

Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology – Kharagpur

Lecture - 40
Regression Analysis (Part -1)

Hello guys; so, in continuation of our discussion relation analysis today is lecture we will talk of regression analysis.

(Refer Slide Time: 00:36)



Last 2 lecture what we have seen between 2 variables or more than 2 variables if there is any relationship if there is an association and if there is an association that how strong is the association how strongly the associated we have considered that in my last 2 lectures. So, now, we will consider now in this lecture will consider more on relation, more on relation means, what I want to say like.

Suppose like if I want to if I know one value I know that there is a relation between 2 variable now, if I know 1 variable can I predict some other variable from that variable. Let me give you an example like if I know that expenses have made on advertising a product from that can I estimate the revenue of the product or if I know the expenses I have done on a producer has done on promoting a movie.

Can he predict the box office performance from the expenses done on the movie or rather another example maybe if I can I want to see if I know the house is in a particular area knowing the area can I predict the price of the house our a particular size of course. So, this if

I want to know this sort of from a relation if I want to know this sort of thing, this is basically we can do this using regression analysis as I told you when discussing correlation is both correlation and regression these are different objective and correlation analysis I tried to find out whether there is any relation between these 2 variables and how strongly they are related what is the strength of the relation.

So, now, in regression analysis basically, if I know one variable from that variable can I predict the value of the other variable. So, this is basically the regression analysis. So, we in this lecture it will be the introductory class for regression analysis and the next coming 2, 3 lectures will complete regression analysis.

(Refer Slide Time: 02:35)



So, now, before talking of regression analysis here I would like to bring to your notice about learning strategies learning there are 2 types of learning concept one is statistical learning another is machine learning. So, you might think why suddenly I brought machine learning because we are talking about statistical learning, statistical learning what is this we have already seen we have done lots of thing methods and the statistical learning.

Basically we try to learn different population parameters and even the correlations also we try to learn from a population remember we have done population parameter rule we are trying to find it from the sample parameter R . So, these are all these are called statistical learning. Now, this something or some other type of learning that is called machine learning I will come to that later.

(Refer Slide Time: 03:25)

Statistical Learning

Usually assumes certain properties of the population from which we draw samples:

- Observation come from a normal population.
- Sample size is small.
- Population parameters like mean, variance, etc. are hold good.
- Requires measurement equivalent to interval scaled data.

First, let us see when we talk of statistical learning always we see there are some assumption I am repeating those assumptions and whatever assumption their population has to be normal which means parent population has to be normal sample size is small, we do not need to consider a huge sample size population parameters like mean variance are whole growth and of course it is applicable for interval scaled data that is when we talk of statistical learning.

(Refer Slide Time: 03:48)

Machine Learning

Input: x_1, x_2, x_3 & Weights: $\theta_1, \theta_2, \theta_3$ | Output: Happy or Sad

$y = f(x) = ax^2 + bx + c$

- Does not under any assumption
- Works well with high volume high dimensional data

Now, when we talk of machine learning machine learning we do not need any such assumption that is what is machine learning? Basically, first let us come to the see the figure here suppose I want to find out the emotion of this people. So what my interest is what to say, predict the emotion of a people by looking at the face of a person, my interest is my objective is to predict the emotion of a people by looking at the face of a person that is what I already given.

The example by giving the amount of money spend on promotion of a movie can I predict my box office performance knowing. What I want to predict rather similarly here, knowing the facial expression can I predict the emotion of the person. So basically, what I predicting means what I need to do is that basically I need to find out the function where $y = fx$ where x is my facial expression, y is my emotion.

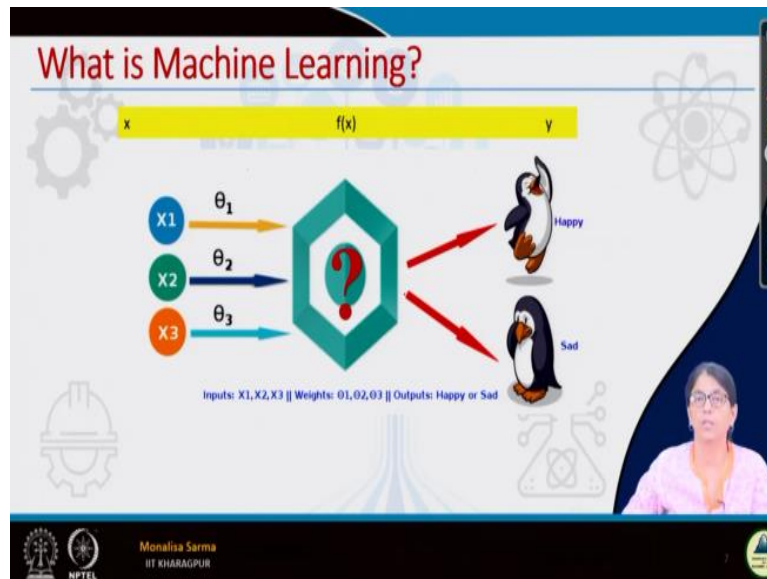
So it has nothing just to find this for me is just finding the function $y = fx$, same thing we do in the statistical learning also, let me just try to find out why so as to effect so, in machine learning. What we do that we try to find out the function $y = fx$ where maybe $f x$ is $ax^2 + bx + c$ maybe our fx is $ax^2 + bx + c$. So, basically here if we our input will be this $X_1 X_2 X_3$ our attributes, different attributes of this person, maybe physical attributes maybe the different pixel information of the person.

Or whatever form May be so, $X_1 X_2 X_3$ are the attribute and then this $\theta_1 \theta_2 \theta_3$ are the different parameters which machine we will find out basically machine learning will find our model which fits this facial expression to the emotion. So, this model maybe $y = fx$ so machine will find a function it will find a parameter once the function is known once the parameters are known.

Then definitely we given a input of a person with a particular facial attribute will be able to tell whether a person is happy whatever the person is sad or whatever it is. So, one advantage of this machine learning is that it does not come under any assumption it works well for high volume and high dimensional data. So, when we have high volume data as well as high dimensional data, so, we were considering till now only 2 dimensional or 3 dimensional and when we have high dimensional data, there can be different categories of the data.

So and that case is this machine learning work is very well. Important point is this learning strategy needs a very large sample data that is the concept we need a very large sample data. So find a model which fits the x to y this $y = fx$.

(Refer Slide Time: 06:34)



So basically to summarize this given x we have to find out the $f(x)$ we have to find out all the parameters. So, now the thing is that why did I talk of learning strategies why did it talk up statistical learning ready to talk on machine learning So, as I have already mentioned this lecture will be on regression analysis. So, regression analysis is very much a statistical analysis tool, we do regression analysis from population by taking a sample it is very much a statistical learning concept.

But at the same time regression and when we learn machine learning, the machine learning usually starts I can very conveniently say machine learning starts with regression analysis. So, regression analysis is like, it is statistical it is a part of statistical learning also is a part of machine learning also. So, that is the reason why I bought this here. So, now, as you could see from the constraint what I have given for both the 2 types.

Regression analysis is possible when I am confined with a very low dimensional data. When there is a high dimensional data definitely statistical learning will be definitely will have to go for machine learning. So, now, we will focus on statistical learning only we will not go for high dimensional data. So, we can very well do it by statistical learning only, as the things are seen just that we will have to if it is high dimensional data instead of statistical learning.

It was very difficult very computationally intensive, so, definitely will have to go for machine learning.

(Refer Slide Time: 08:11)

Relationship Analysis

Example: Wage Data

A large data regarding the wages for a group of employees from the eastern region of India is given.

Attributes:

- Employee's age
- Employee's wage
- Employee's education level
- Calendar year
- ...

Monalisa Sarma
IIT KHARAGPUR

So, now coming to a main focus of this relation of this lecture that is relation analysis, So take an example suppose there is an example we have large data regarding the wages of a group of employees of a particular regions eastern region of India and we have a large data wages of the different peoples are given and this data set as a different attributes. So, what are the different attributes employee's age, employee's wage, employee's education level.

Calendar year by calendar year means at which each year how much salary he got 2022 how much he got 2021, 20, 19, 18 likewise. So, these are the different attributes in this data what we have with these are the attributes of the data.

(Refer Slide Time: 08:50)

Relationship Analysis

Example: Wage Data

Relation 1: Employee's age and wage

How wages vary with ages?

Monalisa Sarma
IIT KHARAGPUR

So, now we may be interested in finding out so, how wages vary with ages. Given there are 2 things how wages vary between ages maybe if I am interested in finding out is there any correlation between is any association between wages and age it is a correlation analysis our

if I am interested in finding out knowing the age can I predict the wage then it becomes a regression analysis.

(Refer Slide Time: 09:21)

A presentation slide titled "Relationship Analysis" with a blue header. Below the title, there is a green box labeled "Example: Wage Data". Underneath, an orange box contains "Relation 2:". A blue box below that contains the text "Calendar year and wage: How wages vary with time?". To the right of the text is an image of hands holding blocks that spell "SALARY". The slide also features a small video inset of a woman in the bottom right corner and logos for IIT Kharagpur and NPTEL at the bottom.

The second thing which may be of my interest is how wages vary with time, we can think in terms of regression analysis is given the wage value given a time can I predict the wage.

(Refer Slide Time: 09:35)

A presentation slide titled "Relationship Analysis" with a blue header. Below the title, there is a green box labeled "Example: Wage Data". Underneath, an orange box contains "Relation 3:". A blue box below that contains the text "Employee's age and education: Whether wages are anyway related with employees' education levels?". To the right of the text is an image of hands holding blocks that spell "SALARY". The slide also features a small video inset of a woman in the bottom right corner and logos for IIT Kharagpur and NPTEL at the bottom.

So, whether wages are anyway related with employee's education level.


(Refer Slide Time: 09:40)

Relationship Analysis

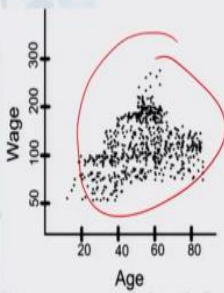
Example: Wage Data


Case 1: Wage versus Age


From the data set, a graphical representation can be drawn:




How wages vary with ages?







Monalisa Sarma
IIT KHARAGPUR



These are different on different relations which I may be interested in finding out. Now suppose I am interested in finding out how wages vary with ages. So given is can I predict wage so first what I do is from the data set, I have done a graphical representation, now see it is graphical representation. This graphical representation is very confusing like, so, is it linear is it monotonic.

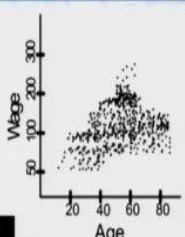
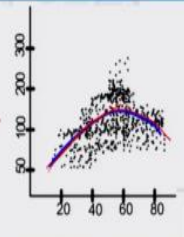
Or what it is nonlinear or what this is what can sometimes come from the graphical it is very difficult to find out what relation does it hold. So, if in the; simplest case when we try to find out the relation between 2 variables. The simplest case is that we try to graph a straight line that is called a linear relationship that is also done if we from one if we try to predict the other than it is a linear regression.

(Refer Slide Time: 10:37)

Relationship Analysis


Case 1: Wage versus Age

From the data set, a graphical representation can be drawn:



→


Interpretation

On the average, wage increases with age until about 60 years of age, at which point it begins to decline.

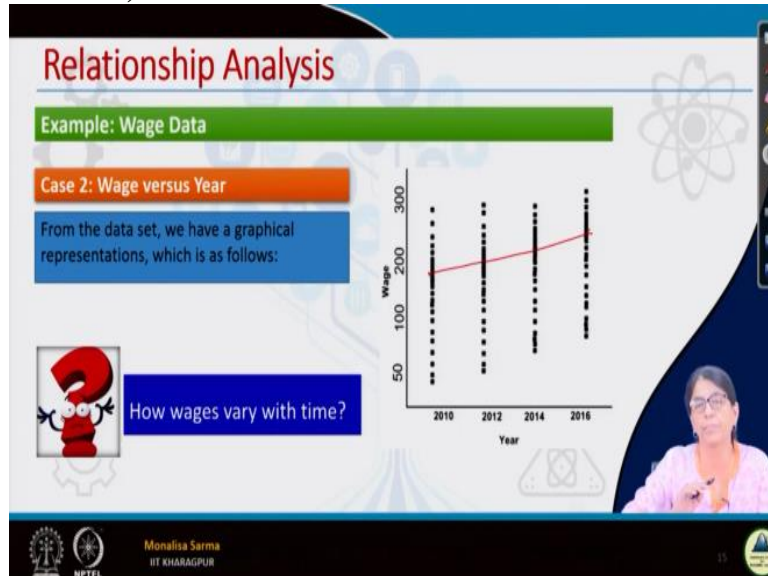


Monalisa Sarma
IIT KHARAGPUR



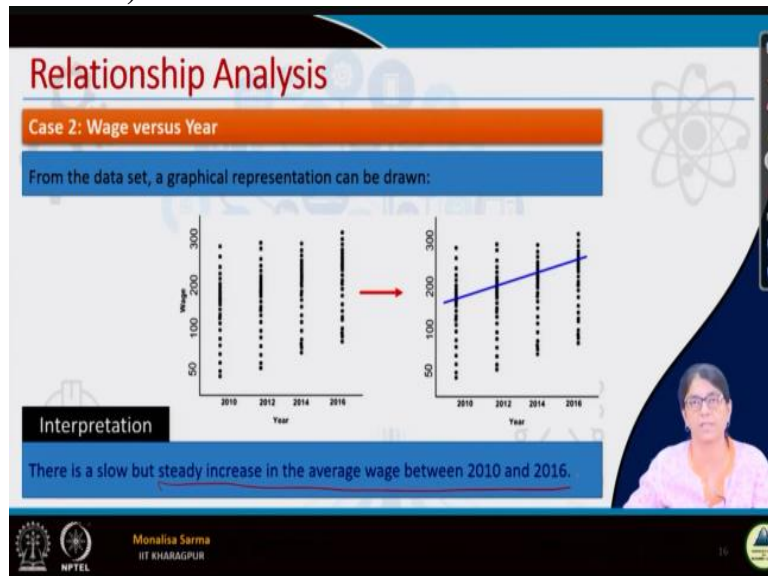
So, now, in this case this does not fit a straight line it basically you can see it is fitting a curve line which at first it is increases then again it is going down then that means it is nonlinear, we cannot fit a straight line here.

(Refer Slide Time: 10:52)



Similarly, wage verses year so, if you see wage versus time, so, if we plot it this way, so, what we get. So, maybe this if I consider this portion, so, this is there is an increasing trend, it is very much a linear relationship it is very much showing linearly it is increasing.

(Refer Slide Time: 11:11)



There is a slow but steady increase in the average wage between 2010 and 16.

(Refer Slide Time: 11:19)

Relationship Analysis

Example: Wage Data

Case 3: Wage versus Education

From the data set, a graphical representation can be drawn:

How Wages vary with Education level?

Monalisa Sarma
IIT KHARAGPUR

Now, again does wage vary with education level, if I just plotted it this way for different education level, if I plotted the wages for different people, I am finding it hard to find out how whether it is depicting as increasing relation or decreasing definitely not decreasing it is increasing relationally but if I find it hard to predict this, I may change this figure and I may draw a boxplot remember a boxplot how we have done.

(Refer Slide Time: 11:45)

Relationship Analysis

Case 3: Wage versus Education

From the data set, a graphical representation can be drawn:

Interpretation

On the average, wage increases with the level of education.

Monalisa Sarma
IIT KHARAGPUR

So, if I am drawing a box plot and this in the boxplot this middle one is the mean. So, if I can see a mean increase in a steady increase in mean as the education level increases.

(Refer Slide Time: 11:57)

So, likewise, from a relationship, we can get different information. So whether wages is any association or both year and education level with the wages it has an association with year and education level that means what? I am trying to find out the association is there any association between these 2 that means I look correlation analysis or I may be given an employee's wage can we predict his age.

That is a different thing I have not done till now, I have done this, this one I have done but this I have not done this sort of things if I want to answer so basically, I will have to do the relation analysis is to find a relationship between the variables, if I can find the relationship between the variables and given one I can predict the other understood. So I will have to do relation analysis to find a relationship in the relationship is given.

That means what is the relationship is given means the function is given $y = fx$ the function is given in then I can find out given x , I can find out a y , here what is y what is x y is the independent variable, x is the dependent variable, this independent variable sorry, x is the independent variable independent variable is also called regression, x is the independent variable right $y = fx$, x can change independently.

So, x is the independent variable that is also called regression. And y is the dependent variable that is also called response, please remember y is called the response variable or dependent variable, x is called the independent variable or regression variable in $y = fx$.

(Refer Slide Time: 13:33)

Question

Suppose, there are countably infinite points in the XY plane. We need a huge memory to store all such points.

Question:
Is there any way out to store this information with a least amount of memory? Say, with two values only??

Monalisa Sarma
IIT KHARAGPUR

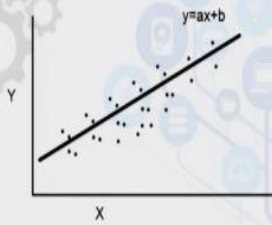
So, to how to find out the; relational analysis, how to do the relation analysis is to find the relationship let us see a simple example. Suppose here, there are countable infinite points in the XY plane, we need to huge memory to store all such points. Suppose we have 5000 points in an xy plane, which 2 attributes x and y , we have total 5000 points. So, 5000 points means we will have to store 5000 x and y in a array we will have to have an array of size 5000 x 2 so this is a huge size.

Now I do not have such much memory, so much memory, is there any way out to store this information with the least amount of memory say with 2 values only. So can I do not have such high huge memory 5000, 10,000 or maybe 50,000 data points, I do not have such huge memory to store this data. So I do not have so much huge memory to store this data. So, can I somehow store this data in a very less memory yes that is possible how this is possible.

If somehow I can find a relationship between these 2 points that means if somehow you can find a function that gives y from x , if I have the function then is the stored a function storing the function means just storing the parameters of the function.

(Refer Slide Time: 14:53)

Answer



Just decide the values of a and b.
(As if storing one point's data only!)

Note:
Here, tricks was to find a relationship among all the points.

Monalisa Sarma
IIT KHARAGPUR

22

Suppose this values what I have given suppose these values fit straight line. Of course, all the points will not fall in a straight line with a slight error, we will be there that we will be discussing later. What is this error how to deal with this we will be discussing later. So if we can somehow find it relationship, suppose here the relationship or model, whatever you call, it is a straight line.

So if I can find out the straight line, finding out a straight line means I have to find out the intercept, I have to find out a slope, so $y = ax + b$ means where b is the intercept and a is the slope. If I can find out this relationship, I can find out the intercept, I can find out the slope, then what happens I need only to store the intercept and a slope and I am done. Given any value of x.

I can find out the value of y, I do not need to store all the values of x and y whenever I need, I just give the value of x, I will get the value of y this is just one example of using the relational analysis to find the relationship.

(Refer Slide Time: 15:54)

Measures of Relationship

Univariate Population							
Temperature	20	30	21	18	23	45	52

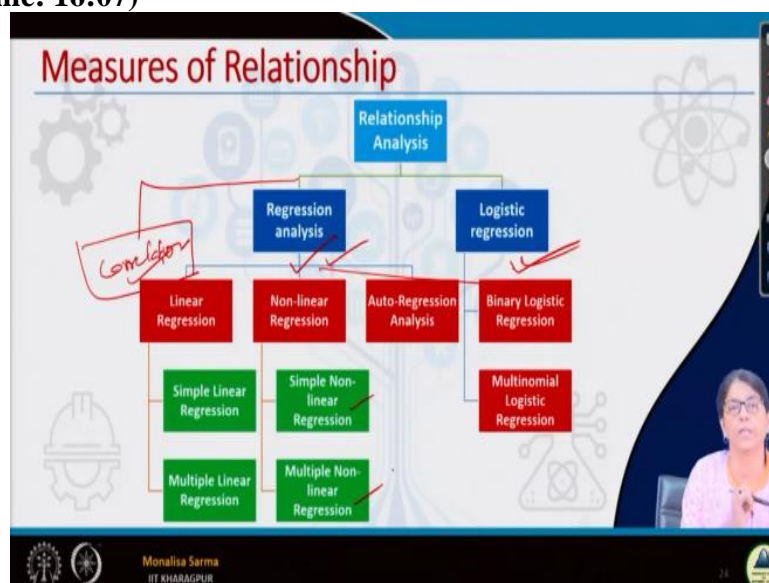
Bivariate Population							
Temperature	20	30	21	18	23	45	52
Pressure	1	1.5	1.05	0.96	1.2	2.5	2.8

Multivariate Population							
Temperature	20	30	21	18	23	45	52
Pressure	1	1.5	1.05	0.96	1.2	2.5	2.8
Volume	20	30	21	18	23	45	52

Monalisa Sarma
IIT KHARAGPUR

Now, when we talk of relationship as I have already mentioned before, we will talk often we will work on bivariate population or multivariate population we will never work on univariate population.

(Refer Slide Time: 16:07)



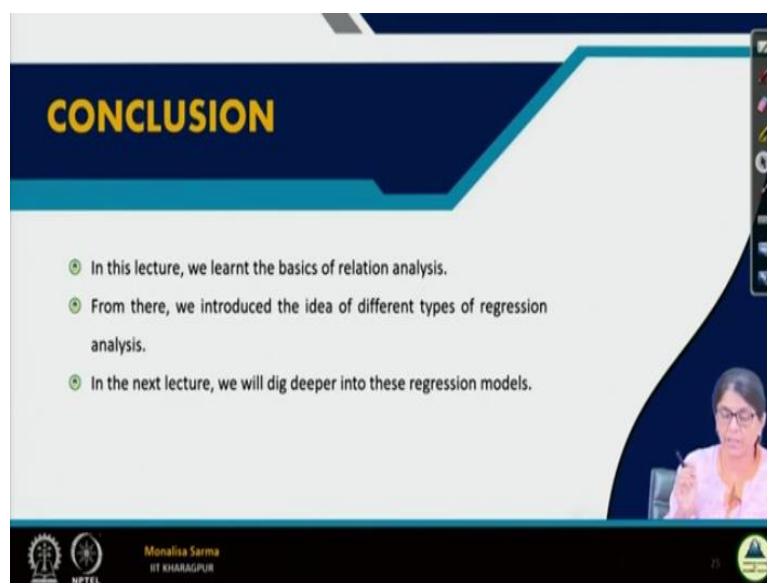
Now talking a relation analysis, last class I told relational assessments there are 2 parts, 2 ways of relational categories of relational analysis that is correlation analysis and regression analysis that is also do not get confused. Let me draw 1 more line here that means I have correlation analysis. So, there is basically regression analysis I am dividing into 2 different categories.

Because this is regression analysis and logistic regressions are 2 different types, where I will not discuss logistic regression now, I will discuss in after 2, 3 lectures maybe. So, relational analysis as I can talk regression analysis and logistic regression with the objective of

whatever objective I told of regression analysis with that objective I can have 2 different regression and logistic regression because correlation is totally different.

So, now, when we talk of regression analysis, there may be linear regression nonlinear regression and auto regression analysis, similarly, for logistic regression, there might be binary logistic regression, multinomial logistic regression. Again for linear regression or nonlinear regression, it might be simple linear, multiple linear regression simple nonlinear and multiple nonlinear equations. So, we will be discussing all this in the coming few lectures.

(Refer Slide Time: 17:34)



The image shows a presentation slide with a dark blue header containing the word "CONCLUSION" in yellow. Below the header, there are three bullet points, each preceded by a green circular icon with a white dot. The text of the bullet points is as follows:

- In this lecture, we learnt the basics of relation analysis.
- From there, we introduced the idea of different types of regression analysis.
- In the next lecture, we will dig deeper into these regression models.

In the bottom right corner of the slide, there is a small video inset showing a woman with glasses speaking. At the bottom of the slide, there are logos for NPTEL and IIT KHARAGPUR, along with the name "Monalisa Sarma" and "IIT KHARAGPUR".

So, basically in this lecture, we learned a basic of relational analysis how we can analysis the relations to find the relationships. From there we introduced the idea of different types of regression analysis. In the next lecture, we will dig deeper into the regression models.

(Refer Slide Time: 17:53)

REFERENCES

- ① The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.



The image shows a presentation slide titled 'REFERENCES'. It features a blue header with the title in yellow. Below the header, there is a list of references. The first reference is 'The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.' To the right of the text is a small image of the book cover, which is orange and black with the title and authors' names. At the bottom of the slide, there are logos for IIT Kharagpur and NPTEL, along with the name 'Monalisa Sarma' and the number '26'.

So this is the reference and thank you.