actually make the note now there is a another of supervised learning problem right we talked about regression, we talked about classification that is a another problem called ranking problem learning to rank right, so it is inspired by information retrieval right but it is used in a variety of other settings right so I am not interested only knowing it is positive or negative okay, that is not the problem I want to know a ranking okay, so where is this appropriate.

So suppose I am yeah, so this is something which people do when they are trying to match protein structures so I will have one structure for a protein right, so and I want to figure out all proteins that look similar to this right which will probably have the same functions that as the

original protein that I'm looking at so that I am not interested in you telling me if there is similar or not similar I actually want you to rank it right, or there is another question.

So I want you to build a recommender system for me right, I want you to predict whether I like a movie or not like a movie right, but I do not want you to end up giving me like or not like I want you to give me a ranked list of movies so okay, this guy is coming in okay these are all his history last movies that he has seen okay, and he is an old guy so these are the set of movies I am going to recommend to him right, but in an order so this is what learning to rank means I am not just interested in yes or no answers but within the s i am interested in ranking order right.

It turns out that one of the ways of improving the performance in the ranking right, is not to look at this precision and recall and Felicity and sincerely in thing it is trying to directly optimize the AUC, trying to improve the things essentially what it means is the more relevant items you are trying to push up in the ranking right, so improving the AUC means what you are essentially making the curve go steeper and steeper right, improving the AUC means you have to make this curve rise steeper and steeper that is why this one will get a higher area under the curve right.

So that they essentially would correspond to making more and more of the positive points come higher up the ordering right, because this is a rank order right so you are trying to push these higher up their ordering so that essentially gives you the ranking effect.

### **IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)  
Copyrights Reserved

# **NPTEL**

## **NPTEL ONLINE CERTIFICATION COURSE**

### **Introduction to Machine Learning**

#### **Lecture 52**

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

#### **Minimum description length & Exploratory Analysis**

So in minimum description length principle the idea behind is very simple right among equally we already looked at it in some form or the other okay, so lasso is one way of thinking of it as minimum description right among all equally good classifiers I would like to pick the one that requires the least amount of bits to describe it right. So the description length should be as small as possible given that it has some acceptable levels of performance right.

So if you think about it so what does this tell us well if the classifier is very complex then I am going to need a lot of bits to describe it right if a neural network with a lot of weights the support vector machine with a lot of support vectors right or a decision tree with a lot of branches, so the more information you have the more detail the classifier is the more number of bits I will need to describe it right.

But then the better that it gets in performance right ideally why would you want to make it more complex only because it is making fewer errors right then you have to come up with some way of trading off this the description of the classifier versus the error it makes right ,you have to have to specify what the classifier is you mean you have to decide on how you want to specified suppose it is number of support vectors right so I will have to tell you what are the individual support vectors you remember support vectors or  $\alpha$  times the  $X_i$ 's right  $x_i y_i$  right.

So I need to tell you what the  $x_i$  are I need to tell you what the  $\alpha$  are for me to specify a SVM completely to you all right. So how many bits do I need for specifying those  $\alpha$  and the  $x_i$ , so the  $x_i y_i$  I can take it as a product and I do not have to describe the way I separately but I need something to describe that and maybe to describe the  $\alpha$  also and or maybe I can describe  $\alpha x_i$  to

you if you can use that somehow to produce the inner product but if I have it there the kernel version of it then I cannot do that right.

So I account pre multiply the  $\alpha$  into that right if it is a linear thing I can do this I can give you  $\alpha$   $x_i$  so there are things to think about it so how do you encode this right, so you want to write a program to implement SVM in end of the day right what is the point in me doing a learning algorithm and then not letting you use it right. So for me to communicate to you how we implement it I need to give you the description right.

And the second part is the errors are there right, so I make some mistakes right and I have to tell you on a training set let us say there is a fixed training set and on that training set I also want to tell you what are the errors I made right the smaller the amount of errors I make the lesser the number of bits  $n$  heat to tell you how many errors I made, makes sense. So for me to make small errors I met need a complex classifier that will need more bits for me to describe the classifier right.

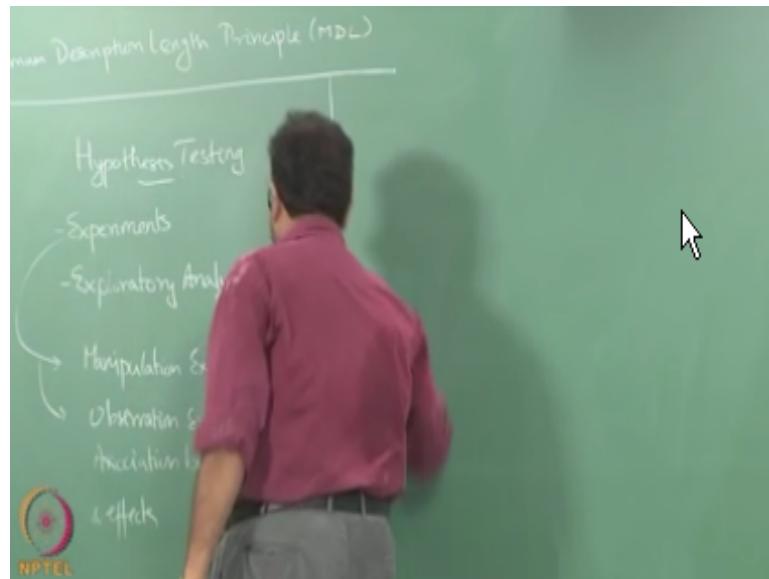
So if you want to reduce the number of bits I want to describe the error I might increase the number of bits I want for describing the classifier and vice versa, I am sorry just to set up a trade-off on equal footing right I am just talking about information on both sides now right. So that is I can make our classified arbitrarily complex right and we keep giving me better and better error right or I can make a classified very simple than it can give me a lot of error so how do I do the trade off ?

Between discrete the size of the classifier and the amount of error I make, so to put these things on an equal footing so that we can compare let people talk about the amount of information required in both cases you need some amount of information here and some amount of information here. So the more information you need here the lesser you need here, so that is the trade-off right.

So that is the idea behind minimum description length right and this huge theory behind it right it is actually a proper Bayesian approach and we talked about Bayesian learning at some point we also talked about ml and map estimates and so on so forth you can show that MDL is actually a proper Bayesian approach and people have derived a lot of complexity measures performance measures based on MDL right.

So I never talked about it earlier that but I think you guys are all ready to now read up on your own about MDL right so the brief introduction I gave should be sufficient right.

(Refer Slide Time: 04:54)



So if you think just one minutes stop and think about how people actually use machine learning right it is very heavily empirical right when doing a lot of math and other things or pseudo math in the class so far right but really at the end of the day when you start using it right it becomes heavily empirical it is actually a very applied subject believe it or not I mean of course you guys are all finding it out now with all the programming assignments but it is actually a very, very applied thing.

So whenever we have these kinds of empirical work right so you have to do our experiments right you really have to do experiments there is nothing like you know an analytical solution to your machine learning problem right. So when they give you a data set right you really have to experiment with the data to figure out what is it that you are going to do so all the theory and everything that we study now is all fine but when you actually get down to doing something you have to run experiments you have to do all kinds of things.

So you have to do experiments you also have to do some kind of some kind of exploratory analysis. So in fact we have not really talked about exploratory analysis at all in this in this course I have to do a lot of different things, so I actually do that whenever I teach my flavor of

data mining I do that so what do you do with external analysis right so there are many things that you have to do first you have to figure out.

So how distributed variables are you know so we have to figure out so what is the range of the variables I give you data right I do not do you do not really know what the data is all about I just give you an the simple form is I give you an excel file at the complex form is I give you like few terabytes of data on a disk right but then let us say I give you a file and then you have to figure out what are the different variables are there right and what kind of values do they take right and what is the variance of these values right or their outliers on these is that some values that I can ignore right.

So the whole bunch of things that you have to core and do some kind of exploration right or that way variables that are important to my prediction maybe know about then we talked about some variable feature selection and things like that but all of this you have to do essentially you have to understand the data before you even think of what is the machine learning algorithm I am going to use.

So if I give you some data you do not just straight away plug it into a decision tree algorithm or straightaway plug it into an SVM right so you have to go around try to understand what the data is all about right. So that is part of it is through exploratory analysis and say a little bit more later but as far as experiments are concerned typically they fall under typically they fall under two kinds of experiments.

So there is the manipulation experiments and observations experiments to who do you think this mean so in observation experiment I basically try to figure out correlations I try to figure out associations between variables right I would make a lot of observations and then I try to say okay if this variable whenever this variable was at this level then the output was at that level right so maybe I can observe the whether, so it is essentially trying to find associations between right associations between factors and effects.

So what do I do in manipulation experiments in manipulation experiments is essentially these are things where I have some control over some of the variables that constitute the experiment right, so typically I set up these kinds of manipulation experiments whenever I want to test a theory a

causal hypothesis right whenever we want to say that A causes B right. So I cannot stop yelled you know from happening right.

So I mean those kinds of that does not make any sense but I can say something like okay, so by learning algorithm A is better than my learning algorithm B right whenever the load on my system is high right. So I can actually make a hypothesis like this that A is better than B whenever the load on the system is high right, no I have made my models I have model A I have model B I want to make a statement that saying that yeah is better than B forget about under heavy load.

I want to make a statement that okay you have a learning algorithm A he has learning algorithm B I want to make a statement that learning algorithm A is better than learning algorithm B. So when will I make such a statement when can he make such a statement haha right, so that is the whole thing that we are going to worry about here I am going to test what I mean what I call.

### **IIT Madras production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

# **NPTEL**

## **NPTEL ONLINE CERTIFICATION COURSE**

### **Introduction to Machine Learning**

#### **Lecture 53**

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

### **Introduction to Hypothesis Testing**

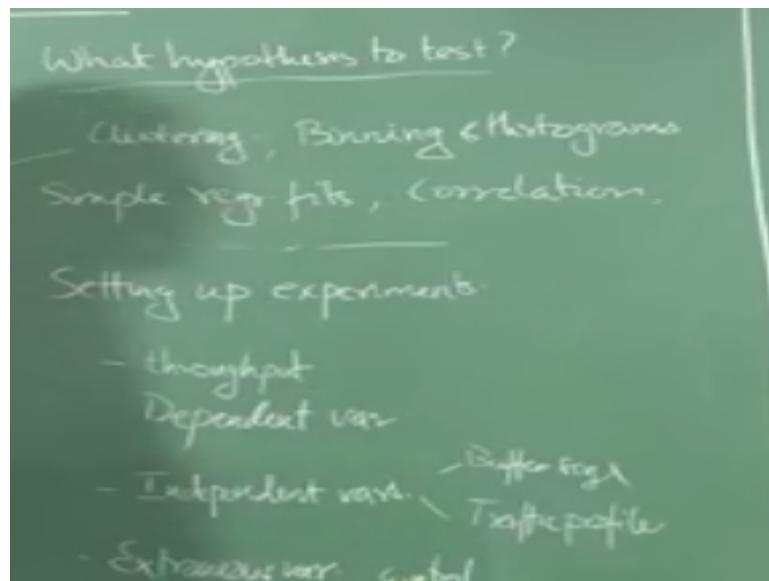
Okay let me stick with whatever I written down there, so let us say that I am trying to build an intrusion detection system right, I am trying to protect my computer network, so I am going to be looking at all the packets that are coming into my network I am going to say okay all these packets are fine these packets are not through the more block them right. I want to build a system like this right and then so I deploy your system in the IIT network for the first 15 days of a month okay, I know it will crash everything will be everything will be malicious or whatever.

But let us say this let us in an ideal world and then it catches like 84% of the malicious traffic okay and then from 16<sup>th</sup> to 38 I deploy his system it catches 87% of the malicious traffic is a system better than yours? Yeah, when can you say that? can you say something more than it depends, that is what we are going to do now looking that so that we are just going to look at a formal way of trying to say how sure you can be his system is better than your system. That is essentially what hypothesis testing let us you do okay.

It looks at the underlying data distribution that you are operating with and it should be able to tell you that okay with some confidence his system is better than your system okay. Typically what we do in hypothesis testing, we set the confidence level a priory okay, so unless with 95% confidence you can tell me that his system is better than his system I am not willing to buy it and I am just going to consider they are all the same system. I need at least 95% confidence for his system to be better only then I will accept it otherwise I am not going to accept. Because there is so much variability in the in the whole process that 95% is something which I can be comfortable with people usually ask for 99 right people, usually ask for 99% confidence because of the, because inherent uncertainty in the whole thing. So that is essentially what we are

going to look at, so how can we set up experiments okay so that we can answer such questions but before that we really need to know what experiments we need to set up right.

(Refer Slide Time: 02:30)



So like I already gave you two examples right I said that your system is better than his system that is first question right it is your system better than his system then the second question is your system better than his system under high load that is intuition direction right so traffic lot of traffic is coming right, so instead of so maybe your system is not very different from this system and the traffic is 10 Mbps right.

The traffic is 1 gbps maybe you start becoming better than you may be yours is a lighter system therefore you are able to respond faster and then his system starts dropping packets because of the heavy load so that could be a question right that that could very well be the thing so but then you have to think about it you have to figure out hey what is happening and then you can basically what you do in such cases is you observe the system like you have to make some kind of exploratory behavior.

And then you can say okay the mean number of packets I let through when it is at 10mbps is the same for both case but then mean but whatever some rough estimate I have seems to be slightly different when it is 1 gbps maybe I should run the more careful test will figure out which is which one is different whether it is with the high confidence whether it is different or not right so

that is a things like and there are other things which I could do right I can say that your algorithm is better when you run it with this parameter setting as opposed to that parameter settings.

When I say the parameters  $\theta_1$  versus when I say the parameters  $\theta_2$  at your algorithm is better when it is  $\theta_1$  versus when it is  $\theta_2$  so there is another question to ask or your algorithm is better than his algorithm when you use  $\theta_1$  so when do you get to these questions so that is where our exploratory analysis comes in right so you have to do some amount of X exploration with the data you have to talk to an expert right who understand this you have to ask you hey by the way will  $\theta$  being  $\theta_1$  versus  $\theta_2$  will it actually make a change to the performance.

And then that gave myself okay yeah maybe so maybe you should not throw out all the packets which are having parameter  $\theta_2$  maybe you should include them right so maybe that does mean that could be something really we can do all kinds of things so some of the simple things you do are well could do clustering right.

So what would clustering help you to find so helps you find how the data is clumped up right when you do clustering you can figure this out right you can figure out whether the data is coming from a single distribution or whether coming from a mixture distribution because you will find different clumps of data corresponding to the same class right so now this is all of you to tailor your classification choices accordingly okay so this is one thing that you could hope to get right in some cases in fact people use clustering to even generate the labels.

So I will give you a lot of data right I do not know I have not labeled the data into anything right but I can do clustering and figure out which are the major clusters and then okay there are three kinds of people in my customer base now I can build a classifier that will predict which of these when a new customer comes in I can make I can have it predict which one which category he belongs to so those kinds of things right.

I want to get some rough idea of the frequency of occurrences of features in my data right I can do some kind of simple bending on the features and I can build histograms that allow me to understand how often something's occur so if the data is concentrated suppose I do this thing and then I find said only some bins in the histogram I have very large numbers that essentially mean seven though my the feature can span a very large range it is only some very small values are actually present in the data.

So these kinds of observations I can make right so this will essentially help me do those kinds of things right and then I can do simple regression fits I can do simple regression fits and figure out if there is any turn to the data already that lets me to figure out whether I should be using a linear classifier or whether I should be doing something else where the data is more complex right and we already talked about correlation analysis you should do correlation analysis for what for throwing away features right we already talked about if the two features are highly correlated you should throw them away because otherwise it will lead to numerical instability in many of your algorithms right so apart from that you can actually use this correlation analysis to figure out what are the kinds of questions to ask as well I think about it right.

So once I know what is the hypothesis to test right once I know what is the hypothesis to test then I have to set up a proper experiment right so I have to set up a proper experiment so here I have to be very careful about right what is the question I am asking and which of the variables in the system okay are important for the question that I am asking which of the variables in the system are important for the question I am asking.

So for example that is stick with our intrusion detection system so I want a good intrusion detection system to have a high throughput right so as in when the things come in I should be able to put it out right it should have a high throughput let us say I want to test the throughput alone. I am not interested in the accuracy or anything. I just want to make sure that the traffic is not being delayed by the inserting the system.

So I can take this throughput I can make throughput as the variable of interest right or if you are looking at classification accuracy I can take classification accuracy as my variable of interest okay this is essentially known as the dependent variable there could be more than one dependent variable that you are interested in okay then I would have many independent variables right it could be the parameter  $\theta_1 \theta_2 \theta_3$  it could be something else right we are talking about throughput it could be something like a buffer size right.

Or if I am talking about classification accuracy it could be a variety of different parameters right so these are all independent variables of interest so independent variables could be something like buffer size traffic profile and so on so forth right and then there might be other variables okay call extraneous variables right, so for example time of day right so time of day can actually affect the network traffic significantly.

But there is nothing I can do about it right and more people are awake in the morning and they will be doing something in the morning and well more people are awake in the night and they will be doing something in the light and more people are less people are awake in the morning or they are in classes okay so that will affect the traffic right maybe that is not something I can control right so I am not going to I am not going to worry about it.

But whatever I do is whenever I do comparison between algorithm A algorithm B, I only do it during daytime or nighttime right so whatever these extraneous variables are I will control for them in the sense that I will make sure that they are the same that even though I cannot independently set it to whatever value I want, I will make sure that they are the same so that they do not affect the outcome of my experiment okay.

So there are extraneous variables for which we should control for, does that make sense? right so these are the things that we should look at right so there are dependent variable independent variables and then extraneous variable which we should make sure you are controlling against okay. Great, there could be other variables in the system right like temperature pressure humidity and all that which does not really affect your network thing this is maybe this if it is very hard people do not less likely to sleep in the night right.

Maybe you should control for that as well do this only on hot days or cold days okay matter has the second does not exist but yeah so that is essentially have to set up the proper experiment making sure you know what are the variables you are paying attention to so I mean all of this is very basic fundamental stuff which should all of you should learn in a proper design of experiments course right.

Once you are set up this experiment right you have to make sure you are avoiding any kind of spurious effects what do mean by few spurious effects the people know the floor effect and the ceiling effect floor effect as in they are close enough so suppose I am setting up an experiment to measure whether his algorithm is better than this algorithm right so and then let us say throughput again let us say take throughput.

And the traffic is flowing in at 10mbps right and your algorithm let us the traffic through a 10 Mbps how can you hope to be to match you can but you cannot beat so I do not know if it is better or not right so both of you can at best achieve 10mbps this is called the ceiling effect so

you might be capable of achieving 30mbps but I do not know that because 10 mbps is all that is there in the system so this is called the ceiling effect.

So likewise the floor effect is at the other end of it right so one of the main so I learned all of this in actually a empirical methods course when I was doing my PhD long time back right but the person who taught it was a very strong believer in avoiding ceiling effects so he used to set question papers which could never be completed in the time allotted for them, so there are no ceiling effects so there is always if you are good and you finish the question paper early mean if you finish 10 question early you are always another question for you to attempt right.

So people typically end up I mean you are the best person ends up finishing about 52 percent of his paper so you can see is an incurable optimist right I mean but he just wanted to make sure that there is no ceiling effect yeah so and likewise there are order effects you know the order in which you actually test things could matter so one example is not exactly an experimentation but very interesting effect that I thought I will mention right.

So when you when you are bargaining we are trying to bargain with somebody so the first thing that you put on the table right actually determines a path in which it is going to go right suppose you want to somebody is trying to sell you something the first thing you should go if you go and tell him okay I will pay a 10 rupees for it now he is going to feel a little bit bad about asking you for 50,000 rupees hey know is that this actually happens in you bargained in Bombay okay.

If you let that guy first give you the money he will say 50,000 rupees now you are going to feel bad about asking him for 10 rupees right so first I went with a friend of mine so he took us to some shop and there was his thing he said like I said what is this thing he said he has 3000 rupees I know I will give you 15 rupees for it okay no they actually bargain with this you know so I ended up buying it for something like 55 rupees.

So all order effects matter right so this is not really that but depending on which order you make measurements in all right yeah I mean there are other examples I can give but I thought this will be more funny anyway so those are things which you should avoid and there is a third thing which you should very be careful to avoid a sampling bias right suppose I want to know whether algorithm A is better or algorithm B is better in playing a particular game okay.

Then I look at the average moves that were taken across games that were one right and then I find that there is no statistical difference between algorithm A and algorithm B both algorithm a algorithm B win in similar number of moves right but I did a very big cardinal sin I made a sampling by said I only picks games which both of them one right, so a1 b1 so essentially this probably are simpler games right.

Both of them won and I am comparing them and so they all won in similar number of both I should actually be looking at all the games at they played right how many they won how many they lost so all of those things I should be comparing so I have to be making sure that the sample on which I am running these experiments or not biased in any particular way it is very important in fact quite often when people do all this phone in surveys and things like that right that is always this criticism of what the doing phone in service.

Like when somebody calls you and say he do ok are you going to vote for Modi or Rahul Gandhi in the next election so you give some answer right so why is this a bad survey to run you asked for me those people who have phone so I am most likely not going to vote so that is a different issue but you are asking people who have phone threat you are essentially skewing your sampling so you can say anything you want it I managed to control for income level so I only ask people who make so much money.

And so on so forth but then that still means you are leaving out a whole set of people with the same income level who do not have phones I mean right so when you could have very low income level and still have phones nowadays right so that does no correlation to having fun maybe it has correlation to how much you waste on the phone but the mere possession of a phone no longer has a correlation with many of the demographic factors.

But still there is something very selective about curly calling people have phones later than India any is any meaningful survey should be done door to door or straight to street and so on so forth there are complaints but you see many of the surveys that people put in all your magazines and things like that are mostly phone in surveys even women in India so in the U.S. it does not make a sense if it does not make a difference.

Because every household needs to have a phone right there the number of people who do not live in households is small in India that is not the case right so these are these are sampling by so this

things enter very often we do not even think about it we do not even think a second time about all the sampling bias that we introduced okay, so I will stop here.

**IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**NPTEL**

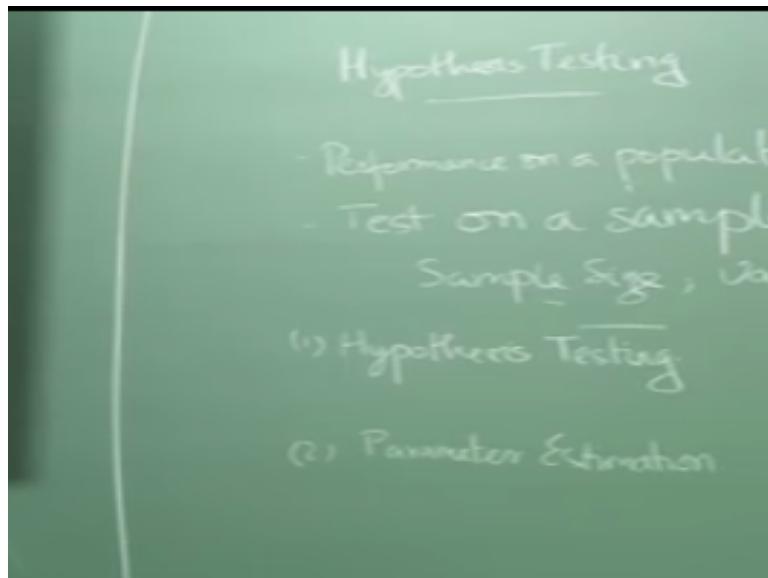
**NPTEL ONLINE CERTIFICATION COURSE**

**Introduction to Machine Learning**

**Lecture-54  
Hypothesis Testing – I**

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

(Refer Slide Time: 00:16)



Right, so last class we were looking at performance measures right and then I started talking about setting up experiments in order to say something about, in order to measure performance of algorithms and why you will want to setup experiments right, and being an empirical subject how experiments are very important right. And so the whole idea behind all of these experiments is that we really want to measure.

So we want to measure performance on a population right, I want to measure performance on a population, but all I get to do is test on a sample right. So whatever is the situation right, whether we do cross validation, or whether you do bootstrap, or whether you just set aside a validation

set, whatever it is, it is a sample that you are testing on. And what I am really interested in knowing is, how will my algorithm perform on the entire population as a whole right.

So I give you this  $P(x,y)$  right, so I want to know, well I do not give you the  $P(x,y)$  that is the whole problem right. So there is this  $P(x,y)$  and I want to know how the performance will be with respect to that underlying sampling distribution. And I do not have axis to that  $P(x,y)$ , therefore I will always be testing on a sample right. But I am really interested in performance on a population right. So what we are doing with hypothesis testing here, this is essentially trying to say that how much can you infer about the performance on the population from the test results on a sample right.

So how confident can you be that whatever you are getting as the test result on a sample is the performance on the population right. So that is essentially what we are trying to here right. So in statistics terminology the test on a sample right, gives you what is called a statistic right. And the performance on a population is in some sense that is a kind of a parameter that what is the average prediction error right, on the entire population.

So that is the parameter that you want to estimate and what you have is, what is the prediction error on a sample okay, that is the statistic okay. So the more common restriction that people can make is average versus mean right. So average is essentially a statistic right, so the mean is a performance thing right, it is actually over the entire distribution right. And you take samples and you take the average of the samples, you use that as the mean of the distribution right.

So we just use it as it is, but that is not correct right. So because when I take a sample average okay, there is some probability that it will be close to the true mean of the distribution right. So the statistic will be the average and the parameter that you are interested in would be the mean okay. So what are the factors that will influence this, how confident you can be about the parameters from the statistics?

Sample size is one, anything else? How? Yeah, no but I am going to take a lot of samples, somebody else said something else. No, no variance, who said variance? Yeah, so the variance of the underlying distribution right, so how variable is underlying distribution. So for that I probably have to compensate for that and I need to take a larger sample and things like that. So the variability also is assured right.

So this is something under my control, this is something that is not okay. So these are the things you should remember right. So we talked about two things that you wanted to do okay. So in the hypothesis testing, so what we are really interested in doing is actually answering some kind of yes or no questions right, I have an hypothesis okay. So my learning algorithm is better than the other learning algorithms. So algorithm 1 is better than algorithm 2, yes or no okay, right.

And I give you an answer, I say yes okay. I also would like to know what is the probability that the answer was wrong okay. So that is essentially what I am trying to do in hypothesis testing. So I will ask you an yes or no question right. So this question usually is of the following form, people have already done some amount of hypothesis testing and have it done something some null hypothesis alternate hypothesis reject one in favor of the other no yes okay.

Yeah people have done the course in terms would know this but apart from that nothing in Electrical signal processing no okay right so the basically is yes or no question will be of this following form right so I will have one basic assumption right which is both the algorithms or the same right and then have an alternate assumption which will say that algorithm 1 is better algorithm 2.

So the question I ask is should I aspect te3h basic assumption or should I reject it but not blindly reject it should I reject it in favor of the alternate assumption that I have right are they equal are is 1 better than 2, right I could also post my alternative question in different way I can say or they equal or they not equal okay so that confidence with which I can answers these two questions will be different for the same data right so 1 case the question was or they equal or is 1 better than 2.

So in other case the question was or they equal or they not equal so in the in both these cases the confidence with which I can answer this will be different for the same data that I have, right so we will see why that is the case as we go along but the questions will be of this form right so yes do you aspect this or do you reject this okay and if I choose to aspect this what is the probability that I was wrong okay.

So I do not want to aspect the something if the probability is to blow I will just basically say I am sorry I cannot say anything that is statistically sound about these two algorithms given the experiments that we have run okay you will give me some data it will say I cannot say something

statistically sound about this given whatever you have told me because the probability of me making errors is fairly large so how large is fairly large.

Yeah but typically I do not want to be even large than 5% okay usually I want to be even smaller 1% right why is that the case because as you will see which we go along we will be making a lot of approximation assumptions so that we can get things in tractable form so given that we are making so many assumptions we would at least the probability of error to be very small so that we can be assume of something good okay something reasonable.

Right so is that fine so we ask yes or no question and you look at the probability that is hypothesis testing the second one is parameter estimation right so here it is not enough for me to answer weather program 1 is better than program 2 right I want to know what is the average performance I program one right let us say it I just looking at running times okay so I have some program that is suppose to crutch a lot of numbers and give some output and I want to look at the running time of this program,

Right so I want to find out what is the average mean running time or the expected running time of the program on any sample given from a population, right but then I only have some 20 samples on which I run this program okay I can take the average of the running tine on this 20 samples but I want to know what will be the running time on any sample I give you from the population right.

So how like right how far away is this estimate on 20 samples from the true mean running time of this program so this is what we mean by parameter estimation right this is why I code sub here okay there slightly different usage here really not at very fundamental level they are not but at least they are very different from the way we have been using it so far right so we have been talking about when you say parameters we have been taking about like wait say in network or the alphas in support vector machine and so forth.

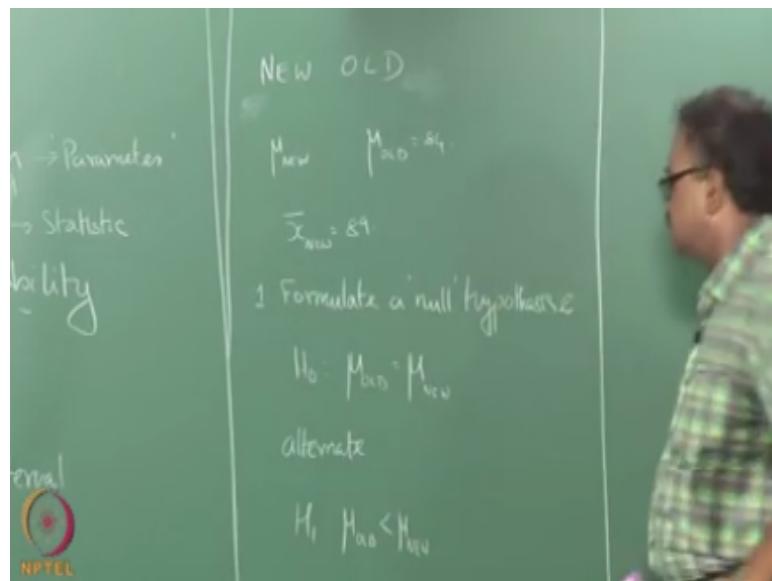
But here when you talk about parameter I actually mean the performance parameters that we are interested in right so the second thing which we want is essentially parameter estimation where I am looking at some kind of a interval right around my statistic, right and I should tell you that okay with some amount of confidence right the true parameter lies in this interval around the

statistics so it is like saying that okay so I run this all my tests on this sample data and I get the performance as say 3.3 seconds okay.

Then I will say it is  $3.3 \pm 0.5$  seconds okay so then the true mean will lie somewhere in that interval okay with the high probability right you can see that the 2 question are related so the first question again says how can I reject can I say 1 is better than 2 right second 1 I am saying no I want to know what exactly is a performance of 1 and in both cases I am looking at some kind of a confidence core of comparing these two does it make sense great I can repeat confidence core.

But I will tell you more about later okay that is the rest of the lecture is going to be telling you about how to get this confidence mission right. So I will repeat something which I gave as an example in the last class right, so let us say I have two algorithms right, I am going to call them new and old okay, so I have two algorithms.

(Refer Slide Time: 12:22)



Okay, this two algorithms new and old okay, so the old one is running for a while okay, the old one I have used the old for a while and I know I am running for a long time right, and I have some measure of how good the old ones performance is going to be right, so I know them mean

performance of the old algorithm because I have been running it for a long time right, and I also know the kind of this standard deviation of the performance because I have to chained a lot and lot of sample let us assume that right.

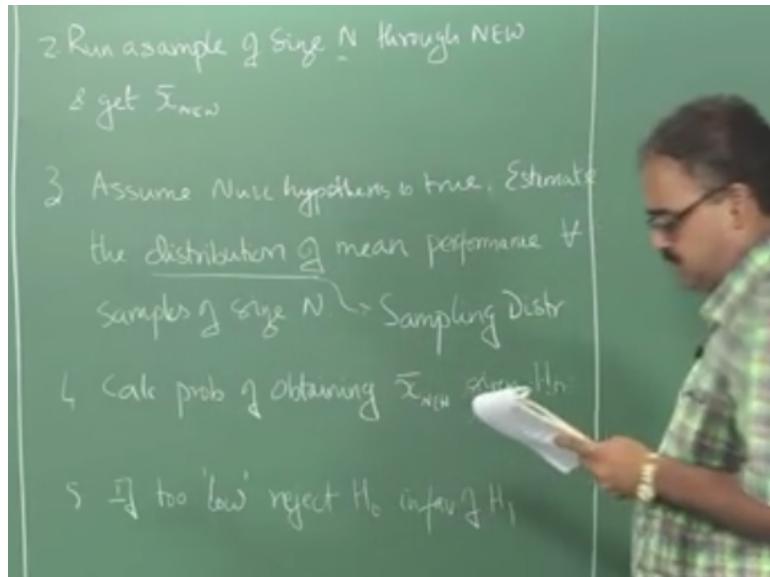
And the example I think I gave in the class so it is on intrusion detection right, I said there is some algorithm that has been running for a while right, and then it gives you some performance it catches and say 84% of all the intrusions right, and then I propose a new algorithm it is runs for like 10 days and it catches 87% of all the intrusions, so it is new one is better than the old one right, is it clear. So the question so the old one has a okay, I have numbers here.

So I do not have fully worked out things the old one has some 84% performance the new one has 89, so it is a new one better than the old right, does not matter this number do not really matter okay, do not get hang up on these just for illustration purposes. So I am going pose it a little more formally now right, so by so this is as we had 84 and 89 I am going to call these as  $\mu_{\text{new}}$  and  $\mu_{\text{old}}$  right, so I am sorry, I should be very careful okay.

And I do not know  $\mu_{\text{new}}$  okay, I only  $\mu_{\text{old}}$  I have some more estimated it to be 84 because I have a lot of experience with the old thing and  $\mu_{\text{new}}$  I do not know, what I do know is a statistic right, so I already  $\mu_{\text{new}}$  is do you know some  $\bar{x}_{\text{new}}$  which is 89, I have a statistic where it run it on some 10 samples and I know that the performance is 89 right. So now what we do is I formulate a hypothesis, right.

So what is the base hypothesis I am going to formulate, seriously I mean all of you have done some probability in statistic course right, okay did you guys did not do all of this, not in the PRP is it, okay I guess it is not a statistic course okay, it is probability and random process there is no statistic in it, okay fine. Because I did it in my very first maths course in under grade and I am not CS student so, right so you formulate a null hypothesis I am going to say  $\mu_{\text{old}} = \mu_{\text{new}}$  okay. Then I am going to formulate an alternate hypothesis okay.

(Refer Slide Time: 16:35)



So I get a statistics okay, I get one measurement of new right, of this right so  $x$  bar new so here is the question is sample of size  $N$  so that is the important thing that we have note here. So next thing I want to really figure out is suppose my null hypotheses is true, okay what is the probability that I would have got a performance of  $x$  bar new on a sample of size  $N$  sorry, this new  $\mu$  hold is less than  $\mu$  new I am sorry.

$\mu$  new is exactly equal to  $\mu$  new that essentially there is no difference in the two algorithms, greater than  $\mu$  new when you are do it okay proposing a new algorithm right at least you are assuming is not better that is a very settle point here right so the question I really want to ask is  $\mu$  new better than new old right. So if new was lesser than  $\mu$  old right what is the question I am asking can I accept the null hypothesis right or can I reject it in favor of the alternate hypothesis.

So that is the question I ask right so  $\mu$  new is actually less than  $\mu$  old as we will see when we go along then we will say that no I cannot reject it right I favor of the alternate hypothesis basically then we have to go back your test basically falls up here your basic assumption was wrong then you have to go back and redo in the test right. So a safer question to ask is new old not equal to  $\mu$  new but that not of interest to you right you really want to establish whether new is better than old or new is worse than old you do not want to know new is different from old.

That is not interesting question for you right so you remember yesterday or the last class I was telling you about it need to be very clear of what is the question you are asking in the experiment right. So running in experiment you need to be very clear out what is that we are looking for in

the experiment right, so for example so I am going to point you next to a really fantastic book on empirical methods in AI right.

Explaining all of these things to you which will usually the statistic book is in a very dry statistical sense right and it very mathematical sense they actually trick real experiments at they run on different kind of machine learning and AI settings right and then talk about introduce this topics very slowly to you right in fact I think I have already did one chapter from this book last class and today we are going to do another chapter from the book.

So I will just want it read it I am not going to the full book do not worry, but I did a course during my PHD I did a course entire course based on that book, so if this is person how quite told you does not when believes very strongly in feeling effects and the course itself was on empirical methods right and so he ask this dialog nice dialog that he ask in the book right so there are two people talking and then one guy say hey what are you trying to do? And he say then the researcher to replies I am trying to run this experiment I want to figure out if algorithm one faster than algorithm two okay.

Then he says how will do how do algorithm one is better than algorithm two, then he say how wily u do this? Then he says this not describing I am going to set up this experiment so that on this data set I will run this algorithm ten times and this data set I am run this algorithm and then I will make this measurements. So then he ask why are refined or why are you doing this experiment?

Again that guy is replied oh! I am trying to do this to figure out if algorithm one is faster than or algorithm one better than algorithm two okay. But then there is a other conversation between two people and they asking hey what are you doing this? Oh! I have this new method I heard this new method for estimating some significant of some biological markers I am trying to figure out whether this is better than that.

Then he says how are you going to do this? Then he goes about describing the experiment setup then he says, why are you doing this? And then he says, oh! I heard that this particular method uses technique X for doing this and therefore that is supposed to be better, so I am trying to figure out whether that assumption there on which this algorithm is based on right is that valid or not?

So there is very certain difference between the two conversation right the first one essentially that guy wants to know it is faster or not right, does not really have any deeper scientific that he is asking right. So in the second case this person actually has some other valid scientific question that they are asking and reducing experimentation as a way of answering this scientific question right and that is the really reason you should do this experiments not just for making measurements for measurement sake okay.

So it Is not directly related to your question but I have just using that as a excuse to talk about the story so you should be very careful about why you are setting up this experiments and what you are alternate hypothesis depending on how we studied of then you have to interpret the results you are getting okay.

Let us move on to point 3 okay assuming that you are null hypothesis is true right how lightly is it that you would have seen this performance this statistics  $x$  bar new right so assume null hypothesis is true and then you try to figure out how well the mean performance be distributed so if I run the algorithm so when you are old that assuming as null hypothesis is true so new or old should give me the same performance if I run this on sample of size  $N$  okay so I take sample one size of  $N$  run it I take a sample two of size  $n$  and I run the algorithm and I take sample three of size  $N$  and I run algorithm and so on so forth.

And for each of this I am going to get some average performance right so how well those averages be distributed right for every sample I draw I am going to get a different performance and how well that performance be distributed right so that is the question that I want to ask so I assume so I have to set up this distribution and that is called the sampling distribution okay so what is the sampling distribution again.

I heard some voice from somewhere I cannot locate who what is the sampling distribution it is the distribution of the mean performance on samples of data okay of the whatever algorithm we are talking about so it is the distribution of the mean of the performance on samples of data of a particular size if I change  $N$  the sampling distribution could also change right I am looking at a distribution right note that the means by themselves do not mean much okay.

So then the use the sampling distribution to calculate the probability of obtaining  $X$  new so once I have a distribution I can figure out what is the probability of seeing  $x$  new under this

distribution right okay so couple of things that we have to decide on here so the first thing is the most tricky part of all the hypothesis testing is how to come up with the sampling distribution right how do you come up with the sampling distribution that is the tricky part.

**IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

NPTEL ONLINE CERTIFICATION COURSE

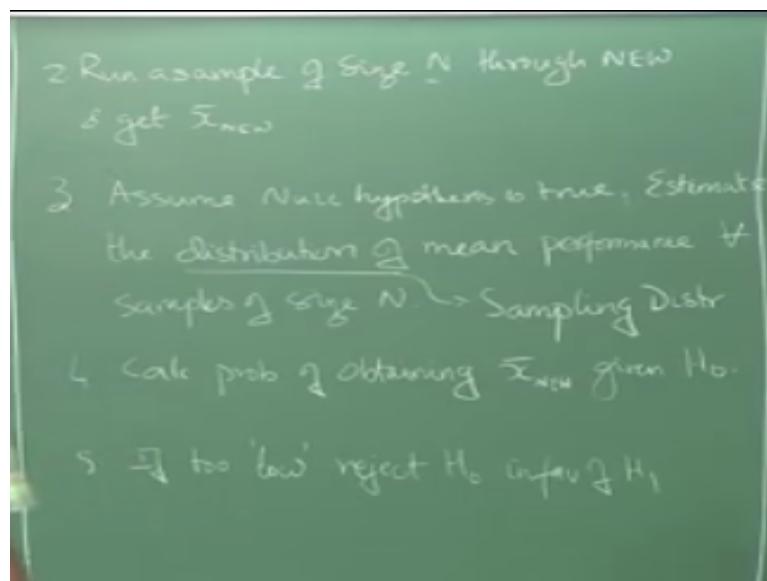
Introduction to Machine Learning

Lecture-55

Hypothesis Testing – II – Sampling Distributions & the Z test

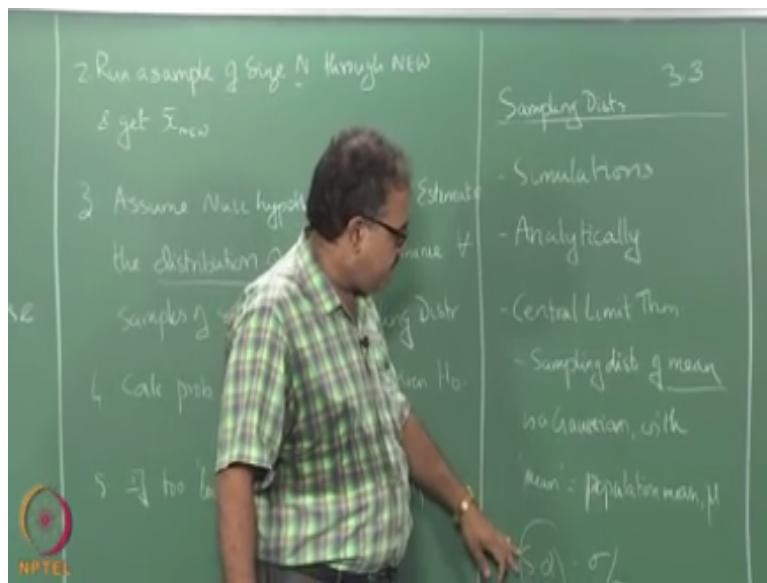
Prof. Balaraman Ravindran  
Computer Science and Engineering  
Indian Institute of Technology Madras

(Refer Slide Time: 00:16)



So I see as people what is your answer to that, well as yeah mostly partly see as people from one, what is your answer to that. Bootstrap right, I can do with some kind of simulations bootstrap does not really give me sampling distributions per say right. Yeah, bootstrap is one way of doing it, yeah you could do bootstrap, but you have to be careful about it.

(Refer Slide Time: 00:57)



So some kind of simulation based methods right, or I can do, I can try to find the sampling distribution analytically provided I have simple enough underlying data distributions and I know something about the underlying distributions, then I can find the sampling distribution analytically. For example, so if I want to look at let us say, let us take as a case, where I am tossing coins okay.

So I toss a coin 20 times and end up with 14 heads okay. So is the coin likely to be fair or not, 20 coins, 14 heads is it fair or not? You do not know, 20 coins 18 heads okay. See all of these are intuiting, how will you make it, now all of you know how to do this right. Now tell me how will you do it formally, think you will setup a null hypothesis where you will say that the probability of the coin coming up heads is 0.5, that is the null hypothesis.

The alternate hypothesis will be probability of the coin coming up which is greater than 0.5 okay. Now what I do is, I run a sample which is 20 tosses, I have found out that it is 14, so the statistics is 14 not  $\bar{x}$  it is not the average, but the statistics is 14 right. This whole process can be done for anything not just for mean right, so the statistic is 14. Now I am assuming the null hypothesis is true, which is 0.5 okay.

And I have to figure out what will be the distribution of the number of heads right. So what is the probability that I will get, 1 head, what is the probability I will get 0 heads if I toss 20 times, what is the probability I will get 1 head if I toss 20 times, what is the probability I will get 2

heads bla.. bla... bla., like that right, I compute all the probabilities right. Now I will look at the probability of obtaining 14 according to this distribution.

And if I do not like that probability right, then I can say that no, no rejection the null hypothesis right. If I like the probability I can say no, no accept the null hypothesis. There is no alternate here, alternate just say this is greater than 0.5 right. So in this case yes, you can accept, see this is why you have to be very careful about formulating the alternate hypothesis, because at the end of the day I am going to say accept the alternate hypothesis rejecting the null hypothesis.

But if the alternate hypothesis was it is less than 0.5 that is not a valid alternate hypothesis to make given the data that you have right. So it is nothing wrong with the whole hypothesis testing process it is something wrong with the where you setup the problem right. So if you are sure if it is higher or lower then there is a different issue, then you can say not equal to 0.5, and then run the test right.

But then it is up to you, if you have to use your understanding of the domain to come up with appropriate alternate hypothesis, there was another question here. Then the, yeah, yeah that is not, I mean if you do exploratory experiments you remember that, I told you in the previous class before you start your actual experiments before you do your actual experiments you do some exploratory analysis of the data right.

When you do the exploratory analysis you will get some idea okay, you start suspecting that okay, this coin is actually biased towards H, and then you will setup this experiment. So this is one very specific statistic that you gather while you are running the experiment, but before you setup the alternate hypothesis you should do some amount of exploration right. So you cannot walk in to an experiment blind about the domain.

So this is the practical issues that you should be aware off, you remember last class I told you very much about the need for exploratory analysis right, so we have to do exploratory analysis before you setup your actual experiment right. Yeah, so that I will talk about, it is just the basic structure I will talk about how about the confidence with that I will come to in a minute yeah right.

So but you know how to do this right, you can analytically you can figure out what is the proportion of heads you will get in 20 tosses right. So all of you know how to do the binomial

and then you can figure out what the probability is right. So the another way of doing it right just to give you another example of that right so in another way is to make use of specific properties of the parameter that you are trying to estimate in fact if you are looking to estimate mean right so we have one big advantage what is that? okay something called central limit theorem so what does the central limit theorem say?

Right, so since I draw samples I draw all independent samples right I draw samples of  $n$  variables I mean the size  $n$  right I draw samples of size  $n$  and these are drawn independently right I do not have any bases of I am sorry any bias from the previous samples I have drawn I am going to draw a lot of independent samples of  $n$  variables so if you think about the performance of the algorithm on this okay it is essentially independent samples I am drawing from the same similar distributed random variable right.

So essentially what central limit theorem tells us is regard less of the underlying distribution from which the data is drawn okay the sampling distribution will be Gaussian right will be normal distribution right it tells us that the sampling distribution will be a normal distribution and anything else exactly so the mean of the sampling distribution will be the mean of the population from which the samples are being drawn right so central limit theorem tells us that okay.

So I am call it the population mean as  $\mu$  right and the standard deviation is  $\sigma$  is the population standard deviation right so the thing to note is that this does not depend on the underlying distribution right regard less of what the pollution distribution is right the sampling distribution will be a Gaussian the mean will be the same as the population mean right and the standard deviation will; be  $\sigma / \sqrt{N}$  where  $\sigma$  is the population standard deviation right standard deviation will be the population  $\sigma / \sqrt{N}$ .

Right so the larger the sample size the narrower the sampling distribution does it make scene so the sampling distribution will always be centered around the population mean right so one thing which we can control is how wide is the sampling distribution right so if  $n$  is small then the sampling distribution is wide n Is large the sampling distribution is narrow okay.

So unfortunately we only have a central limit theorem for mean right so this is what said the sampling distribution of the mean okay so I just want to stop once and just sorry if it is a getting

too reparative just once more to emphasize what I mean by sampling distribution of mean what does it mean to many means remains me of Kamala Hassan movie distribution of the sample means of the data so that is what I mean by sampling distribution of means's okay is that clear.

Right because I have seen people get confused and give all kinds of different interpretation of that so I have the sample data so I have taken I readily sample data of size  $n$  from the population right and at compute the mean of this samples and distribution of that means is the sampling distribution of means I see great so this standard deviation of this sampling distribution is also sometimes called the standard error of the mean is the standard deviation of the sampling distribution.

Okay so empirically right we can say that  $n$  greater than equal to 30 indicates  $n$  is large enough right your standard deviation standard error becomes small right the standard error becomes small what happens so any sample of size  $n$  I can take and estimate the statistics and I am more or less correct with the very high probability I will be correct so that is what it means. But okay, you have to be careful about what is correct.

Okay, so couple of savages here so if you are, if your population has a high standard deviation so what you have to do, you have to make your sample size very large so that your standard error comes down right, further if you variance of your thing is very high so underlying population is very high right, that means you require very large samples right, so what you should be thinking at about at that point is to see if there are some other way you can step up the test, right.

So you should not come to a point where I need millions of sample just to reduce my standard error. I am sorry, yeah is it that is what I am saying you have to think of some other way of going about doing this rather than just saying that increase the number of samples right, so what would be other ways of doing it is to try to bin data right, so if you bin data what happens is so small variations in the data will go away right so something like 3., 2.3, 3.4, 5.6 all of them will can be bin to say 3.5 right, that means a small variations will go away so lesser amounts of noise right.

So the variance will also drop a bit, so you can do things like this they can do some kind of noise reduction techniques, try to do is see if we can reduce the variance in the data without actually dropping anything important so that is curial, right so that is a kind of things you have to try so that is one caveat that I wanted to say.

(Refer Slide Time: 14:15)



So the next we look at is specifically using the sampling distribution of mean right, and come with something call the Z test okay, the people know about the Z test right, okay, so again okay let us do another example here okay, so I am giving you, you know how to solve a certain kind of problem right, so you have been train to solve of solve these problems for you know years and years and shuffle like that, like you do in your JEE preparation kind of things right, and then I know how much I would expect student to score on a specific set of problems, right.

Then suddenly I find that a new kind of problem comes up right, and the students are taking longer to solve these problems right, let us say that the students takes a unit time to solve problem traditionally now they are taking say 2.8 times that to solve the problem right. So my claim is that this new problems are harder than usual right, okay how will I verify that right, so again I am just going to walk you through this setting up this hypothesis right.

So basically I am going to start off with saying that so I have the easy problems, I have the hard problems now null hypothesis is there are no different okay, I am going to say they both take unit time to solve, right. So on the second is the alternate hypothesis right, so the easy problems take lesser time than the hard problems to solve, is it fine. So what I know from previous data is that okay, just some number say the actual numbers are less important but just the process here, right.

And then so what now as it is said set is up I take 25 hard problems I ask the students to solve it and I get my okay, so we are looking at mean so the sampling distribution is going to be

Gaussian right, what will be the mean of the sampling distribution. You do not have to do that yeah okay, so now what I am going to do is to a little trick now I have the sampling distribution right so I need to know what is the probability of seeing 2.8 under the sampling distribution.

So the mean is 1. the standard deviation is 0.19 right, and there is 2.8 I need to know what is the probability of c 2.8 under the sampling distribution right. So I can do that lot of you know how to find the probability here right so this is mean is one and this is so many standard deviation above the mean right.

So the probability is actually very small okay, but you can also do this in a straightly both convenient fashion so you basically points and something call the z score let us say essentially assuming that you know sampling distribution is going to be 0 mean and unit variance so what you do this? This is called the standard normal so the standard Gaussian so I am going to convert my sampling distribution in to standard Gaussian and then try to find the probability of 2.8 in that standard Gaussian right.

So essentially the z score is right so that will be the z score so essentially I take my actual statistic subtract the mean from that right and then divide by the variance, so this case will be unit variance. that gives me the 0 mean right. So it is 9.47 so essentially what it says is this guy is 9.47 standard deviations above my mean right so what is the probability of that happening? Very, very small right so what we do? Well we do not know that, so en we do is essentially looking at some standard values that we do now right.

What is the probability of something lying greater than  $1.645 \sigma$  above the mean, this number you should by heart it if you know at some point 1.65 is your friend right, so implies the probability of that happening is yeah, so basically what it means is right so this side is 95% right so if we take the Gaussian take this is mean right take  $1.65 \sigma$  above the Gaussian right the area to the curve to the right of that is 5% of the total area right.

Likewise right area here is 5% right so it is number we need to know, so why you are interested in 0.0 to 5.2 tail right sometimes we might what do? Not look at greater than we might want to look at not equal to right. Now I have make the, my hypothesis was the second set of problems was harder than the first set of problems, if my hypothesis had been if the second set of problems are different from the first set of problems then I will have new is not equal to  $\mu$  old right.

So in such cases I should have been looking at 0.196 because I could go on either side right so I have to be make sure that my statistic is greater than either +0.196 or lesser than -0.196 for me to be sure that I will be wrong only 5 % of the times right. So in the olden days we actually use to have a z table that use to tell you for one side a test what should be the z statistic you should look at for 2 sided test what is should be z statistic which you should look at so on and so forth.

If you people have actually looked at clacks table ever you actually have a Z table as part of the clacks table and this is essentially what the Z table is telling you,  $\mu_e$  is the, the easy task we had right that is under the null hypothesis that is the population mean yeah. It is 9.47 right I mean 9.47 is way, way higher than my 1.65 right so therefore I can reject the null hypothesis right at a confidence level the way we straight this like this I can reject the null hypothesis at a confidence level of at least .05 so that means that probability of making me error in rejecting the null hypothesis is less than .05 so the conclusion is that yes it is harder.

So I accepted the alternative hypothesis right in rejecting the null hypothesis and the probability of making me an error is less than .05 okay yeah it could say that but confidence level means something else okay so that I why the statisticians are very careful about the conclusion that they will draw from this.

I will write down the conclusion okay you can say p level or p value right but the sensity the p value of something means the probability of P being wrong in that conclusion okay so probability is wrong in that conclusion given all the assumptions I have made see this is the reason why we want a very high things right we are assuming our sample size is large enough for central limit theorem to apply right.

And then whole bunch of things that we are doing right the point is here right there is no notion of the accuracy here so the probability of making me an error in concluding that H0 is not write is 5% okay so if you want the whole answers of confidence is 95% I recommend you to read the book so it is not you cannot really say confidence level is 95% right.

So I can just say the probability of error is I will come back later end tell you the confidence when I talk about confidence intervals at this point I have just leave it out right so therefore we can reject the null hypothesis so this is good right yeah this is Z test yeah and in the null hypothesis they are the same right.

There is a whole idea right under the null hypothesis what the sampling distribution is? That is ht we are trying to under the null hypothesis is being true what is the sampling distribution is what we are trying to find right so they are the same level when null hypothesis is true right so typical p value is that run these things alright fine so if we are looking at more critical domain like medical domain and I would expect the p value of .001 right not.005 however so you have to be more careful really sure what, what kind of.

### **IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

# NPTEL

## NPTEL ONLINE CERTIFICATION COURSE

### Introduction to Machine Learning

#### Lecture-56 Hypothesis Testing III

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

So what do you do when so here we assume that we knew the population standard deviation right so what if you do not know the population standard deviation, sorry go for the t-test do you can still do the z-test? You try to estimate the sample standard deviation okay you can estimate the sample standard deviation instead of knowing the true population standard deviation assume that the sample standard deviation is correct right. And then go ahead and do the z-test result okay so that is essentially what you do.

But what you do even the mean is unknown can you say something useful if I do not know anything about the mean of the old systems is that is that even a question that you can ask right you can so you can just say that a I am going to hypothesize that the running time on this new problem should be two standard units and under is 2.8, so can I say that the new problems are things that will with confidence okay.

Can I say that they will take more than two standard units to compute and I can make an assumption about what the mean should be of what I think is the baseline case and then you can compare against the assumed mean right so you can still use the z-test I we do not be in a hurry to abandon the z-test because it is still useful right, so you do the t-test let any questions about this I mean so if you do not have the standard deviation just run some samples and deviation tests estimates.

Clip code you do not do that so essentially what would you have to do in that case is okay what is the probability that these ten different measurements I made gives me I mean I would have generated all these ten measurements from this right then I my sampling distribution becomes slightly different so I have a set often samples that I have drawn right and what is the probability

that these ten samples will turn up exactly this fashion can you imagine how horrendous that sampling distribution will be right.

So if you can if you have an easy way of computing the sampling distribution which you could write you can because with all this simulation based ideas you can set up arbitrarily complex sampling distribution the reason we have to stick to a simplified simple terms the sampling distribution because that is what central limit theorem gives us right so if you are happy to do this in a simulation you can set up the sampling distribution using simulation.

Assuming you have access to ways of generating many samples from the underlying data right so if you are just doing it bootstrap then you run into problems right because the sampling is no longer independent there maybe do some large  $n$  samples they are repeatedly sampling from that so the sampling is no longer independent and you gotta dress for that right but if you truly have a way of sampling from the underlying data right.

You can set up the sampling distribution you wanted right so people understood this question this question was why did I have only one  $\bar{x}$  right why cannot a sample  $\bar{x}$  on multiple sample so why cannot I just compute this on multiple samples and figure out right so the answer is yes you can but what do you have to do is you have to find out suppose you do this 10 times right now I will have ten numbers you have to find out what is the probability that I could have drawn all of these ten numbers under the old under the null hypothesis right.

So if I can find that out okay now I will need a different quote unquote sampling distribution for this right so if I have a way of constructing that sampling distribution then I can run the test and one way of constructing the samples sampling distribution is through simulations just keep drawing many samples sets of sets of ten samples right and then and then look at the distribution of those and then try to form industry formula estimates from that right.

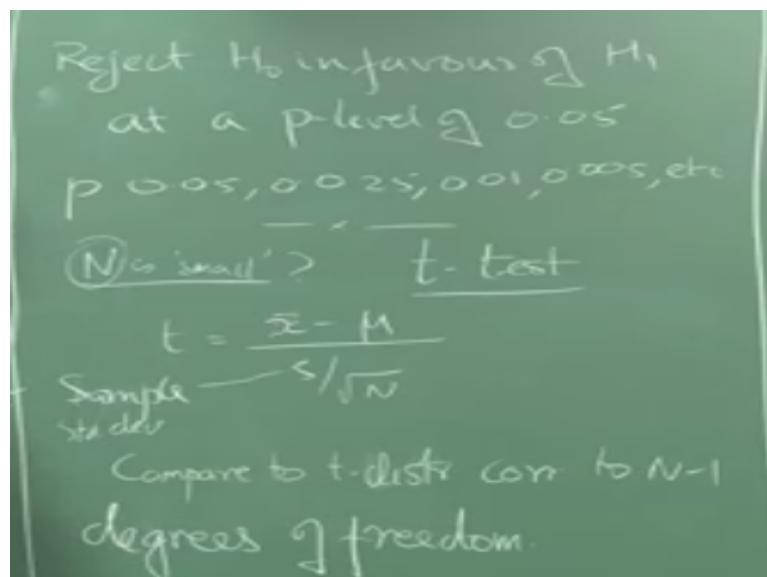
Yeah that is what I said that so you have to actually draw samples from whatever samples that you have right estimate the sample variance right so in fact if I do this right if I have this  $\bar{x}$  as 2.8 okay I can estimate the sample variance of that also let and assume that variance is the variance of the population, so instead of using  $\sigma$  here I will use the sample variance here right and divided by  $\sqrt{n}$  and use that as my denominator in the Z statistic okay.

Okay this is something which have forgot to mention sorry about that in all of the things that I am talking about today right the assumption is that the means are different and I am testing for

the difference in means but I am assuming that the standard deviation is the same across the new and the old right that is why I can estimate the standard deviation on the new data and I can still use it as the population standard deviation right.

And we are assuming only that the means are different right so there are the whole class of statistical tests that you can run if you assume that variances also are significantly different and you want to estimate the variance right and this broadly fall under the class of algorithms known as anova, anova sent for analysis of variance that is what I am not going to get into anova methods just the usual case we are assuming means are different right.

(Refer Slide Time: 06:45)



So what if n is small how small is small so that goes to work 4:30 today how small is small I am just getting starting with this test will have another hours of us material how small is small less than 10:30 okay no we are talking about the N here okay n yeah integers piece yeah right n is small so if you should think about it right it turns out that the central limit theorem works fine only if the sample sizes are reasonably large right.

Suppose my sample size to say five but my sample size is ten right then it is no longer clear that I can use the central limit theorem so the sampling distribution might not really be Gaussian it turns out that the sampling distribution is slightly different version of Gaussian they are heavier tailed right there is more probability mass in the tails then you would have in the Gaussian no, so

the Gaussian is actually of a specific form right  $e^{-(x-\mu)/\sigma^2}$  line so this is not so for the same mean and  $\sigma$  values this will actually be flattered okay.

So this distribution is called the T distribution or more correctly the students T distribution okay. So people know why it is called the students T distribution in a way okay. So that is a person there is a very famous statistician whose name no escapes me but he used to work in a brewery in England you know the place where they make whiskey and things like that right.

He was one of those people who was in charge of making sure that the whiskey that was being produced were of the, was of the same quality where there is not too much variance in the in the visible in the quality of the whiskey listing so there is not too much difference in the alcohol levels and things like that right and so he came up with all kinds of interesting statistical tests for figuring out known is a serious thing it is a serious application I mean infact something which will people pay you for right.

I mean for solving assignments in this class nobody is going to pay anything right but then so he was actually doing all of these things and he published serious mathematical articles based on this but if people knew that somebody from breweries publishing these articles they are not going to pay much attention to it so he wrote under the pseudonym of student right so it is called student's T-distribution.

Because the author of the paper was student his name was student a pseudonym of student that was called students T distribution right so people want to know more all of this kind of history very interesting about history of statistics right so that is this book called the lady tasting tea so this is actually a very serious book and I recommend it to people if you are interested in knowing more of history of mathematics and stuff like that it is amazing so apparently there was this English lady who claimed that she could tell the difference.

If milk was poured into the tea or if Tea was poured into the milk okay and of course she happened to make this statement in a gathering of scientists and so on so there are a couple of statisticians who then ran one of the very first documented a case of what is known as a double-blind test okay they did not tell her what was happening right hey started giving a tea right there somewhere someone behind the screen was sitting there in some cases they were pouring milk into T some cases you are pouring tea into milk and giving it to her.

And then apparently the lady identified this correctly some X percentage of times right now the question is what she doing it by chance or what she truly able to tell the difference between milk being poured into the Tea or Tea being poured into the milk was a very valid scientific question right so they came up with significance test.

I have, go read the book right and it is actually that is this is history if this is true history right so instead of the Gaussian we use the student's t-distribution right so the thing is students T distribution is not a single distribution it is a family of distribution I just do one thing here but this is not truly just a single T distribution it is a family of distributions right one for each degree of freedom that your setup has, right.

So just like I had the Z statistic I am nothing very different right is the same thing as the set statistic the T statistics is exactly the same thing as the Z statistic except that well here I use the population standard deviation by  $\sqrt{n}$  here I am using the well sorry the standard since the sample standard deviation by  $\sqrt{n}$  right big difference right. Right this is what I was telling you so you could use the same thing in the Z statistics also right now you do not have to move to tea.

The reason you want to move to tea is if n this small right now what you do here in the z-test you compare it with the Z table right in the tea test what are you going to do tea table let us sit all right but you have to be careful about which tea table which row in the tea table that you use because there is one row for each number of samples right suppose you have n samples you have to look up the row corresponding to n-1.

If you have n samples you have  $n - 1$  degrees of freedom okay. So that is essentially what it is so one thing about the T distribution is that it assumes so one thing about the T distribution is that it assumes that the underlying distribution from which the samples are drawn right the population distribution is normal right. So earlier we are having a sampling distribution where we did not have to worry about the underlying Sample population distribution regardless of the underlying distribution we knew the sampling distribution was normal.

But in the t distribution it assumes that the underlying distribution is normal but it turns out that in practice it is extremely robust when you can run T tests on arbitrary distributions right and still it gives you reasonable answers provided remain is not too skewed or anything right and so most

distributions that you would likely see in practice right the T the T test gives you reasonable answers okay so you can use them.

So moving on I have okay so very roughly let us look at it this way right so suppose I give you the mean okay and I give you  $n - 1$  sample you can construct 10 sample cannot you, for a given mean I give you  $n - 1$  samples we can construct the N sample right so that is roughly that so you have only  $n - 1$  free things that you can set in the system so but the  $n^{\text{th}}$  one will be determined so that is what it means the  $N - 1$  degrees. That is a more formal definition of it but roughly I mean intuitively this is what the thing is so how many independent factors that you can set in the system.

**IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

## NPTEL ONLINE CERTIFICATION COURSE

## Introduction to Machine Learning

## Lecture-57

## Hypothesis Testing IV – The Two Sample and Paired Sample t – tests

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

(Refer Slide Time: 00:17)

Two Sample t-test

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$\bar{x}_1 - \bar{x}_2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} \quad N_1 + N_2 = 1$$

$$\sigma_{\bar{x}_1 - \bar{x}_2}^2 = \sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2$$

Okay so in the two-sample t-test the question that we are going to ask is I am going to take two different samples right I am going to take two different samples and I want to know if both of these samples came from the same distribution or not and they have some underlying distribution from which I have done two samples I want to know if both of these samples came from the same distribution or not right.

So it could be the distribution of errors right so I could actually draw two sample I can say I am going to run algorithm one okay I am going to get this many errors like ten I am going to run algorithm one not at ten different times right so does it remind you of something would think of

something like 10-fold cross-validation I run algorithm one on using 10-fold cross-validation I get 10 different numbers right.

I run algorithm two using 10-fold cross-validation I get 10 different numbers right now what does what do I mean by the question do they come from the same distribution that means that if I run this algorithm one again and again and again and again and again I am going to see some distribution over the errors right if I run algorithm two again and again and again I am going to see some distribution over the errors right or these two distributions the same right.

So the question I am asking is I have algorithm one I have algorithm two or the errors similarly distributed that means there is no statistical difference between algorithm one algorithm two okay so that is what we is a kind of questions that we would like to ask right, so two sample tests t-test allows us to do that compare means of two samples to see if they are drawn from the same population or different and again remember when you are talking about same population or different we are only asking the question of their means same or different assumption we are making is the standard deviation at all same right.

So null hypothesis is yes they are drawn from the same distribution so  $\mu_1 = \mu_2$  right and the alternate hypothesis is  $\delta$  for a change okay let us do a two-tailed test so this is called two tail because I am going to look at both ends of the distribution right so the greater than or less than were called one tailed or single tail because we are looking only at one end of the distribution. So what I am really want now is to look at the look at that right.

I want to look at  $\bar{x}_1 - \bar{x}_2$  and what it should be zero if the null hypothesis is true right so I am going to have I will compare it with a zero-mean Gaussian or a zero mean yeah so zero mean T distribution right so with some number of degrees of freedom but I need to really compute the T statistics right so the t statistic look something like this in this case right, so I am going to look at so this is zero mean right.

So  $x_1 - x_2 - 0$  right divided by the variance so how will I compute the variance right, so the variance of the difference is actually the sum of the individual variances intuitively that makes sense right so here is we will do something these are these are details okay there are nothing to get hung up about right so I basically had to estimate this variance but how will I do this variance

I can do one of two things I can take the samples have drawn right under algorithm one I can estimate  $\bar{x}_1$   $\sigma^2 \bar{x}_1$ .

I can take a look at the samples I drew hundred algorithm two and I can estimate  $\sigma^2 \bar{x}_2$  so I can do that independently and I can get this variance and then what I do, I can plug this in here and I can get away with it right but the problem is not really problem that is a small advantage that I can take care what is advantage or what is it what is they can do I am assuming that the variances are equal right so what did we do earlier when we had a situation where we had this thing and you assume the variances are equal people remember that we did something called a pooled estimate right, so the pooled estimate what you do is you essentially look at the variance across the entire population right and we compute the variance so you can actually do a pooled estimate right. It is my  $\sigma^2$ , right.

So how many degrees of freedom is going to have so this essentially I will plug this in here wait I will plug this in here and they will compute my T statistics right once I computer the T statistics like I said these are all details if you understood everything so here so far to hear everything is fine here we are just computing the variance this just looks little complex where is nothing but this computing the sample variance by using a pool of estimates right.

So now how many degrees of freedom I am going to have here we talked about the last time also  $n_1 + n_2 - 2$  right that is  $N_1 - 1$ ,  $N_2 - 1$  so it is  $N_1 + N_2 - 2$  so the number of degrees of freedom is  $n_1$  plus so you take this T statistics look up that table and figure out for whatever p level you want right so that is basically it so this is called the two-sample t-test and it is very useful when you want to compare performance of two different algorithms on some sample that has been drawn right you remember the example I told you right.

In fact the nice thing about the two-sample t-test is I do not really need to do 10-fold cross-validation on both algorithms let us say one algorithm is significantly more expensive to run than the other so I can do a five-fold cross-validation on one on a 10-fold cross-validation on the other because I am not expecting the  $n_1$  and  $n_2$  to be the same here right but the variance is going to be higher right.

If you think about it so the variance will be higher if the samples are very different right because the  $n_1$  samples I run on the algorithm one right on the  $n_2$  samples on which I run algorithm two

if they are different sets of samples then if I look at the pooled estimate of the variance the variance will be higher right because there will be some underlying variance because of the change in the samples itself and if I run the same algorithm again and again on the same on different samples I am going to get variants I am running different algorithms on different samples.

So the variance will be larger so what will that mean so naturally my T statistics will become smaller right so the larger the T the more by p value can be right so the T statistics might become smaller if the variance is larger so in some way I can get rid of at least some of this variance right so we do something called the.

(Refer Slide Time: 10:29)

3.3

Paired Sample t-test

$N-1$  degrees of freedom

$H_0: \mu = 0$

$H_1: \mu \neq 0$

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}} \quad \hat{\sigma} = \sqrt{\frac{s^2}{N}}$$

So Paired the sample t-test so what does this mean so I am going to run algorithm one and algorithm two on the same sample right so if I am going to take ten different samples I will run both algorithm one and algorithm two on the same set of ten samples right instead of running them on different samples right, ideally if you have the control over how the sampling is done and how the experiments are done then you should run Paired sample tests okay.

The two sample tests is appropriate only when somebody gives you the performance on different samples a priori right they do not allow you to sample and run the algorithm somebody says okay I have might I have the have access to some 15 samples I have run my algorithm on it here are the 15 performances and you can do whatever you want on the on samples that you draw will not tell you what the samples I drew is okay.

Then you can run your algorithm on 10 different samples that you draw from the same data and then you can compare the two then you do two sample T – test but if you have complete control over what you are doing then you do paired t-tests right paired sample T test and so essentially what does this mean it means the following right suppose I am doing 10-fold cross-validation so what do I create these 10 folds right.

People know what the folds are right so I am dividing them I will do some stratified sampling or whatever it is I create these ten folds I keep them I write them on to disk so whenever I run a algorithm and there is going to read the folds from the disc and done it I am not going to regenerate the foals every time I run the algorithm so that would mean that for every fold I will have results from both algorithms, so in fact this is catching on so much in the machine learning community now that for many of the newer data sets that are being published okay people are actually publishing the folds on which they run the experiments, so that you can also run them on the same Folds.

So you do not you do not generate your new folds and start running because then the comparison becomes little diffy right you can use the same false at a Rand experiments on therefore you do not have to repeat their numbers I can directly compare it with their numbers and I can report right so that is that is why people are actually publishing the folds also right, so when you do pair sample t-tests what you are doing is here what are you doing here you are taking the mean of X 1 and the mean of X 2 and comparing it against zero right.

In this case what you can do I can take the difference because I am running it on the same sample right so I can actually the difference of the performance now makes sense on one sample here instead of averaging the samples and taking the difference I can first take the difference okay and then compare it to a zero-mean distribution right, so instead of instead of having a lot of excess and then getting  $\bar{x}_1$  I am going to have lot of excess lot of  $X_2$  and then I will get have a lot of  $x_1 - x_2$ .

And then I will take a  $x_1 - x_2$  the whole bar okay and then compare it to a zero-mean distribution so that is what I do in paired sample tests and so this is going to have  $n - 1$  it is going to have  $n-1$  degrees of freedom right my  $H_0$  is right so what is this  $\mu$  so this is the difference of the means when I saw the difference of the performance right so that is 0 right the mean of the difference of the performance is zero across many samples that means they are the same as my null hypothesis right.

And or I can do  $\mu > 0$  which case which case depends on which one I am subtracting which  $x_1 - x_2$  if we say  $\mu$  greater than 0 that means  $X_1$  is better than  $x_2$  but there should be some something from the data that supports your alternate hypothesis, so this basically the standard stuff right so you do this and well if  $\mu_0$  then this is 0 this is just  $\bar{x}/\sigma^{\wedge}$ ,  $\sigma^{\wedge}$  is the samples standard deviation by  $\sqrt{n}$  and we haven m minus 0.

So this is actually a lot lower variance because you do not have any problem generated variance you only have the variance due to the performance of the algorithm right the samples themselves are exactly the same so that gives you a much higher T estimate than you would get if you run the two-sample t-test okay, so we explained all about all of these things too but almost all packages that you can use right have all of this built in so you can do t-test z-test whatever it is you want you can run.

You do not really have to worry about the internals of it right you just need to specify the sorry the P level you just need to specify the p level you what is it a apart from center before a given test you have to specify what is at p level right what is the P level that you are looking for so if you say I want a P level of 0.001 atleast right then some of these could actually reject it saying that no I cannot reject the null hypothesis at a level of 0.001 or something it could come back and tell you, okay.

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**NPTEL**

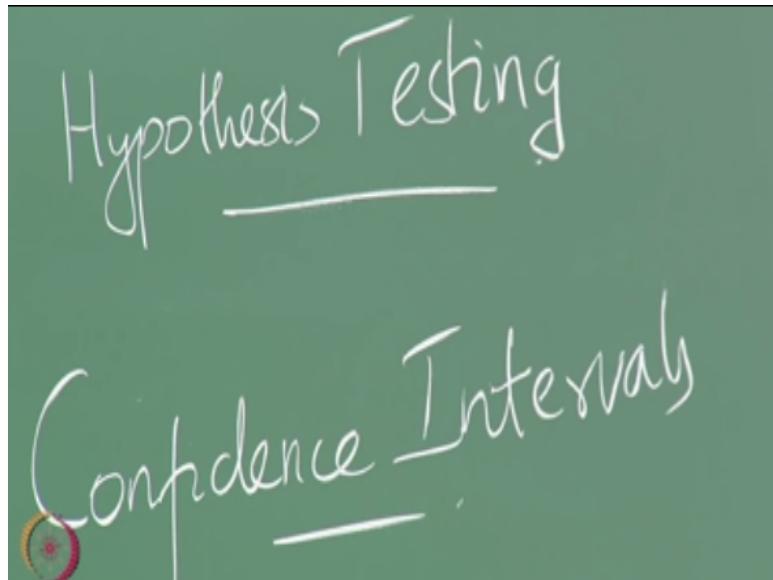
**NPTEL ONLINE CERTIFICATION COURSE**

**Introduction to Machine Learning**

**Lecture-58  
Confidence Intervals**

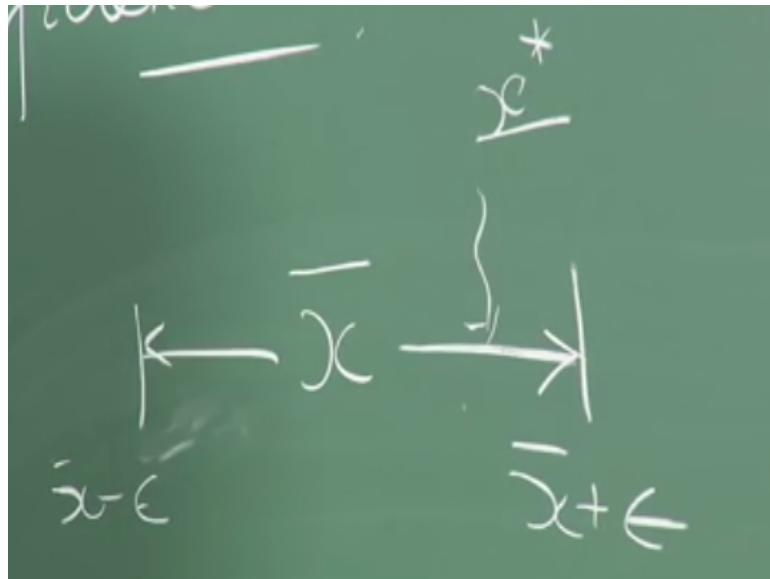
**Prof. Balaraman Ravindran  
Computer Science and Engineering  
Indian Institute of Technology Madras**

(Refer Slide Time: 00:17)



Confidence intervals right so we talked about confidence intervals all right so essentially the question I want to ask is the following right. So if you think about it what we are doing is right we are trying to estimate some parameter some performance measure by looking at some statistics that we compute on a sample of size  $n$  right, so that is basically what we are trying to do be trying to measure some performance as on a statistic in a sample of size  $n$  right.

(Refer Slide Time: 01:18)



So I am doing this repeatedly right suppose I have done this with one sample and I have some number let us say  $\bar{x}$  okay let us say that is the average error or average whatever okay I am giving you some performance measure  $\bar{x}$  right so in what fraction of samples of size  $n$  right I draw this  $\bar{x}$  and then I will give you some interval around  $\bar{x}$  so right and I will give you some interval  $\bar{x} - \epsilon$  and  $\bar{x} + \epsilon$  right.

So and there is some true performance measure I do not know called  $x^*$  right so ideally I would want to give you this plus and minus  $\epsilon$  such that  $x^*$  lies somewhere in this interval right so I give you dot  $\bar{x}$  I give you  $\bar{x} \pm \epsilon$  says that with a high probability I want my  $x^*$  to lie within that interval okay, so in fact the confidence interval essentially the amount of confidence you have in this interval essentially means the following okay.

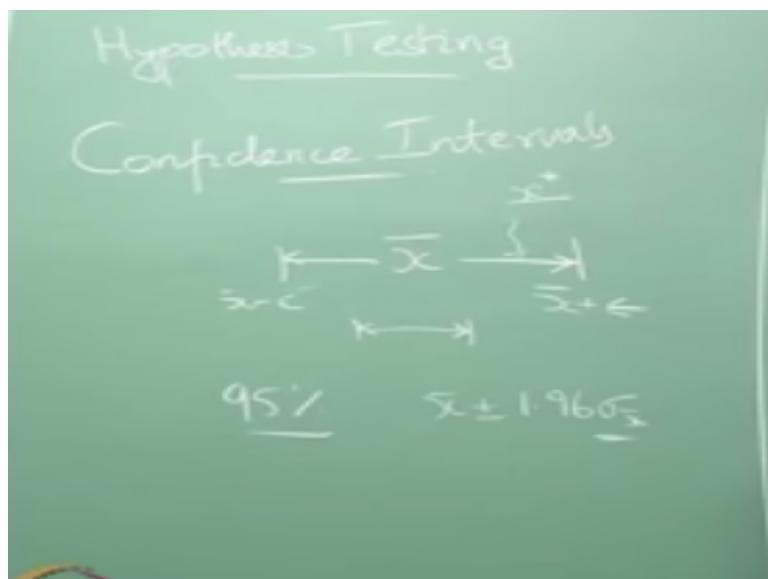
In what fraction of the samples of size  $n$  that I draw suppose I keep drawing samples of size  $N$  and I tell you that I have give you a ninety-five percent confidence interval right so what does this mean exactly 95% of samples of size  $n$ ,  $x^*$  will lie within the spar would lie within this star would lie  $\pm \epsilon$  ( $\bar{x}$ ) right so right you understood what I say right in 95% of the samples of size  $n$  right  $x^*$  will lie within  $\pm \epsilon$  ( $\bar{x}$ ) okay.

Is this is the same thing as saying that with ninety-five percent probability the  $x^*$  is within  $\pm \epsilon$  ( $\bar{x}$ ) no, why? Could I am talking about samples of size  $n$  here right so depending on my sample size may sample is very large right then possibly this will approach that probability I am talking about samples of size  $n$  okay so whenever I give you a self-confidence interval remember that it

really does not mean even though people often mistake it for the probability of  $\bar{x}$  being within  $\epsilon$  ( $x^*$ ) is really not the case.

What it really means is if you repeat this with samples of size  $n$  right in ninety-five percent of the samples  $X^*$  will lie within  $\epsilon$  ( $\bar{x}$ ) okay some kind of an assurance right but not suppose I want to reduce the confidence interval what does it mean, sorry what does it mean to reduce the sampling I mean confidence interval reduce the  $\epsilon$  right I want to reduced  $\epsilon$  I want to make it smaller right so you say reduce the confidence interval I really mean that okay I want this as opposed to that right. If I want to do that what is the best way to do it increase the sample size  $n$ , right.

(Refer Slide Time: 05:25)

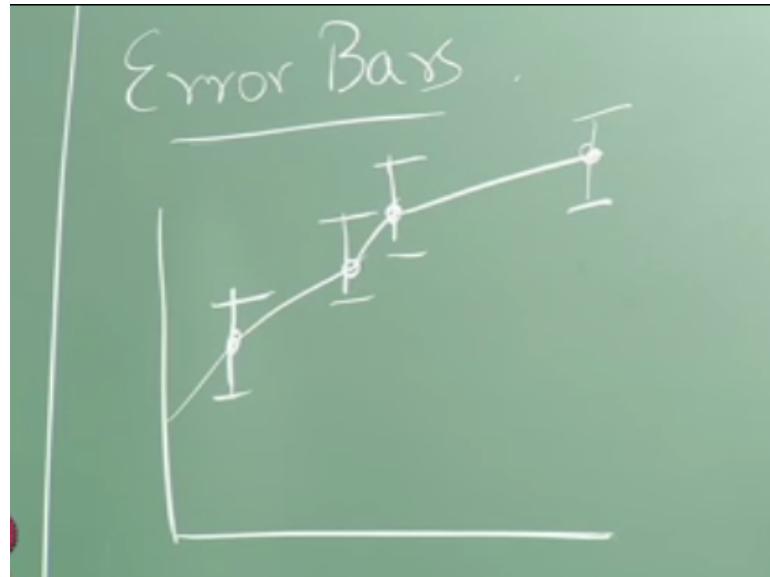


So if I want a ninety-five percent confidence interval right so what would that be, we look at some magic numbers in the last class increase from this is  $z$ , so we cannot use the standardized thing here because we really need to give actual values here right so we cannot use the standard normal Gaussian so it has to be  $1.96 \times \sigma_{\bar{x}}$ , so it is 1.96 because it is 2.5 that side 2.5 this side right so it is 1.96 right.

So if I want a tighter confidence interval right then essentially I have to reduce the  $\sigma$  right. This is assuming that your end is fairly large right if a  $n$  is very small then you have to use the T distribution you cannot use 1.96 lot to use the corresponding statistics from the appropriate entry in the tea table so appropriate table find appropriate row in the tea table corresponding to the

degrees of freedom right so typically for  $n > 20$  right. You can use even something simpler 1.96 or even two times  $\sigma$  is good enough so related to the confidence interval right.

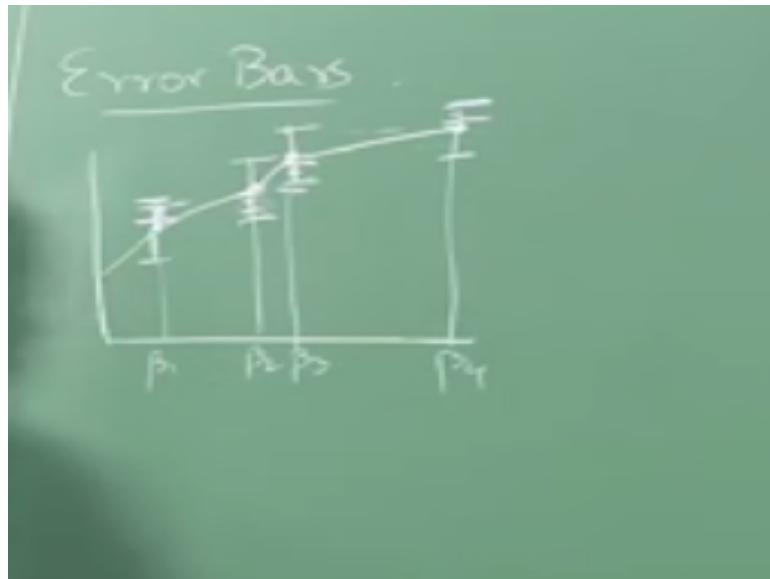
(Refer Slide Time: 07:24)



You also have this notion of eradiate what error bars are what is that somebody said something remember standard errors what is it standard error yeah but what is it really it is see it is the variance of the standard deviation in the sampling distribution right that is what we call the standard error so error bar are essentially things that you plot around your estimates so that it tells you what is a what is the variance that you are likely to see in the estimate that you are getting.

So typically what you do is right so you make some estimates and then you try to make some plot right I am varying some parameter then I say okay and that is how my performance varies right so instead of just plotting these points and trying to draw a curve right what I had like you to do is essentially give me a error bars around that right so each of this point I would have run an experiment I am varying some parameter here right and I am looking at the performance right some parameter I do not care what it is and looking at the performance. And in each of this point I would have run an experiment right.

(Refer Slide Time: 09:12)



Assuming they are all experiments around the thing right each of these outer run some experiment and for each of those I can give you the standard error right so I plot these error bars now the question that you have to ask is okay from let us say I have these values some values  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ ,  $\beta_4$  right so I have just run it at some intermediate points I have these curves right, so can you tell me if  $\beta_2$ ,  $\beta_3$  is there any difference in the performance between  $\beta_2$  and  $\beta_3$ ?

Not really because my error bars overlap significantly and I cannot be sure if there is a difference between the performance of  $\beta_2$  and  $\beta_3$  just on evidence of this curve alone right what about  $\beta_1$  and  $\beta_2$ , No right  $\beta_1$  and  $\beta_2$  also I cannot say that that is actually a difference what about  $\beta_1$  and  $\beta_3$  barely  $\beta_1$  and  $\beta_4$  surely what about  $\beta_3$  and  $\beta_4$ , not really right so I mean so yeah in the means are different if I just gone by means I would have probably said that  $\beta_4$  gives me better performance and  $\beta_3$ .

But on the evidence of the experiments that you have run so far I cannot conclude that because the error bars significantly overlap right so this is why whenever you are running empirical series you are always supposed to plot these error bars if you just give me an average performance right it is not at all clear so if I am comparing two things then I can run your two t-test and so on so forth this gives you a rough idea of which of the performances are actually different.

So  $\beta_1$  and  $\beta_4$  are certainly different that  $\beta_4$  is certainly better than  $\beta_1$  so for other things evidence is kind of shaky sorry, good they could so the way for you to verify this is now go run more experiments with  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$  try to see if you can get a better estimate because as you know that

so with the now the true estimate could be anywhere in this interval right so rather with ninety-five percent of the cases that to estimate could be anywhere in this interval so what I do is I run more experiments so if I run more experiments with  $\beta_1$  I might actually see that my mean shifts here right and my confidence interval becomes much narrower.

But now remember this is not the same confidence interval as it before because this is a confidence interval of a larger sample size look you cannot directly compare these two it is a confidence interval of a larger sample size so I might actually rerun the experiment and this might be the values I end up with this if there is a better color that I run the experiments again with a lot more data and we can see that now things are little clearer.

So  $\beta_1$   $\beta_2$  there is really no difference right they are the same alternatively  $\beta_2$  could have moved up  $\beta_1$  could have moved down and in could anything could happen and this giving an example here where no  $\beta_1$  repeated or almost like more likely to be the same and  $\beta_3$  is certainly better and  $\beta_4$  is certainly better than all the other three that this could happen one potential scenario another potential scenarios this could move down this could move up right.

So it could whole thing could change right essentially what the error bars tell you is what kind of conclusions can you draw from the experiments you have done so far could very well be that it is enough for you to find out which is the best  $\beta$ ,  $\beta$  for seems to be the best interms of the experiments that you ran even the first time around but if you want to produce a ranking among the  $\beta$  you will have to rerun the experiment.

The only conclusion you can make from the previous experiment that you had was that  $\beta_4$  is probably the best  $\beta$  best value forbidden if that is all you are interested in finding out you can be happy with that experiment but if you want to produce a relative ordering of the parameters then you will have to be more careful so that is essentially the use of error bars.

### IIT Madras Production

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**NPTEL ONLINE CERTIFICATION COURSE**

**Introduction to Machine Learning**

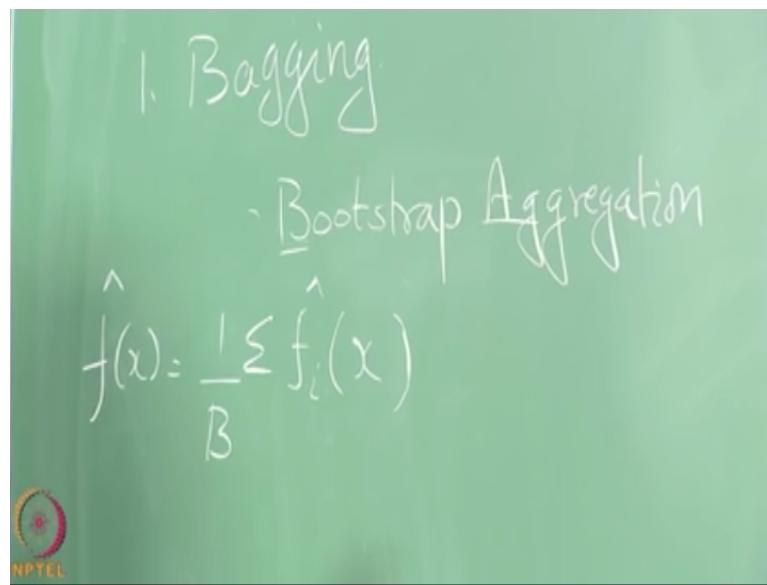
**Lecture-59**

**Ensemble Methods- Bagging, Committee Machines and Stacking**

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

So we will move on to another topic which is essentially ensemble methods so what do people do in ensemble methods is that instead of using a single classifier or regressor you tend to use a set of them right in order to make the same prediction typically these end up improving some performance or the other of the classifier right statistically speaking more often than not they end up in reducing the variance of your classifier right but that also ends up giving you better empirical performance at the end of the day right.

(Refer Slide Time: 01:17)



So we are going to talk about several approaches for an ensemble methods I will start off with the one that is familiar too familiar to all of us called bagging right so, so what is bagging why did I say similar at all of us so bagging stands for bootstrap aggregation okay do not ask me how

they dread bagging out of boots bootstrap aggregation right but the idea is very simple all of you know what bootstrap sampling right select mean about bootstrap sampling.

So what I am essentially going to do is I am going to create you give me one training set of size n I am going to create multiple training sets of size n by sampling by replacement and then I am going to train a classifier on each of those sets I am going to train a classifier on each of those sets and how will I combine the outputs of the classifier sorry I can do a majority vote or average what sorry cannot end up blowing majority vote right.

So average what average if you can if your classifier is going to produce probabilities for the class labels right I could do some kind of a weighted average of the probabilities the classifier is just going to give me one or zero right I end up essentially doing majority vote okay does it make sense so the idea is very simple backing idea is very simple I am going to produce lot of classifiers right so I am going to call them  $f_i$  right so it could it could be it could be regression as well.

So it does not have to be classification right the situation I just take an average of the outputs of all the classify so each  $f_i$  is trained on one bag which I have produced from the from the original sample like this is another back derivation from the word left bagging it so it is bootstrap aggregation but then each of those bootstrap sample you produce is sometimes called a bag right so if I produce B bags then I eventually average by I mean be to get me the prediction.

And if I am doing it for classification I can produce majority vote on average the probabilities okay the few things to note so backing reduces variance right so in, in effect it ends up giving you better classifiers normally then what you would get by training on a single sample of the data right or producing a single classifier it is particularly useful when you are dealing with unstable classifiers right it can take an unstable classifier and produce something that is more stable right that is just a fallout of reducing variance right it can take an unstable classifier.

And produce something more stable so one thing that you have to be careful about when you are bagging is that if you bag bad classifiers the performance can become arbitrarily worse something that has a classification accuracy less than .5 less than or equal to 0.5 two classes sorry each when you change the data on which you train the classifier right you are going to end

up with a different classifier in sets of data yes you could as well if you want to but it is a good point if you initialize two different variables different values for the parameters.

You introduce an additional source of variance there but the you could you could there is nothing stopping you from doing that just that you have to be careful about how we do the analysis if at all you are doing a variance analysis now do be careful about how we do the variance analysis right yeah so by that he brings have a good point so in some of the ensemble methods that we talked about right the ensemble is will typically be of the same kind of classifier okay.

The only way we are distinguishing one classifier from the other is by training it on a different data right except for one approach which I will tell you later where typically if we use end up using different kinds of classifiers okay so data would be the same but the classifiers would be different but this is one of our aggression but anyway soon I am not using different variables it is like suppose he is using a neural network right so you need to have an initial starting point for the weights right.

So if I use a different random starting point what so that was this question should I use the same random starting point or should he is a different starting point right and even then going back to your question right so you think about it this way right so right we are talking about  $f(x)$  instead of that think of it as right so this  $hi$  will give me whatever features I want from  $x$  even if I want to run each classifier on a different subset of the features it will just be that that will get enrolled up into the classifier.

I can still do the averaging if I want right that is not an issue but that is not the question us asking anything else on this okay so if you throw a bad classified into the mix right your performance can become arbitrarily bad so that is something that you have to guard against okay so bagging is a very, very intuitive very simple thing and a couple of more practical things about bagging is that it is you know what they call embarrassingly parallel you know you can run how many other instances of training on bagging you want at some of the other ensemble methods.

We talked about are going to be inherently serial in nature right so allowed to run one after the other right suppose you are looking to run this on many large data sets right so doing bagging is kind of easier because you can run it because one, one bag like or classify trained on one bag

does not depend on a classifier trained on the other bag in any way right so they can be trained independently and trained in parallel.

So that is the first thing okay next thing you talk about something called committee machines okay this is nothing big okay all it says is it read a lot of different classifiers let us say we have some glass or something no and all with all the individual classifiers performed well on test eight or do they just have to learn the training do they have to generalize well or so each classifier you typically train it using whatever is your normal training procedure right.

So if normally you would expect it to generalize well on the test data right so you would want to produce classifiers that generalize well on the test data right but that is a call to be made I mean if you do not want to test each in every classify sometimes people just tested the what they call the bag the classifier right the combined prediction alone is something that they test they just train each classifier on the data that is given right.

And then they test the combined classifier there are multiple reasons for wanting to do that so one is that typically the classifiers that you use in bagging or not very powerful classifier right so the chances of the mover fitting or low so you do not really try to do a validation on the test set to make sure that it is not over fit and things like it because the classifier itself is not very powerful classifier and then you just go ahead and test it on the tested on combined classifier on the data right.

So why would want to test the combined classifier on the data you will want to know whether you should produce more bags and think like that right so the nice thing about the bagging is that because you are using at any point of time you are only using a weak classifier to fit the data right and not if we classified but not necessarily a you know very strong classifier to fit the data the chances of you over fitting is very small even if you increase the number of classifiers in, in the bag even if increased number of bags and I can do this for 10,000 samples 10,000 such bags right.

And I still want to overfill the danger of over fitting is no more than training it once right so that is a nice thing about bagging I can keep making the classifier I can reduce the variance in my estimate more and more but I am not getting into any danger of over fitting right so that is a nice

thing so the other thing is committing machines did have to really think about anything you do not even have to think about how oh my god how do I paralyzed this thing that is this typically.

I mean it is a term that people use in architecture and they are trying to think of parallel computing so things like though this is embarrassingly parallel so I can do dude that whatever levels of parallelism I want and things like that maybe I am misusing the term but, but yeah but it is really easy to paralyze right you can just want it on different sample separately yeah so what is embarrassing about it I mean why do you even have this whole parallel computing field to study something that can be parallelized.

So easily so I am really embarrassed to be working in parallel computing and just making it up okay, okay committing machines can I want the committee which this any other questions okay so computing machines is very simple idea so I am going to train I have given a data set and I am going to train a lot of different classifiers on, on the given data right and then I am going to combine their outputs right based on some kind of weighting mechanism okay so what could be the weighting mechanism.

I will try the neural network whatever it I trained many, many different classifiers right and then I have this set of classifies that have already been trained right and I have to combine their output how do I go about doing this in there are many ways in which you can combine their output right so I am just taking this classification from the textbook elements of statistical learning and not that I completely agree with it so in committee machine.

Suppose I have M classifiers the weight is assigned to each classifier is  $1/M$ ,  $1/M$  so I treated classifies as being equal right so that is called a committee machine and I have many different classifiers as the outputs of all the classifiers I am going to give each one of them an equal weightage or equal vote right so I call that a committee right then we go on to something more interesting called stacking no badge me says the same classifier same algorithm but trained on different samples of the data right in committee machine.

It is the same data but trained on different algorithms right so I have a three and I could have never let works I could have anything right or it could be says it could be neural networks with different number of neurons I mean I am not saying that it has to be a completely different algorithm it is the different classifier it could for different settings of the parameters and so on so

forth right and so starting the stacking is like committing machine so I have many, many different classifiers right but what I am going to do is instead of arbitrarily assigning a weight to each of the classifier what will I do what can I do.

I could do that but with stacking what do I do I learn the weights right so that is a natural thing to try and do right so I have the prediction that is made by each of the classifiers right I go ahead and I learn the weights so another way of thinking about stacking is the following so I use each of these classifiers okay to generate a feature for me so this way it is called stacking so I have a set of classifiers right they all output some it could be a probability vector or it could be just a class label or whatever at the classifier one comes and tells me okay.

I think this data point is class 1 plus if I to comes and tell me I think this data point is class to the classified three comes in tell me I think the data point is class 1 and now what will happen is i will my input to when next machine learning stage will be class 1 class to class one right and now again it is a machine learning algorithm now I can run whatever machine learning thing I want it could be linear regression because I am interested in finding weights right so doing some kind of regression seems to make sense right but then you know problems with regression classification all of you know.

That so you might want to use it for classification or you might want to use some other method for classic you might want to use logistic regression for classification whatever it is right but then the inputs to this stage or they are these outputs of the first stage of classifiers and they try target is the same target as the first stage the same class level right so one way of thinking about it is like stacking these classifiers one upon the other so I first have some set of classifiers they produce features for me right the features are essentially what the class what they think are the class labels right.

(Refer Slide Time: 18:13)

3 Stacking

$$f_i : (x_1, \dots, x_p) \rightarrow \mathbb{R} | G$$

$$h : (f_1(x), \dots, f_M(x)) \rightarrow \mathbb{R} | S$$

And then I learnt to combine the features I learn a predictor based on these sets of features so that is another way of thinking about it makes sense then make sense okay right let us take a classifier FM either I there are some people are actually saying they did make sense I am trying to make it easy more explicit let us take a classifier some  $f_i$  right so it basically it operates on at  $X_1$  to  $X_p$  right I just going to give me something right it is going to give me real number or, or some, some class label okay.

So that is basically the, the function you see there does classification or regression or whatever r so now what I am saying is I am going to Train another H right that is going to take as input right it is going to take this  $f_1$  to  $F_M$  as input so  $F_1$  is the first level classifier I have m of them right and then h is going to take  $h(x)$  is going to take  $f_1(x), \dots, f_M(x)$  as input and it will produce whatever is the thing I am looking for real number or write so if you go and look at the structure of h right.

(Refer Slide Time: 19:34)

$$h: \left( f_1(x), \dots, f_M(x) \right) \rightarrow \mathbb{R} \text{ or } S$$

$$\beta_1 f_1(x) + \beta_2 f_2(x) + \dots + \beta_M f_M(x)$$

To make it explicit let us say I want head  $h$  to be a linear function right so that will essentially mean that  $H$  will look something like right it just going to look something like this so this is essentially saying that okay I am taking the outputs of all this classifies I am combining them in some kind of a weighted fashion the same way I tried any of their face yeah the same training data yeah the same training data that we had gives for  $f_i$ 's in use the same training data for  $h$  may be, may be not depends on the kind of classifier that you are using right.

I mean  $H$  is a completely different training algorithm right oh I can see your confusion okay so my initial training data is going to look like this I do not know okay initiate training data is going to look like this, this is my  $X$  and that is my  $+1$  so corresponding to this that will be a training data for  $H$  which will be  $f_1$  of this guy so now I have only two elements here but there is  $f_1$  of this and this is  $f_2$  of this and the same plus one comes in here right so I can I can do this so the dimensions do not have to be the same wait so this is stacking.

So stacking is very powerful method and in fact you can do all kinds of weird things with stacking in fact these weights that I am learning right I can make them functions of  $X$  as well what does it mean what does it mean if my weights are functions of  $X$  type depending on where in the input space the data is coming from when I might want to trust one, one output more than the other way suppose it should say the top left quadrant of my input space then I trust  $f_1$  and  $f_2$  maybe a little bit but then if it is in the top right quadrant.

Then I trust both  $f_2$  more than  $f_3$  less or something like that I can actually do that also so with stacking this function can be arbitrarily complex that is why I did not want to write the linear thing first because it will bias you into thinking about simple linear weighted functions but this  $H$  can be arbitrarily complex so if you think about it in fact we are doing something like this in the neural networks right so the first layer it gives you features are complex feature some, some, some hyper plane is being learnt in the first layer itself and it produces a complex feature and the second layer takes all these complex features.

I have produced and it learns to produce a final output right the only difference is the first layer is not trained in this way right the first layer is not trained directly using the training data it is trained using the back propagation error or whatever is the training algorithm you use it is not directly trained using this data so that is the difference right but we already looked at things like this right these are all some kind of general additive models they are called additive models fine so any questions on this can we take less pay directly as it affects so or are you meaning that training like this just simplifies you are the way you are doing it basically all, all your Plus page can be linear.

But a combination but any combination of linear classifiers you will be able to explain much more complex cream using stacking the basic idea is, is that when my classifiers need not necessarily be linear classifiers see the thing is so any of the classification algorithm that we are looking at right comes with two own biases in terms of what are the class of functions it can fit and so on so forth right and it could very well be that across the entire input space the function is so complex the final function.

I want to learn it is so complex that no individual classifier can actually fit it or if I try to fit it with a single classifier and it end up with something that has too many parameters so when you do this kind of this layer wise training so the I can get by with whatever I know a simple classifiers in the first stage first stage right I could use decision trees it need not necessarily be linear so addition trees are simple enough I do not have to grow it all the way out that I can stop at some point I can use decision trees.

I can use neural networks whatever I want as my choice right and then later on try and combine the outputs you could you could given how much success people have had a deep learning I would suspect that if you if you work on this carefully yes you could have multiple levels of

stacking people do that I mean it is not that it is not that the reason it is called stacking is because people actually did multiple levels so it could do multiple levels of this but then the question comes how do you group these things and so on so forth right do I do I run it on every give all my first level classifies as input.

To all my second level classifiers or should I group them somehow to a form a hierarchy so those kinds of issues will arise as long as you can address them sensibly then you can go ahead and do multi-level stacking is one thing which I wanted to mention sometimes when people want to run you know competitions and so on so forth they do something very clever they do not expose the actual data to the competitors they give you the outputs of the first level classifiers okay.

And then the actual class level they do not tell you what the features were that they measured they give you the outputs of the first level classifiers and then they give you the class labels and then now all you need to do is train your second level classifier take the first level classifies output as inputs and train the second level affair and see what we can do with it in fact it ran for a couple of ways.

I do not know I am not sure they are still running there is an ensemble learning competition which essentially does then this also allows you to have some amount of data privacy rights I do not have to release X but I am releasing some amount of simple functions computed on X okay and then you build whatever classifier you wanton so it is hard for me to reverse engineer because I do not tell you what F is even a very tell you the output of F I do not tell you what F is I do not tell you what X is so it is very hard for you to recover.

Because you cannot essentially compute F inverse right so, so that is another nice thing that you can so the reason that there are so many approaches for doing this is because there is no one clear winner under all circumstances way so yeah so it depends so in fact like I said so stacking is something that you can use under a variety of circumstances you can even use it under cases where you cannot do bagging how so I mean I can use stacking when I do not want to have you want to give you access to data right so that is one case in other cases the data set is small enough.

That bagging does not make much of a sense on that right so it is not really truly representative of the underlying sample and I am not really sure we want to do that in which case I can use the

different biases given by my multiple classifiers as my that is my variation right to my ensemble so there are different ways in which you can do this next thing when I want to talk about which is more interesting thing yeah why yours and I run like I train them on a single data set and then I get the percentage of points misclassified and then I normalize them through all the classifiers like say 3% 5%.

I normalize them to come for each of the classifier you know what is a percentage of data points I got wrong okay learning this when  $\beta$ 's are not a function of X oh that is one way of finding the  $\beta$  here nothing wrong with it like when betas are not the function of X so how else would be you distribute the  $\beta$ 's that is one way of estimating  $\beta$  I mean what you mean missing how else would you read this remove it I can think of many ways of doing that right take the classifier that has a smallest error and give  $\beta_1$  to that and make everything else.

But why would I even want to give weights to classify switch which give me higher error than the lowest error okay give me an answer why the possibility of them you having a better chance no it could be making errors in different things so I might have a one percent error another guy might have a 3% error but he might actually be capturing that 1% correctly the one that I make the error on he met the other classifier might get it correctly so I do not want to completely neglect other classifiers also so that is the reason why you have to go about trying to be more clever about the weight assignments.

I can do this proportional to the errors in fact there is another, another more Bayesian approach for doing this I can look at the likelihood of the classifier given the data and I can assign the weights proportional to the likelihood the higher the likelihood the higher the weight and I can do some kind of normalization on that so that my  $\beta$  sum to 1 and I could do that as well instead of just looking at the error the error would be a frequentist way of looking at it a Bayesian way of looking at it.

We look at the likelihood of the data and do this there are many ways in which you can derive the data and so stacking this takes it to the extreme say okay fine I am not going to worry about a specific algorithm I mean I am just going to let it learn from the data directly someone right so yeah there are pros and cons and whole bunch of things so yeah empirically if you want to validate it finally do cross fraction and then test against all of these ways of generating the joint classifier right.

But then analytically also you can try to analyze some of the variance of the combined classifier and you will end up hitting the wall at some point that says that it depends on the characteristics of the variance of the data so that is basically what that I mean we are entering territory where you can come up with whole bunch of things like ensemble methods like a thousand papers or their research papers out there which I proposed lots and lots of variations right so think of something and before you try to publish it do a literature survey you will probably find it there ok so that is, that is, that is how crowded the space is.

**IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

# NPTEL

## NPTEL ONLINE CERTIFICATION COURSE

### Introduction to Machine Learning

#### Lecture-60 Boosting

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

Okay, so with one of the most popular and most some sense mind-blowing thing with the non-thermal method space right, so boosting the original boosting work original analysis of the boosting work comes from theoretical computer science community not necessarily from an empirical machine learning community right, so, therefore, they were that they looked at having some oracle that had a probability slightly greater than 0.5 of being correct right. Then they try to see how you can better predict somebody who is just above 0.5 okay.

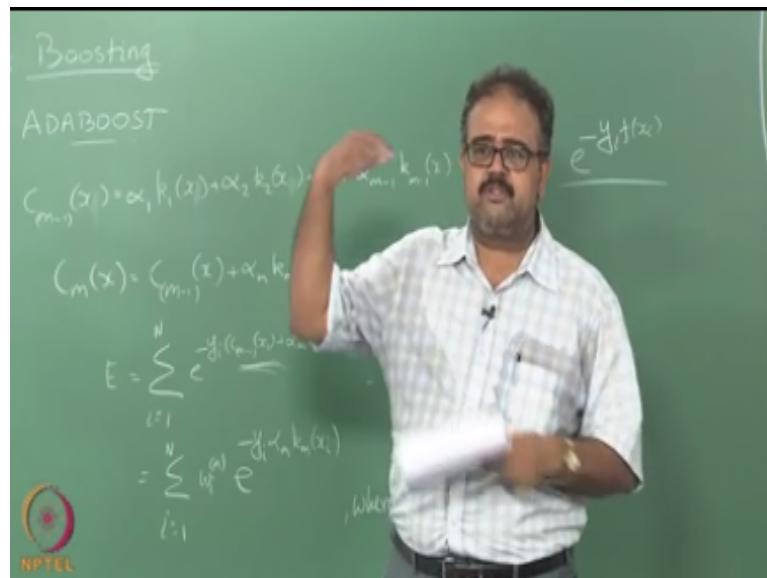
By combining many, many search Oracle's right I can keep improving my accuracy of prediction arbitrary close to 1 right, so that is the amazing part right I start with each predictor as the accuracy of 0.5 plus some epsilon just better than random that I can combine a lot of them and produce something that as accuracy close to 1 right, so this was a very big result that came out earlier and so we are going to look at some kind of simplified version of it, so they remember the goal that distinguishes I mean the main thing that distinguishes boosting from the other methods is that boosting is inherently serial okay.

So boosting is going to build this ensemble classifier in an incremental fashion right, where at each stage I am going to try and explicitly reduce the error produced by the previous stage, so this is something that you have to keep in mind you cannot write it just cannot come up with some ensemble method and call it a boosting method and I have seen that happen in many papers that I have reviewed people just write something that has multiple classifiers in it, so it is boosting, because they have read somewhere that boosting is a very hot area and people papers in boosting get accepted, so they come up with any classifier and an ensemble method and call it boosting, boosting has this very specific property that at every stage right, you add one more

classifier to the existing ensemble right, and this is done in such a fashion as to reduce the error produced by the classifier up till that point okay, makes sense.

Sorry, you get the choice as to what to add next right, so that is that you choose it such a way that you minimize the error that they have not at least you reduce the error okay, so not necessarily minimize. However, you reduce the error of whatever has happened the prediction till that point, okay, does it make sense, so that is essentially what boosting is sometimes you can think of it as error boosting, sometimes they call it error boosting and so on so forth.

(Refer Slide Time: 03:33)



The one very popular and one of the original boosting algorithms is called ADABOOST okay, so let us. It is going to I am going to put up a tutorial for you guys to refer to okay I will use the notation from the tutorial so I will not translate it to the notation in the textbook, okay, so when you read the textbook you have to do the translation yourself, so that is one of the main problems when you have too many different disciplines contributing to the same field right, machine learning has people from computer vision, people from statistics, people from AI and all other disciplines contributing to it and each one of them brings their notation to the mix, right.

So it becomes harder to keep track of everything, but currently, the I output is necessarily complicating my dress, okay, so I am going to denote by  $C(m-1)(x)$ th stage classifier okay, that is obtained by basically adding the outputs of all this individual classifiers so  $\alpha_1, \alpha_2$  to  $\alpha_{m-1}$  right, so I am going to add up these are the weights and  $k_1$  is a classifier that I added in the first stage

right,  $k_2$  is the classifier added in the second stage and so on so forth and  $k_{m-1}$  is a classifier added in the  $m-1$  stage okay.

And then basically I want to produce that okay, yeah, the rest of the class, so there are a couple of things which I should point out here one of the most obvious ways of doing this forget about Erebus one of the most obvious ways of doing this is to say that okay I am going to take this guy right, look at the residual error you know I can think of this as a prediction problem right, and look at the residual error of the predictor right, and then train a classifier  $k_m$  to minimize the residual error, right.

So what will be  $\alpha_m$ ? Essentially how to make sure that this whole thing is along the direction of the residual, so we talked about this earlier right, when where did we talk about this, forwards stages we ask stage-wise or stepwise, stage-wise, so when we talk about stage-wise feature selection we talked about something similar right, so you could think of something along the same lines here instead of thinking of selecting features right, I am just selecting classifiers right.

So I can just take the residual error of  $c_{m-1}$  and then use that to train  $k_m(x)$  and then add it here right, in fact this can be one, it can be one does not matter because the  $k_m(x)$  will actually align itself in the direction of the residual so I can just add it here so it is fine right, so that is a simplest way to do this thing and it is actually a good way to do it if you are doing regression let that make sense and I can take this as they can take the residual error and then train my  $k_m$  to actually go in the direction of the residual.

So I can actually do this, so you can get a boosting like algorithm for regression just by training it along the direction of a residual right, but when I am doing classification that is not necessarily the right thing to do so people come up with different kinds of loss functions and then they try to improve the classification, so the loss function we look at is the exponential loss, so people remember the exponential loss, I talked about it when we are doing SVM's is exponential loss okay,  $e^{-y_i f(x_i)}$ .

So we looked at the exponential loss earlier so we will essentially continue with that, so I will sum over all the training points right, so that is the exponential loss for the  $m^{\text{th}}$  stage classifier right, people agree with me on that, so that is essentially what I wanted to write right, now

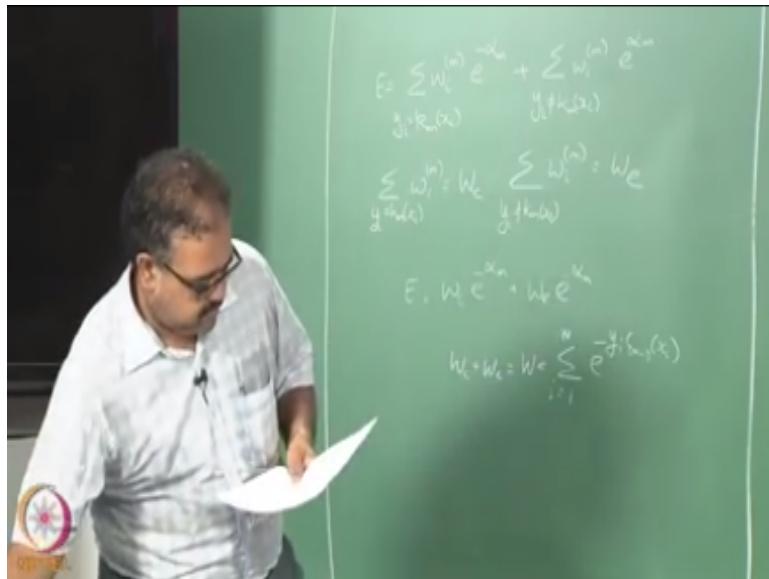
expanded the cm. I have written it as this expression in the bracket here. Yet, can people see me at the back I see not me, but I am kind of hard to miss right, that makes sense okay great.

So this thing we already know right, so there is no control we have over that that thing we already know that is given to us all we need to find is  $\alpha_m$  and  $k_m$  right, so I am going to rewrite this as what is that, that is the last function we are going to be using like that is exponential loss function so for classification we looked at, if you remember we looked at the different loss functions and when we looked at hinge loss right, and I said exponential loss is one of the loss functions and this is how we defined it so essentially I am using that exponential loss function here.

And I mean this was not the way ADABOOST was originally derived okay, ADABOOST was derived in a completely different way and later on about five years after they publish ADABOOST they kind of discovered the connection between this kind of stage ways modeling right, forward are additive stage wise additive modeling and exponential loss they said okay, I can do forward stage wise modeling with an exponential loss function I end up with ADABOOST that connection was discovered five years later.

But now almost always people except in the theory community, in the machine learning community is always introduced like this okay. So where  $w_i^m$  is sorry, m the same thing I wrote here, so the weight of the  $i^{th}$  data point at the  $m^{th}$  stage right, the weight of the  $i^{th}$  data point at the  $m^{th}$  stage is essentially  $e^{-y_i c_m l(x_i)}$  right, so what does this mean what is exactly this expression if you think about it, it is a loss I have incurred on that point  $x(i)$  up till the  $m-1$  stage right, that is essentially the right just the loss that I have incurred on the  $i^{th}$  data point up till the stage  $m-1$  right.

(Refer Slide Time: 14:03)



Okay, now I am going to break that sum up into two components, so do you think of these two components, no, no, no this is varies say  $e^{am}$  so when will I get  $e^{-am}$  when I am correctly classified it, when I will get  $e^{am}$  when I am misclassified it so these are all the data points such that right you are the correctly classified data point is all the miss classified data points right, is intuitively you can see where we are going with this, so what is the best classifier that I can find at the  $m^{\text{th}}$  stage.

Well, the best classifier can find the one for which this  $\sum$  is empty the right, and get everything correct classified correctly, so that is the best classifier. But now increase the cracks right, remember our classifiers are all weak classifiers, and exactly that is a basic assumption we are starting with right the classifiers are all weak classifiers I can do only slightly better than random, so I have to get nearly half the data points incorrect, right.

So which half should go here which half should come here, and then we can move one data point from to here that here to make it better than half, so which half should go here which half should come here intuitively you tell me, which will incur less penalty, what is small half, it is half man what is small half that will be what clear me, can be more clear as to what is small means, no that is a valid way of interpreting small half tell me,  $w^m$ 's right, so all the  $w$ 's that have a large value should come here because they get  $e^{-a}$ .

The  $w$ 's ensemble small value should go there because they get multiplied by  $e^{am}$ , so what are  $w$ 's is a small values the ones that I have correctly classified up till the previous point  $w$  is the large value are the ones that I have incorrectly classified up to the previous point, so at the  $m^{\text{th}}$

stage what I should be looking at is try to get the data points which I misclassified from the previous stage, try to get them correctly as many as possible right.

So that is essentially the intuition behind ADABOOST, so at every stage what you do is you try to look at the previous stage see which are the data points you misclassified it tries to get them correctly in this stage, right. It is okay if you make mistakes on data points that you have correctly classified till the previous stage why is it, okay. And because those classifiers can adjust for it okay, then we will look at how we will do this again right.

So I am going to call, so it is all the weights of all the data points I got correct at  $m^{\text{th}}$  stage right, likewise weight of all the data points I made a mistake on at the  $m^{\text{th}}$  stage right, so then I can write my  $e$ 's simply as okay, so if you think about it the value of  $\alpha$  really does not matter in my choice of  $k_m$  right, regardless of the value of  $\alpha$  right, regardless of value of  $\alpha_m$  whatever argument I gave you this no holes right, the idea is to see how much of the weight, the weight you can push to  $W_c$  right, and how less of the weight you keep in  $W_e$ .

I mean there is total of weight right, there is some total  $W$  right,  $W$  is a constant okay,  $W_c + W_e$  is a constant, the goal is now to see how much weight you can push into  $W_c$  as posted,  $k_m$  will be the classifier that rise as to my  $W_c$ , so how we do this well you can use you can classifier they images that can assign based data point, we discussed very briefly in session the case right, we can assign ways to data points and you can essentially multiply the error that you make on a data point by the corresponding weight, right.

So the error that you make you multiplied by the corresponding weight so that you can use weighted minimize other ways of doing this, so one way people see what I am saying over  $k_m$  right, you see what you are supposed to do to get your  $k$ , the  $k_m$  is such that maximum weight goes into  $W_c$  they are splitting your  $W$  into two parts and depending on what data points are making mistakes on right, the data points you do not make mistakes on contributed  $W_c$  the data points you make mistakes on contribute of  $e$ .

If you want to see how much larger you can  $W_c$ . We basic that is the classifier you have to find, before that you use some kind of a payment method, so one way of achieving this is do the following you are saying weights to all the data points now what you do, if you go and sample some of these data points according to their weights, create a new training set by sampling from

this data points are given things according to the weights, so what does this mean points for which the weight is higher you get to sample more often into this data sets, points for which the weights are very low I do not even appear in the data set, right.

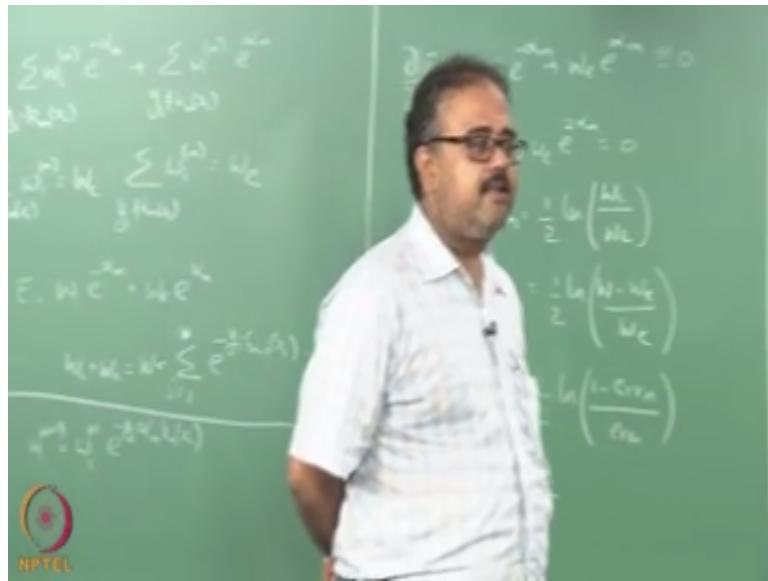
So the points appears multiple times in the data set then when you are trying to minimize the training error you are likely to get a point correct, so instead of using a directly using a weighted training algorithm people simulate that by sampling from the data weights okay, so what has happened unfortunately because of this I change of tends to compact by bagging and boosting in the minds of people and if you look at some of the data mining text books especially some of the earlier data mining textbooks exciting and boosting we needed will described in a very similar fashion right, what do you do in bagging whenever you add a new classified.

So in the older textbooks how they describe is that what you do in bagging is every time you generate a new sample you sample uniformly right, with a replacement right you with sample replacement and boosting the differences every time we generate a new sample you use the prediction error from the previous thing there is the only difference between bagging and boosting right.

But operationally if you think about it, so there is the only difference between bagging and boosting, but then boosting is inherently serial and then there is this error minimization property right, but that never comes across and people just tend to think of boosting as bagging with the different sampling distribution right, whether it is incorrect at the fundamental principles of the two things are very different okay.

So we have found  $k_m$  now right, so we all know how to find  $k_m$  you do some kind of weighted error minimization you find  $k_m$ , so what is next, what is next we need to find  $\alpha_m$  right, see regardless of what value of  $\alpha_m$  you choose the minimizer is the for  $k_m$  is the one that gives you maximum weight into  $W_c$ , correct. But then having chosen a  $k_m$  I now have to choose an  $\alpha_m$  that gives me the error detection, so how do you go about doing that.

(Refer Slide Time: 25:28)



In fact we can do our, so set is equal to 0, so  $\alpha_m$  is essentially  $1/2 \ln 1 - \text{the error rate}$  it is error rate is essentially the weight of the data points on which you are making a mistake divided by the total weight right, so this is for the km classifier alone right, We is the data points on which the  $m^{\text{th}}$  classifier alone makes the error not  $C_m$  but km correct, so that is what we divided these things into right, so this is the thing where km makes error, so essentially that so the data points on.

So essentially, it tells you how good the classifier if the classifier is really good right not just on the data points that you are interested in but on the entire data set. If the classifier is very good then the weight will be high that is the classifier has an error of 0 what will happen rate will be infinity because the only classifier you will need right you have a header of 0 on all the data points why do you need other classify just that one is enough right.

But then suppose it has a very high error, error close to 1 where it will be 0 okay, so depending on how good the classifier is this way it will vary okay, and then anything else that you have to do I have found km, I have found  $\alpha_m$  what do I have to do, I have to change my  $W$ 's now for the next stage right, so what is my  $W_i$  is  $e^{-y_i C_m - 1(x_i)}$  right, so now it has to become  $e^{-y_i C_m x_i}$  so what is the right best way to do that, this multiply the existing  $W$  by  $e^{-y_i C_m x_i}$  right, does it make sense after you have done that you come here okay, I do not erase that part right. So because you need the  $\alpha_m$  here for your update, so once you find the  $\alpha_m$  you come back here and change the weights of

all the data points by this amount okay, as it makes sense so that is a plain simple version of ADABOOST okay.

So in fact we can show that the exponential loss function is closely related to the deviance right. An equally popular version of boosting called logic boost exists, where we use the deviance the logistic function right, the log odds function that we used for logistic regression you can use the same error function and then derive all the update rules that we just did for the exponential loss function you can do the same thing for the logit function the log odds function also. You can come up with similar update rules okay.

So the recent ADABOOST is so popular is because it deals such very simple updates right, if you think about it all the computation you do is okay, you find a classifier that minimizes this weighted the error right then you come back and compute this  $\alpha_m$  and then you go back and change the weights and then repeat until you are happy with the performance of the total classifier right, and both with basting I mean bagging and boosting the command both here start basting things anyway if you do both decision trees are very popular classifiers for this, okay.

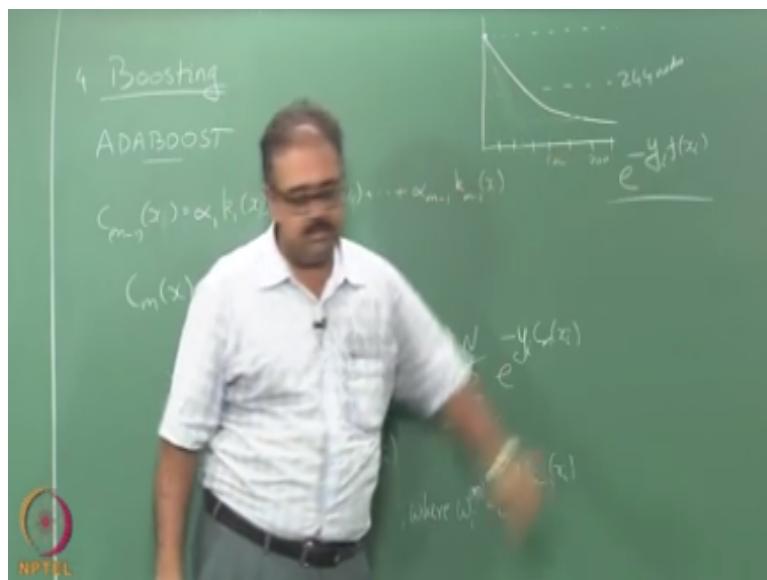
In bagging it seems to make sense right, why you want to bag decision trees they are notoriously unstable, so if you want to, if you bag decision trees you get more stable estimates, why would you want to bag, why do you want to boost decision trees, are they weak classifiers. Exactly, so what do you do with decision trees you can do the most extreme thing you can just have one node, just have the root node right, one node what can you do with one node decision tree.

Yeah, that is somewhat like linear right yeah, so somewhat likely linear I agree. However, people call it decision trees right, so one node decision tree because of the way I choose which feature I pick right, I will use information gain or Gini index or one of those things okay, I will at least take 50% classification otherwise I would not even split right on that 50% will I will be better than 50%.

I will be better than random even if I split on one node right, so I will split on one node and or maybe if the performance is too weak I can perform I can do a two level tree okay, these are called decision stumps, I do not build a full tree, but it is like chopped off at a very close to the root right, so one not two levels of the trees. However, they are v-classifiers and they take very

little time to estimate and I can do many, many, many of these very quickly essentially what I do is, I boost these decision terms okay. In fact, there is one result in the book if you look at it right.

(Refer Slide Time: 33:38)



So I do not remember the exact scale on the y-axis. However, the x-axis is the number of levels of boosting that they do right, and so on so forth, the number of levels of boosting that they do so 100, 200, 300 and so on so forth, so a single stump it gives some performance level at that height okay, just one strum the best single stump gives you a performance there. They trained it on the full data, and they get a performance here, and this is like a 244 node tree, 244 nodes is a fairly complex tree they built and that is the performance that they get right.

And then they did boosting the start here obviously with a single node right, and then they do boosting, and then they find that the tree the performance just keeps improving as I do ADABOOST. Remember, and these are all single node trees okay, they are all single node trees and so essentially by the time they reach 100 that means they have only 100 nodes basically.

They are way better than the 244 nodes that you get with a singletree right, and they reach 244 notes they are like more than twice as good as the single tree they built with 244 notes.

Because the objective function you are minimizing is something very, very different right at every stage, you are changing the function and you are focusing your efforts on actually getting to the harder parts of the space right so that essentially it is little magical. This is more dramatic is it something like this but look at the book crippled for the exact figure right, so this is amazing posting is very powerful. I talked briefly about the random forest in the next class which is you do not even have to do any decision making right, you do things randomly but then do a lot of them right that is also very powerful.

So random forest is not a boosting technique by the way the random forest is a bagging technique right, but then that is also very powerful.

**IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)  
Copyrights Reserved

## Introduction to Machine Learning

Lecture-61  
Gradient Boosting

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

(Refer Slide Time: 00:16)

$$\begin{aligned}
 &\text{Gradient Boosting} \\
 &\text{GBDT} \\
 &T(x; \theta_j) = \sum_{j=1}^J \gamma_j I(x \in R_j) \\
 &\theta_j = \arg \min_{\theta_j} \sum_{i=1}^N L(y_i, \theta_j) \\
 &\text{Boosted Tree: } f_M(\omega) = \sum_{m=1}^M T(x; \theta_m) \\
 &\hat{\theta}_m = \arg \min_{\theta_m} \sum_{i=1}^N L(y_i, \theta_m + T(x; \theta_m))
 \end{aligned}$$

Right, so I want to talk to you about the interesting idea known as gradient boosting. So all of you remember what boosting is about right. What is boosting? What is bagging? Boosting yeah, so boosting is specifically a stage-wise process where at every stage you try to boost the classifier from the previous state says that the error is minimized right, an error is reduced not necessarily minimize, but the error is reduced right.

So that is a characteristic of boosting, so at every stage, you have to look at the errors from the previous stage, and you are trying to reduce that right. So we looked at AdaBoost right, one of the most popular boosting algorithms. I then told you that AdaBoost uses exponential loss and

that it is related to the logistic loss. So you can use the logistic loss function and derive a boosting algorithm called logit boost right.

But it is very similar properties to AdaBoost. However, AdaBoost is more popular, especially from an analysis point of view and things like this because it is not nice properties, okay. There is yet another approach to boosting that is gaining a lot of currency recently. It is called a gradient boosting, at not well recently would mean in the last decade or so, compared to AdaBoost, which is several decades old right.

So sometimes they even call it gradient boosted decision trees right, gradient boosted decision trees because you use this specifically in conjunction with trees right. And in fact, in many applications now, gradient boosted trees are getting hard to beat right. And so, we just reintroduced some notations that you might have forgotten right. So  $I$  is an identity function, which is 1 if  $X$  belongs to  $R_j$  is 0 otherwise right.

And so, this is summing over all regions, so  $R_1$  to  $R_J$  and  $\gamma_j$  is essentially the output I am going to produce if  $x$  lies in  $R_j$  right; this is the regression tree thinks. So what is my  $\theta$  here? It is all the  $R_j$  is the specification of the  $R_j$ , and the  $\gamma_j$  is for each of those regions. So that is my  $\theta$ . And typically, we pick some loss function if it is regression; it is going to be squared loss and then right.

So I look at the loss incurred when the actual output is  $Y_i$ . The output I am giving you is  $\gamma_j$ , so for all data points  $X$  that belongs to a region  $R_j$  the output will be  $\gamma_j$ , so this is essentially the loss there and sum this over all regions, and this is the rectum just recapping the decision trees for you right. And then we looked at greedy methods for finding  $R_j$  right, and given an  $R_j$ ; we knew how to fit  $\gamma_j$  right, so given that we looked at some greedy search methods right.

So you can do, you can do boosting with trees also just like you did boost with other classifiers you can do boosting with trees right. So I have  $M$  trees so essentially it is taken some of the output of all the  $M$  trees that gives me my boosted tree right, remember that I mean this is not a single tree okay it is now a forest, then I have a collection of trees a collection of trees is a forest right.

So it is a forest, so I do that, and the difference here is, can people at the back see this right. So this is essentially when I find the parameters for the  $m$ th tree right, so I am going to look at the

classifier or the predictor that is formed by the first  $M-1$  trees right. And then I am going to find that tree okay, whose output I will add to this predictor right, and you search for computing the loss right.

So for every data point in my training data right I look at the way I look at the output produced by the  $m-1$  stage 3, I look at the value that is added by the  $m$ th tree, so this is the output produced by the  $m$ th tree I look at the value added by the  $m$ th tree to that right. And then I will compute the loss function okay make sense yeah.

So the basic area is that every point this is just forward stage wise addition that we like whatever we did before introducing boosting right that is exactly that. So now this becomes boosting because I am explicitly trying to figure out what the residual error is from here okay, and trying to adjust for the residual error using my tree the new tree I am learning right. So when will it be the residual error, when it is a regression task, and squared error is my metric right.

So and the loss function is squared error right, and I am so trying to solve the regression problem right, then essentially what I will have to do here is take the residual error. So whatever the tree does not explain them-1 stage tree whatever that does not explain, so that error I will have to explain using this right. So if you think about what we are doing here, so you will first build one tree to predict your output as best as possible right.

The predictor function as best as possible will build a tree, then what you will do is okay you will take the residual of that, build another tree that predicts the residual as well as possible and add the output to this okay. And then take the combined thing find the residual of that build the third tree which will predict the residual and add it back to this and so on so forth, you just keep doing this right.

So that is essentially what boosting increase means, so you still not come to the gradient boosting part okay. Finally, it will look very similar to what I am telling you now, but we have still not come to that part yet right. So as with regular decision tree learning given the  $RJ$ 's right finding the  $\gamma J$ 's is easy given the  $RJ$ 's finding, the  $\gamma J$ 's is easy. But the problem is finding the  $RJ$ 's, in general, it becomes a little tricky finding the region's becomes a little tricky because I have to take into account the other tree's output also right, in general, right.

But I am talking about the squared error. It is very easy because things nicely decouple right when I am talking about a squared error. I do not have to worry about  $FM-1$  after I compute the

residual right. The residual could have been generated by any classifier any regress righty; it does not have I do not even have to worry about the fact what generated the residual was a tree right, I do not have to worry about it, all I need is just the residual.

The residual then becomes any function right, so with squared error boosting becomes just like learning a series of decision trees, nothing special about it. But if you have other kinds of loss functions, then we will have to worry about how to accommodate it, but at least in this case of squared error loss okay.

(Refer Slide Time: 10:24)

Squared error loss: Pick the tree that best predicts the residual

$$y_i - f_{\text{true}}(x_i), \hat{y}_{ij} - \text{average error in } R_{ij}$$

For 2-class, & Exponential loss fn - AdaBoost on Trees

---

Differentiable loss fn

$$L = \sum_{i=1}^N L(y_i, f(x_i))$$

$$\hat{f} = \arg \min_f L(f) \quad f = (\hat{y}_1, \dots, \hat{y}_N)$$

So that is essentially your target function right, and what will be the  $\gamma$  hat that you will need just be the average residual error in the Jth region right. Any questions, so in fact, there is another case where it becomes simple, which is essentially, so for two-class problems and exponential

loss functions, what you think we get, it becomes the same as doing ADABOOST with trees okay, the two-class problems.

But, so it turns out, there are tricky things here right, so if there is if it is a multi-class problem okay, then things do not decouple as nicely. So if you have two-class problems and your loss function is exponential loss okay you can show that this is essentially the same the computation that we are trying to do here right that minimization everything that we are trying to do here essentially reduces to the same solution that you get if you did the ADABOOST derivation on decision trees okay.

But these are the two cases where this thing simplifies that, for example, if you are trying to use deviance as a loss function, then things do not decouple this easily okay. So these are things that you have to keep in mind, so this is one part this is essentially telling you how to do boosting with trees the regular way okay. So let us look at something else now, so I suppose I have some differentiable loss function some loss function which is which I can take the derivative of..

So if I want to do, so if you want to take a numerical approach to optimize this kind of a loss function typically what will I end up doing I will start with some guests for a solution right take the gradient of the loss function for the parameters at that solution point we will number gradient descent right. Then I will compute the gradient, and then I will move in the opposite direction of the gradient then I will move a small step in the opposite direction of the gradient go to a new place and compute the gradient again and then move again and so on so forth until I converge to the right answer right.

So if you think about what this is doing this is something like okay take the initial solution, okay, then I add another solution to it which is essentially the gradient times something then I add something more to it, so I will add something so essentially the solution I am computing a sequence of additions that they have done on the basic solution I started with right.

So even though one way of thinking about it is at every point, I give you a parameter vector, but the parameter vector itself is composed of a sequence of additions. So I can think of it as first starting with initial guess for my parameters, then adding something more to it, then adding something more to it, then adding something more to it, and adding something more to it, till I come to the final answer right.

So that is one way of thinking about it, so let us try and write this down a little formally. So now I am going to see what I can do about this  $f$  is for the time being ignore the tree constraints I will come back to the trees later, time being let us ignore the tree constraints right. So what I need is, so just when I am trying to do this numerically, I am just operating with a single data set right.

So when I say  $f$  what I have looked at is a point in  $R^N$ , so what does that mean let  $f$  means okay what is the value of  $F$  at  $X_1$ , what is the value of  $F$  at  $X_2, X_3, X_4$  up till  $X_n$  so when I impose constraints on  $F$ , then I will be restricting the kind of vectors I will see. But in general, when I am talking about  $F$  in this context, I just mean like an  $n$ -dimensional vector right. So you can think of it as a point in  $n$ -dimensional space right.

So typically, what you do is you start with some solution right  $F_0$  you start with some solution let us call it  $H_0$  right. So you can think of it like this I start somewhere here that is my  $F_0$  right. And then, I compute the gradient and move in the opposite direction right, so I take a small step in this direction, so I come here that gives me a new set of parameters right. So this is  $1 \theta$ , this is another set of  $\theta$ , and this will give me another  $F$ .

But instead of saying that this will give me another  $F$ , so I am going to say that okay this is one  $F$ . I add something to it right, so that gives me the second  $F$ . So what I am computing in every step is the amount that I add to the previous solution to derive my new solution okay. So I am calculating  $\theta$  and  $F$  here, so what I have here is  $\theta$  corresponding to every  $\theta$  I have here there will be an  $F$  corresponding to every parameter setting I will have that will be output vector  $F$  when I change  $\theta$  this values will change right.

So when I am here I have one solution right, so when I want to go here that essentially means that whatever  $F$  vector I have here I will have to change each of those coordinates by some value so that I will end up here right, those values I change the coordinates by it is my  $H$  vector right, is it clear what we are doing here.

(Refer Slide Time: 18:58)

$$\begin{aligned}
 f_0 &= h_0 \\
 f_m &= \sum_{m=0}^M h_m \quad h_m \in \mathbb{R} \\
 -\text{Steepest Descent} \\
 h_m &= -\rho_m g_m \\
 g_m &= \left[ \frac{\partial L(y_{m-1}(x_i))}{\partial F(x_i)} \right] \\
 \rho_m &= \arg \min_{\rho} L(f_{m-1} + \rho g_m)
 \end{aligned}$$



So, what is the normal mechanism by which we will do this right So, I have been using the same example so far right that steepest descent will pop up right even the other way of doing this optimization right. But steepest descent is the one that we are all familiar with the one that I have been using as an example here so far right. Since I have not chosen any arbitrary parameterizations to form a  $\theta$  right for the  $F$ , I have not chosen any parameterization  $\theta$  or anything right.

So the parameters of  $F$  are the output set each one of the input points to see the way I characterize my function  $F$  is looking at okay, what will be the value of  $F$  at  $X_1$ , what will be the value of  $F$  at  $X_2$  and so on so forthright, I do not have any other parameterization for it. So instead of finding your  $\delta L / \delta \theta$  you find that I am writing it as  $\delta L / \delta F$  okay. So  $F(x_i)$  is essentially the output of  $F$  at  $X_i$  and what is  $F$  here, it is  $F_{m-1}$  because I am determining the  $M$  stage I am looking at the  $m-1$  guess for my function right.

So the steepest descent direction would be saying that okay, so  $G_m$  is the direction in which you have to move because that gives me the direction in which the or rather  $-G_m$  is the direction I have to move. After all, that gives the direction of steepest descent. And  $\rho_m$  gives me the step size I have to take in that direction how larger steps I can take in the direction. So how is the flow I am determined you should look very familiar right?

This is exactly how we did the ADABOOST derivation. Now people are thinking about it ADABOOST derivation like this yes, we did go back and think about it okay. So very similar not exactly the same thing, but very, very similar the exact the same steps that we did right we first found out which way we have to change it right. Then we found out what the step size should be, and the way we did it was okay, I have already had a classified till  $m-1$  stage okay, what should I do at the  $M$  stage.

To minimize the error, so this is exactly what we did, the idea behind each of the steps are the same. The mechanics might have been slightly different right. So once I get this, then I do okay, right is it clear people are doing so far right. So whatever we are trying to do right, is nothing, so there are two different parts here okay, if you people are getting confused. Hence, the first part here talked about boosting trees okay, the second part here talks about taking some differentiable loss function and trying to do some kind of a stage-wise process on it okay.

I just took your normal gradient descent procedure and told you that you could think of it as a stage-wise process, I guess as we did with boosting, we can think of it this additive model right. So whatever you will learn here right, so now the thing is how we connect up the two will somehow account for that later. I do not want to erase anything from the board because what we are doing right now is connecting the two parts right. So I do not want a erase anything from the board, so you can see both that and this well we are looking at this right.

(Refer Slide Time: 23:48)

$$\tilde{\Omega}_m = \arg \min_{\theta} \sum_{i=1}^n (-g_i - f(x_i; \theta))$$

Fit a tree as close as possible to gradient desc.

<u>Reg</u>	$\frac{1}{2}(y_i - f(x_i))^2$	$y_i - \frac{1}{2}(x_i)$
$ y_i - f(x_i) $	$\text{Sign}(y_i - f(x_i))$	
$\boxed{\text{Hart}}(x_i)$	Demand	$I(y_i = \sum_k) - P_k(x_i)$ ( $k^{\text{th}}$ component)
$m$ )		
	$\hat{f}_{lm} = \arg \min_f \sum_{x_i \in R_{lm}} L(y_i, f(x_i) + \hat{v})$	

So far, so  $G_m$ ,  $G_m$  is some kind of an unconstrained maximal descent direction I do not have any constraints on  $f$  for anything else right what is the maximal descent direction and essentially get  $G_m$ . So now what we are going to do is to say that hey, all of this is nice, but I would like some parametric forms for what I am doing right; otherwise, things become too complicated. So what I am going to do is I am going to fit a tree.

So what I want to know is  $G_m$  right, so if you think about it, so I need to compute  $G_m$ , and instead of doing this in this arbitrary unconstrained form, I am going to build a tree that approximates  $G_m$  as closely as possible okay. So you should note something here what is it you should not, so for all this while I have been very carefully writing  $L$  for the loss function well, I am trying to keep this as generic as possible right.

But here I wrote a squared error loss function because it does not matter what the problem that I am trying to solve is, it does not matter what this loss function is okay. After all, what I am trying to solve now is trying to approximate a vector I trying to approximate a direction right by a tree essentially I am always solving a regression problem here right. So if you think about it, Gim is going to be some kind of a vector right. Gim will be some kind of a vector. All I am trying to do is predict the value of that vector component right.

So I am just doing regression regardless of whether my original problem was a classification problem or a regression problem or whatnot, I could use any loss function here, and I could use any loss function here this is the crucial difference you should appreciate right, for the actual or solving the problem that I could be using a different loss function here okay. But when I am

building the  $M$ th stage decision tree, all I will be doing is regression, because all I need to predict is what is that particular gradient descent direction for that input value  $X_i$  right.

So this is what I am going to do, how am I going to go about doing this. Well, it depends on what this loss function is okay, so is this loss function, so if the problem I am solving is a regression, this is what I am solving, this is the squared error loss function okay, this is not this okay, this is that. So if this  $L$  is squared error okay, then what do I get here is essentially what is this, give this is  $G_m$  essentially the gradient of the loss with respect  $F(x_i)$  right, the gradient of this for  $F(x_i)$  is essentially the residual  $Y_i - F(x_i)$  right.

So this is basically –  $G_m$  if you would think correct. So now, what happens if I am doing regression with squared error loss function, and I am trying to do this gradient boosting right. So I am trying to build a new tree that predicts the direction of the gradient what do I end up doing, I end up predicting this is the residual right. So I end up predicting the residual and here what we said if you are making the squared error loss pick the tree that best predicts the residual.

So that was derived from just the basics of boosting regular boosting. Well, here I am talking about a technique that can do boosting on trees regardless of what is the underlying loss function right, but it does boosting using trees okay. So that is a cool thing about gradient boosting right, so you always are solving a regression problem as far as the tree is concerned, and solving regulation problems using trees is very easy right. Your solving regression problems using a tree is always regardless of this loss function.

If you remember when we are deriving this boosting update, I said squared error, and for two-class exponential loss, the boosting form is easy right. But now if you busy doing the gradient formulation of it okay regardless of what you are doing with the loss function, you can still do boosting with trees. So that is why gradient boosting decision trees have become very popular now because you can do all kinds of cool stuff with it.

So what about this right, and suppose you are doing classification let us take deviance as the loss function right, we will remember deviance, we looked at deviance multiple times right, and turns out that. So, if the  $i$ th class is  $G_k$ , then it will be 1 minus that the data point  $X_i$  is in class  $K$ , the probability of data point  $X_k$  belonging to class  $K$ . So it will be 1 minus that, and if the actual class is not  $K$  right, then it will be a minus probability of  $X_i$  belonging to class  $K$  okay.

So this is like the  $i$ th component of it, so  $i$ th component of my  $G_i$  okay. So again, what I have to do, I have to take this expression to plug it in here and do regression again, I will take this expression to plug it in here and do regression right. So all you need to do is figure out what is the derivative of your loss function with respect here  $F(x_i)$  right, and then once you find out the derivative, you just do the regression for that for each stage in your decision tree okay. So what will be my  $\gamma_j$  this is  $\gamma_j$ , such  $\gamma, \gamma_{jm}$ , what will be  $\gamma_{jm}$ ? So earlier, we said it would be just the average residual error in  $R_{jm}$  right.

So, in this case, it is going to be, so once I have found out what the actual regions are right, once I have found out what the region. So this will give me the regions right, so once I find out what the actual regions are, I will do the following. So what have we done here, okay? So earlier when we are finding out the, when we are building decision trees right, so the way we found out the regions right.

If you remember the way we found out the regions were we postulated a split point right, and then for that split point we figure out what is the best  $\gamma$  in both halves of the tree right. Then we took value for that, and then we kept looking at all the possible split points right, the splitting variables and split points for all the possible combinations and for each one of them we evaluated what the resulting residual error would be right.

And based on that, we pick the split point, so here we are doing something different when we are splitting picking the split point right, we are going to pick the split point such that the residual error in predicting the gradient is minimized right. I am only predicting the gradient here; the residual error in predicting the gradient is minimized. But when I finally decide what the output that I am going to give in each of the final regions right is.

So, in the regular decision tree building, I would have already solved the optimization problem right by the time I reached that point. So I know what the solution that I have to give is, and it is the one same thing which I use for splitting criteria. But in this case, when I finally give the outputs, I am going to look at the loss function right of data points that fall in that region look at what already exists, which is  $F_{m-1}(X_i)$  right.

And look at what I need to add to bring up the output so that the loss is minimized. Whatever is the loss function, so when I am doing the loss function here I will no longer be using the squared

loss I will be using one of these, I mean this is quite a loss I could be using the absolute gloss or I could be using deviance or whatever is the measurement that I want right, I will use this loss function the loss function I use there, I will use that here in order to figure out the outputs okay.

So you let this sink in a little bit, so it is a pretty cool idea right. So I have somehow come up with a mechanism where I can use decision trees in a very powerful way, because finding regions. At the same time, I do regression is very easy with decision trees because we know how to use squared loss and there are lots of tricks and optimizations that you can do when you are searching through with square loss right.

Now we looked at some of them, but there are many other things that you could do so what I am going to do is for all my tree growing part right I am just going to use the regression trick right. I finally have to give an output at that point I will use whatever is the true loss function I want right, that is why it is called gradient boosting. So I use the gradient at every point to boost my performance right.

So the way I fit my, so if we think about it, this might not necessarily be an ideal way of doing things, why is that right. So if I want to predict  $\hat{\gamma}_j$  that truly minimizes forget about the  $R_j$ 's right, that truly minimizes the loss function, I might want to split the space differently right. But I am abusing the gradient information to split the space, then whatever splits I get whatever regions I get by using the gradient information, I am using the same regions in order to reduce the error also right.

It is fine as long as I am unconstrained right, as soon as you put in the constraints of trees it is not entirely clear that the tree that you want for getting the representing the best  $\hat{\gamma}$  is the tree that you want for representing the best GM it is a very, very subtle point you have to think about it a little bit. But more often than not, it turns out to be fine, okay, but there is no guarantee that the best three for predicting GM is the best three for representing your  $\hat{\gamma}$ 's okay.

That is two slightly different things, but still, it turns out to be fine right. So all of this discussion collapses if you move to L being the residual, I mean the squared error right, L is squared error everything looks the same you already solve it here is an easy thing to do. What I want to point out is that it is essentially the same square error mechanism you can build boosted trees for any loss function that you want, it can be regression, it can be classification.

So whatever it is, you can build a decision tree. So one thing that you have to be careful about is that you do not overfit the things any time right. So you have to be careful about not overfitting the data because you are only working with the n data points always right, you can build a very complex tree that will try to overfit the gradient for just the training data right.

So that is not a good idea, so try to keep your tree down the complexity of your tree down. Quite often people choose the size of the tree a priori. And then, you might end up adding more trees than necessary because he chose too small a tree, but at least you will avoid overfitting right. And so that is it for gradient boosting.

**IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

# **NPTEL**

## **NPTEL ONLINE CERTIFICATION COURSE**

### **Introduction to Machine Learning**

#### **Lecture-62 Random Forests I**

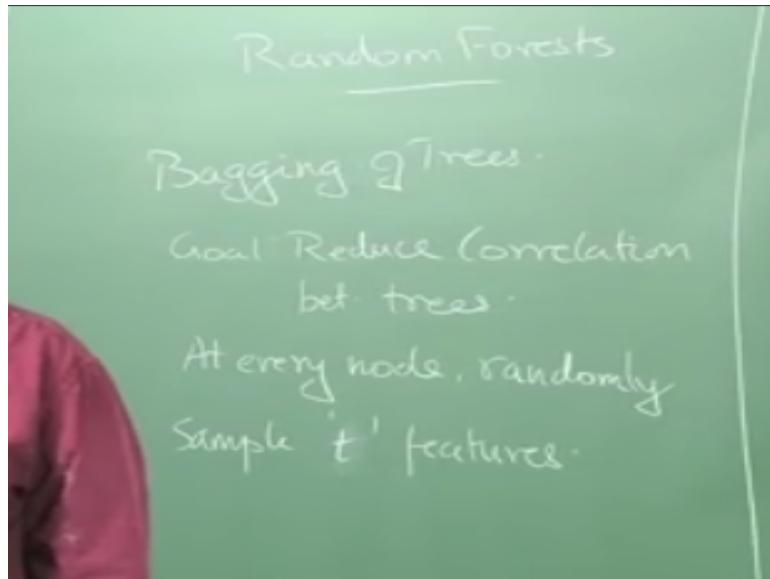
**Prof. Balaraman Ravindran  
Computer Science and Engineering  
Indian Institute of Technology Madras**

So now we know that trees are great candidates for boosting as well if you are using gradient boosting right. However, trees are great candidates for bagging as well as right. What is the important property in bagging that we talked about? What does bagging help us reduce variance, right? So bagging allows us to reduce variance. You can show that the reduction in variance is highest if the classifiers that you are building okay, or not correlated right.

So I am building many, many classifiers, and the classifiers are predicting the same output right. So if the classifier parameter set I am estimating or somehow if we can make them uncorrelated right, then the reduction in variance is maximum it kind of intuitive right. If the classifiers are very correlated, there is no point, they are not different classifiers right, and they will give me the same output, so the variance will be high.

So if you can somehow make the classifiers uncorrelated, then the reduction in variance is high right. So there is a particular relationship between the amount of correlation between the classifiers and how much you pay in terms of the reduction variance right. So I am not going to derive that, I just point it out to you, so if you want, you can look it up; it is there in ESF right. And if you think about what we are doing with bagging right, we are taking one data set right, and you are sampling with replacement from that right.

So the probability of the trees that you are generating being correlated express rather high, the probability of the trees that you are generating being correlated is rather high. So it can become up with some way to reduce the correlation between the trees you are constructing right.  
(Refer Slide Time: 02:17)



I am going to be doing bagging right, but the goal is to reduce the correlation between trees. The people who came up with the random forest had a very simple idea for doing this right, and you start doing bagging as you would normally do okay. So you have your data set, then you create a bag by sampling with replacement from that data. Now, when you start building the tree on this data set, so what do you do at every node right, sample some P features from your feature set, and use P for the regular feature description right..

So let me use a different, so we have a total of P features right your data points come from some RP space, you use randomly sample some T features from that P features okay. Find out which is the best split point, which is the best split variable among these P features alone, split the data go down to each of the subsets repeat the same process, sample, and other T variables right not necessarily this joint, and just sample okay, sample again T more variables okay.

And then, try to find out which is the best split point among these T variables and keep doing this. So what does this get us to see if you had worked with the same data set right, which I mean if I am just done bagging at the root level it is highly likely that each one of the bagged trees would have picked the same attribute right, just because you have sampled it again right? So it does not mean that the very predictive attributes will get discarded.

So at the higher levels of the trees, it will look very similar right. But now you are getting rid of that I said okay. I had chosen randomly. I have chosen T variables, and only from them, I am choosing the best variable; therefore, I am reducing the chances of the trees looking similar. We

can show that this leads to a significant reduction in the variance in the bagged estimate, and random forests are performing very well.

Random forests are competitive with gradient boosted trees in some applications and vice versa, right. So boosted trees are better than random forests in some applications, and random forests are better than boosted trees in some applications. And because sometimes, till sometime back there are very efficient random forest libraries and people use random forests a lot right. But now there are also very nice libraries available for gradient boosted decision trees. And therefore, try everything and see which works right.

**IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

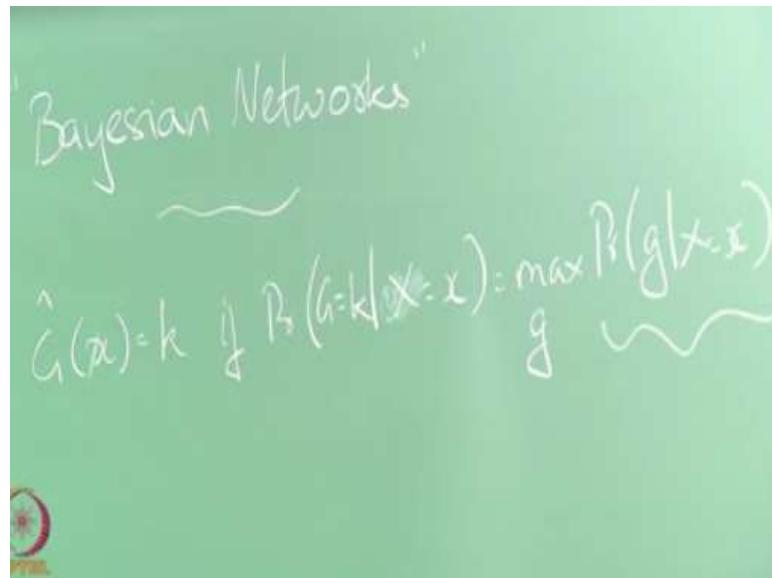
Copyrights Reserved

**Introduction to Machine Learning**

**Lecture-63  
Naive Bayes**

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

(Refer Slide Time: 00:17)



We mentioned the Bayesian classifier or the Bayes optimal classifier in the very first class. We talked about the nearest neighbour methods we probably have one 'X' and how do you know what its class is. The probability of the class given a data point because you have only one data point in your training and therefore we took an average over a region and so on so forth. We will use that estimate over a region for finding the probability. So, this is how we motivated what KNN right the K nearest neighbour classifier right.

$$G(x) = k \text{ if } P(G = k | X = x) = \max_g P(g | X = x)$$

(Refer Slide Time: 02:15)

$$P(g|x) = \frac{P(x|g)P(g)}{P(x)}$$

So now I am going to take a slightly different tack right, so what we want we want the probability of right 'g' given 'x', we want the probability of 'g' given 'x' right, So we have our friend the Reverend how many of you know that Thomas Bayes was an ordained priest okay, so we have Reverend base to help us outright, so is essential. Right so then we have a lot of quantities that we can estimate this, you can estimate this from data. How will you estimate that from data?

### Bayes Theorem

It describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It is expanded as,

$$P(g|X = x) = \frac{P(x|g)P(g)}{P(x)}$$

$P(x|g)$  – Class Conditional

$P(g)$  – Class Prior

$P(g|X = x)$  – Posterior Probability

$P(x)$  - Data Prior

Okay, we will come to that, Can you estimate this from data? It is the fraction of a particular class you can do that, or you can make assumptions about the class densities, and you can estimate the parameters of those from the data and so on so forth. That is fairly straightforward to estimate right, so what about this guy? But in general, if you wanted to do the max, yes if you don't want to do the max then this becomes a question is how do I go about estimating this right.

(Refer Slide Time: 04:19)

Bayesian Networks

$$P(x) = \frac{1}{Z} \sum_g P(x|g) P(g)$$

$$P(x|g) = \frac{P(x|g) P(g)}{P(x)}$$

$$P(x) = \sum_g P(x|g) P(g)$$

$$= \sum_g P(x|g) P(g)$$

So there is one way of doing this how do you do that, so you can assume that you have a is what is the probability of 'x'. That essentially some of the numerator for all possible 'g' so I get my denominator so this I can do so all we need to know is how to estimate that right, so can you estimate it?. Yes, it is the distribution of data points given the class. But the main problem we will face is the sparsity issue.

The  $P(x)$  can be estimated as,

$$\begin{aligned} P(x) &= \sum_g P(x|g) P(g) \\ &= \sum_g P(x|g) P(g) \end{aligned}$$

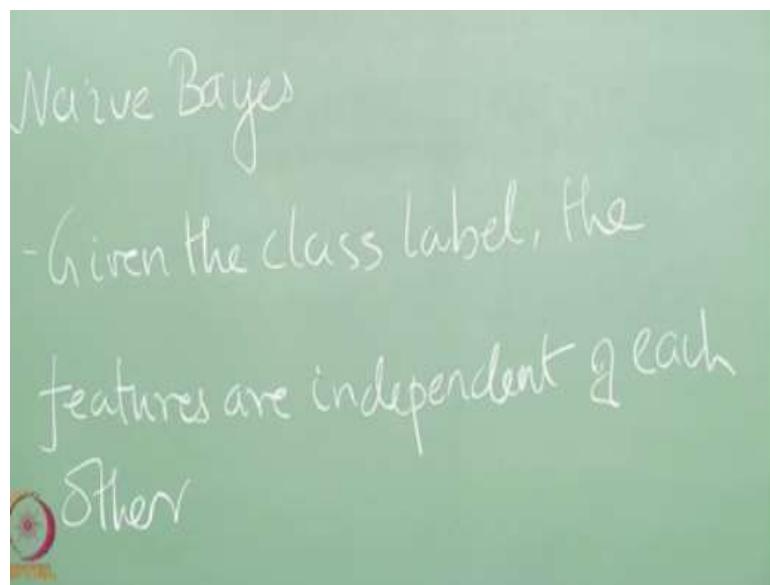
Often enough we might get one sample here one sample there and so forth. This will not cover the entire data distribution to get a good estimate of the probability distribution. This is especially happens in really high dimensional spaces. Suppose we are assuming X comes from some  $R^p$ . If  $p$  is a hundred-dimensional thing  $p$  is 100, so if X is a data point in  $R^{100}$  right, so what is going to happen data points are sparse in this vast space right. This makes the estimation hard, so we have to make assumptions about the distribution.

This is called as class condition distribution because their probability of x conditioned on the class right. So sometimes you also call them what likelihood yeah I thought somebody would say

likelihood before anything else. Still, people are just keeping quiet, so this also called the likelihood but is the class conditional distribution right so what is the difference between likelihood and class condition distribution, so likelihood is a function of 'g' right I remember I kept repeating that multiple times when we looked at likelihood when I am conditioning it on some parameters  $\theta$  okay. Still, it is a function of  $\theta$ , so 'x' is the same.

But when I am talking about class conditional distribution I am talking about the probability of X okay it is an a function of x conditioned on g okay good, we will just go back so the most the simplest assumption that we make.

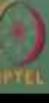
(Refer Slide Time: 08:13)



To get this to be tractable is called the Naive Bayes assumption right, so what does the Naive Bayes assumption tell you it says that given the class label the features are independent of each other right.

(Refer Slide Time: 09:17)

Others

$$P(x_1, x_2, x_3, \dots, x_p | g) = \prod_{i=1}^p P(x_i | g)$$


So what does this tell us it says that if I have the probability of that, I can write this as this is called as Naive Bayes assumption right. Once I do this now it becomes very easy to for me to estimate the parameters. In how many data points has "x\_i" taken a particular value of that particular class. First segregate all the data points by class and then in class one in how many data points did "x\_i" take the value of zero in class one how many data points did "x\_i" take the value of one and class two how many in the how many data points did "x\_i" take the value of zero and "x\_i" I take the value of one so on so forth.

Do you want me to say "x\_i" takes the value of 0 "x\_i" takes a value of 1, "x\_i" takes a value of 2, "x\_i" takes the value 3. I know that just takes a lot of time so just making it binary so that it is easier for me to speak right, so it could be anything right, so "x\_i" could be real value this, for example, our always our setting is our R power p right in that case what do you do some binning you could look good that is one very valid option even though many textbooks do not recommend that but not that they do not there is not that actively not recommend it just they do not even talk about it okay.

But that is actually a valid option, and I will tell you why in a minute but the usually recommended option is to have some kind of parametric form for these marginal distributions these are conditional marginal. If you think about it these are marginal if so this was the Joint Distribution

right then this is just the marginal that is the conditional Joint Distribution, so this is a conditional marginal so for the conditional marginal they ask you to assume some kind of parametric form and the usual one that they suggest is a Gaussian right.

Based on the Naïve Bayes assumption the conditional joint distribution could be written as,

$$P(x_1 x_2 x_3 \dots x_p | g) = \prod_{i=1}^p P(x_i | g)$$

$P(x_1 x_2 x_3 \dots x_p | g)$  – Conditional Joint Distribution

$P(x_i | g)$  – Conditional Marginal of the "i" th feature

(Refer Slide Time: 12:00)

Others

$$P(x_1 x_2 x_3 \dots x_p | g) = \prod_{i=1}^p P(x_i | g)$$

- If  $x_i$  is discrete valued, then we a  
 discrete dist.  
 - If  $x_i$  is continuous, then we a Gaussian

Usually, this is some kind of a Gaussian form for this conditional marginal right can you read it at the back can okay you can read it okay fine now because the thumb didn't stay up long enough, so I was not able to make sure which direction it was you just did something like that and right but then what is the problem in using a Gaussian all of you know the problem just think for a minute or lesser okay not a big issue there is a much bigger issue. Too much inductive bias but the wrong kind of inductive bias why is it wrong what do you know about the Gaussian.

Sorry, close so what does it mean the Gaussian is unimodal? Suppose there are two different values of "x\_i" which are separated right which is very probable right for this particular class if you use the Gaussian what will you estimate the most probable point as, is there mean okay, so if we say 3 is very popular very likely to occur and 5 is very likely to occur your Gaussian will say 4 is the most probable value for x given the class, so you do not want that to happen.

So that is the reason I said binning makes sense, but then the problem is you have to find the right kinds of bins because when you are using the discrete distribution, there is no I mean it can be multimodal right I mean it is no notion of like you unimodality there right so I could have one output having very high probability another output having very high probability I could have like 10 different outputs having high probability is everything else having low probability it could be anything right.

So there that I do not have to worry about these, but then the problem here is hard to find the right binning, so that is a whole set of lectures by themselves, right so how do you bin your input variables there are many ways in which you can bin input variables in there you have to keep coming up with clever tricks depending on the application you have and so on so for this is actually not trivial right and you could do that.

(Refer Slide Time: 15:32)

$$p(x_1, x_2, \dots, x_n | g) = \prod_{i=1}^n p(x_i | g)$$

- If  $x_i$  is discrete valued, then use a discrete dist.

- If  $x_i$  is continuous, then use a Gaussian OR  $B(n, \lambda)$  | Mixture Gaussian

But of course, the Gaussian is just a simple example if you know that the data is going to be multimodal right what should you be using? The mixture of Gaussian and not a single Gaussian. Naive Bayes to work seems to be a very simplistic assumption is even called Naive Bayes even though I have never heard used in any papers or any literature so T. Friedman says that it is also known as the Idiot Bayes algorithm.

So Naive Bases sounds a little better more sophisticated right yeah do you think Naïve Bayes will work well. It did work pretty really well right weren't you amazed you didn't know how simple it was at the time compare it against SVM's did not we compared against the SVM's should so it turns out that the let us say for examples like text classification where you are the data dimensions are inherently very high right it is incredibly hard to beat Navie Bayes you know it looks like it is something so simplistic right.

I mean look at the assumption you are making I have a lot of texts I want to classify them as politics or sports okay that is one very simple problem that is out there is a standard problem that people use for text classification is a standard data asset I want to classify a news article as being sports

or politics. And what I was saying here is given that it sports the probability that I will see cricket is independent of the probability that I will see football am.

Not only that the probability I will say cricket is independent of the probability i will see say Dhoni in the document sounds like nonsense right well if you are talking about Indian media yes right, but in general it seems very surprising, so that is because we are trying to assign all kinds of semantics to what is happening and but the algorithms that we are trying to use whether it be SVM or anything else or really not into the semantics of these things right, so we are only worried about it because we have all these others the knowledge base superstructures that we have built.

And that we are trying to look at the data through that right at the end of the day if you look at it is more a question of things occurring together co-occurring and in a very large document space, so the probability of any words co-occurring right is kind of diminishing right, so if I do not know it is a document about cricket right and if i see the word Dhoni then I will say okay now maybe the probability of me seeing something with actually a sports-related term goes up because I did not know whether it was a sports document or a politics document before I looked at it right.

But given that I know it is a sports document and the probability of me seeing this anyway be higher it is not going to change appreciably because the word Dhoni appeared you see the reasoning if I had not known anything about the document because the word Dhoni appeared in it right the likelihood that it is a sports document goes up right well given the nature of Indian cricket the probability that is politics has not gone to zero yet right so but if I had known that it is already at the sports document knowing that okay it is Dhoni the words Dhoni appeared in it does not appreciably change the other probabilities okay.

So that is essentially the idea here because the number of words is very very large if there are only like ten possible words or ten possible values these features can take then there will be appreciable change even if I know that value for one feature but each of those words can take something like 10,000 different outcomes. Hence, it ends up becoming a pretty good approximation right when you are doing it in practice if you remember I told you something like KNN would be bad if you are in 100-dimensional space or a thousand-dimensional space.

And imagine text is a very large dimensional problem right, so what will be the typical dimension for text classification yeah so naively modelling it you can get 24,000 right you can do some kind of feature reduction and so on so forth try to reduce it to smaller state space. Still, it is pretty large right 10's of 1000 dimensions so if you try to do KNN in that things are not going to work that well. However, people still do that right so all these cosine similarity-based retrieval systems and other things that people used to do in the past but all nearest neighbour kind of techniques.

And not that they do not work but something like Naive Bayes when you want to do classification works tremendously well in this you can see this I'm challenging you to go try it out against SVM right you will find that maybe there is like a couple of percentage difference right and the math is so much simpler there are a few things which I should point out about Naive Bayes one of the biggest advantages of Naïve Bayes is that I can have mixed-mode data. I can have some of the attributes being discrete-valued some of the attributes being continuous-valued and everything.

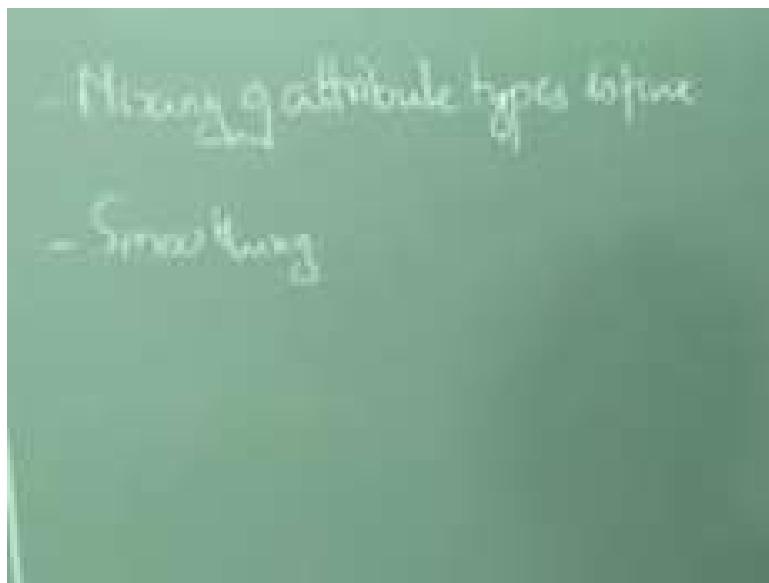
And it is very it is fine for Naive Bayes so what other classifiers can you say that? Trees right anything that is based on trees are also pretty robust when it has kind of mixed and you do not even have to do any kind of normalization of those attributes also. I can just keep each attribute one thing can run from one to a million other one can run from 0 to 0.1 is all fine right when I am looking at other more numeric classifiers in the sense of the kind of computations that they do they do distance-based computations and things like that.

There I have to normalize things a neural networks if something goes from one to a million. Another one is going from 0 to 0.1 the feature that runs from one to a million will overpower the feature that runs from 0 to 0.1 right I should Ideally be normalizing everything to 0 to 1 right so those kinds of things I do not have to do in Naïve Bayes it is all clean. Nothing the second thing I do not have to do is any kind of feature encoding normally i do not have to do any feature encoding I do not have to convert this into some kind of code.

So I do not have to take red and convert it into some code, code words so that I can feed it into my neural network wait how will I feed right into my neural network you have to encode it somehow right whether you use the RGB value for the colour or whether you use some other encoding for

the colour we will have to do something about it right I do not have to worry about all of that thing this the same thing in addition trees as well right.

(Refer Slide Time: 23:45)



So that is one thing the second thing I want to talk about is how do you handle missing values right if 'x' is continuous if 'x\_i' is continuous we are all fine right we are anyway using a Gaussian the Gaussian has an infinite supports, or something else which you have never seen in the training data comes will always assign some probability to it but what about discrete value things right, so I have trained you I have trained my classifier by looking at data right that has red, blue and green.

In that thing and in the test time somebody comes along with yellow, I can assign it a probability of 0. It says different so what if I do not have all the 'x\_i' one more example I mean things like neural networks would cover that right I mean something I have not seen beforehand as long as have some encoding for it will just take care of it I do not need to have seen it in the training phase we do not know "x\_i" is going to come or not say we have to somehow account for all the unseen "x\_i".

So there are multiple ways in handling this in Naïve Bayes one thing which you can do is you can just ignore that attribute do not multiply it in and make it zero just ignore it look at all the attributes you do know the probabilities for okay and then multiply that is a problem with it what is the problem sorry you are assuming the probability is 1 right, so you will be overestimating the probability, so you will have to come up with some mechanism by which you will normalize that right if you are using lesser features that they come up with some mechanism by which you are normalizing them right.

So that is something which will have to think about the other one is called smoothing right smoothing is essentially similar to what [Student] was saying earlier is that you assume that everything is everything that you could see right has occurred at least once in the training that means that you will give it some probability at least you will not make it zero and will also take care of this overestimating problem you will not make it one it will make it like 1/10,000 or something that will be a very small probability that you are signed to all the unseen values in the training data.

So when it test time at least you will not assign zero value to that data point right if everything else is very probable except that the colour is yellow right so you will not make the probability 0 the probability will get depressed significantly. Still, it at least not go to 0 okay, so this is one thing the problem is smoothing is the following if there are lot of values that you do not see I suppose that there is a color I see only four colours in training. Still, my actual colour spaces 256 right so I mean right or my actual colours space is 64,000 okay, and I see only four colours in training what will happen.

I am just talking about practical issues here if I do smoothly in such a situation what happens. It means you are training data is messed up, so you have to think about it. Still, if you use smoothing blindly in this situation what will happen exactly you will smooth the heck out of your probability distribution right you are taking this probability of one you are dividing it among 64,000 guys, so everybody should get a count of one at least right suppose I have 10,000 training points right the best account I can hope for is if all the 10,000 of the same colour it will get 10000 by 64,000.

No it will get 10000 by 74000 each one of the 64,000 I will count at least once. It 10,000 has happened so 74,000 it will be 10,000 by 74,000 will be the probability for the colour which occurred in all your training data points okay that is a really small value for the colour so smoothing you have to be very careful when you apply to smooth right so if there are too many unobserved values and smooth it blindly like so you will essentially lose all the information in the training data you will have to come up with other mechanisms like [Student] was pointing out there is something wrong with your training data first right.

So you have to go back and try to fix that see if we can generate more representative sample so that is those are the things that you should look at, we know how to do  $P(g)$  right, so we know how to do  $P(x)$  once we know how to do this and this we know how to do by doing this so parameter estimation is taken care of and this all the ancillary things.

### **IIT Madras Production**

Funded by

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**NPTEL**  
**NPTEL ONLINE CERTIFICATION COURSE**  
**Introduction to Machine learning**

**Lecture-64**  
**Bayesian Networks**

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

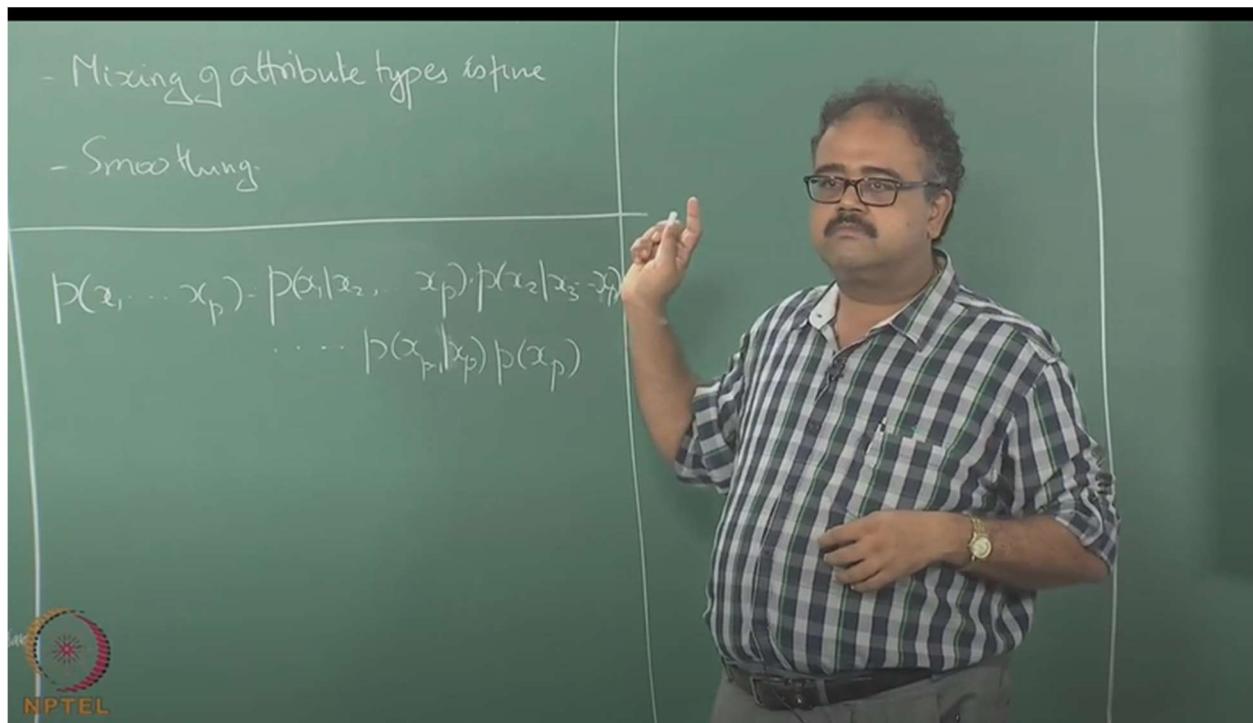
One of the things that you could do along the lines of this independence assumption is trying to be more nuanced about your independence, so what do I mean? So just do not make this assumption that everything is independent of one another given the class right so think of something like this I want to look at this joint distribution  $X_1$  to  $X_p$ .

Consider the Joint Distribution

$$P(x_1 \dots x_p)$$

So I am going to say things like okay, I am going to write something work I am not stopping. So what I am saying here is given  $X_2$ ,  $X_1$  is independent of everything else if you know what the value of  $X_2$  is  $X_1$  is independent of everything else.

(Refer Slide Time 1:55)



So I can write the probability of  $X_1$  to  $X_p$  as probability of all conditionals. To find out what is the probability of  $X_1$  given  $X_2, X_3, X_4$  up till  $X_p$ . Is like trying to find the probability of  $X_2$  given  $X_3$  I mean either some arbitrary ordering of choice and  $X_1$  to  $X_p$  right, it could be any other ordering now the probability of  $X_2$  given  $X_3, X_p$  and so on so forth.

The Joint distribution can be factorized as

$$P(x_1 \dots x_p) = P(x_1|x_2, \dots x_p) P(x_2|x_3 \dots x_p) \dots \dots \dots P(x_{p-1}|x_p) P(x_p)$$

Now I am going to tell you that, okay this is you can always add a conditioning 'g' if you want right this makes my life easier if you do not put everything conditioned on 'g', right. If you want, you can do that as well okay. How likely that variable like  $X_1$  depend on value taken by other system variables. Especially, if I am going to have 30 and 40 variables how likely is that  $X_1$  is going to depend directly on all the other 30,40variables in the system right, is it not going to happen right.

(Refer Slide Time 4:02)

he  
each  
 $x_1 | g$   
but a  
is also Gaussian.

- Mixing of attribute types is fine
- Smoothing

$$\begin{aligned} p(x_1 \dots x_p) &= p(x_1|x_2, \dots, x_p) p(x_2|x_3, \dots, x_p) \\ &\dots p(x_p|x_p) p(x_p) \\ &= p(x_1|x_2, x_3) \cdot p(x_3|x_6, x_7) p(x_4|x_5) \\ &\quad p(x_5) \cdot p(x_6|x_7) p(x_7) \end{aligned}$$

NPTEL

In reality, the dependency might be simpler than above,

$$P(x_1 \dots x_p) = P(x_1|x_2, x_3) P(x_3|x_6, x_7) P(x_4|x_5) P(x_5) P(x_6|x_7) P(x_7)$$

So what will happen is? Let us say that this is equivalent to say it is something like the probability of  $X_1$  given  $X_2, X_3$ , probability of  $X_3$  given  $X_6, X_7$  the probability of blah, blah, blah right up to get another example. Oh, well just okay right, so maybe my system is like this, so what does it mean? So,  $X_1$  is dependent only on  $X_2$  and  $X_3$  given  $X_2$  and  $X_3$ ,  $X_1$  is independent of all the other variables in the system. Likewise given  $X_6$  and  $X_7$ ,  $X_3$  is independent of all the other variables in the system, right. Given  $X_4$  and  $X_5$ , well-given  $X_5$ ,  $X_4$  is independent of everything else.

And  $X_5$  is independent of everything else just by itself right it is independent of everything else and  $X_6$  depends only on  $X_7$  and  $X_7$  is independent of everything else. It is just one way of writing it right, whenever I say  $X_6$  is dependent on  $X_7$  I can always flip it around and say  $X_7$  is dependent on  $X_6$  I am not talking about causal directions here I am not saying that  $X_7$  causes  $X_6$  right.

The probability distribution can be factored in the form of  $X_6$  given  $X_7$  into  $X_7$ . Otherwise I can also do it as the probability of  $X_7$  given  $X_6$  the probability of  $X_6$ . Just keep in mind there is nothing sacrosanct about this way just a convenient way of representing.

Does it make sense? Like I said if you are worried about the classification scenario, you can add conditioning on "g" everywhere. This kind of factoring things is more powerful than just using it for classification. And you can use it for learning about any probability distribution okay does not have to necessarily be about classification you can use it for representing any probability distribution okay.

So one way of I mean this looks a little hard to track right, one way of specifying these kinds of conditional independence relations is to use a graph, so what will I do in this case I will have a graph that has seven nodes. So one node corresponding to each feature, right. What are the features here? More generally the features here are random variable say  $X_1$  is a random variable that will take the value in whatever range  $X_1$  can take and so on so forth these are all random variables.

So I am going to have it. Is there something that is missing here? So I have connected the graph. So I have  $X_1$  okay, so  $X_1$  depends on  $X_2$  and  $X_3$  so I will put arrows.  $X_3$  depends on  $X_6$  and  $X_7$ . And  $X_2$  depends on  $X_4$ ,  $X_4$  depends on  $X_5$ , and  $X_6$  depends on  $X_7$ . So, this graph structure right gives me the dependency or independence/conditional independence relation I wrote in that expression that right makes sense.

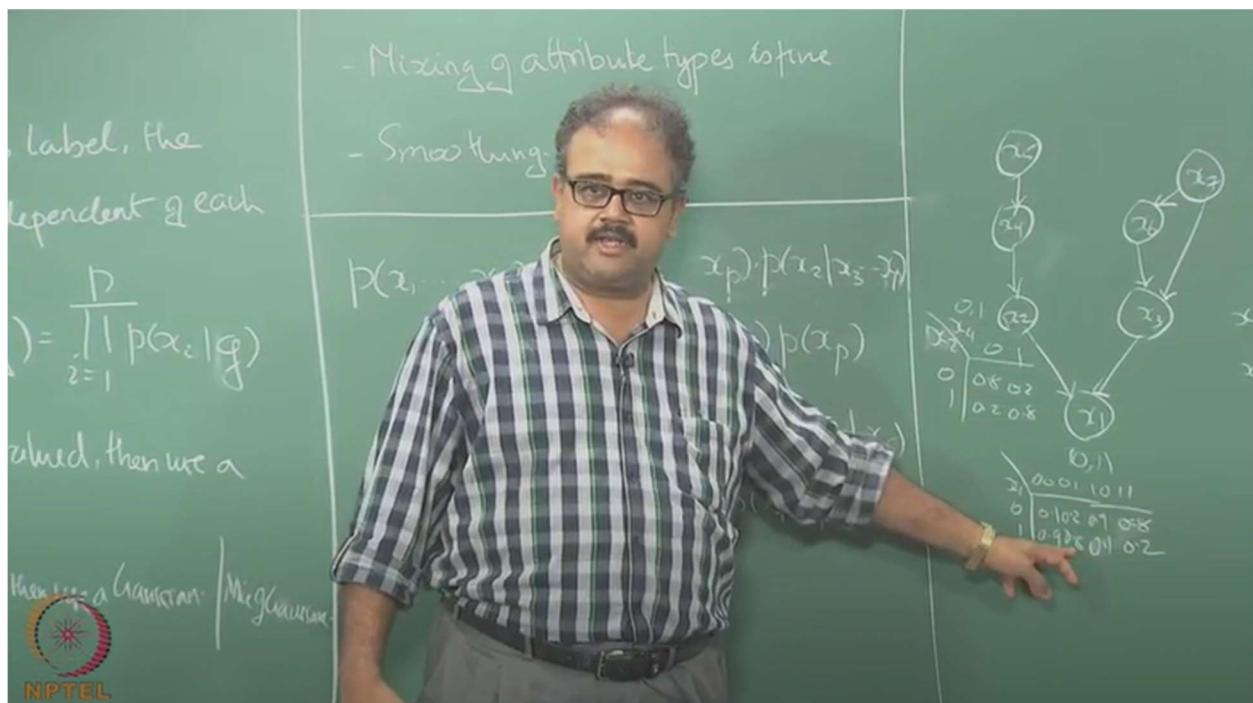
So if you remember, when I was talking to you about the interpretation of conditional independence that I said if you do not know what the class is then the Dhoni and cricket might become dependent. But if you know what the class is the Dhoni and cricket are independent, right. I mean occurrence of the words right I was telling you about that at likewise right if I know what  $X_2$  is right  $X_4$  and  $X_1$  are independent right is it clear if I know what  $X_2$  is then  $X_4$  and  $X_1$  are independent.

But if you do not know what  $X_2$  is then  $X_4$  and  $X_1$  become dependent, so what do I mean by that if I know something about  $X_4$  then I can tell you something about  $X_1$ . So if that is a little confusing

we will try to make this concrete let us say this can take values 0 and 1 and this can take values 0 and 1 not that is not confined to binary things right, Boolean things what makes it easy for me to write.

So let us say that the probability of something like this. So  $X_2$  copies  $X_4$  with a high probability right and likewise, so that is  $X_2 X_3$  so I will have to write a table like this for  $X_1$  right and yeah. So basically it says that when  $X_2$  is zero, probability of  $X_1$  being zero is low and the probability of  $X_1$  being one is high. Likewise  $X_2$  is one the probability of  $X_1$  being zero is high the probability of  $X_1$  being zero is low that is what I am saying that okay, so now if I know what  $X_2$  is right. Let us say that I tell you that  $X_2 = 0$ .

(Refer Slide Time 12:05)



Now the fact that  $X_4 = 1$  and if I say  $X_2$  is 0 then you know that the probability of  $X_1$  being one will be high, right. Regardless of the value of  $X_3$  because that is the way I have written this thing down. But if I know the value of  $X_3$  also then I will know that okay whether it is 0.9 or 0.8 right. Now if I tell you that  $X_4$  is 1, it does not matter because the only way  $X_4$  will give me any information about  $X_1$  is through  $X_2$ , but I know  $X_2$  is already 0. But suppose I do not know that  $X_2$  is 0 I, but I tell you that  $X_4$  is 1 right immediately what do you know that the probability of  $X_2$  being one is higher right. Therefore, the probability of  $X_1$  being zero is higher, right if I had not told you the value of  $X_2$ .

But if you are told you the value of  $X_2$ , I say that okay  $X_4$  is 1 right. Still, there is a small chance that  $X_2$  can be 0 right so  $X_2$  can become 0 in which case the conclusions you can draw about  $X_1$  completely changes is a very dramatic example, but this is not always so dramatic. Still, the point I am making is because of the way I have drawn these arrows right if  $X_2$  is not known to know  $X_4$  will tell me something about  $X_1$  if  $X_2$  is known the knowing  $X_4$  will not tell me anything more about  $X_1$ .

So everything that I can know about  $X_1$  by knowing  $X_4$  I already extracted by knowing  $X_2$  is it clear so this is this whole idea of conditional independence and why this kind of graphical representation helps us right. So knowing  $X_4, X_3$  right not  $X_2$  I know only  $X_3$  but not  $X_2$  still does not disconnect me from  $X_4$  right because the paths are very different so  $X_2, X_4$  can still leak influence  $X_1$  if I do not know  $X_2$  but know  $X_3$  is it clear okay, I will come to that right.

So this is the initial setup right, so what these kinds of graphical models do or rather this kind of these are called, Bayesian network right sometimes call sometimes called Belief network and then sometimes called a Bayesian belief network okay. In literature, we will find all the terminology you will find a Bayesian network, Belief networks, and Bayesian Belief Networks. And so in the Bayesian network is a DAG it has to be an acyclic graph and because if it has cycle in it, you are basically messed up, right because the semantics of the thing right so we will talk about a graph representation which does not have any arrows even right which is an undirected graph.

So there are undirected graphs you can start talking about cycles we will come about come to that in the next class. But when you are talking about a directed graph representation right it has to have no cycles because it has cycles then  $X_1$  depends on  $X_2$   $X_2$  depends on  $X_3$  and  $X_3$  will, in turn, depend on  $X_1$  therefore what will happen? This thing will get completely messed up. So you cannot write out a factorization like that right.

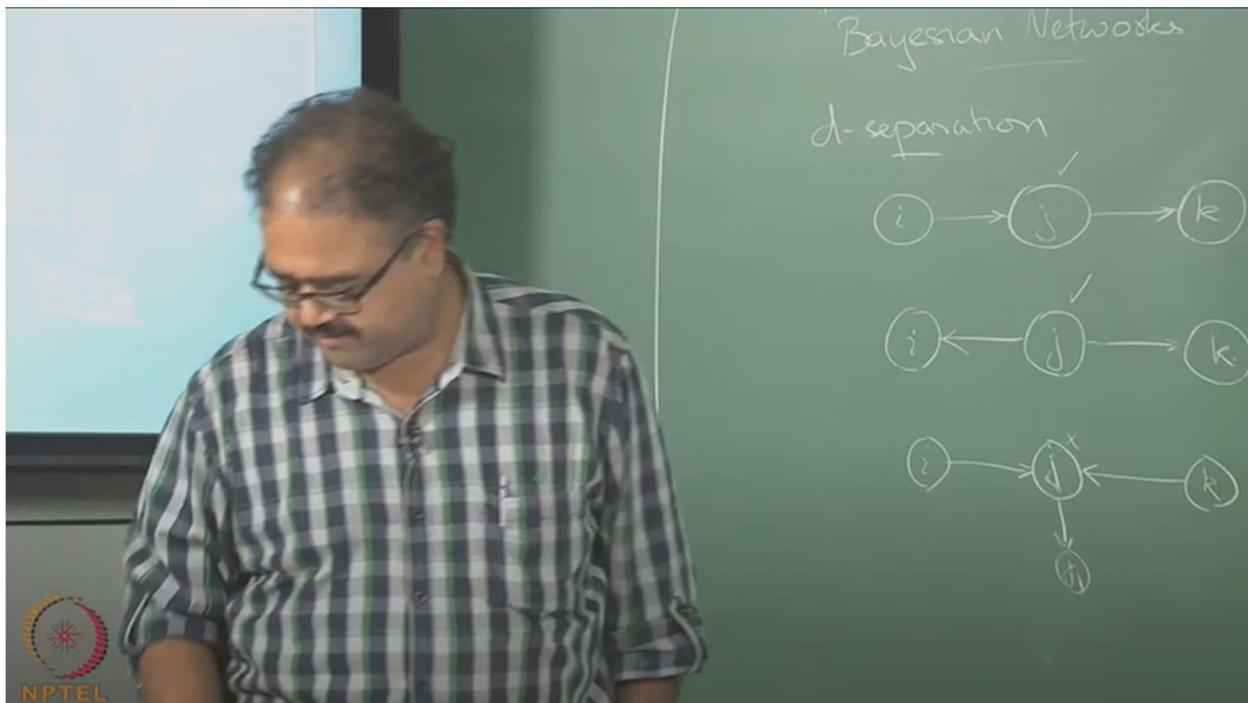
So one way of thinking about it as a set of conditional probability distributions and each node is going to have a set of the conditional probability distribution.  $X_1$  is going to have a distribution which gives you  $X_1$  given  $X_2, X_3$  likewise  $X_2$  will have a distribution associated with it which will give you the probability of  $X_2$  given  $X_4$ . So if I take the product of all these conditional probability distributions, I will recover the Joint Distribution of those variables okay.

So that is the semantics associated with it right take the product of all this conditional probability distribution I should recover the Joint Distribution of all the variables so if you are going to have cycles then that property will no longer be satisfied. So, we do not want cycles in this case right. And what is this here? So it is a DAG where each node is a random variable, okay and each edge represents a conditional dependence great.

So because of the nature of the graph, we are drawing right, so the graph encodes a lot of separation rules so what we mean by separation rule it tells me that  $X_1$  is independent of  $X_4$  given  $X_2$  right. So I would say that  $X_4$  and  $X_1$  are separated by  $X_2$  right, so likewise can you say something about  $X_6$  and  $X_1$  separated by  $X_3$  what about  $X_7$  and  $X_1, X_3$  right what about  $X_6$  the would  $X_6$  separate  $X_7$  and  $X_1$ ,  $X_6$  will not separate  $X_7$  what about  $X_6$  will it separate  $X_7$  and  $X_3$  no, right.

If there is a directed path from  $X_j$  to  $X_i$ , any node along the path will separate  $X_j$  and  $X_i$  provided that is the only path. If there are multiple directed paths it has to appear on all of those directed paths. Else you have to have a set of nodes, these two nodes together will separate  $X_j$  and  $X_i$ . So you will have to select one representative from each of those directed paths then it will be separated. Because we will have to consider directed edges here this is not called separation it is directed separation or d-separation come on, obvious right, so d-separation.

(Refer Slide Time: 28:31)



So there are 3 d-separation rules very simple d separation rules okay, j d-separates i and k that is a rule we already saw okay. So, what do you think about this hmm if I know j, i and k are independent if I know j, i and k are independent, right they are separated if I do not know j they are connected, so knowing j okay separates them. Likewise knowing j will it separate i and k here? Yes again think about this right suppose I did not know j right, but I know i let us say something like this right.

So I know that  $X_1$  is has a value of one I know  $X_1$  has a value of one right and I know that  $X_1$  will be 1 right, only when or rather I will know that i is one with a high probability if j is 0 right, so if I know that i is one then I know that the probability of j being 0 is higher as soon as you know the probability of j being 0 is higher then I will know something about k right because there is a direct influence from j to k.

But if I knew j I know that j is 0, I do not care what i is, i can be anything but knowing i will not tell me anything more about k than I get by knowing j right. So, in this case, knowing j separates

i and k, knowing j separates i and k is there anything else that we need to worry about any other combination. Sorry, i,j,k all the way yeah anything else is substantially different convergent say.

So what do you think? Knowing j, actually something else, it does, not knowing j separates i and k knowing J connects i and k think about it knowing J connects i and k because let us go back here right. So I know that  $X_1$  is 0 I know that  $X_2$  is one, no, I know that  $X_2$  is also 0 right I know that  $X_1$  is 0 and I know that  $X_2$  is 0 then what about  $X_3$  both are 0 the probability of  $X_3$  being one goes slightly higher right. Because this is the case where both zeros occur with a higher probability. So if I know  $X_2$  is 0 I do not know anything about  $X_1$  I cannot say anything about  $X_3$  does not matter in fact,  $X_1$  and  $X_3$  are independent, and  $X_2$  and  $X_3$  are independent if I do not know  $X_1$  right. If I know  $X_1$  then  $X_2$  and  $X_3$  become connected let that make sense is right. In both these cases knowing j separates in this case not knowing j separates. It is slightly stronger, I can look at any descendant of  $j_1$ , knowing any of the descendants of  $j_1$  will end up connecting i and k. As soon as you know the value of  $j_1$  I can make an inference about j and now that will help connecting i and k.

So these are the 3 d-separation rules so it is great to see the d-separation rules it did not talk about the values of the probabilities. It is just a representational thing so I can plug in whatever values I want all I am saying is just from the structure of the network I can tell you something about the separation properties right. So the actual probability values could come in later the values I use there was only for illustration purposes did not necessarily be that right this is the structure of the network itself tells you that what are the separation properties okay. Any questions on this? This is clear. I can give you a very large graph right and ask you okay Are A and B separated if I know C, D, and E okay, what should you do then? Sorry, you have to find out all the paths directed and undirected between A and B because. So, I have to look all paths between A and B and figure out the connection. We have C, D and E are the variables that are known and all the other variables are unknown. Then find out whether knowing those variables disconnects the path between A and B,. So you have to apply the third rule for the unknown variables. If all the paths get disconnected between A and B then you say that A and B are d-separated by C, D and E so this is a kind of analysis that you can do to make sure that you understand your system properly right.

So one of the original motivations for proposing these kinds of belief networks it is kind of DAG representation for the variables was to study causality was to figure out causal relationships. These kinds of networks are also called as causal networks. Typically when you talk about causal networks you do not associate conditional probabilities. It is just you that A causes B kind of relations. Don't worry about the probabilities in that in this setting.

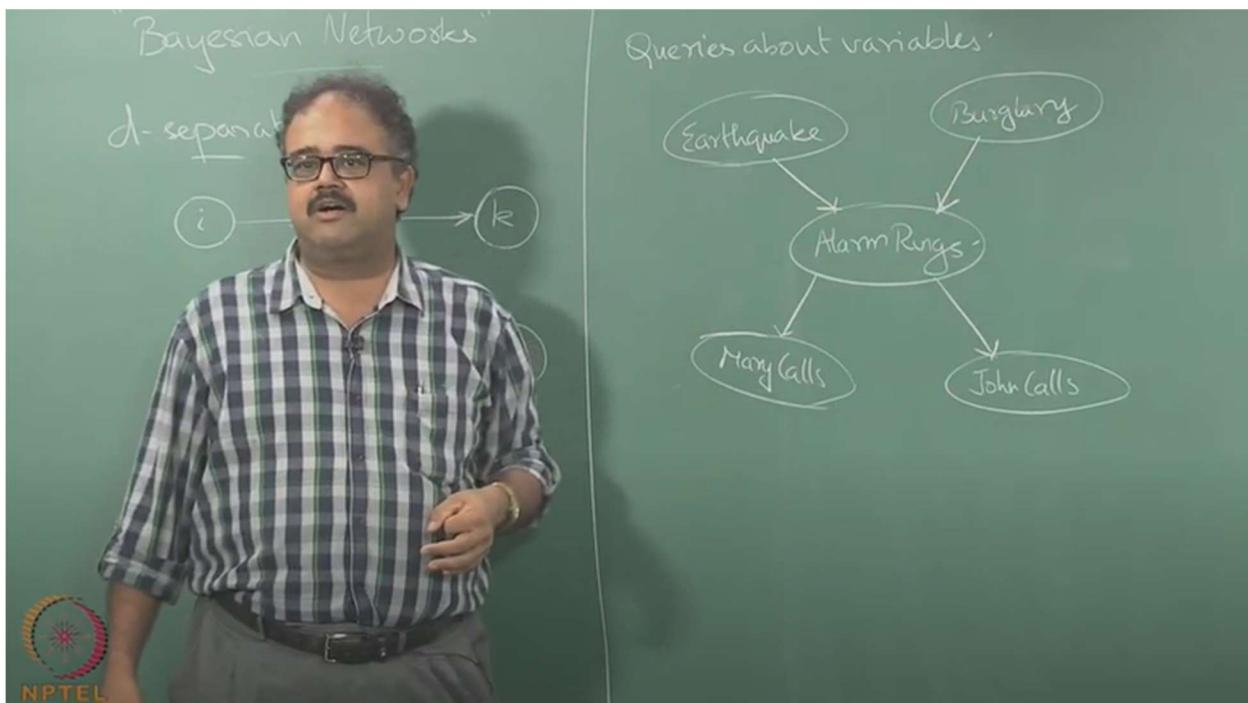
So the same representation can be used for representing causality also let A causes B right so that kind of relationships can be represented using the same representation but in general, when you are using this as a Bayesian network you do not imply any causality, is something which you have to keep in mind when you are using it in practice right. So you are not implying any kind of causality right when you are using this direction this does not mean that I believe that  $X_2$  causes  $X_1$  right.

When you are using it as a causal Network model, yes, okay when you are putting in an arrow here that means you have thought about it and you believe from the physics of the system or whatever it is that  $X_2$  actually causes  $X_1$  and it turns out that when you are trying to do this learn this graphical structure by just observing the system looking at the data and trying to infer this kind of graphical structure between the variables the most compact structure that you can derive will turn out to be the one that corresponds to the actual causal nature of the system right.

If you do it incorrectly, you will end up adding a lot more spurious dependencies between variables. That if you are doing it in the correct causal ordering, then you will end up having a much more compact graph than you would if you are doing it at Willy-Nilly way right. What is the use of doing all of this?

So essentially we are interested in answering queries about variables right, so no class, no lecture on graphical models or Bayesian networks is complete without you looking at the earthquake network at least once okay the very, very popular network for historical reasons. So you have a burglar alarm in your house okay, and the alarm rings okay the alarm could ring because of two things right it could ring because there is a burglary it could also ring because as, okay.

(Refer Slide Time: 34:01)



So this network was originally made up by Judea Pearl okay is one of the early pioneers in the study of causal networks and belief networks and so on so forth. Judea Pearl lived in LA in California he was not thinking of wild animals he was thinking of something else remember what I call this network, okay. So probably the most two common occurrences in California earthquake and burglaries, fire alarm know it is a burglar alarm I am not interested in the fire alarm I am interested in burglar alarms okay.

And it turns out that Pearl had two very nice neighbours right who would call him at his office and tell him, hey your burglar alarm rang okay with some probability, so this is so if the alarm rings Mary or John will call Pearl in his office and tell them that hey your alarm rang. So, you can think about the causal directions here right so the alarm will be caused by the earthquake or burglary right and then Mary will call, and John will call if the alarm rings came both of them might call, or none of them might call because they are all probabilistic things right.

So now I can ask questions like this. Mary called me and said that there is she thought she heard the alarm okay. So, what is the probability that the alarm rang? Both Mary and John called me and

said both of them thought they heard an alarm, what is the probability that the alarm rang? I know there was an earthquake in my place, but both Mary and John did not call me what is the probability that alarm rang? They are dead, is it? Good point.

Few Queries to Earthquake Network,

$$P(\text{Alarm Rang}|\text{Mary Call})$$

$$P(\text{Alarm Rang}|\text{Mary Call}, \text{John Call})$$

$$P(\text{Alarm Rang}|\text{Earthquake})$$

No, but if that is going to happen, I should have had an arrow like this. Since they do not have the arrow I am going to assume that it is not likely that is what I am saying right if the earthquake is going to directly influence John and Mary's behaviour whether they are the question of their mortality or not or other things right, so I would need to put an arrow directly between earthquake and Mary just assume that Pearl and Mary live in different earthquake zones right there is this one small fault line which will only shake Judea Pearl's house and go away go on guys this is illustrative example do not take in too, much too hard.

So the point here is I can ask all kinds of queries on variables on this right I can ask even other things like that hey Mary called, what is the probability that a burglary happened? Right, Mary and John call, what is the probability that a burglary happened? Things change or not, change or not you know everything to answer that question. Yeah, it changes, it will change because if both Mary and John call then my belief that alarm rang goes up right if it believes the alarm rang goes up then whatever belief I had about burglary happening will also automatically go.

Now we know why these are called belief networks right so when I say Mary call then my belief on whether the alarm rang or not changes right base I have some belief on alarm ringing. Hence, I think okay if nobody calls the law must not run if Mary call cell I will flip this backward right here this will be the only probability of Mary given the alarm, right. So this is the probability I will have here but now given that Mary was one what is the probability of A being one and given that

Mary and John or one what is the probability of A being one and given the probability of A being one what is the probability of burglary happening right.

So all of these things I can do all kinds of reasoning about, about the system based on just this whole model that I am learning right, so I can ask all kinds of questions I can ask questions about joint distributions so what so Mary call okay, what is the probability that the alarm rang and that John will call? it kind of redundant probability but still you can try to ask those questions like this right or I know that the alarm rang somehow I know that the alarm rang what's the probability that Mary and John will call me so I can ask all kinds of questions I can also conditional questions I cannot join probability questions they can ask conditional probability questions right.

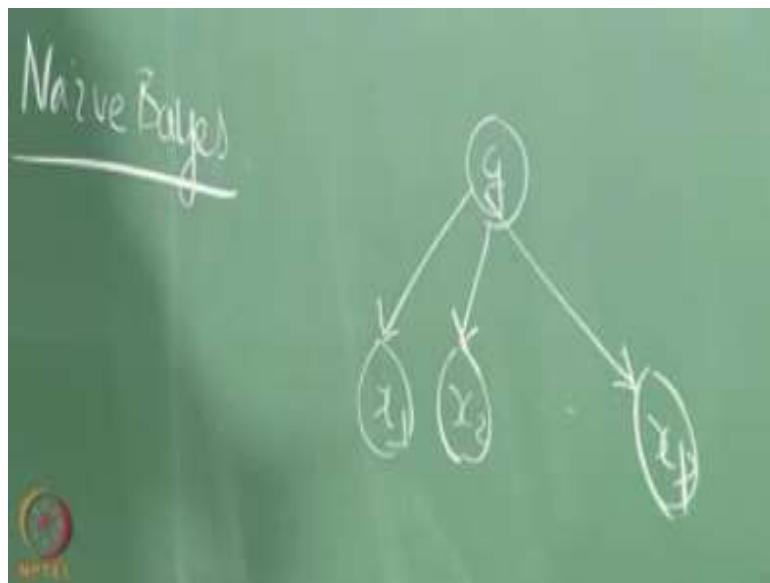
And if you think of this as classification problem I can ask questions about okay I know these five variables what is the probability that this is class one, you got around our Naïve Bayes problem right I said what if you observe some variables whose values I never see before how will you estimate the probability I can still do that I can just assume that the variable is unobserved. I can estimate the probabilities that will give me give me valid answers that given that I know A, B, C what is the probability that class is one. I might have another D, E, F, G, H, I, J, K, which I have not observed that is okay.

So I can ask all these kinds of questions that given partial data now I can ask questions about classification right, Or given class labels I can ask questions about conditional class densities right the given that it is a document on cricket right how often will I do not know let me not pick on any cricketers. Given that it is a document about football how often will the word Ronaldo and goal occur together in my document right yeah if you leave it to my son he'll see you say the probability should be 0. But I know this is almost religion right the camps, anyway right.

So that is the whole idea right so these kinds of questions these kinds of queries. I asked about these variables, so we call this problem as the problem of inference on the graphical model is the problem of inference on the graphical model essentially is to figure out all these conditional probabilities or the marginal probabilities that we are interested in the right. So we looked at Naïve Bayes right.

So can you think of drawing the Naive Bayes assumption as a graphical model every node is independent, every node is independent, is it so that it will be like this is that my graphical model. No, what is the Naive Bayes assumption given that class they are independent, so where should the class be? The top or the bottom and of course I can draw it wherever question is the direction of the arrows if you let me draw it at the top.

(Refer Slide Time: 38:36)



So people tell me how the arrow should go down, right. You will be surprised how many times people draw the arrow up? The reasoning is the variable values are the one that causes the class to happen right. So if  $X_1$  is this  $X_2$  is this  $X_3$  is that then they should all be influencing the class variable that for the arrow should go up well that is a fairly valid argument it is just that it does not capture the Naive Bayes assumption that is a different kind of assumption that you are modelling there right.

So each variable is somehow affecting the class; this is essentially the opposite of Naive Bayes. Okay instead, is essentially a complete model right, so since all the variables are influencing the class that if in fact if you think about it like given that class what happens in that case? All the variables get connected; it will be this case right. So if all the arrows were going up it will be this case so if  $j$  was known all the variables get connected right in this case its opposite of Naive Bayes

that given the class all the variables are dependent on one another that is the assumption if you draw the arrows upwards the arrow should go down and up and down its relative rate arrow should go away from the class node.

**IIT Madras production**

Funded by

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

# **NPTEL**

## **NPTEL ONINE CERTIFICATION COURSE**

### **Introduction to Machine Learning**

#### **Lecture 53**

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

#### **Undirected Graphical Models- Introduction & Factorization**

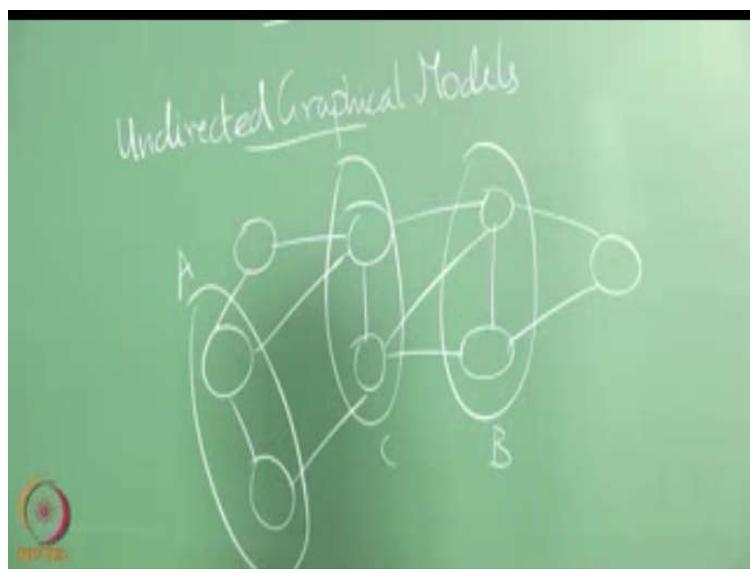
So we continue looking at graphical models right so we looked at belief networks we also said that we would call Bayesian networks Bayesian belief networks. And then we looked at the concept of d-separation, we looked at what d-separation is, and we also discussed what is the question of inference is.

What is the question of inference? I will give you certain observations and ask you questions about the conditional distributions. For example, given that Mary calls what is the probability that there was an earthquake? So, what are the questions? I am asking in the earthquake case they were all marginal. I have the Joint Distribution of Mary, alarm, John, earthquake and burglary right these five variables, so the whole system is specified by a joint distribution over these five variables. But I was asking you a question but a marginal. What is the probability that earthquake happened? okay that is a specific marginal I am asking a question about.

So typically inference questions will be about marginals or conditional marginals right so conditional marginals will be when I am given some observation given that Mary Call what is a probability of an earthquake, so that is a conditional marginal right. But, if I just ask a question, What is the probability of an earthquake in California? I can still do that then I can take this entire network right marginalize out of everything I can tell you what the probability of an earthquake happening is is. That is not very interesting because it is already I give you a marginal as one of the components in specifying the network.

All I can do is just look up the earthquake probability distribution I already have given you the marginal this one of the things that specified right. I can ask questions like okay what the probability that Mary will call is? , What is the likelihood that Mary is going to call me today? I can only marginalize over all other variables, and I can give you that answer. This is essentially queries, on some kind of marginals or conditional marginals it could be joint distributions as well it could be joined to marginals as well.

(Refer Slide Time: 03:10)



In the sense that what is the probability that Mary and John both will call me today is a joint distribution over a subset of the variable, so it is some kind of a joint, marginal rate. Hence, instead of looking at the full joint distribution, so those are the kinds of queries I am asking, so we look at the directed networks. So, today we look at using undirected graphs. Look at undirected graphs and. So I am going to call this set of nodes as A where these two nodes are A these two nodes are B just asking you, grouping the random variables here.

So as before each node is a random variable just like we had in the directed case each node is a random variable. The edge denotes some kind of dependence between the two random variables so in the directed case you could confuse an edge for a causal relation you could keep the edge as

representing a causal relation right but here. I removed the arrow direct there is no direction here, so it is just some kind of dependence between two variables right. The edges in an undirected network encodes the notion of conditional independence. Just like the edges in an directed network encoded some idea of conditional independence.

Yeah, so there is a subtle difference between the two right, so I am not going to get into it because it is subtle. Still, the class of conditional independence is are different right there some that you can represent using directed networks. There are some that you can represent using undirected networks in most cases, you can choose whichever you want. Still, there are some cases which are more convenient to go one way or that right. So I do not want to get into the discussion at least not in this class it is for the class next semester if people want to get in to do that right.

For any path from node in to node in B. I have to pass through some node in C. So any path from A to B I will have to pass through some node in C correct. Then I would say the nodes in C separate the nodes in B from the nodes in A right last time. We had this notion of d-separation like we have to write three different rules d-separation is nice for making up fun questions for exams right but.

But it is a little confusing right, so you have all these arrows going to take care of the notion of separation here is easy. It says is if there cannot be any path from A to B which goes through, which does not go through C okay now like I put a double negative there, every path from A to B goes through C right in which case then you say C separates A and B okay. The d is gone here right; it is just simple separation.

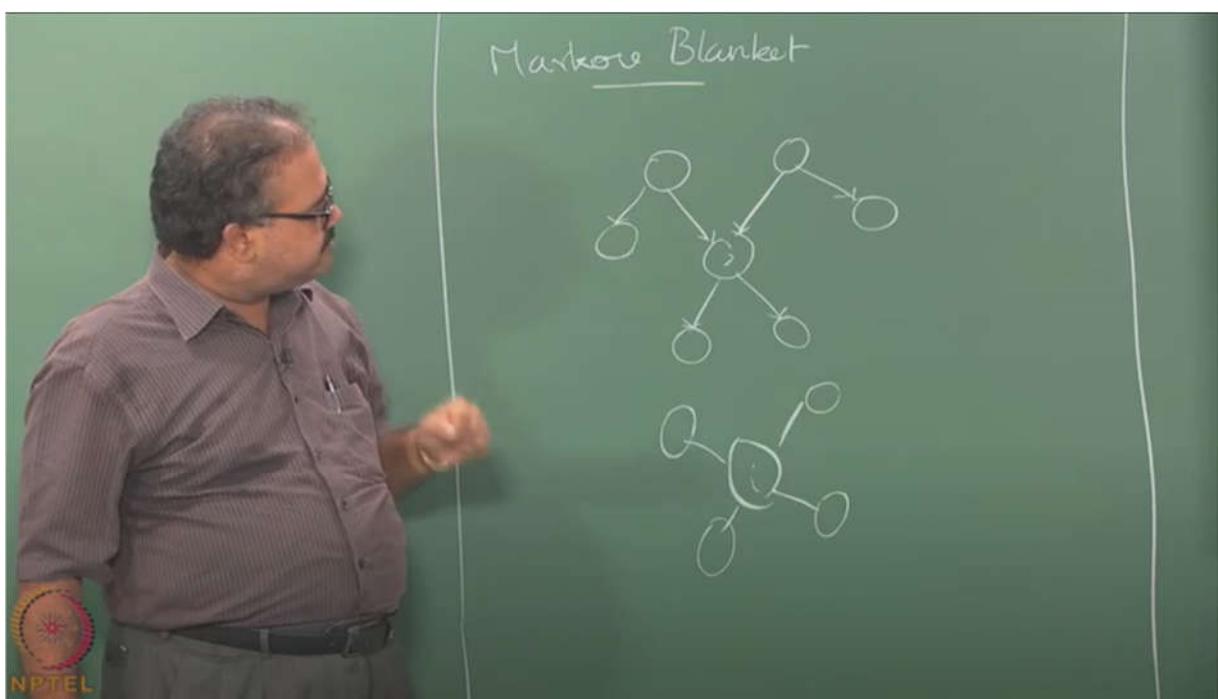
That make sense, so that is the simple enough let so this is what we started we started with this notion of separation that is encoded in a directed model so in the undirected case I am telling you that separation is very simple like we had d-separation we have separation here

So, next, we have to think about something else right if you recall when we started the discussion of directed models. We started by looking at factorization of the Joint Distribution. We had a very complex joint distribution over, many variables, so we started looking at some kind of factorization

of it, and from there we constructed the network right. Hence, the graphical model inherently encodes a factorization, right.

So, this all this separation business ties indirectly to the factorization right. So the d-separation gives you the rules for the factorization. So, directed model it is very easy to write down the factorization you just look at the conditional distributions. To write down the factorization here we do not have such kind of a one-way implication here right, I am just saying something links the two variables together right so how do we go about doing the factorization.

(Refer Slide Time: 11:56)



Right, so there is something there is a concept called the Markov blanket.

And I forgot to mention when we did directed models right. Hence, a Markov blanket is essentially all the variables right that could potentially influence a given node right so given the node the parents can influence the node correct. Then, anything else the children can influence the node right anything else siblings can influence the node right sibling yeah right.

So that is essentially the Markov blanket over a node right, so I have a node. I take node  $i$  right so this is essentially the. If it does not know the parent, then this node can influence this one right if I do not know the parent this can influence it is right here so this is essentially the Markov blanket of  $i$ , right so what will be the Markov blanket in undirected case.

Same, it has to directly connect the neighbours of  $i$  in the undirected case is just the neighbours of  $i$ . So, these are connected to other nodes I do not care right so they cannot influence me except through this, yeah. I am looking at direct influence so that node cannot directly influence me right, so if I do not know this guy.

Okay, this one so if I know this node right these two get connected if we know this node these to get connected right so, so essentially that is it so if I know this guy okay then I get connected to this guy so if I just condition if I say okay I know this right.

Okay, so, so that is the Markov blanket of this  $i$  node  $i$  likewise here if I know these four guys then that is it nothing else will influence my  $i$  right so it will cut off from everybody else right these four nodes will separate  $i$  from the rest of the network.

Right and here I need to know all of these guys right before I am cut off from the rest of the network right so I will.

You do not need to know the Markov blanket right just an aside that I wanted to tell you, no no you need a Markov blanket for other things but not for this lecture or this course right. Hence, the Markov blanket is very important because you need it for making the inference. I will clarify the thing, but so essentially the idea behind the Markov blanket is once you know all of these nodes right. You are separated from the rest of the graph.

Yeah so if you know the children, then these two get linked up right. No, if I know the child but then the probability on  $i$  right so if you remember the third d-separation rule we had right. If I do not know the child, then these two are independent. If I do not know the child these two are independent information can go from one to the other but if you know the child, then these two get linked up.

(Refer Slide Time: 15:22)

Consider two nodes  $X_i, X_j$  such that there is no direct link between them,

Assume the following conditional distribution,

$$P(x_i, x_j | X \setminus \{x_i, x_j\})$$

where  $X$  is set of all nodes in the graph,

We can simplify the conditional distribution because it is conditioned on every other node,

$$P(x_i, x_j | X \setminus \{x_i, x_j\}) = P(x_i | X \setminus \{x_i, x_j\}) P(x_j | X \setminus \{x_i, x_j\})$$

The Joint Distribution could be factored as,

$$P(x_1, x_2, x_3, \dots, x_p) = \psi(x_1, x_2, x_3) \psi(x_2, x_3, x_4) \psi(\dots) \dots \dots$$

So for  $X_i, X_j$  there should be no factor containing both of them as variables. Since they are not directly connected.

Flip it around, and we could say that, there should be individual factors for all maximal cliques.

So we will go back to the factorization now.

Consider two nodes  $X_i, X_j$  such that there is no direct link between  $X_i$  and  $X_j$  right so something like this right. So, this could be  $X_i$ , and that could be  $X_j$ , so there is no direct link between them right. Let us say that the set  $X$  denotes the universe of all variables right so it is  $X_1$  to  $X_p$  so set  $X$  denotes all the variables and if I condition the Joint Distribution of  $X_i, X_j$  on everything other than  $X_i, X_j$

Right, so what can you tell me about this distribution? It should factor out. So, people haven't encountered this before. That is set difference, from  $X$  it removes  $X_i, X_j$ . Conditioned on that they should be independent right condition on the rest of the variables what essentially I am saying is okay  $X_i$  and  $X_j$  I am conditioning it on everything else. So every path between  $X_i, X_j$  has to go through well everything else right, so I have conditioned like that they have to be independent, so that is the basic assumption I have right the factorization I am assuming here is that.

I mean the conditional independence thing is if all path from A to B pass through C right then A is conditionally independent of B given C, so here my C is everything. Other than  $X_i, X_j$  this has to be true if you think about it now when I write my factorization right when I am going to write me what I mean by writing my factorization I am going to take my probability of.

I am going to take this joint probability and are going to write it as some factor one on some set of variables.

Right so like this, I am going to write it out like this I this is what I mean by factorization right in the directed case. I also did the factorization except that my Psi had a very specific form my Psi would have been something like what is the probability of  $X_1$  given  $X_2, X_3$ , so that is a factor. Psi could also be the probability of  $X_3$  given  $X_4, X_5$ , which is another factor. Like this, we could have multiple factors in the undirected case. But, have to figure out what these factors are going to be in the undirected case.

So we want this to hold right we want this conditional independence to hold so what can you say about the factors.

If there is a factor that directly connects  $X_1 X_i$  and  $X_j$ . There could be some assignments of some, kind of derivation of these factors which will connect  $X_i$  and  $X_j$  so I cannot say that  $X_i$  and  $X_j$  will be independent right regardless of what values I am assigning to the factors. If there is a factor that has  $X_i, X_j$  in them, then I cannot be guaranteed of independence for all assignments to those factors. The only way to ensure that this independence will hold is if I guarantee that  $X_i, X_j$  will not appear together in any factor.

So, I cannot have a Psi function here in the set of factors that I write I cannot have a Psi function that has both  $X_i, X_j$  as arguments is it clear, the reasoning clear if  $X_i X_j$  appears in any one of this Psi together right. Then I can assign some use to the Psi such that those two get connected right so to make sure that I can never do that that this conditional independence holds for all possible distributions I can write.

No factor should contain  $X_i, X_j$ . So I am going to take this intuition right and flip it around if there is no edge between  $X_i$  and  $X_j$  then no factor should contain  $X_i, X_j$ . Right if there is an edge between  $X_i, X_j$  there should be a factor that contains  $X_i, X_j$ . To encode the dependence between  $X_i, X_j$  there should be a factor that contains  $X_i, X_j$ . If there are edges between  $X_i, X_j, X_j, X_k$  and  $X_k, X_i$ , you can put in factors for every edge. Alternatively a more compact way of doing this is to put in a factor for the clique.

Consider a set of three variables which are fully connected. We should add factors for all three variables because of three-way dependence. Instead of adding three different factors, use a single factor with all of them together. So flip this around and then say that for.

You want me to repeat what Psi represents because I did not even say what it represents, but I will tell you what it represents. Still, the thing is it is just some factor right in directed models rights Psi was the conditional probability.

So essentially I am taking one complex function right I am taking this Psi of  $X_1$  to  $X_P$  right okay so I am taking this function. That is a very complex function I am writing that out as a product of many smaller functions. Such, each of these Psi is one such function.

I didn't get what you are saying everything else is a constant and then yeah, okay. When I am conditioning on every other variable, that means I am essentially assigning a value to every other variable. So all the other factors will reduce to some fixed value and then I will have only  $X_1$   $X_2$  left even then my choice for  $X_1$  would depend on the choice for  $X_2$ , so that is one way of interpreting.

So people got his way of looking. He says it fixes all the other values right so let us assume that Psi, Psi one has  $X_1$  and  $X_2$  and make sure  $X_1$ ,  $X_2$  does not appear in anything else just for simplicity sake right so fix all other values so everything else will reduce to a constant right. So it will be Psi  $X_1$   $X_2$  and constant. The probabilities for a value of  $X_1$  will depend on the value of  $X_2$ . To make it independent give the probability for  $X_1$  as 0 when  $X_2$  is 1. So I can assign the same probability there is a very specific assignment that can make it look independent so something like this right. Right, so what is the probability that  $X_1$  is 0.

So I am conditioning it right so the conditional probability I can say that okay so this is 0.2 this is 0.8 right now  $X_1$  is independent of  $X_2$  even though I have a factor Psi  $X_1$ ,  $X_2$  but  $X_1$  is independent of  $X_2$  right. Still, I have to be very careful about actually assigning the values to the factors right. So what we are trying to do is give the factorization. Still, regardless of what function we choose for Psi one right, it will be it will end up being independent if a condition on everything else it will be independent.

So I can put in whatever random numbers I want there okay I still want  $X_1$  to be independent of  $X_2$  so that cannot be the case if there is a factor that has  $X_1$   $X_2$  in it.

For each clique you have a factor in this yeah you put in a factor so I mean you are designing how the factorization should be right. So I am just saying for each clique you, you include a factor.

So whether  $X_2$  is 0 or  $X_2$  is one it does not matter right so what is the probability of  $X_1$  equal to zero given  $X_2$  equal to 0 right. So that is equal to the probability of  $X_1$  equal to zero given equal to 1, the probability of so it is independent no I said it is conditional right.

No, no this is writing the conditional I am only writing the conditional. I am only writing the probability of  $X_1$  given  $X_2$  I am not writing the Joint Distribution.

So each column should be a valid distribution, row right, no no each column should be valid distribution row not need not be.

Row need not be why I am writing the conditional right I am writing the probability of  $X_1$  given  $X_2$  so why should the rows be able to, joint PDF should be valid. It is not Joint I am just written the conditional

So whatever so I to get the joint I need to define a distribution over  $X_2$  right so which I am not then I just looked at one factor here so you could have a factor over  $X_2$  if you want, but I have not done that, yeah.

For the directed graphs right so you can easily write down what the factorization is right so I have some probability Joint Distribution remember all of these graphs represent joint distribution over  $P$  variables right  $P$  is hard.

Let us make it  $N$  okay I am going to make it  $N$  so having probability as  $p$ . The dimension as  $p$  is a little confusing for me, so I have  $N$  variables whether it is a directed graph or whether it is the undirected graph so what I am trying to represent is the probability of Joint distribution over  $n$  variables right. The whole idea of going to this graphical models is that I do not want to define this whole  $N$  squared minus 1 number of values for specifying my probability distribution right.

So I want to say that know it really in a probability distribution is not so complex they are not  $N$  squared independent values in my distribution. Hence, there is some kind of factorization that will happen. So I am truly trying to figure out which are the independent values that I have to specify

so that I can get the entire probability distribution, so for that, I am finding out what is the right way to factorize my probability distribution.

Sure there can be more than one factorization by the way why what is the confusion right. So if you think about even the directed models, you can always think of flipping the direction of the edges right and writing another factorization. Right so it is nothing to it there is nothing very sacrosanct about this the reason we choose this factorization is that it is easy to handle these things, so that is the only reason right.

And so in the directed graph, we knew how to find the factorization, so we are we look at the conditional independence. So in the undirected case right so we need to come up with some way of finding what the factorization is right. Hence, what I am saying is if there is no edge between two variables, they should not appear together in a factor.

Right if they appear together in a factor, then there is a way of assigning values to the factor right. This is a factor in a conditional in a directed graph this could potentially be a factor in a directed graph.

And so, for example, something like this.

So this could be it right but no, even though I have written this dependence factor I have assigned values here so that the dependence does not hold right, but I can assign something else here right. So then this equality is broken, so if I have a factor that has both X1 and X2 in it when I write this factorization, then I can assign some values.

So when we are talking about these factorizations what we are looking for is a representation that regardless of the numerical values, you eventually end up assigning to it preserve the independence relations that you are looking for.

They might introduce additional independence relations but at least the ones that your guarantee should be there in the probability, so when the graph guarantee is something the graph structure

guarantee some independence relation okay. The factorization you give should guarantee that independence also so if I put in a factor that connects  $X_1$  and  $X_2$  when there is no edge between  $X_1$  and  $X_2$ , I can no longer guarantee that.

Right so any other questions I know it is always a little tricky when you move to undirected graphical models right, so director graphical models are all very easy. Everyone gets it the first time around, so undirected graph demons are a little confusing. But it is good to look at them early on. A lot of techniques that we use for inferencing are common between directed and undirected models. There might difference in implementation, but algorithm wise it is almost same.

So let us spend a little time and try to understand the undirected model.

Condition no we should not have a factor  $\Psi_i$  that connects  $X_i$  and  $X_j$  that have both  $X_i$ ,  $X_j$  as an argument these are independent only in the case yeah. But I can assign values that is what I'm saying I can always assign values to that functions  $\Psi_i$  such that they become connected.

See the whole idea of giving a factorization is regardless of what you do with that  $\Psi_i$  that you can do whatever you want with the  $\Psi_i$ . However, I still want to guarantee the independence so as soon as I put in  $X_i$   $X_j$  in the same function, I give you the freedom to do something with that function to introduce the dependence, so we do not want to give the freedoms to.

Great so I am flipping this around since we say that if there is no edge, there should not be a factor and I am saying that if there is an edge there should be a factor right in fact I want to go a little further and say that if there is a clique right that should be a factor associated with the clique.

So better.

Let suppose I have a graph like that so I have four random variables right so what should be the factors here.

I should have something that connects.

1 2 3.

Or something that connects 1 2 3 4 is it

I should have something that connects 1 2 3 4 why is not a clique right, so 1 4 is not connected.

Therefore I do not want a factor that has 1 and 4 in it at the same time I cannot have a 1 2 3 4 factor so it can be 1 2 3, 2 3 4.

All I need to consider are maximal cliques. 1,2 is a clique 2,3 is clique 3,4 is a clique and everything is each edge by itself is a clique. You did not give me factors for all the cliques you gave me factors for the maximal clique. It turns out that if you introduce factors for the maximal cliques you essentially have the same representation power as having factors for all the cliques in the graph.

So, in this case, is that and that so we have only two factors okay.

Is it clear?

**IIT Madras Production**

Funded by

Department of Higher Education

Ministry of Human Resource Department

Government of India

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

## NPTEL ONLINE CERTIFICATION COURSE

## Introduction to Machine Learning

## Lecture-66

## Undirected Graphical Models - Potential Functions

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

(Refer Slide Time: 03:26)

$$p(x) = \frac{1}{Z} \prod_c \Psi_c(x_c)$$

↗ Potential fns

$$\Psi_c(x_c) \geq 0 \quad \forall x_c$$

$$Z = \sum_X \prod_c \Psi_c(x_c)$$

$$P(X) = \frac{1}{Z} \pi_c \Psi_c (X_c)$$

$$\Psi_c(X_c) \geq 0 \quad \forall X_c$$

$\Psi_c$  – Potential Function.

$$Z = \sum_X \prod_c \Psi_c(X_c)$$

Z- Partition Function (Normalizing Constant)

We have ‘Z’ because as opposed to directed case potential function could be anything and not restricted to probability.

So we sometimes call these. Sometimes call this  $\Psi$  as potential functions right, so  $\Psi(C)$  is a potential function associated with the clique  $C$  right and  $X_C$  is the set of variables that are participating in clique  $C$ .

Right, so this is a product over all of these cliques okay there is a small problem here right, so we have to make sure that whatever we are writing is a. Probability function right is how do you make sure of that. Right, so you basically have some kind of a normalization factor where  $Z$  will essentially be whether the integral or the sum or whatever if you're looking at.

Looking at discrete values we have been talking about binary so far right so if you are looking at binary value at the variables essentially this will be sum over all values that  $X$  can take right. Suppose you have  $n$  binary variables the sum will run over  $2^n$  entries right sounds like a really bad idea right anything where ever you're summing over  $2^n$  elements seems to be a bad idea.

And it is so the most the biggest difficulty in using undirected graphical models. so why did not you have this in the director graphical models. We chose the factors cleverly, right the factors were chosen to be conditional distribution, so when it took the product it was it is guaranteed to be a distribution but here we have no such restrictions on the  $\Psi$  right this is what makes it even more confusing I have no restrictions on the  $\Psi$  can be anything right so I can go run up till three million I do not care right  $\Psi$  can be any function right  $\Psi$  can become negative oops can it. No, right the only condition I have is  $\Psi$  has to be non-negative yeah okay that is a good question this  $\Psi$  have to be positive or non-negative right.

So it is okay for me to have zero probability for some configuration right. So  $\Psi$  can be so  $\Psi$  non-negative then that is all the condition I need, so the only condition I need is that so for all values that  $X_C$  can take so  $\Psi$  has to be non-negative right.

So  $Z$  this also sometimes called the. Partition function which is the terminology that comes from physics right, so I am not going to get into the explanation of it but sometimes also not a separation

so if you reading up something and somebody mentions partition function okay essentially the Z that they are talking about. So that makes it a little tricky right, so we are saying that your. This thing is not restricted in the interpretation. Right here, so I see is not restricted interpretation it can be anything and as long as you can do this normalization will get a probability okay, so I am going to write down a very powerful theorem.

(Refer Slide Time: 06:36)

$$\Psi_c(x_c) > 0$$

$$\Psi_c(x_c) = \exp\left\{ -E(x_c) \right\}$$

Hammersley-Clifford Theorem

If,  $\Psi_c(X_c) > 0$

$\Psi_c(X_c) = \exp\{-E(X_c)\}$

$E$  – Energy Function

Right, so the Hamersley Clifford theorem or the Clifford Hamersley theorem it says that for any probability distribution right for any probability distribution that is consistent with this kind of a factorization over a graph right. So any probability distribution that is consistent with this kind of a factorization over graph right the condition here is a little stricter, so the condition says that.

It cannot be 0 right so the condition says it has to be positive not non-negative so what the Hamersley Clifford for theorem says is if a probability distribution okay that is consistent with this

kind of a factorization that is means of such a factorization exists okay then that probability distribution can also be expressed by using factors. of this form.

So that makes our life easier, so then now my energy function or what they the E function is called the energy function all of this comes from physics right, so you will see all this energy and other things here energies and potentials you will see that so the energy function right can be anything, now so as no restrictions can be negative it can be positive right well as long as it is real I suppose not complex right, so if the energy function can be anything so this is known as the.

So essentially what the Hamersley Clifford theorem tells us is that so if you if you write your probability as a product of Exponentials right offset these kinds of factors then there exists a graph representation where this kind of a factorization can be obtained in like ways if you are able to write a factorization like this on a graph then you can have it is expressed as a product of Exponentials right, so each factor is an exponential so essentially my probability will be the product of exponentials right, so this is actually a very powerful result because it allows to simplify a whole bunch of computations right I do not have to consider any arbitrary form for my  $\Psi$ .

Yes immediately you see this right I do not have to consider an arbitrary form for  $\Psi$  it is just an exponential okay, now I have to consider an arbitrary form for my energy but now that we started talking about it as energy, so we can start applying our intuitions from physical systems.

Right, so what should be a state with a high with a high probability. Exactly we know that right the state with the high probability should have low energy so when you start looking at the data right I find that configuration  $X_c$  which is most popular most prevalent in the data right and I am going to assign that the least energy and I do this for every clique.

The power of nomenclature, so I call it energy and now everybody understands what the graphical model is doing right, So if the energy is very high so it is going to be e power minus the energy so the probability is going to be low or rather the this factor will be low right and this  $\Psi_C$  is going to multiply herein your numerator right therefore the probability that you assign will be low right if

the energy is very low the e power that will be high relatively right and therefore the probability you will assign will be higher.

So that is essentially what we are going to do right, so as far as undirected graphical models is concerned so how do you decide what these energy function should be it depends on your The prevalence of that particular configuration in the data right, so what will be the energy of X1 is 1 and X2 is 0 and X3 is 1 okay how often did the combination occur in the input right and what should I do with that?

Nothing I can just use the count as the energy right because energy is unrestricted do not have to worry about normalizing it or anything right., I just use the count as the energy for that counting the data I can use that as energy, so the higher the count the higher the energy and therefore more prevalent the more probable that configuration will be sorry higher the count okay then one by the count is energy then use one by count as energy right, so the higher the count the smaller the energy and then the more likely the configuration will be right sorry yeah,

So that is the easy enough to do all right so I wanted to look at before I go on to do some inference thing I want to look at one or actually two popular graph (inaudible) yeah. So this can be zero so that can be at most one yeah. Anyway we are interested only doing probabilities here right,

So that is not that is not that much of a much of a problem right. Then again that is a beauty of the Hamersley Clifford theorem that is essentially it right it tells yeah it is fine you still can represent any probability distribution you want by looking at as this product of exponentials right. So that is the thing now since all of us now understand the undirected graphical models I start with the simple undirected graphical model right, and.

(Refer Slide Time: 15:49)



Right so this kind of a simple lattice like structure right so undirected graphical models are also sometimes called Markov random fields just like directed graphical models are also called Bayesian networks right undirected graphical models are also called Markov random fields right so people if you have heard of the term Markov random field somewhere right so one of the most often new structure in the with Markov random fields is like it is kind of a lattice structure okay.

So what this is really tell you it tells you that this variable okay is independent of everything else in the network given the four neighbors right then this variable is independent of everything else

in the network given only these two neighbors I mean these lattices can run for like you know, 32 cross 32 or sometimes 256 cross 256 people typically use these kinds of lattices for modeling images.

Okay, so it is random you agree with the right, so all of these are random variables right so I have this collection of random variables okay it is called Markov because in this particular case right given the immediate neighbors right I am independent of everything else right so if you think of what the Markov assumption is in a probabilistic models right so it is a stochastic model with Markov assumption says that given the immediate predecessor your independent of the past right, so that is a normal Markov assumption right so given the immediate predecessor you are independent of the entire in the past so here what I am saying is instead of the predecessor because there is no notion of direction here so I am since there is no predecessor here, so instead of that I am saying given the immediate neighbors I am independent of everything else that is why it is called Markov right.

So let us take  $X_i$  right so  $X_i$  is independent of everything else given all right so now what people do is they try to use this for by making all kinds of predictions. Right, so I am interested in labeling every pixel in an image, so I have a big image right I want to label every pixel in an image give me an image labeling task.

I want to label every pixel either foreground or background and this is the guy standing there or is it the tree behind him right so I want to label it as foreground or background right, so now it is a two label task right so each of these random variables will take one of those values what are the values it will take it will take whether it is a foreground or is a background okay.

Now here is the additional assumption I am going to make I am going to make the assumption that the value of the pixel I am going to see. The value of the pixel I am going to see depends on whether it is a foreground pixel or a background pixel right nothing else.

The value of the pixel I am going to see it depends on whether is a foreground pixel or a background pixel and nothing else right now essentially what I will do is. I am going to assign more random variables right. Each one of them stands for the individual pixel.

Each one of the stands for an individual pixel so now what will I do is I will observe these pixels right I initially will observe the pixels, so what will be my potentials here what are the  $\Psi$  how many how many  $\Psi$  do I need one for each edge right. So that is the maximal clique here I cannot do anything better than that so for every edge in this graph I will need a  $\Psi$  but for every edge in the graph I will need a  $\Psi$  right.

So what I will do is I will observe these pixel values okay, so some something some values I will observe okay then what I can do is I can figure out what should be this level of label of this pixel right just with this knowledge alone right because of there is this potential right I can kind of convert that into a potential on the node alone.

You see that because I have observed the values for the pixels right I can essentially take that entry from that thing right, so there will be one column associated with that pixel value right so

People giving me blank stares we are talking about a function of two variables right so I call this  $X_i$  will call this  $Y_i$  the pixel value  $X_i Y_i$  so I will tell you what  $Y_i$  is right then what will have we left with. A function on  $X_i$  right if I tell you what  $Y_i$  is I will be left with a function on  $X_i$  correct so I can convert given an observation right I can convert these potentials right into potentials on  $X_i$  alone.

$Y_i$  is a part of the graph but the way use it will always be that I am given the  $Y_i$  so here is an image give me the labels which is the foreground which is the background, so I will always know this  $Y_i$  right so given the  $Y_i$  I can convert this edge potentials into, node potentials.

Right, so this essentially from now I from the function of  $X_i, Y_i$  it will become a function of  $X_i$  alone okay so if you look at many such I mean graphical model applications right, so you will actually find that they will always reduce this into node potentials and edge potentials.

So it looked like they are defining something called a node potential okay which is like a potential function on single variables and then they will be defining edge potentials which are potential functions on pairs of variables okay in reality something like this would be happening for you to assign node potentials okay the node potential are essentially marginal some kind of information you have about the marginal's okay, so in this case I am telling you how the marginal information comes.

It is not complete it is not the complete marginal okay given the pixel you can make some guess of work what the  $X_i$  should be so that is my node potential so I already can reduce all of these edge potentials between a node and a pixel okay between a label and a pixel I can reduce it to a single potential on the label right. So now having done that right what can I do so that is there will be some potential for a label here okay there will be some potential for label here.

Right and there will be some potential for these two labels happening together. Right, so essentially this is telling okay this is a background okay what is the pro likelihood this is also a background if this is a background what is it likelihood this is also a background okay if this is the background what is the likelihood that is a foreground.

So like that right so for each of these edges so I have where the edge can change where the label can change that information I have, so finally when I assign the final labels to this so what will I do I will find that configuration of labels okay that gives me the lowest energy. And essentially that would mean that this potential should be. Low right if I suppose this I say is background and this I say is foreground then when I say label is B here and label is F here that entry in the potential function should be low.

Right people get that, so that entry in the potential function should be small so like that I need to do this for all the pairs here so it is a very hard problem so because it severely constrains so I have consider all possible pairs and I have to figure out where we low entry occurs across this pairs, so for example so if I say this is background and this is foreground and this is B and F and this gives me a low value right but I say B and B for this and this gives me a high value right. I you can possibly turn this into an F right but then I the whole thing might come around and then this might

get changed back into an F B right so then I might go around or not have to figure out what I see right potential to pitch this and that is the inference problem, so inference problem is really hard in the undirected case right so in so much so that when you have loops like this right when you have loops in the graph like this exact inference is impossible to actually give you the right answer is impossible like quite often we end up giving some kind of an approximation right. So where can you give exact answers when there are no loops right so undirected with no loops is a.

Tree right so on trees you can give exact computation but as soon as you have any loops in that then you have to do some kind of approximation there are some very special cases but we will not go that right you just want you to get an intuition of what the individual factors would mean right. So how will I determine what these factors are okay so let us look at this right I want to look at. I want to look at the that right so what am I going to do.

What is it really it is? Oh I want to assign a low energy to the configuration that have occurs most often in my data right so what I will do is look at the label data so I look at each pixel's label right the label data will tell me which pixel is foreground which pixel is background right then I look at for these two pixels in the image let us say I have only a three by three pixel image right but for each of these two pixels I look at them right, I will figure out okay how often was this foreground and this also foreground how often was this foreground and this background how often was this background and this foreground how often this background and this background, so all these four things I look at just count that from the data and then I will set the energy to be some inverse of that count so the one with the largest count will get the smallest energy right it pretty simple right, so like that I can do this for each and everything right. So that is just one thing so the second thing I have to do is then start looking at this thing.

Exactly that is what I think that is what is particle sorry you are looking at my back you can see from this I said there should be a potential function for  $X_i$  and  $Y_i$  as well right so what that what I will do I look at the data again so look at okay when the pixel value was this okay what was the label right so I will do that I will do the co-occurrence information okay now things should start looking a little fishy to you guys I mean the pixel can have a lot of values depending on how I am encoding my color or brightness or something like that right.

So that I could end up having a very large distribution there itself right so even if you assume that my pixel is going to have 256 levels of brightness right so for every value so it is to 512 probability 512 counts that I have to make right.

512 counts looks suspicious right will I have data to actually make accurate estimates of 512 individual counts well I could if I have very large volumes of data but typically what you do is you don't for  $Y_i$  you do not do the explicit counts like this right you try to learn the factor  $Y_i$  by some kind of a parameterized function right.

So you could use logistic regression right so figuring out what the what is the probability of  $X_i$  given  $Y_i$  right that logically regression can tell you right, so you encode your pixel using whatever things you want right so you can look at you can look and even more funny thing so you can do funky things you can make this a function of all of these pixels right so because you are moving away from your Markovness, but once you start thinking of doing a logistic regression you can come up with very powerful classifiers right.

So typically the  $Y_i, X_i$  probabilities right or other the  $Y_i X_i$  factors are learnt in a different manner they just do not do the maximum likelihood estimate you do some other thing that typically the most popular choices using logistic regression right you can do other things but then the  $X_i X_i$  potentials you can learnt using maximum likelihood estimate simple maximum likelihood provided they are small enough okay so that is basically how you train this Markov random field and it turns out that they are pretty powerful in terms of working with images and in a very wide variety of setting right and people use MRFs a lot right. And there are variants of it which is called conditional random fields so people use those also tremendously so.

Very popular and powerful classifier and training it can be a pain right, so I just give you a simple example right value just did the counting and stuff like that when you have a very large model right very large graph right training it can be a pain but then people because the data sparsity right, so I had to look at all possible combinations of all variables right and so it becomes a little tricky and so you have to come up with the clever ways of training the models.

Okay they say it is particularly this part of it is painful right not  $\Psi$  not completely (inaudible) just that the inference processes is hard like you are selling your head you might you may keep going in circles it may take a long time for you to actually converge to a probability and so on and so forth yeah.

So there exists a proper assignment to this so that yeah so but the finding it is hard yeah Clifford Hamersley let tells me that no I am not worried about loops right as soon as there is a clique there is a loop.

Clique loose but there is only one potential right, so it is fine I will be doing the inference on that potential alone I would not get into this loop business so the loop business started off saying that okay I will make one inference one assignment here one assignment here one assignment but if this had actually been a clique right then I will know what is the combination of assignments to these four variables that has the lowest potential, I would have just done that right so I would not have to runaround chasing the things so that is the right so.

But I still have to do the chasing run because this has other things it is involved in right so the only way I can ensure that I do not do the chasing around this if I have one connected graph a complete graph right one if a complete graph then of course there is no chasing around and then what so what is the difficulty.

There is no factorization I am back in the system as well not have worried about the graph right, so if I have a complete graph there is no factorization. So but that is only way I can ensure that I will not get into.

### **IIT Madras Production**

Funded by  
Department of Higher Education

Ministry of Human Resource Development  
Government of India

[www.nptel.ac.in](http://www.nptel.ac.in)

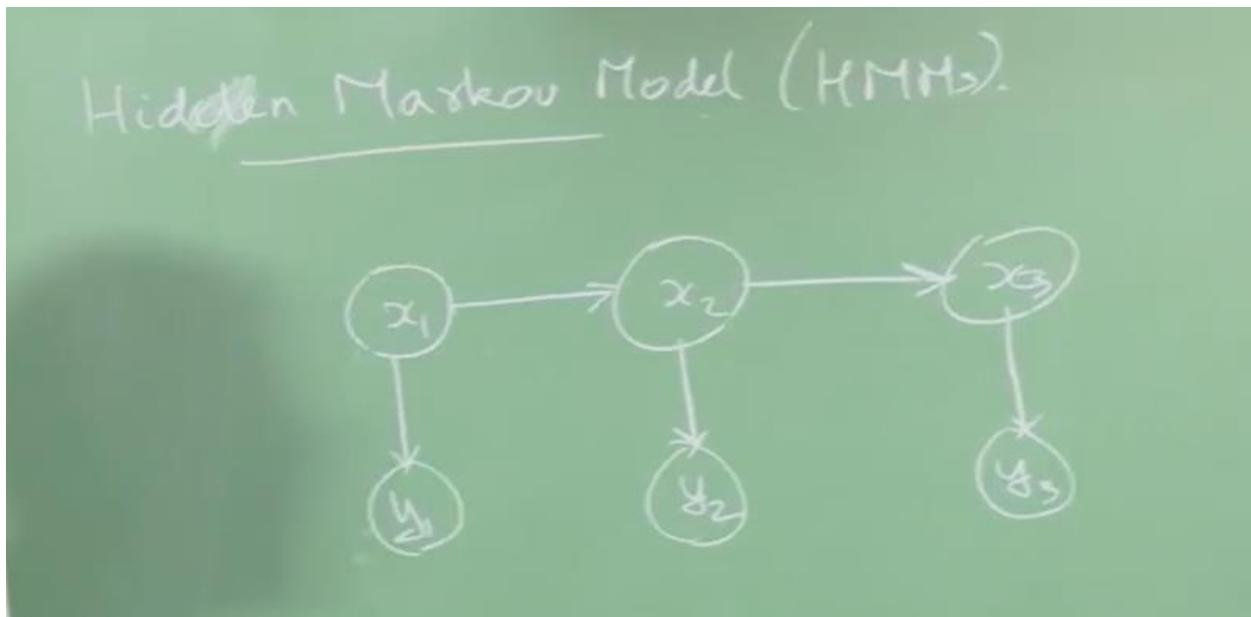
Copyrights Reserved

**Introduction to Machine Learning**

**Lecture-67  
Hidden Markov Models**

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

(Refer Slide Time: 00:16)



Right so people might have come across HMMs in other contexts right, and you might be wondering what does he mean by saying it is a graphical model, but it is a graphical model right so here I will stick with a concrete example, I am going to have a sequence.

Right, so I am going to have a sequence of random variables right I am going to assume that it is Markov in the traditional sense, so what would that mean.

It is left to right, and hence that is the earliest point in time. Okay, so there is no time it is a sequence right. This is the first element in the sequence okay the second element in the sequence the third element. Because it is Markov so this is this will be dependent only on the previous one. So knowing this value makes this independent of anything that came before right, so this is essentially the graphical model version of Markov right just have a chain directed graph, so it is sometimes called a left-to-right model. So this is the Markov part right. I am also going to assume that. I have a set of observations that I make.

Right, this could be like the pixels we are talking about right, so I have labels for the images right labels for each pixel on the actual pixel value right. So likewise in the hidden Markov model so I will have random variables which could be labeled right, so these are labels. Right and these are the entities that are being labelled. So give an example of such a situation they would find.

Media is a little tricky because you also have the spatial dimension to it right, so the text is why I asked how many of you are in NLP other day. So text right I might want to do say part-of-speech tagging. And I want to take each word again I want to assign a part of speech to the world right all I want to assign each word I want to take each word. I want to say that whether the word is part of a phrase or not part of a phrase. For example, I would need to figure out the United States of America is a phrase .

Is that so is that some way of doing this automatically right, so people have come up with ways of doing there are many things there are tasks in the NLP called chunking, so people have come across chunking, no, okay? So chunking essentially says that I am going to take a piece of text right and break it up into some meaningful chunks it could be noun phrases verb phrases whatever right, so but some chunks right, so that is called chunking right. There is another task called shallow parsing right so shallow parsing essentially breaks it up into phrases right but does not look at the structure of the phrases I just want to give the phrasal structure of the sentence. Hence, there are many tasks which people do this right, so people use hidden Markov models a lot in text right and the other place where people use hidden Markov models a lot is in speech right because speech is inherently a. Speech is inherently a forward process, so people use that a lot in speech. You can use it in videos except that you are individual random variable will now

become something that covers the entire spatial dimensions. This becomes a little more complex hidden Markov model right.

So great, so I am going to call these after that as X's. Well, these are words right. This could be words, and these could be whatever label you want to send to them you can say that okay so is it part of a chunk or not part of the chunk what is it this is part of a name of not part of a name so I could do all kinds of labels.

That is the assumption we are making to make the model easy. Okay, so yeah we can relax this assumption there are models that the more complex model that does this specifically the conditional random field that I mentioned when I was talking about Markov random process. It relaxes the assumption of this dependence.

Right, so this Markov assumption is still there right. Still, this dependence is relaxed, so that is typically what you are looking for I can look at any word in the sentence and then I can label the seventh word so that is that assumption is this is like right now I have to look at only the seventh word and the sixth label I cannot look at the sixth word I had to look at the sixth label and the seventh word right, and that is all the information I have.

So one thing to note here is a hidden Markov model essentially says that your  $X_i$  give rise to the  $Y_i$  right does not go the other way. This gives also gives ways to other problems later on. So the  $X$ 's are essentially you have labels, so you do not know the labels you don't see the labels but whatever you do not see right that part of it is Markov okay.

And whatever you see is or the labels okay and they are just given influence only by the I am sorry. Whatever you see are the words then they are influenced only by the labels that are the assumption you are making so very strong independence assumptions right. Still, it turns out that it works like Naive Bayes works right so this also works in many situations right, but then, of course, it also does not work in a lot of situations, so people have come up with other models I just wanted to introduce you to HMMs.

I do not get to see  $X_1, X_2, X_3$  and I only get to see  $Y_1, Y_2, Y_3$  and  $X_1, X_2, X_3$  I will have to guess. I mean you can also think of this as a hidden Markov Random field if you want, but they do not use the terminology. Still, hidden Markov models are something that is used quite often, and there are inference techniques which people have specially honed for HMMs right. Still, it turns out that the same things work on many of the graphical models that you will see you will see in practice right, so they have all kinds of things they have an algorithm called the Viterbi algorithm which essentially tells you what the probability of the labels given the text, no it does not give you the probability of the labels given the text it gives you the MAP estimate and it gives you the most probable label sequence given the text.

But then I can use it on any kind of this kind of an inference process right, so I can use a Viterbi algorithm for any kind of a MAP inference process instead of looking at the probability distribution over  $X$ . If I want to answer the query which is the most probable configuration over  $X$  this is the MAP estimated we looked at that right maximum likelihood MAP and then the full Bayesian inference which cases me the full distribution right so far we have been talking about knowing the full distribution. Still, you also want to do the MAP query right. So that is essentially what Viterbi will give you right, so we are talking about estimating a potential between  $X_i$  and  $Y_i$  right, so we are talking about estimating a potential between  $X_i$  and  $Y_i$ . Hence, the question is  $\Psi(X_i, Y_i)$  independent of  $i$  with regardless of where in the pixel the label occurs right sorry we're in the picture the label occurs right is the relationship between the pixel the label the same right.

When I say that the  $\Psi$  is independent of  $i$  that is what it means right that need not be the case right suppose on the edge if a particular pixel value occurs it might have a higher probability of being a background right I mean it is very edge, you typically not going to frame a picture so that the foreground goes to the very edge rate. Still, then that might happen, but the problem is supposed I have a  $256 \times 256$  image then I have to learn a classifier for every one of those positions or to learn a classifier that gives me the probability of  $X_i$  given  $Y_i$  for every one of those positions.

And that is a very large problem right; it is a very hard problem. Hence, what we typically assume that  $\Psi$  is an independent of  $\Psi(X_i, Y_i)$  is independent of  $i$  so I am going to estimate the same model across the entire image right so of course, you can try to be more clever I mean if you know some things about the about the task we can be trying to be more clever and say that okay I am going to break it up into four classes of  $\Psi$  okay for  $X_i$  and  $Y_i$ .

I will break it into four classes I will use one class here rather than the bulk of the image and so on so forth you can do things like that but yeah but typically without any prior knowledge you just assume that it is the same all along all right. This has to be different right, and if that's also the same all through then, you are we can use one pixel and predict the whole image the label for the whole image Yeah, do we end up doing that you could do that also yeah  $\Psi(X_i) \Psi(X_{i+1})$  right whatever that could also be taken as being independent of  $i$ .

You lose more modelling freedom that way, but you could do it that way as well yeah depends on how much effort you want to spend in doing the building the model yeah.

### **IIT Madras Production**

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

**NPTEL**

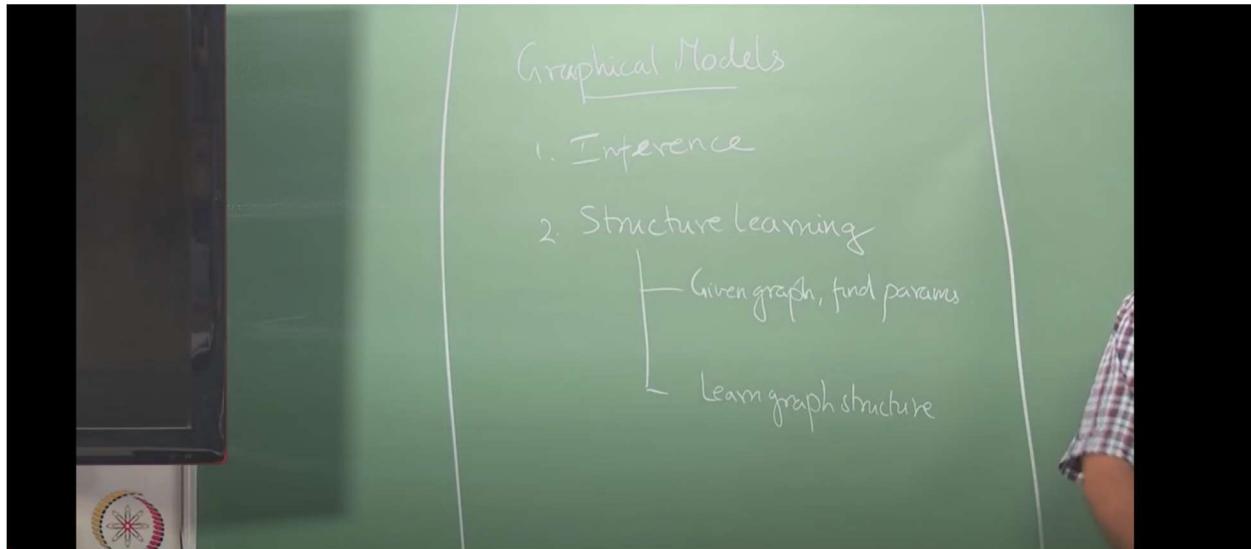
**NPTEL ONLINE CERTIFICATION COURSE**

**Introduction to Machine Learning**

**Lecture-68  
Variable Elimination**

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

(Refer Slide Time: 00:14)



Right, so we are looking at graphical models we looked at both directed and undirected models right. And I said the thing of interest was, so two things are of interest. The first thing is.

Given a model right, how do you do inference using the model right? So what is the inference question, inference question is trying to answer queries on marginal's right. So I give you a very complex joint probability distribution, I want to know what is a probability that there is an earthquake, yeah that is not a very complex system.

So what is the probability that there is an earthquake right, I can also ask for conditional marginal's given that John called what the probability that is a nice case is. So these are things we looked at right, so it turns out that this itself is a hard problem and for large graphs, you will have to come up with ways of approximating given this right. So I will kind of motivate why there is a hard problem in a minute.

And the second problem that we are interested in is, What was the first problem, sorry, exactly. So what could be the second problem? No, find the model then, how do you derive the model right, but you are close right, so how do you find the model right, from the raw data may I will give you training data, I will give you a lot of data how do you find the model right. So the Bayesian network structure learning itself is a hard thing.

So the simple problem is even in the structure learning they split it into two things right. So I should probably put this down, the first problem is inference.

Right, so here there are two components with, so given the graph right. So I will give you the graph find the parameters and in the directed case that would be finding the conditional probability distribution. So once I give you the graph I know what the conditional probability distributions I need are, I can just go to the data count and find it out.

And in the undirected case what it would find the potentials, so given the graph find the potentials right. As soon as I give you the graph, you know what the potentials that you need to estimate right are. So you will have all these edge potentials, you will have node potentials, and you have clique potentials, so you will know what the potential that you are estimating right is, and you just go and estimate the potentials right. So this is essentially the learning problem given.

And the second problem would be to find a graph right. So one of the things you should look at finding in trying to find a graph essentially you would need to find that graph structure right, that supports them in conditional independence that is present in the data, it is directed graphs or undirected graphs whatever graph structure you are learning. So you have to infer what is the conditional independence that is present in the data, and you have to find a graph that will support that. So essentially you will have to, there are many ways of doing it people start off with a

completely connected graph, and then they start knocking off edges right. And then you can do some kind of cost complexity pruning as you do in decision trees right. So you could have a much more complex graph that then you can try to prune things down so that you can do a tradeoff between the number of edges you have. So the variety of algorithms that people are proposed for graph structure learning. So this part is the easy right given graph find parameters are easy, how will you do that just like conditional probability distribution estimates right. So you can very easily do that for directed graphs just counting.

Look at the data, see how many times Mary called when there was an earthquake right, or when the alarm rang how many times Mary called and then you can essentially fill in this conditional probability right. So those things we can do a straight forward right. But learning the graph structure is a little bit involved to get into that because it is a lot of, you know a lot of structure that we have to build in before you can. So I am going to now go back to inference, so the inference is the interesting part right.

So let me start off with an example right. So I am taking this example from. So for a long time, we did not have a really good book on graphical models, and then Koller and Friedman wrote this over complete book on graphical models I mean it is like it has everything that you would need to know about graphical models and more right.

So it is like this huge stone right, but it is a fantastic book, it really is a good place to start right. So why I am saying it is a good place to start this, this is still a very active area of research right, probabilistic graphical models and every year newer techniques, newer breakthroughs keep coming up. So it is like, it is not like you can write a book and say okay everything you need about graphical models is captured in the book right so because it is still evolving field. Right, I am going to draw a really large graphic here.

(Refer Slide Time 7:52)



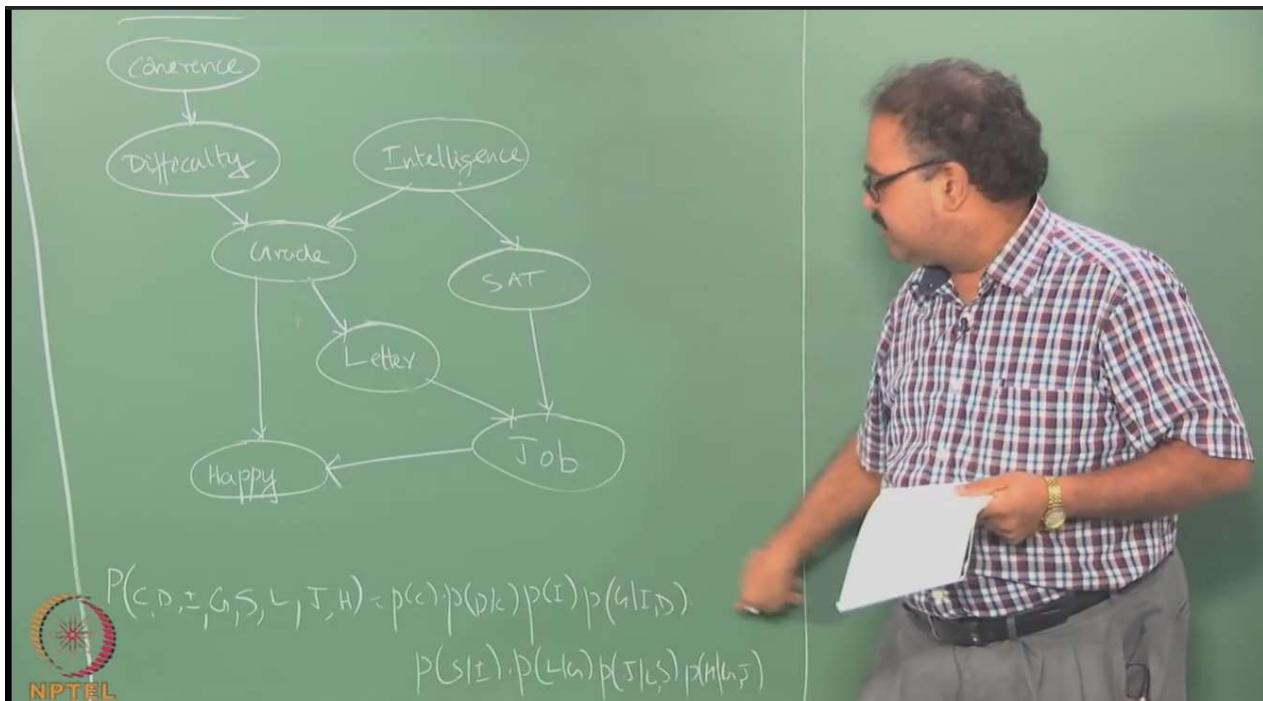
Okay, it is a small thing which Daphne Koller came up with to capture some fraction of her interaction with students right. So depending on the difficulty level and the intelligence of the student okay, the student will get some grade in the course right. And the difficulty level of the course depends on how coherent the teacher is right.

So the coherence influences the difficulty level, okay, and then the difficulty level intelligence influence the grade right. And so depending on whether the student got a good grade or not in the course, the teacher might give him or her a letter right letter of recommendation if the grade is bad, then the probability of getting a letter is very small, as the grade is good the probability of getting a letter is very high right, even there that happens.

And whether they get a letter of recommendation from the teacher or not right, it influences whether you get a job, and whether you get a job and whatever grade you did influences whether you are happy or not right, this is like, so sometimes you might be very happy for having done very well in the course even if you even though you do not find a job.

I am just giving you the structure here because this is sufficient for us to talk about some of the difficulties in the inference process right. When you are actually solving problems in this, you would need the probabilities, but we are not going there right. So they just give you the structure. Suppose I am interested in answering a.

(Refer Slide Time 10:02)



Let me write this out now, probability C, D, I, G, S, N okay.

So you people see what I have written if you cannot, you can write it from the graph directly, so you do not really need me to write this out. So right, so this is a probability of coherence times, the probability of difficulty given coherence times, the probability of intelligence than the probability of great given intelligence and difficulty, so on so forth, I have just written out the joint distribution you can just look at the graph, and you can write out that yourself easily right.

The Joint Probability distribution of the variables can be factorized based on the above graph,  
 $P(C, D, I, G, S, N) = P(C)P(D|C)P(I)P(G|I,D)P(S|I)P(L|G)P(J|L,S)P(H|G,J)$

$$P(J) = \sum_{L} \sum_{S} \sum_{G} \sum_{H} \sum_{I} \sum_{D} \sum_{C} P(C, D, I, G, S, L, J, H)$$

$O(2^7)$

What is the probability that a student in this universe will get a job; in this universe, I mean the universe is captured by this way. So what is the probability that person will get a job right, so what will you, how will you go about doing this essentially this will be. Okay, right, so if you think about it is essentially the order of  $2^7$  computation if everything is Boolean right.

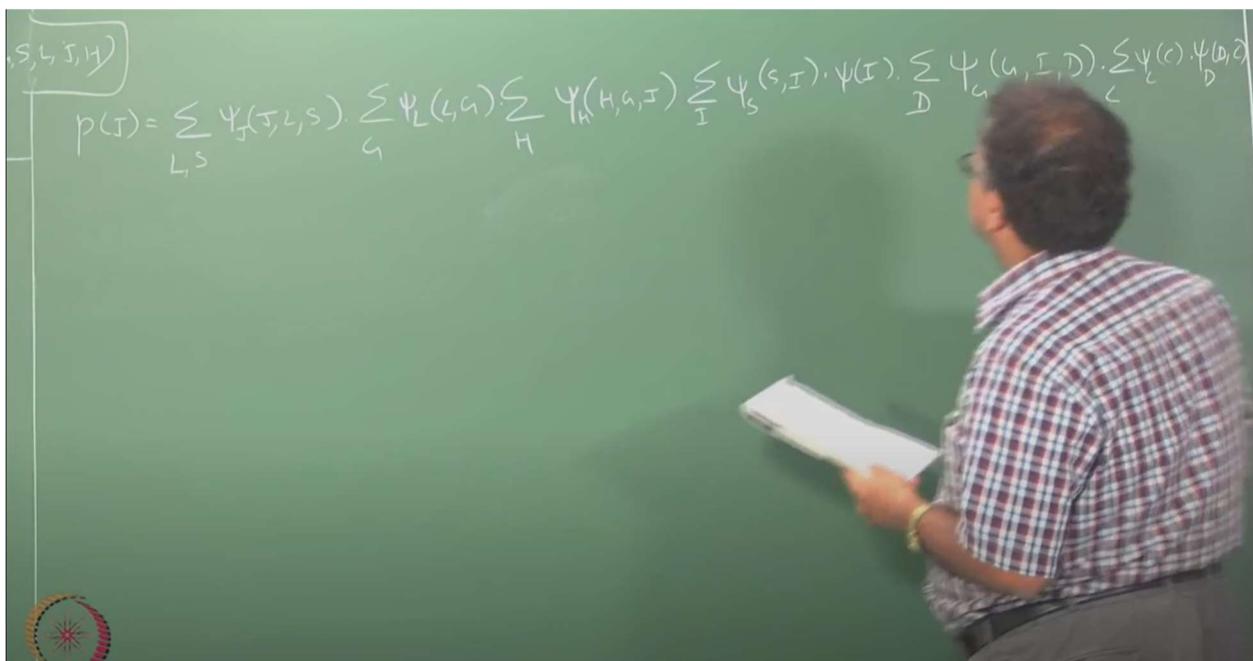
$$P(J) = \sum_L \sum_S \sum_G \sum_H \sum_I \sum_D \sum_C P(C, D, I, G, S, L, J, H)$$

Assuming everything Boolean, we will have  $O(2^7)$  computations.

So it looks odd right I mean so it means running this over the entire table running the summation over the entire table is not correct. So the whole idea of us making inference was doing this factorization was to make this computation simpler right. If I did not have the factorization right, I essentially would have had to do this computation. So yeah, so this is some set of running over this very large table.

So now what we are going to try and do is try to make the summation simpler by pushing in some of the sums right, pushing it into the maximum extent possible so that what I sum over okay, as a smaller table as possible right. Right now and all my seven sums are running over the entire joint distribution right, I want to rearrange this in such a fashion that each sum runs over as smaller setup as possible right. So how will I do that? Just for the same question here

(Refer Slide Time 17:30)



We can simplify the above summation form of  $P(J)$  as,

$$P(J) = \sum_{L,S} \Psi_J(J, L, S) \sum_G \Psi_L(L, G) \sum_H \Psi_G(H, G, I) \sum_I \Psi_S(S, I) \Psi(I) \sum_D \Psi_G(G, I, D) \sum_c \Psi(c) \Psi_D(D, C)$$

$\Psi$  – Factors of the graph.

In directed graphs,  $\Psi$ 's will be conditional probabilities.

So I have moved from the conditional distribution to the potential formulation right, but you know what this means this is essentially the conditional distribution here so you can actually think of that having been represented as an undirected graph. Also, we can use the same technique that I am doing here even with undirected graphs right.

So that is the point I am to make that point I just switched over from this notation to this notation. So in this particular case, these factors happen to be conditional distributions, but they could be factors that you get from here. So in which case you probably have to have some kind of normalization going here right. So if you are going to use this as an undirected model then you have to have some normalization to take care of, so is it correct, so what I have written is correct right.

So the notation I am doing here is essentially this it takes in J L S as arguments again returns a distribution over J. So that is what the J here stands for, so it takes J L S as arguments and returns the distribution over J right, or some function over J, this takes L&G has arguments okay and return something over L. So that is what this is right, so this is essentially the probability of J given L, S or something like that the equivalent to that in my potential notation. So that is the thing I am marking here okay, is it clear?

So now you can think about it, so the C runs once over only those two tables they are small tables, so C has just one in two entries in it right  $\Psi(C)$  will have only two entries in it right.

Whether the teacher is coherent or the teacher is not coherent right. And  $\Psi(D, C)$  will have how many interests in it. Four entries in it right, how many independent entries in it.

Yeah, two okay it will have only two independent entries, not three right. Because given the course is not, given the teacher is not coherent, what is the probability it is difficult. So automatically 1 minus that gives me the probability it is not difficult right, even the teacher is not coherent what is the probability is difficult and 1 minus that gives me the probability that it is not difficult right. So I only have two parameters, so you can see that I am reducing the parameters tremendously.

So here what would I have had, I would have had  $2^{8-1}$  parameters right, the full joint distribution right if I specify  $2^{8-1}$  parameters and 1 minus the sum of that will give me the last one. But here look I have tremendously cut down, so this has one parameter, this has only two parameters right. So likewise this is going to have.

Four parameters that are for every combination of I, D you are going to have one possible outcome for the other is 1 minus that right. So for every combination of I, D you need to have one parameter so you will have four parameters, so likewise here you will have one parameter again, here will have two parameters so like that, so you are reducing, if you take the product is much, much smaller than the  $2^{8-1}$  that we had right. So that is the power of doing the factorization, so the number of parameters you need for specifying the joint distribution comes down significantly.

You can start pushing the sums in so that this sum runs over only if the small number of elements right. Likewise, this sum runs over a small number of elements and so on so forth, and then I can complete the entire joint distribution right. So this kind of an approach right where you push the sums in is known as—variable elimination.

(Refer Slide 22:10)

$$\begin{aligned}
 P(J) &= \sum_{L,S} \Psi_J(L, S) \\
 &\stackrel{\Psi_L(L, S)}{\sim} \sum_I \Psi_H(H, I, J) \\
 &\stackrel{\Psi_H(H, I, J)}{\sim} \sum_D \Psi_G(G, I, D) \\
 &\stackrel{\Psi_G(G, I, D)}{\sim} \sum_C \Psi_C(C)
 \end{aligned}$$

$$P(J) = \sum_{L,S} \Psi_J(J, L, S) \sum_G \Psi_L(L, G) \sum_H \Psi_H(H, G, I) \sum_I \Psi_S(S, I) \Psi(I) \sum_D \Psi_G(G, I, D) \sum_C \Psi_C(C) \Psi_D(D, C)$$

Variable Elimination: A exact method of inference,

Using this method, we can reduce  $P(J)$  further,

$$\begin{aligned}
 \tau_1(D) &= \sum_C \psi_c(C) \psi_D(D, C) \\
 \tau_2(G, I) &= \sum_D \psi_G(G, I, D) \tau_1(D)
 \end{aligned}$$

We can continue performing this creation of  $\tau_i$  factors with estimated parameters finally to estimate  $P(J)$ .

Right, so for small graphical models okay, this is a good way to do inference right. It is not an approximate way of making an inference; it is an exact way of making inference right it gives you the same result as you would have gotten if you had summed over the entire distribution okay. So it is called variable elimination, and so the advantage is like I said they have an amount of

computation that you are doing you will be minimizing right. So how much computation would you be doing, what will be the maximum, what will be the largest table that you are summing over?

Exactly it depends on how much you are able to compress the things and how much is actually able to eliminate the variable. So the more variables are supposeing variable eliminates the faster will be your computation right. So think about what you are doing here, the first step is marginalizing over C right.

So I am going to say that you marginalize over C right, and you end up with a factor over D right I am going to call it some  $\tau_1(D)$  right so what will  $\tau_1(D)$  look like.

That is  $\tau_1(D)$ . Right, next what do I do what I am I marginalizing over D right, so this guy this whole thing I am marginalizing over right I am going to call that factor  $\tau_2$  and what will be a function of G and I?

So I keep doing this next I'm eliminating I, so what we will end up with the factor over G & S right yeah. Then what we will end up with I will have this guy as it is right am eliminating H right, so there is no H here so  $\tau_3(G, S)$  will continue propagating beyond this point right but I will also introduce a new factor called  $\tau_4$  which will have you can see that right at this point I just trying to try for you to get an appreciation of what the computation is happening right at this point you will have  $\tau_3$  you will also have  $\tau_4$  right.

So when you compute it till  $\tau_3$  you have eliminated you eliminate a  $\tau_1$  you eliminated  $\tau_2$  because you have rolled up everything into  $\tau_3$  but we do  $\tau_4$  you are not able to eliminate that right so  $\tau_3$  is still carries on to the next level. Now we eliminate G, so what you get at  $\tau_5$  so eliminate G, so we will have J left-right we will have S left, and L will now get added here finally you will get depending on what order you do this thing in. Here I will first sum over S I get this then sum over L, I get my P(J), okay so this is essentially the or how you will be doing the elimination.

So as and when you are doing the elimination, and you are creating this new factors so what we should be thinking of is it is as if you are adding a new potential it is as if you are changing the graph right. So when it did this right well, I did not let us certainly really add anything new D is already there right what about this. So now I create an edge between G and I and not a big deal G&I already existed right but what about this now I create a potential between G and S right. So when I come to this point, so it is like I am adding a. Another connection between G and S right so likewise anything else is happening. Anything else. J & L is already there JSL, JSL.

Right, I need to have a clique for me to have a potential JSL I need to have a clique, so I am essentially like adding an edge between S and L right. So you can think of the way we are doing this is essentially like we are making this larger some of these potentials are making larger and larger right.

So, in this case, it turns out that luckily none of the intermediate steps that we are creating makes a large table right none is nothing is larger than any of the existing tables right so we could choose a bad elimination ordering I can choose a different order.

So here the order we chose was C, D, I, H, G, S, L okay, so that is the order in which we eliminated the variables started off with the right-hander at C, D, I, H, G, L okay, suppose I did this.

(Refer Time Slide 28:00)

$$\begin{aligned}
 & \sum_{\mathcal{G}} \Psi_L(L, G) \Psi_H(H, G, J) \Psi_G(G, I, D) \\
 & \sum_I \Psi_S(S, I) \Psi_I(I) \tau'_1(L, H, J, I, D) \\
 & \tau'_2(L, H, J, D, S)
 \end{aligned}$$

Order of Elimination:

$$P(J) = \sum_{L,S} \Psi_J(J, L, S) \sum_G \Psi_L(L, G) \sum_H \Psi_G(H, G, I) \sum_I \Psi_S(S, I) \Psi_I(I) \sum_D \Psi_G(G, I, D) \sum_c \Psi(c) \Psi_D(D)$$

Consider a new order of elimination for the same graph,

$$G, I, S, L, H, C, D$$

To eliminate G, we have to all factors that have G,

$$\tau'_1(L, H, I, J, D) = \sum_G \Psi_L(L, G) \Psi_H(H, G, I) \Psi_G(G, I, D)$$

We have a 5-way table if we eliminate G first.

Now we eliminate I,

$$\tau'_2(L, H, J, D, S) = \sum_G \Psi_S(S, I) \Psi_I(I) \tau'_1(L, H, I, J, D)$$

Likewise, we can eliminate other factors. This shows how the order is important in variable elimination.

They start off eliminating G right. So I can sum over G, and I have to put in all the factors that have G in it right, so what were the factors that had G in it I will sum over G I will do right from this side right so I will have  $\Psi(L) \Psi(H)$  when I summed over G over all these factors. So now I am going to create my new  $\tau_1$  right so I will call it  $\tau'_1$  so  $\tau'_1$  will be a function of everything in

that that is not eliminated right. So G has been eliminated so what it will be so L H J I D ouch. Now I created a 5 a table there. By choosing to eliminate G first right, I have created a 5 a table so that is a large table and now I am going to sum over this.

So now will be summing over a table which has  $2^5$  entries right so that is a bad thing right so next one what I have eliminated next try to eliminate I next so what I will do that so I will have  $\tau'_1(L, H, I, J, D)$  is there any other factor that has I  $\Psi(I)$  this doing right and what will this do it eliminate I right but it will add S to the factor so my  $\tau'_2$  will be a function of L, H, J, D, S now I have another five-factor table, in fact, this is the worst possible elimination order okay to give you the really bad picture right that is the worst possible elimination auditory then I eliminate S.

So what do I do in that case well I add JLS also to the mix right a large JLS also to the mix I will eliminate S that but J and L are already there in the factor so, in fact, this will come down so my T 3 will have only L H J d because I eliminated s right then I will eliminate L right nothing a new gets added that the only thing that is left out to is C right yeah so by the time I come to C, yeah so everything else will get eliminated.

So finally I will be left with a factor that contains only D&J and then finally eliminate D so what will happen when I eliminate S we are done to yes okay what happens on eliminating L I will end up with a factor that has HJDS then what happens if we eliminate H I will end up with a factor that has JD, JD what L right no L is already gone eliminate H will just end up with the factor that has JD I have a factor that as JD I will also have the this is the C's the last two factors will still be there the  $\Psi(C)$  and  $\Psi(DC)$  that those two factors will still be there right everything else will get eliminated, and then what I eliminate C that means those everything all those factors will get eliminated I will be left with a factor that has only D&G.

And finally, eliminate D, okay but what I had done along the way is that I have created a big clique herewith five variables in is right, so if you notice as we went along so even though this looks like a clique of four variables okay it was never created as a clique of four variable said that at best I only did a clique of three variable just two different cliques of three variables it looks like a clique of four variables but we never generated the clique right but in this case we actually generate a

clique of five variables so it can become very large right. So it turns out that the complexity can be related to this the complexity of running inference on this graph can be related to the size of the largest clique you generate along the way right,

So, these kinds of edges that we generate like this right are called fill-in edges yeah. This one G&S this one yeah also eliminating I right. So when you add this thing I mean well I did not want to erase everything but When you add this fill-in edge that essentially when you remove that so this not really a clique.

So this is not really a clique this is only a this the edge is the maximal clique, in this case, so your question is I do not have a potential that says I GI and S right, so when we did the original ordering we never did a G, I, S potential that is because what you pointed out. So I was eliminated, and therefore we only have GI was already existing right we have a potential corresponding to G, I, S in the beginning.

No why should we have one corresponding to G, I, S no we do not need one corresponding G, I, S, so we do not need one corresponding to G, I, S, so you do not need one at all in the inference also when this fill-in edge is added that those things are not there right. So we only have to worry about those fill-in edges which actually leave you with a clique is what I am writing the size of the largest clique. In the elimination, ordering is called the induced width of that ordering.

**IIT Madras Production**

Funded by

Department of Higher Education

Ministry of Human Resource Development

Government of India

[www.nptel.ac.in](http://www.nptel.ac.in)

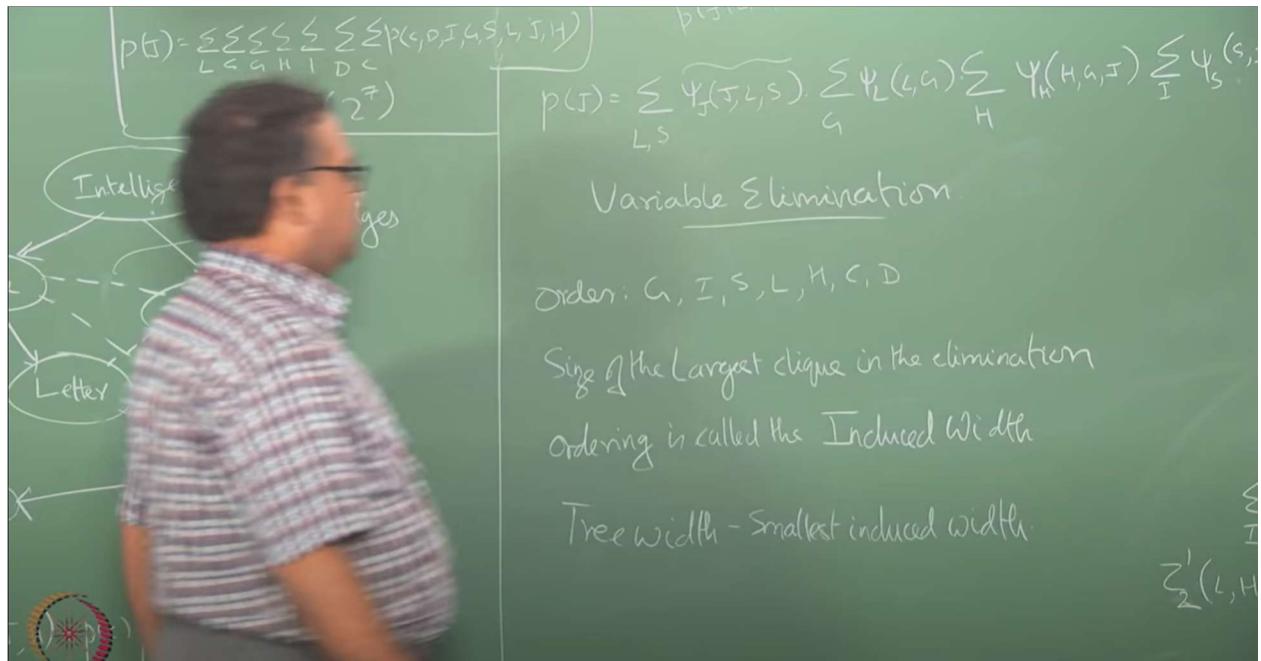
Copyright Reserved

## Introduction to Machine Learning

Lecture-69  
Belief Propagation

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

(Refer Slide Time: 01:31)



And we have a concept called the treewidth of a graph, treewidth of a graph which is the minimal induced width, the treewidth of a graph is the minimal induced width so what do I mean by that so across all possible variable elimination orderings that you have right, so you can find out what is the induced width for every elimination ordering right, and there will be some order that gives you the smallest induced width that is your treewidth.

The complexity of Variable elimination is

$$O(K^{\text{Tree Width} - 1})$$

Tree-Width – Smallest Induced Width

Where k is the number of values for each random variable, so, in this case, we assume k is 2, so it will be the order of  $2^{\text{Treewidth}}$  right. So what would be the treewidth for a tree? One or two depending on how you count tree width some people count treewidth that is the size of the clique -1 okay, in which case it will be 1 right if you count as the tree width as the size of the clique it will be 2.

Because I can eliminate it from the leaf to the root right, at no point we will add a larger factor so every time I will be just collapsing one edge at a time right, no point will be adding a larger factor so if I eliminate it from the root to the leaf then that might be problems right, is some arbitrary ordering, but the smallest thing is to eliminate from the leaf all the way back to the root so every time you will be just removing one edge, right.

So if you think about it the smallest elimination ordering for us here right, also started off with the kind of like the single node hanging off here. Right, so you have to eliminate C first if you eliminate C at the end then you ended up adding some other nodes along the way right, so getting rid of here somehow coming from the outside inward right, or going from inside out was a bad idea, so that is essentially well. So for trees, variable elimination is great right, because it does things kind of in the best possible way you can expect it right.

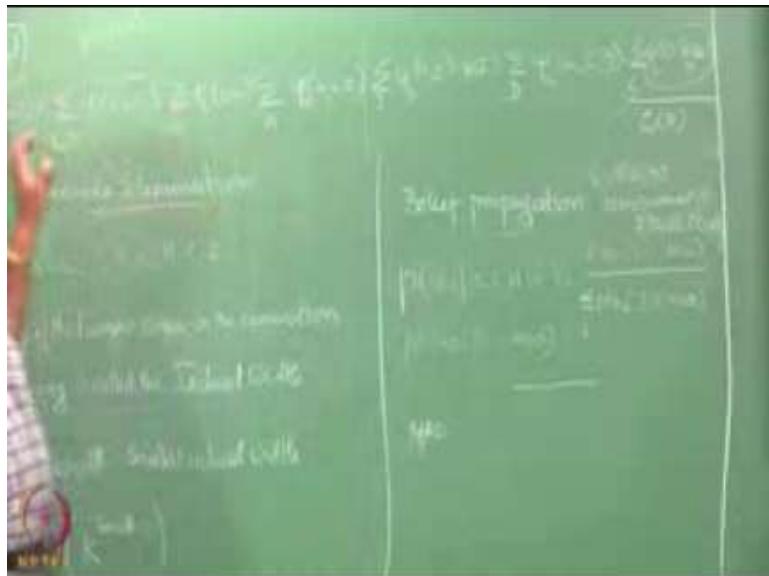
But still, there is a problem, what is the problem?

I asked you to find  $p(H)$  right if I had asked you to find  $p(H)$ . Right, you basically have to redo this computation all over again right, so many of these tables that you computed internally could actually be reused that if I want you to find  $p(H)$  in fact.

Up till this point, everything can be reused; in fact, some of this computation also could be reused right, inappropriately modified form right.

So but then you end up doing everything all over again right if I asked you to do  $p(H)$  right, so there are more efficient techniques where you can keep caching these things away right, the most popular of this is called.

(Refer Slide Time: 05:26)



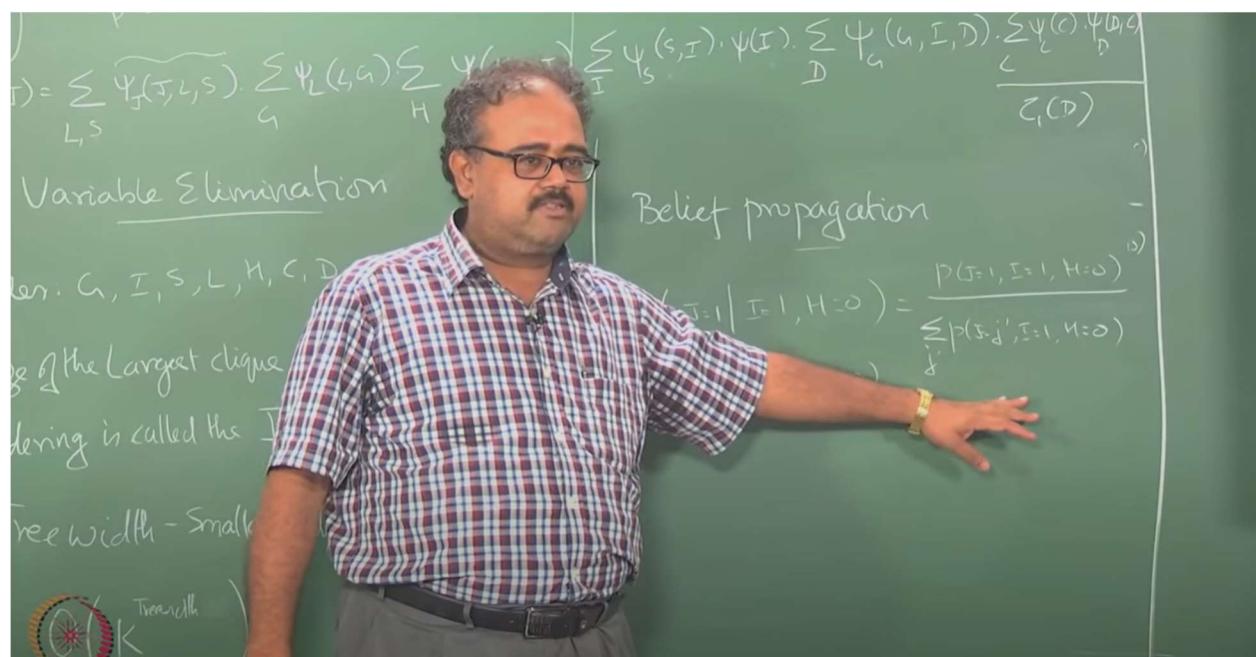
Most popular of this is called belief propagation right, wherein you have some kind of an incremental way of computing these  $\tau$  factors right, that you have by passing what is known as messages between the nodes.

Okay, and the nice thing about belief propagation is it allows you to reuse a lot of the computation that you have already done for answering different marginal queries, okay any question so far.

So I can answer any marginal queries you can see that right, so I basically have to sum out all the other variables and have to find out now appropriate variable elimination ordering right, and then do it.

So the trick here is finding the right ordering, so I gave you the right ordering right, but in arbitrary graph finding the ordering is actually NP-hard right, finding the right order is NP-hard so you just have to do the best that you can and tree it is easy you can immediately see in trees you can go from leaves to the root but in an arbitrarily structured graph right it is hard to find out what is the right ordering, okay.

(Refer Slide Time: 09:01)



Other Queries,

$$P(J=1 | I=1, H=0)$$

$$P(J=0 | I=1, H=0)$$

We can compute it as,

$$P(J=1 | I=1, H=0) = \frac{P(J=1, I=1, H=0)}{\sum_j P(J=j, I=1, H=0)}$$

Great, so what about queries like, what is it probably that I let us say get a job, given that I am intelligent, but I am not happy. Right, and also what is the probability that I do not

get a job. So this is essentially a conditional marginal line right, so I condition on some variables, and I want you, I want to know the marginal right. So we know that everybody here is intelligent so once we figure this out if  $J=1$  is lower than  $J=0$ .

What should you do, be happy somebody said that yeah, so be happy yeah that is it, that will actually put you in we do not know we have to evaluate the marginal for that right, that does not guarantee your job but being happy at least well leaves you happy right.

No, that is one of the things for getting a job right, you want to be happy, so if you choose to be happy already it becomes irrelevant, anyway so what do you do for this.

Right, so this is essentially I can just compute this as.

Right, it is essentially ratio of two marginal's, but this is one marginal, so this is another marginal what is this a marginal over, it is marginal over  $I=1, H=$  the marginal probability of  $I=1, H=0$  right, so this is a marginal probability of  $J=1, I=1, H=0$ .

So essentially I will have to eliminate all the other variables I will be left with one table, and from that table, I can read this value right, so if I eliminate all the other variables I will be left with one table that has  $J, I$  and  $H$  as the entries in it right, and I can just read off the entry corresponding to  $J=1, I=1, H=0$  right, and this one again is another marginal so once I know how to compute marginal's I can also answer questions about conditionals.

So last thing then we will stop there will be the end of graphical models for now right. Suppose I am not interested in marginal distributions right, so this is another inference query I mentioned this in the last class right, the second kind of inference queries we would be interested in our MAP queries right, so why would we be interested in MAP queries?

In fact, many of the classification and other things we will be talking about we are only interested in MAP queries quite often right, so I will give you some image that I want you to label the image, and I am interested only in the MAP estimate of the label right, I want you to give me a label I do

not want the distribution over the labels I want you to give me a single label so in which case I need the MAP estimate.

So which label is the most probable according to the posterior right, so that is essentially what I need so for finding the MAP estimates so what do we do in this case. So what I am going to do is once I decide this kind of an ordering right, I am going to replace the sum here with a max.

I am going to replace the sum with a max. Right, so I am going to say look this is the max over C right, of something right so these probabilities I will compute then what you do is, when I do this max all over and I finish the computation right, I will get some probability right, the probability is the MAP probability, the probability of the most probable point.

How do I recover the most probable point at this computation is only for a probability right, so when I say do a max over C what does it mean for every value of D you will be entering one value right, you have a  $\tau_1(D)$  right, so  $\tau_1(D) = 0$  will be that probability right, for which is maximal you look at. Essentially you will have when you finish this computation right, so you will have some factor that is called a  $\tau'(C,D)$  when you finish taking the product will have some  $\tau'(C,D)$  right.

So  $\tau_1(D)$  will essentially be something like this  $\tau_1(D) = 0$  will be max of  $\tau'$  of,  $\max(\tau'(0,0), \tau'(1,0))$  right, this is what I mean by taking the max, so my  $\tau_1(D)$  will be the max of okay when  $D=0$  and  $C=0$ ,  $D=0$   $C=1$  now these two entries whichever is the largest I will put that as  $\tau_1(D)=0$  likewise for  $\tau_1(D)=1$  I will take  $\tau'(0,1)$  and  $1,1$ ) whichever is larger I will put that there, okay.

Now I will be eliminating D here right, so for each value of  $\tau_2(G, I)$  right, I will figure out which is maximum across  $D=0, D=1$  right, I will put that in my  $\tau_2$  entry, so likewise I keep going until I finally get a product right, and now how will I recover the actual point now I have found out which is the most, what is the probability of the most probable configuration, how do I find out what is the most probable configuration.

I keep track of which one gave me the max right, so I keep track okay here for  $\tau_1(D) = 0$  did I get the max from  $C=0$  or did I get the max from  $C=1$  right when I had  $D=1$ , did I get the maximum  $C=0$  or did I get the maximum  $C=1$ , so I keep track of in every stage I keep track of which entry gave me the max right, and then once i finish the computation I just go back and read out the max entry so that essentially gives me the MAP the probability of the most probable point and also this most probable point, right?

So if you have a tie, then you can choose 1, so it will give me at least one of the most probable points okay, so that is essentially how you do MAP estimates and yeah, so it is rather bad because it is exponential in the largest factor, but it is useful for small graphical models, because most of the other methods have significant overhead in setting up the entire process, right. Suppose you want to do belief propagation you will have to set up the data structures corresponding to the messages and it is a little bit of overhead terms of computing, right.

If you have a very small graph like the earthquake graph okay, the earthquake graph, you can graph you can make an inference by. Inspection right, just look at the graph and make inference you do not have to even do any computation right, so some slightly larger graphs like this right, so where you have to actually do some computation you can do variable elimination it is fine. But when you start talking about running it on images right, so we told, I told you right we have like a lattice like structure one node for each pixel in the image, and then I want to run this on a 256x256 pixel image, right.

Then you really need some help right, and then such cases variable elimination is not the right thing to do right, because the first case the treewidth can be large right, for that and so there are more efficient ways of doing it and even belief propagation right, in this kind of directed acyclic graphs right belief propagation is actually an exact algorithm right, even though it is pretty efficient it is an exact algorithm so can still be time-consuming. So people look at answering queries in an approximate fashion, so I will not be able to tell you what the exact probability is, I will not be able to compute the exact that same MAP probability right, but I might be able to give you the MAP is the actual point that has the highest probability.

But if you ask me what the highest probability I might not be able to give you that accurately is, so those kinds of approximations we are willing to take so that we can make inference efficiently.

**IIT Madras Production**

Funded by

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

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

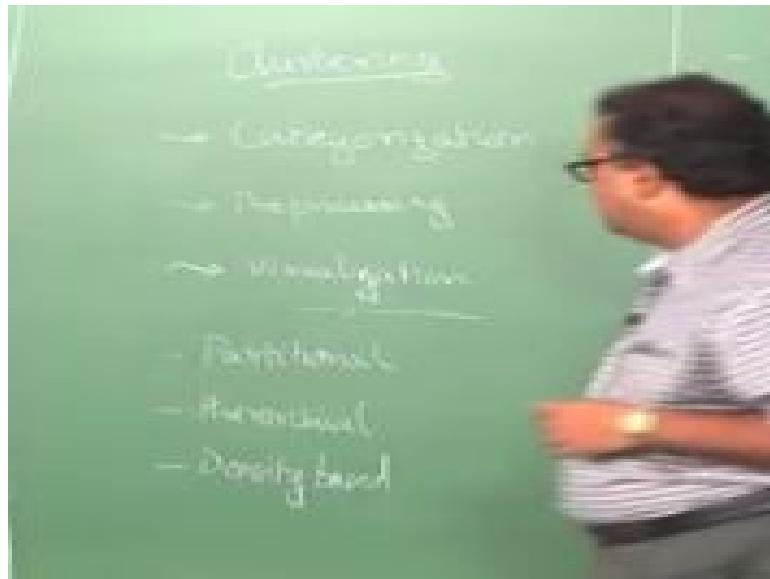
**NPTEL ONLINE CERTIFICATION COURSE**

**Introduction to Machine Learning**

**Lecture-70  
Partitional Clustering**

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

(Refer Slide Time: 00:17)



Clustering so far you know what clustering is right, so we did in the very first class we looked at clustering, so all you know what clustering is and essentially is the idea is to group together data points that are similar right, so what you are essentially trying to do is find a partition the simplest clustering problem is stated as follows you want to find a partition of the data points such that the similarity between the points are belong to the same cluster is maximized right and the similarity between points that belong to different cluster is minimized right.

So that essentially the thing and so if I give you a set of  $n$  data points right and I ask you to partition into  $k$  clusters, right how many clusters are how many different clustering are possible so when I say a clustering it is a set of  $k$  cluster okay so how many clustering's are possible if I give you  $n$  data points and  $k$  clusters and before condition that none of the clusters should be huge empty number right, so nice question to ask in exams but it is a huge number, huge number of clusters right.

So it is just impossible for you to exhaustively search through all the possible clustering's and then come up with the one that is best right, so inherently clustering is all clustering algorithms we will look at are all some form of an approximation or the other right to, to the actual base solution, right in fact some statistician so consider clustering a ill-defined problem right and do not dive into solving it you know, okay that is a ill post problem I will not solve it right so thing is it is a real problem.

People do not so all kinds of a place where we look at clustering so there are two main things that we do it clustering right, so the first one is as a machine learning as a task as a data mining task in itself right, I am interested in producing clusters right so I am interested in producing clusters right so this is some kind of a what we will call some kind of categorization I want to take this data point data is set of data given to me and then want to categorize them into different groups right in it of itself.

I am interested in doing clustering right clustering is also very valuable as a pre-processing tool right, so why would I want to do clustering as a pre-processing to so I can take a very, very large data set if I have a cheap way of clustering it right I can cluster it reduce it to a few data points right I can take a say 10m no dataset.

Like 10m for 10m items and then I can say I am going to cluster it into 10,000 clusters right, pretty large let we take a long time to do it but then if it has only say 10, 000 representatives of

this 10m data points so I want that to sample 10, 000 data points from this 10m but I wanted to it in such a way that they are as representative of the data is possible so what I do is I use clustering right and so 10, 000 is a large number of clusters right so each of those clusters is going to have few 1000 points right.

Not very large right so then I can just go and pick out one the representative for each one of these as suppose to sampling directly from a 10m node space, okay so that gives me some kind of leave here right so pre-processing, the other place where we want to use clustering right and any other things we can think of our clustering is useful, exactly right so for visualization again cluster is something there is very useful right instead of just looking at a large table of data something like that if I looked at it pictorially and I show you that okay here are one set of data points at belonging to one group here is another set of data points and stuff like that then it makes it lot easier for you to understand the structure of the data that you are looking at right.

So clustering allows you to visualize the data and understand any kind of special structure or structure in the feature space right when you say special structure does not mean that it is actually 2D space right I mean structure in the feature space. So is there any kind of structure in the feature space that you are able to understand that right. So it is a very valuable visualization tool and also very valuable for you to understand something about your data right.

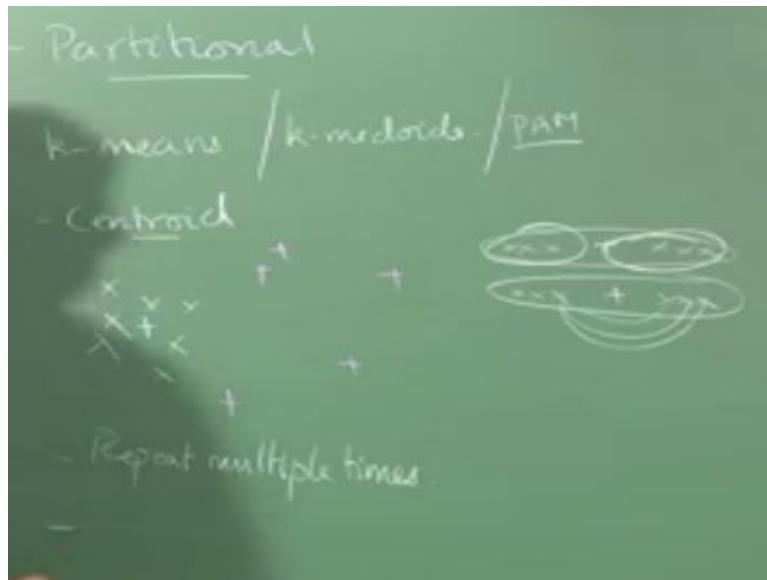
So in fact when I talk about categorization right so in and of itself it can be the problem that you are solving or it could be something that you do as a people are seeing step before you go and actually solve the problem, so I have been interested in finding out how many classes are there in my data so I do not know I am not be given class labels a priory okay I do not give class labels a priory but can I tease out class labels right can I say that the data contains seems to be coming from three different distributions right and then I am going to say each of those is a separate class condition distribution, and I can assign a class label to all of those distribution rights.

So it is like I am going to look at the data try to understand that and then see okay these are people who are likely to finish a course right these are people who listen to all the lectures but not write the exam I do not know what to be so but we not look at the data at here okay. So this kinds of things which you could do that are what I am saying, clustering is just not a one short okay give me the data here, or the clusters in you are done with it in fact quite often it is not even called clustering it is called cluster analysis because you stop with the clustering alone you actually have to go and figure out what the clustering is telling you.

Classification in some sense is much easier in some, so you are given something you are basically returning the labels and more often than not you are done right but clustering it is usually a step on the way to something else okay right. So there are many ways in which you can do clustering so the most popular of this okay are called partitional approaches right so partitional clustering and I tell you what is the others are I get to them on okay.

So the first thing is partitional clustering then hierarchical and then density-based, and they are not really disjoint classifications right, but typically methods get the index to the end of this three and then there are many, many other smaller things raise, so somebody wants to get a paper return the over I propose something that is neither partitional large hierarchical and then they will propose a new algorithm and thing like that, so they are to sell their papers and not necessarily individual classification right. However, these are the three main classifications right and so look at each in turn right.

(Refer Slide Time: 08:55)



Partitional clustering things K means, everyone knows what K means as of right, so why do we call them partitional clustering? Clustering is partitioning the data right so why I am calling this as partitional clustering. So I am essentially these are methods which search through the partitions directly right the final partitions that I want right they search through the partitions directly, so that is why they are partition clustering methods okay.

So suppose to hierarchical clustering methods right which does not says through the space of partitions on the entire data set right they first try to do it into two groups right they do not search through all the partition suppose I am interested in K things right they do not do all the K they do not do the K clusters right they can start off with two clusters and then split the two clusters into four and so on and so forth okay.

They can go down into there, they do search directly in the space of k partitions right while k means exactly does that right this when what did you do in k means I am sorry so what you are doing in k means you start off with k guesses for the centres of the clusters right so there are few things that we need to know so the first concept here is the centroid so what is the centroid well this et center of cluster right so essentially you take the average along each coordinates right to suppose I have 10 points it belongs to the cluster it is in a five-dimensional space.

For each dimension, I take all the ten coordinates take the average right so. Finally I will end up with the single data point, so that is the centroid right suppose I have some set of data points like this right so the centroid could be somewhere there now that I am exactly computing the centroid, but it will be somewhere there so right sure so what is a how do you find the perfect solution, yeah so there are many ways of cleverly initializing this right we talk about vanilla k means right so the many ways in which you can look at a clever starting point right.

So the other things which people do they what they do is they start off with 10% of sample of data and then they repeatedly run clustering on that and figure out which are the good centroids for the 10% of data and then that use has an initialization for clustering the whole data right. So the idea there being rather than repetitions of very small when you are doing the 10% of sample of whole data so this is the one way you do it the other way of doing it is to do this kind of an initialization where you try to move to the further corners.

But then that has it one problem right so essentially it will make it more sensitive to outliers right because you are trying to put your centroid the beginning centroid at the edges of your space right so it will make it a more sensitive outline, so those there are issues. Yeah that is k means another's so I do not think he meant that I think he meant something else we will come to that right so might there are I mean many heuristics again so I think the current came in variation champ is I think k – means ++ of some thigh right.

So that gives you a very good initialization, and then your drawing a clustering from there and works well right but k – means is an approximate thing so when you look at EM right so you actually see a more well-founded derivation for k – means so right now I am just going to introduce it to you as a heuristic right but later on when you look at EM right the canonical EM problem that you solve right first canonical EM problem that you look at will be K – means.

The variant of k usually means we call Gaussian mixture modules but it is essentially like k means okay so what you do with k – means is that you first pick centroids for all your clusters right a randomly right of course I have k clusters then I will just choice k centroids at random right that could even be like that right choice k centroids at random and then I will have my data points so what I do is I assign each data point to the centroid that is closes to it right I say in each data point the centroids that is closes to it okay now I forget the old centroids.

Now I have k groups of data point's right and for each of those groups I try to find the new centroid okay this is the actual centroids the first once I started over of centroids where really not centroids but we keep we still call the centroids anyway so in the just to keep the terminology uniform right now I recomputed the centroids then I go back and assign each data point to the centroid that is closes to it.

Yeah whatever distance measure you are using right so you are in some RP space right remember our data points come from some p dimensional space so in that space whatever is nearest you will you could use Euclidean you could use whatever distance measure you want it choose the appropriate distance measure right so k – means or any of these distance-based computations that you do right works best okay any of the distance based computation you do works best if everything is real value right not even integral right.

So everything is real value that works best if you are going to categorical attributes you will have to think of a different distance measure that you will have to define and I do not know if may I can talk about 1 or 2 things that people use for categorical things but the most popular by far is using some kind of Jaccard similarity, right so I have lots of categorical values that the thing can take and then I say okay how many of those it actually is similar on right so how many dimension I have may actually agreeing with the other one on so if I have the same value in that categorical dimension then I will say one if I do not and the I will say 0 and the I will try to find how many I agree on right.

And that could be one measure but then defining a centroid there is little tricky right so categorical variable may centroid might come with a value of you know red 0.5 where I mean right 0.5 times red + 0.5 times green right so what would that mean no do not add up the colors please I mean I said brown or something I do not know what red and green would up with but do not know that does make sense right so we have to we, very vary about using k – means when you have categorical attributes.

Yeah so what I mean by probability vector like one of n kind of encoding right 1 hot encoding right yeah you could do that but then really what would be the mean centroid value for that 1 hot encoding, yeah that is a interpretation issue so you could still try k – means I mean you do not really have to interpret what the centroid means unless your returning the centroid as a representative point right if I am returning the centroid as a respective point then you get into problems.

So just going to do clustering on it you can go ahead and do crusting on it you do not have to really interpret the value of the centroid right so there are ways of handing this when you have categorical attributes right so k – means is the simplest way then there is something called k medoids right which kind of gets around this whole issue of having to generate an artificial centroid right so instead of centroid it essentially uses a the equivalent of median right so the mean cannot be not actually be a data point but the median is always a data point right.

So likewise medoid I always a data point so it gets around this interpretation issue so computes the centroid and takes the data point closest to the centroid as the representative so the medoid right and the you also have other kinds of things which is called partitioning around medoids right so we do partitioning around medoids you work with data points as the representative of the clusters right and you do not ever generate an artificial point you always choose a another data point as a centroid I will describe that more right.

So those in such cases you do not have to worry about this you know meaningless attribute values being generated but you still need a distance measure so you will have to come up with some kind of distance measure which categorical attribute so if you use 1 out n or 1 hot encoding or anything else you still use some kind of a Jaccard similarity right so then do that okay.

Right so going back let us finish up the simple k means algorithm right so what you do with k means is so once you have done the assignment to the centroids if you forget the centroids estimate new centroids and keep repeating this until you no longer make any changes right sorry done?

No, no labels is necessary, there are no labels, there are unstable is learning from right. So if you have labels and other things, I mean depending on what your application is you might want to look at labels right, that is what the scameans is concerned, the question is the following. And you start with initial guess for the centroids right, I assign data points to the closest centroid right, recompute the centroids, then reassign the data points the closest centroids, I keep doing this until the centroids no longer change.

The question is have you done? Consider a, yeah, okay. Al right, how many clusters you want me to put them into? Three clusters I am done, you are getting close. So the probability of recycling is really small yeah, yeah whether other things which I can show is slightly more dramatic right, okay. These are my two sets of data points. Obviously, they are in two clusters, aren't they?

These are the two initial centroids I start off with okay. So what are the two clusters I will get? Okay, and I compute the centroids for those, where will I end up with, and I reassign the data points to the centroids, I will end up with the same thing right so because of my bad initial choice for centroids, I do end up at a point where the centroids do not change anymore, but it is a really, really bad clusters.