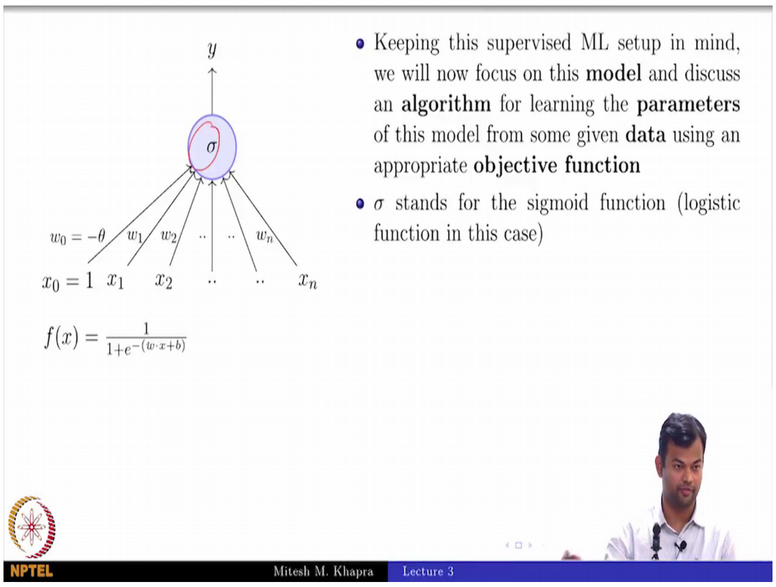
Deep Learning

Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 3.3
Learning Parameters: (Infeasible) guess work
Lecture – 03

And we will move on to the next module, where you will try to learn these parameters, and initially we will try to learn them by guesswork, and I will show that that is actually infeasible; that is why we need a more principled approach, ok.

(Refer Slide Time: 00:26)

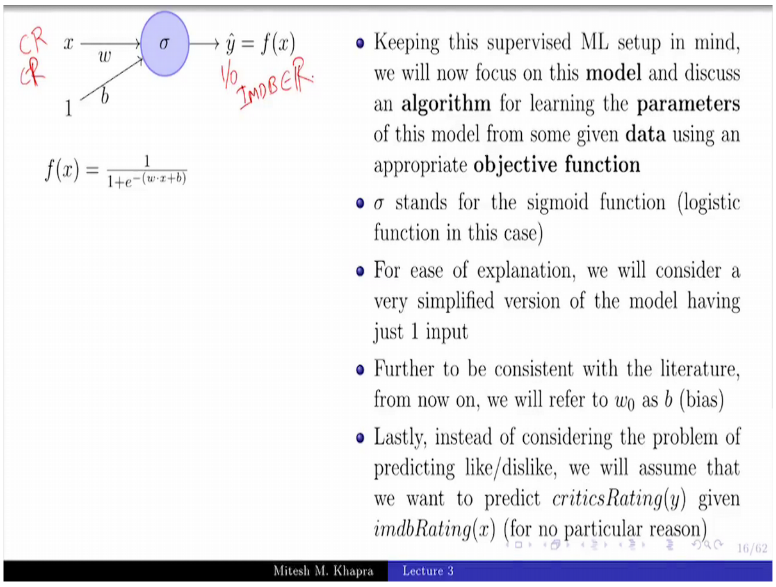


$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$

- Keeping this supervised ML setup in mind, we will now focus on this **model** and discuss an **algorithm** for learning the **parameters** of this model from some given **data** using an appropriate **objective function**
- σ stands for the sigmoid function (logistic function in this case)

So, we will keep the supervised machine learning setup in mind and now we will focus on this model, and discuss an algorithm for learning the parameters which are w and b, right? Given some data using a giving appropriate function objective function right. So, that is what we are going to focus on, ok.

(Refer Slide Time: 00:47)



$f(x) = \frac{1}{1+e^{-(w \cdot x+b)}}$

- Keeping this supervised ML setup in mind, we will now focus on this **model** and discuss an **algorithm** for learning the **parameters** of this model from some given **data** using an appropriate **objective function**
- σ stands for the sigmoid function (logistic function in this case)
- For ease of explanation, we will consider a very simplified version of the model having just 1 input
- Further to be consistent with the literature, from now on, we will refer to w_0 as b (bias)
- Lastly, instead of considering the problem of predicting like/dislike, we will assume that we want to predict $criticsRating(y)$ given $imdbRating(x)$ (for no particular reason)

16/62

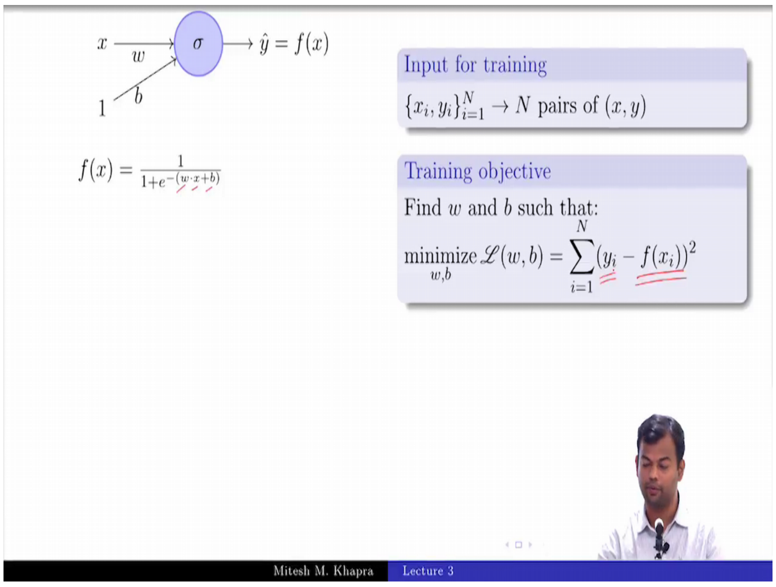
Mitesh M. Khapra Lecture 3

Now sigma here stands for the sigmoid function, the logistic function in this case when this sigma is actually the logistic function. And now I am going to simplify this further. So, that it helps us to do a better analysis, I am just going to consider the case; where I am just in one input and the bias, ok. And also following the normal terminology in the literature, this w naught from now on I am going to call it b. Because that is the normal convention b stands for bias, ok.

So, I have 2 parameters w and b, which I need to estimate ok, and this is my model for the movie example. And the other change which I am going to make is instead of deciding whether I like or dislike which is 1, 0; the setup that I am going to work with is that I am giving the critics rating, and I want to predict the IMDB rating, right. So, I am given a real value, and I also want to predict a real value.

For no particular reason this just makes life easier for me for explaining a few things, but the same thing or the same algorithm would also hold if you add a binary output right, and you will see that later on in the course, ok. So, here is a setup, clear? We just have 2 parameters w and b, and we are going to assume that y belongs to real numbers it is a IMDB rating and x also belongs to real number; it is a critics rating, ok.

(Refer Slide Time: 02:05)



$x \xrightarrow{w} \sigma \rightarrow \hat{y} = f(x)$, $1 \xrightarrow{b} \sigma$

$f(x) = \frac{1}{1+e^{-(w \cdot x+b)}}$

Input for training

$\{x_i, y_i\}_{i=1}^{N} \rightarrow N$ pairs of (x, y)

Training objective

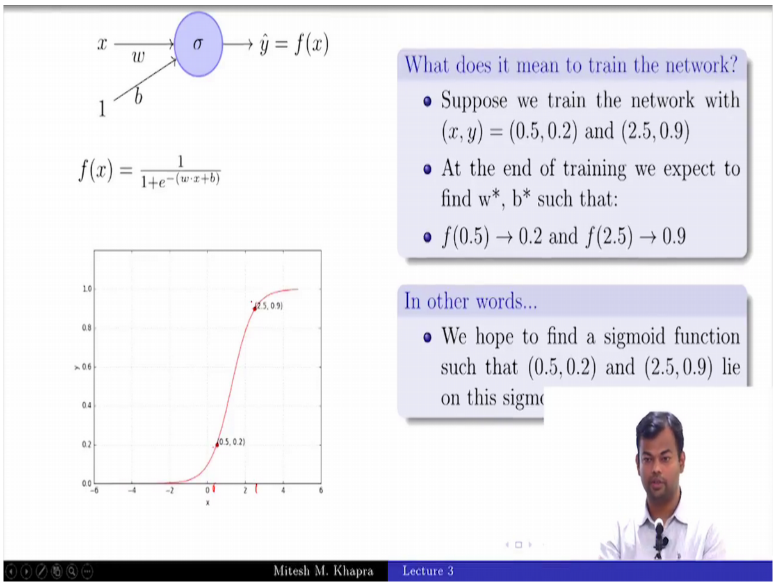
Find w and b such that:

$$\underset{w,b}{\text{minimize}}\ \mathscr{L}(w,b) = \sum_{i=1}^{N}(y_i - f(x_i))^2$$

Mitesh M. Khapra Lecture 3

Now, let us see what we are given as training is a set of points, we are given some n training pairs, and now we understand what this means; that means, for a lot of movie I am giving the critics rating and I am also given the true IMDB rating for them. Of course, in the 2 variable case this does not make much sense, but just bear with me, ok. And now the training objective is such that whatever my function predicts which is a function of w, x and b that should be very close to the true output that I know, this is the function that I want to optimize ok. Now let me ask you this.

(Refer Slide Time: 02:37)



$x \xrightarrow{w} \sigma \rightarrow \hat{y} = f(x)$, $1 \xrightarrow{b} \sigma$

$f(x) = \frac{1}{1+e^{-(w \cdot x+b)}}$

What does it mean to train the network?

- Suppose we train the network with $(x, y) = (0.5, 0.2)$ and $(2.5, 0.9)$
- At the end of training we expect to find w*, b* such that:
- $f(0.5) \rightarrow 0.2$ and $f(2.5) \rightarrow 0.9$

In other words...

- We hope to find a sigmoid function such that $(0.5, 0.2)$ and $(2.5, 0.9)$ lie on this sigm

Mitesh M. Khapra Lecture 3

I am trying to tell you that I am going to give you an algorithm for training this. Network suppose I have trained this, with 2 data points, 0.5 comma 0.2 and 2.5 comma 0.9, right. At the end of training, I will give you some values of w and b. Let us call them w star and b star; these are the final values of w which I have given w and b. What do you expect from these values?

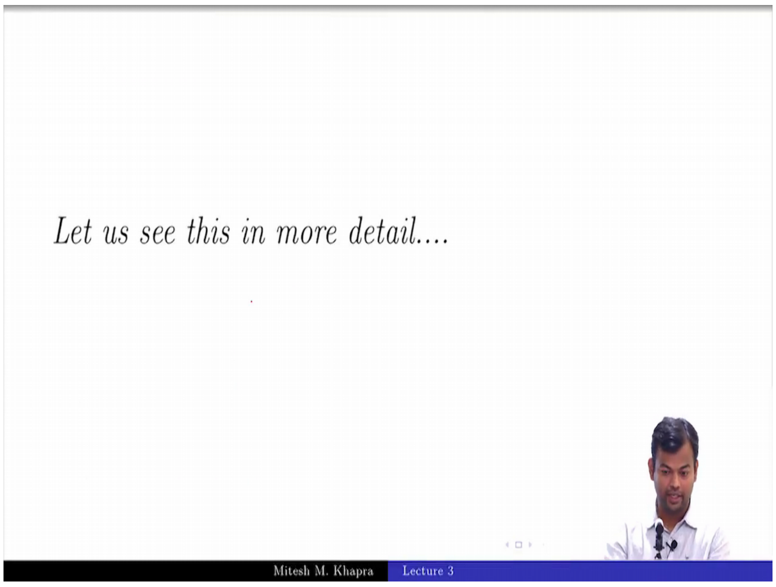
What do you expect at the end of training if I say, ok; now, the network has learned, what do you expect? You are still going to the test case I am just talking about the training still, ok. We expect such that what happens if I plug in at the end of training, if I plug in the value 0.5 here what should happen?

Student: (Refer Time: 03:21).

0.9 right? So, this is what you expect at the end of training, if you plug in the value 0.5 it should be very close to 0.2 the output and if you plug in the value 2.5, it should be very close to 0.9 right.

This is exactly what you expect and this is what training means ok, fine. In other words we hope to find a sigmoid function such that these 2 points lie on that function. Can you imagine a geometric picture for this? What would happen actually? How many if we can imagine it? Ok, how many of you get it now, right? This is what will happen right. So, you will get a sigmoid function such that these 2 points lie on that fair ok, and that exactly means that when I plug in this value, I will get this value and when I plug in this value I will get this value right. So, that is what it means, ok.

(Refer Slide Time: 04:09)

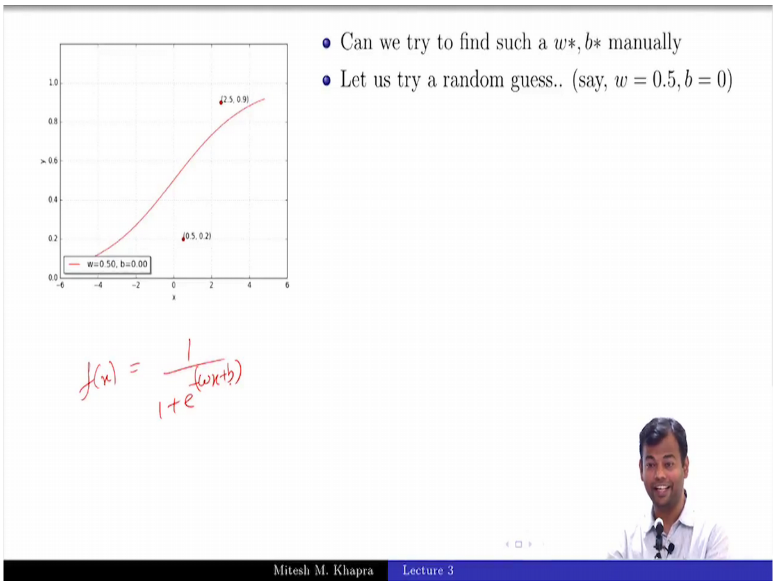


Let us see this in more detail....

Mitesh M. Khapra Lecture 3

So, let us see this in more detail.

(Refer Slide Time: 04:10)



- Can we try to find such a $w*, b*$ manually
- Let us try a random guess.. (say, $w = 0.5, b = 0$)

$f(x) = \frac{1}{1+e^{-(wx+b)}}$

Mitesh M. Khapra Lecture 3

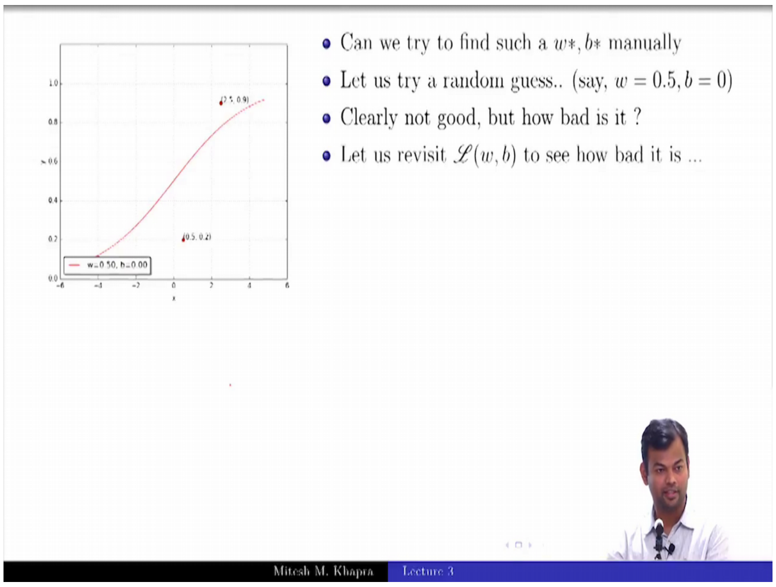
And now what we will do is our quest is for this w star and b star, I will try to find this manually, I will do some random guesswork and try to find this because I do not have any clear principle algorithm for finding it as of now. So, I will just use some guesswork. So, I will give my initial guesswork as w equal to 0.5, b equal to 0 for no reason, I just picked up some values right. And this is what the function that I got, what does this mean? This function an error; so the sigmoid formula should be here we should have this sigmoid formula here.

So, is this a good are you happy with the solution; if I give you?. Are you happy with this solution? Is this good, bad, ugly? Has to be something, bad ok, we will not call it ugly, ok. So, why is it bad? It is not passing through those points, ok, I will ask you a question how bad is it? Can you assign a number to it? We are always good at qualitative stuff, but quantitatively can you tell me a number, how bad is this? Can you tell me a way of finding how bad this is? I already told you in detail how to find that how bad it is.

Student: (Refer Time: 05:25).

The loss function; right?

(Refer Slide Time: 05:27)

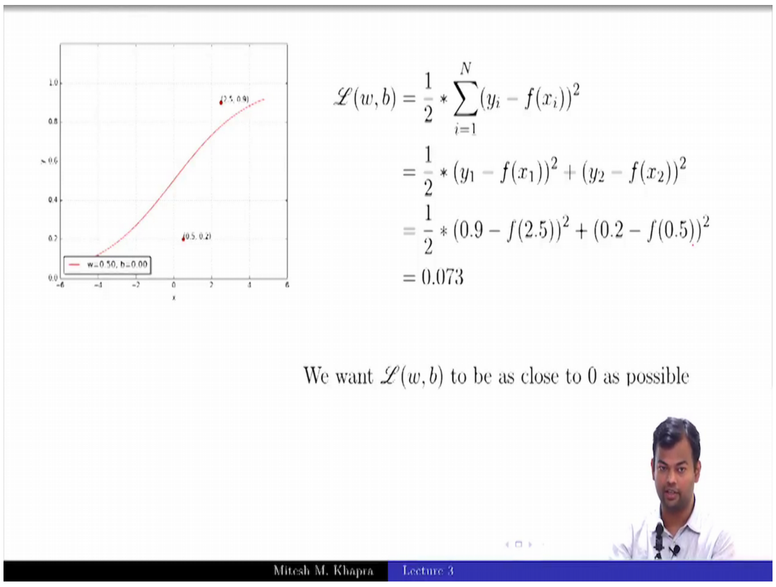


- Can we try to find such a $w*, b*$ manually
- Let us try a random guess.. (say, $w = 0.5, b = 0$)
- Clearly not good, but how bad is it ?
- Let us revisit $\mathscr{L}(w, b)$ to see how bad it is ...

Mitesh M. Khapra Lecture 3

We have the loss function, let us see that again and see if we can find out how bad this is ok.

(Refer Slide Time: 05:32)



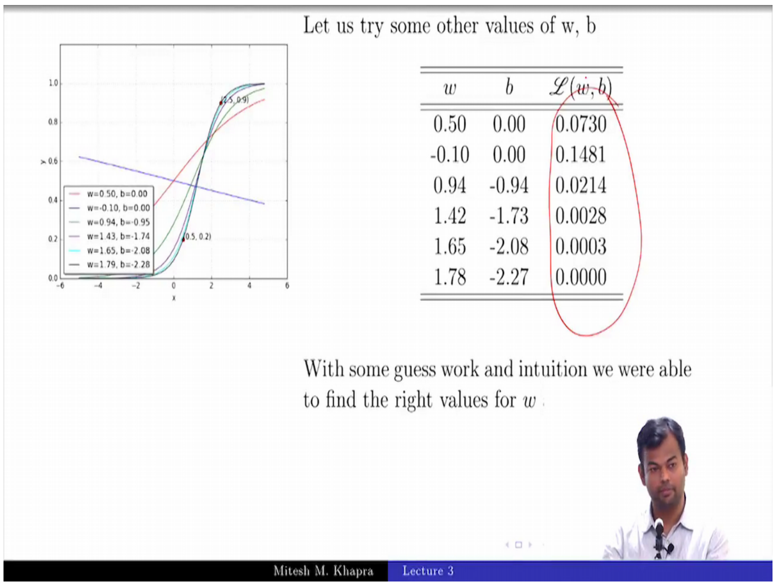
$$\mathscr{L}(w,b) = \frac{1}{2} * \sum_{i=1}^{N} (y_i - f(x_i))^2$$
$$= \frac{1}{2} * (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2$$
$$= \frac{1}{2} * (0.9 - f(2.5))^2 + (0.2 - f(0.5))^2$$
$$= 0.073$$

We want $\mathscr{L}(w,b)$ to be as close to 0 as possible

Mitesh M. Khapra Lecture 3

So, this is what my loss function is, and I have 2 data points, I will just expand it out, fine? Now I will plug in the values I know this is 0.9, and I will compute the value of f 2.5, I will plug in this and I will plug in this and this is what I get. So, this is how bad it is. What did we actually expect it to be in the good case? 0. So, this is not 0, this is 0.073; so now we have a quantitative handle on how bad this is, ok. So, let us keep this in mind and let us try to continue guessing. So, we want the loss function to be asked close to 0 as possible; we are not there yet.

(Refer Slide Time: 06:14)



Let us try some other values of w, b

w	b	$\mathscr{L}(w,b)$
0.50	0.00	0.0730
-0.10	0.00	0.1481
0.94	-0.94	0.0214
1.42	-1.73	0.0028
1.65	-2.08	0.0003
1.78	-2.27	0.0000

With some guess work and intuition we were able to find the right values for w

Mitesh M. Khapra Lecture 3

So, then I make a different case, I say let me try minus 0.10, 0.00 what happened now? Is it now good, bad, ugly?

Student: (Refer Time: 06:23).

Now let us call it ugly, right? So, it is worse and how do I know it is worse? Because I plugged it in to the loss function, and I got a value which is greater than the value at which I was. So, I clearly know this is bad, ok.

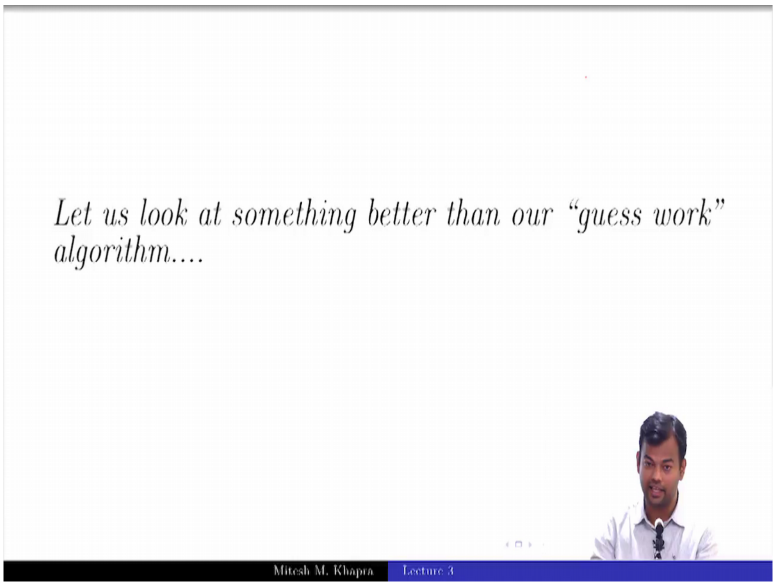
So now this is how my mind is working, right oh I as far as w was positive things looked at least I was close to 0 in the first decimal. Now when I made it negative that does not look good; so, let me just keep it positive and keep increasing it right. So, I saw 0.94 and I also tweaked the b of it I have done complete random guesswork, right? Now what happened? Good, bad, ugly?

Student: Better.

Better ok, now what will you do? What would your next case would be? Make w even more positive; perhaps that would help, and be even more negative and so on.

I can continue in this manner and actually get close very close to the solution. So, I can do this guesswork and find these values. But it is still an educated guess, right I am not guessing in the dark, this is what is helping me drive towards those guesses, and I am just looking at these values and making an educated guess right. And that is the educated guess which I took that probably making w even more positive would help, but this is still brute force in a sense, right this is not something that you would want to do when you have 100, 1000 parameters and so on right, and 1 million data points and so on, right.

(Refer Slide Time: 07:42)

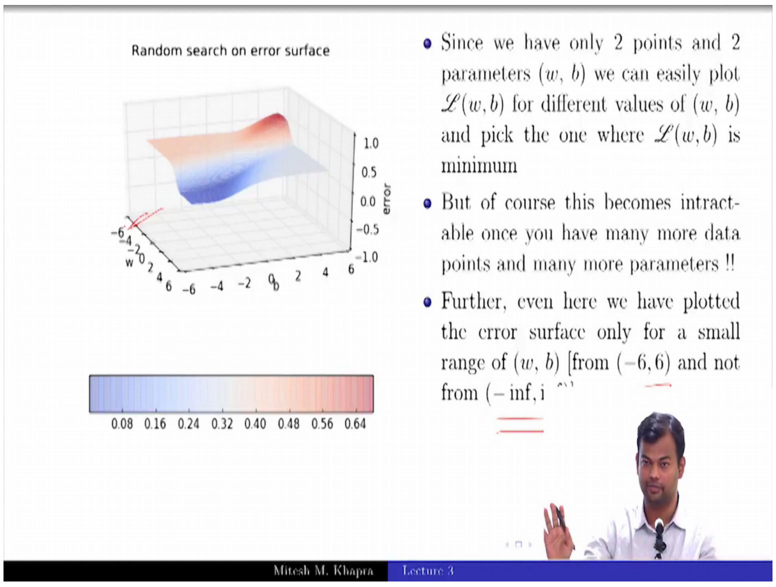


Let us look at something better than our "guess work" algorithm....

Mitesh M. Khapra Lecture 3

So let us look at something which is better than our guesswork algorithm ah. So, we are not there yet actually on the next slide I am still going to talk about the guesswork algorithm.

(Refer Slide Time: 07:53)



- Since we have only 2 points and 2 parameters (w, b) we can easily plot $\mathscr{L}(w,b)$ for different values of (w, b) and pick the one where $\mathscr{L}(w,b)$ is minimum
- But of course this becomes intractable once you have many more data points and many more parameters !!
- Further, even here we have plotted the error surface only for a small range of (w, b) [from $(-6,6)$ and not from $(-\inf, i$

Mitesh M. Khapra Lecture 3

And eventually we will get to something which is better than the guesswork that, ok.

So, since we have only 2 points and 2 parameters, what I can do is I can take all possible values of w and b, right that is what I was trying; I was picking up some values of w and b. Why just pick some values of w and b? I will pick all possible values of w comma b, right? And I will fix the range, I cannot fix pick it from minus infinity to infinity, but I will pick a range I will say from minus 6 to 6, let me try all values of w comma b compute the loss and plot it, right? Let me tell something about this error function, because this is going to stay with us for quite some time.

So, what you see here is something like a flying carpet? This is colour coded, red is bad, red are the places where the error is high. Blue is good, blue are the places where the error is low, darker the shade of blue, lower the error, darker the shade of red, higher the error. So, in particular if I look at this point, what has happened is, I have taken the corresponding value of w comma b right, which is say minus 4 comma minus 1, right something like that.

I have plugged that value into my loss function, and I got this as the loss function. This has the loss value, and that is what I have plotted for all values between minus 6 to plus 6 and minus 6 to plus 6, for w and b. So, everyone understands how I have constructed this error surface, ok, fine.

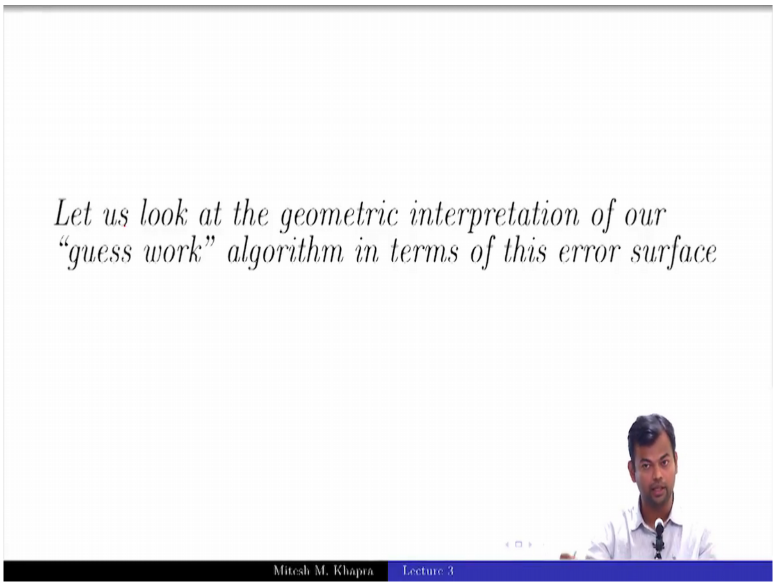
Now, this of course, becomes and now what I can do is once I see this error surface, I know how good this is the point where I need to be, this is the darkest ah shade, and this is where the error is the lowest. So, I can just pick a w comma b value which lies there. This is fine, for this toy example where you just have 2 parameters. But this becomes untractable, once you have more data points and many more parameters.

And that is what happens in most real world applications, right. So, this is not a feasible way of going about things, right? And here again note that I have only taken the range from minus 6 to 6, I do not even know what will happen if I have to look at all values of w comma b, right maybe there was something outside here, right which was even more lower error or something, right. So, I do not really know that.

So, I cannot really use this. So, I need something better than this plotting the error everywhere and finding it order; that is pure brute force. Or surrogate to this was the guesswork algorithm, but which is again something we cannot do for if you have large number of parameters.

So, everyone gets this? That this is a way of finding the solution, but this is not feasible, right that is the only point I am trying to make ok.

(Refer Slide Time: 10:24)

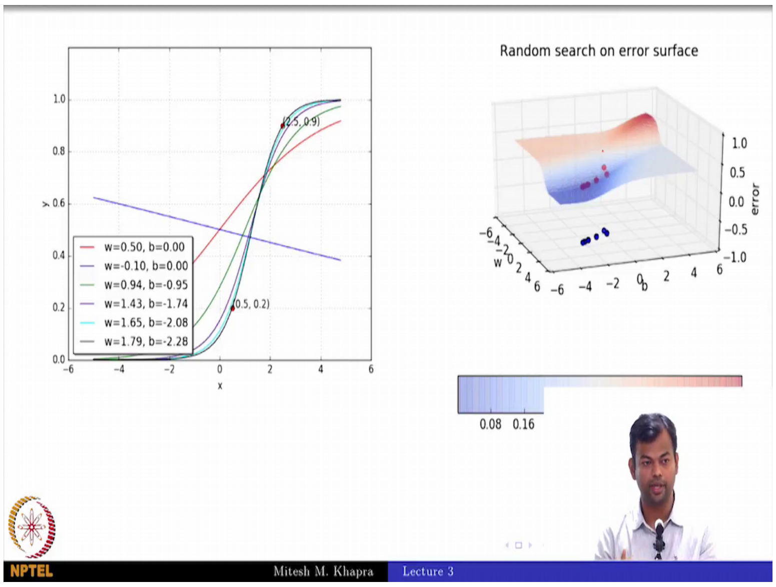


Let us look at the geometric interpretation of our "guess work" algorithm in terms of this error surface

Mitesh M. Khapra Lecture 3

And we look at the geometric interpretation of what was actually happening in the case of the guesswork algorithm with respect to the error surface, ok.

(Refer Slide Time: 10:33)



So, I had chosen some values of w comma b, the first value that I chose actually gave me an error of if you remember it was some 0.073 or something like that right. So, that is the point then I decided to take a very random guess, and my error actually increased. So, you see that I am actually climbing up on this error surface, I have gone from a slightly darker shade of blue to a lighter shade of blue right. And then I corrected myself and then kept moving in a direction where I was going towards the darker and darker shades of blue.

So, what I was actually doing is; I was trying to traverse the error surface and land up in the good regions which were the dark blue regions. Now what I want to do is I want an algorithm which will allow me to do this in a principled manner which is neither brute force nor guesswork ok. So, that is where we learn that module.