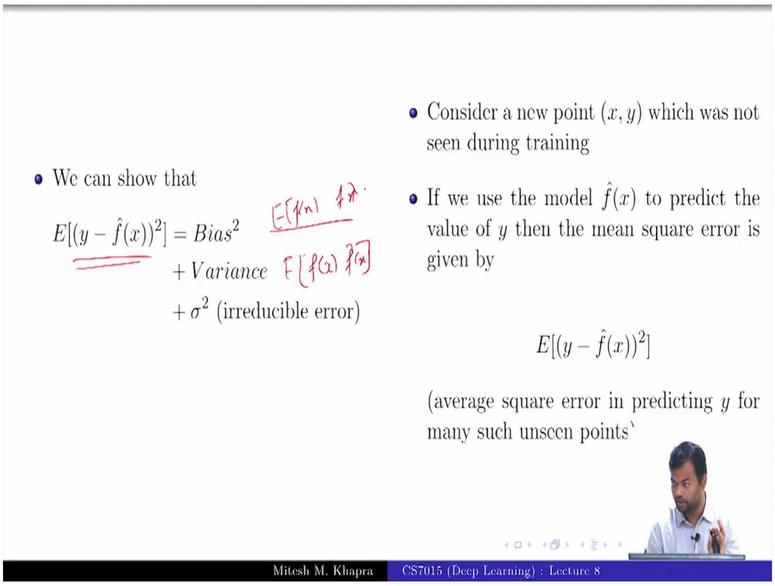
Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module - 8.2
Lecture - 08
Train error vs Test error

So, we would start the next module where we will talk about Training Error versus Test Error and before that we will see this Bias Variance Tradeoff.

(Refer Slide Time: 00:20)



- We can show that

$$E[(y - \hat{f}(x))^2] = Bias^2 \quad E[\hat{f}(x)] \ f x.$$
$$+ Variance \quad E[\hat{f}(x)\ \hat{f}(x)]$$
$$+ \sigma^2 \text{ (irreducible error)}$$

- Consider a new point (x, y) which was not seen during training
- If we use the model $\hat{f}(x)$ to predict the value of y then the mean square error is given by

$$E[(y - \hat{f}(x))^2]$$

(average square error in predicting y for many such unseen points`

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

So, now what have we done so far in these complex models and the simple models, we have trained them using the dash data training data and what are we interested in always a test data. I already know; what was the oil amount of oil mined from the training data locations that I was given and I am not interested in predicting those. I am less interested in learning those, so that if you give me a new location, I should be able to do the right production.

So, I am always interested in the test data, ok. So, now consider a new point which is not seen during the test data and there are several such points that you could see. Now, if you use the model f hat x to predict the value of y, then the mean square error is given by you get this. It is just the expected value of this squared error that I will get. So, what is the randomness here? Y expected value because the x that I am going to feed attest time is going to vary for each of these different xs, I will get a different error.

So, hence that is a random variable. Do you get that? So, please focus on these things, right. I mean just do not take a formula for granted. Just see what is it trying to see. So, whenever you see an expectation over something, always question; what is a random variable here? So, what is the random variable here? It is the squared error loss. Why is it random? It is because it changed the input x. You are going to try it over a multitude of test examples. You will take 1000 text examples, 10,000 text examples and so on. For each of this, you will get a different squared error. That is the randomness. So, you want to see what is the expected value of this or very loosely speaking the average value of this. Now, it turns out that this.

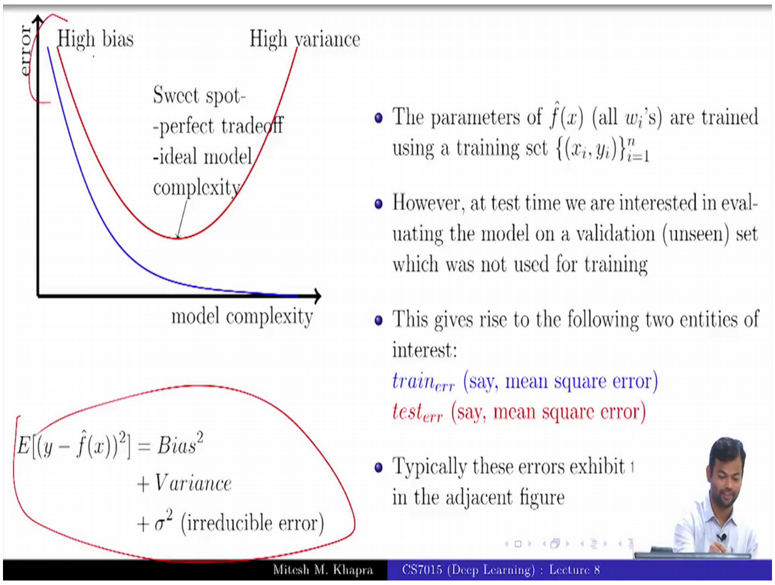
Now, just try to remember that this is also some expectation and you had the terms f x and f x hat here. This also had some expectation and term f x and f x hat and so on, right. If you do not remember the exact formula it is ok, but you do remember there were some expectations inside and the terms f x and f x hat whether they are. So, this is just simple. You are dealing with a minus b the whole square on the left hand side.

If you if you open it up here in some terms, you will get this. So, you can show that the mean average or the expected square error on the test data is actually the bias square plus the variance. That is a small amount of irreducible error. You can go back and work this out and actually the proof is given here on the link, but I hope you get the intuition. You have this a minus b the whole square.

If you open it up and rearrange the terms, you should be able to get this. Now, what does this tell you? What happens if the bias is high? The squared error is going to be high. What happens is if the variance is high, it is going to be high. So, that is why you do not want a very high bias, you do not want a very high variance also. You want this sweet spot in between where the bias and variance are just about optimal. You get that? That is why there is a tradeoff between bias and variance.

You cannot rely on simple models which have high bias; you cannot rely on complex models which have high variance. You want something in between.

(Refer Slide Time: 03:18)



- The parameters of $\hat{f}(x)$ (all w_i's) are trained using a training set $\{(x_i, y_i)\}_{i=1}^{n}$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:
  $train_{err}$ (say, mean square error)
  $test_{err}$ (say, mean square error)
- Typically these errors exhibit t in the adjacent figure

$$E[(y - \hat{f}(x))^2] = Bias^2 + Variance + \sigma^2 \text{ (irreducible error)}$$

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

Now, the parameters of f hat x remember that they are trained using the training data which consists of these end points that you have at test time. We are interested in evaluating the model on a validation set which was different from the training data. This gives rise to the following two quantities. One is the training error which you deal with at dash time training time. That is the error that you are trying to minimize, but a test time you have a different error which is the test error and that is the error that you care about.

Typically these two errors exhibit a certain trend. Do you know what the trend is? Now, on the x axis, I have model complexity and on the y axis, I have error. As a model complexity increases, what would happen to the training error? It will go to almost 0. That is exactly what happened from the linear function to the polynomial function. This is how it will behave. As the model complexity increases, what would happen to the validation error? It will decrease up to a certain point, right because you are still not over fitting on the training data. Your answers are still generalized.

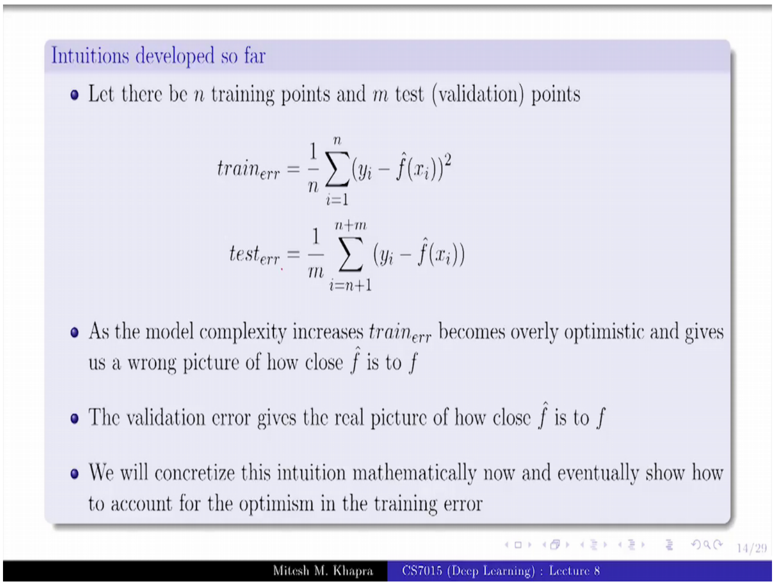
So, you had this degree 1 polynomial degree 25 polynomial. If I take in something in between, then probably this is where I would have ended up with the training error and that would not have been too bad for the test error. You see this. Now, you see I will mark two points two regions rather one of this corresponds to high bias. The other one corresponds to high variance. Tell me which one is, which do this? I cannot understand. So, let me ask this is?

Good. So, you see that there are; these two extreme and we want somewhere to be in between, ok. At least you get the intuition behind this fine, and you are looking for this sweets pot which is the perfect tradeoff between the bias and the variance, right. So, now everyone gets why there is a tradeoff and how this relates to model complexity and therefore, we are looking for the ideal model complexity. How do we achieve the ideal model complexity? Well, we cannot really. Ideal is ideal, but we try to do this using dash. What is the title of this lecture?

Student: Regularization.

Regularization, I will try to use regularization to achieve this, ok. So, let us formalize this a bit more, and remember that this curve is actually because of this equation that you see, high bias. You will be in this region. I am actually inserting it, fine.

(Refer Slide Time: 05:50)



Intuitions developed so far

- Let there be n training points and m test (validation) points

$$train_{err} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

$$test_{err} = \frac{1}{m}\sum_{i=n+1}^{n+m}(y_i - \hat{f}(x_i))$$

- As the model complexity increases $train_{err}$ becomes overly optimistic and gives us a wrong picture of how close $\hat{f}$ is to f
- The validation error gives the real picture of how close $\hat{f}$ is to f
- We will concretize this intuition mathematically now and eventually show how to account for the optimism in the training error

14/29

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

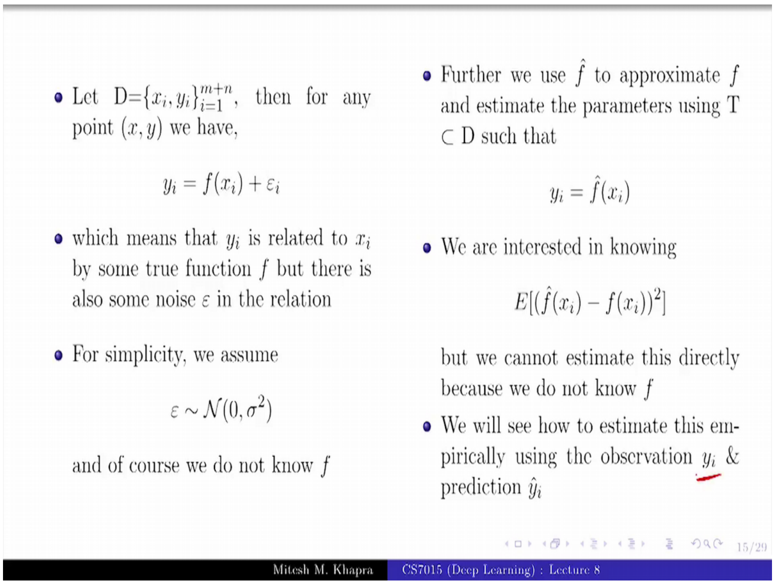
So, the intuitions that we have developed so far is that if there are n training points and m test points. Then we have a train error which goes over the training points and we have a test error which goes over the m test points.

So, I am just taking a total of n plus m points. The first n is training the next, last m is test. Now, as the model complexity increases, what happens to the training error, it becomes very optimistic and gives you a very wrong picture of how close the predicted function is to the true function whether it makes you feel that you have done a perfect job, this you have actually discovered the true function, but that is not correct. It is giving you a false picture of that. Therefore, we should always look at the dash error.

Student: Validation error.

Validation error: so now you see that why you always do this train validation and test. Plate test is unseen. You try to optimize on the training error, but you should always tune for the validation air. Your optimization algorithm is going to take in the training error. It is going to be very optimistic, it is going to try to drive to 0, but you should look at the validation error and try to see that you are not over fitting on the training data even gets this intuition.

(Refer Slide Time: 07:09)



- Let D=$\{x_i, y_i\}_{i=1}^{m+n}$, then for any point (x, y) we have,

$$y_i = f(x_i) + \varepsilon_i$$

- which means that y_i is related to x_i by some true function f but there is also some noise ε in the relation
- For simplicity, we assume

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

and of course we do not know f

- Further we use $\hat{f}$ to approximate f and estimate the parameters using T ⊂ D such that

$$y_i = \hat{f}(x_i)$$

- We are interested in knowing

$$E[(\hat{f}(x_i) - f(x_i))^2]$$

but we cannot estimate this directly because we do not know f

- We will see how to estimate this empirically using the observation y_i & prediction $\hat{y}_i$

15/29

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

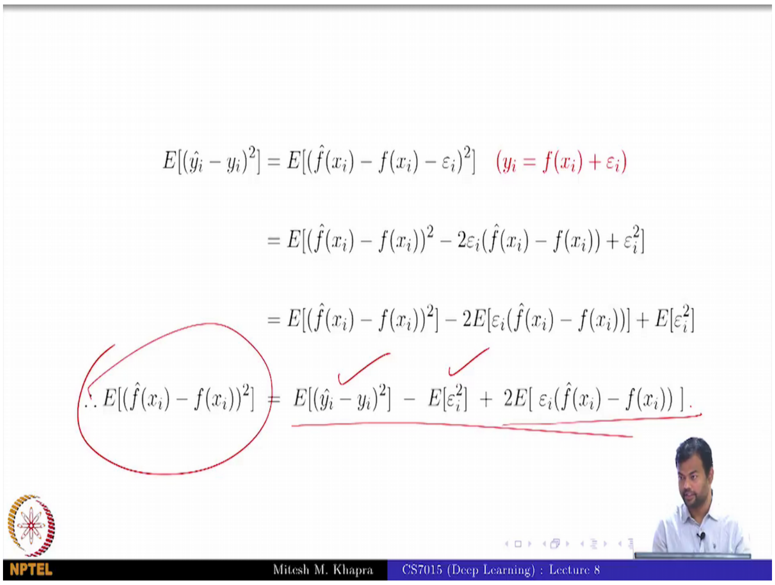
So, now this is all intuition. We will have to concretize this mathematically. So, that is what we will do now. So, that d b these training test points that we have, we know that this relationship holds. We do not know what f is, but we know that this relationship holds. So, what am I trying to say here that we know that there is a true relation between y and x which is given by the function f, but I am also willing to admit some noise that may not be a very neat function, but a small noise might exist.

That is the epsilon i, and I am going to assume that epsilon comes from a normal distribution with zero means. So, on average the noise is going to be 0, but there is a small feeling. Everyone gets this. This is a true relation, but I am willing to admit some noise in the relation, fine and of course, we do not know if we never know, right. Now, going by our paradigm where we have these five components, we use f hat to approximate f f hat. We will have some dash which will I try to learn from the training. What is this dash parameters, right which will try to learn from the training data? The training data t is a subset of your total data which is thus those endpoints, right and we are interested in knowing this quantity. This is what we are actually interested in. Can we compute this quantity? How many of you say yes? How many of you say no? We cannot. Why cannot we compute it?

We do not know. So, why cannot you raise your hands if you all can answer in chorus? So, we do not know what f is raised and how do we compute this quantity right, but what do we actually know. So, now we are going to see something which is true expectation and something which is empirical estimate expectation. How many of you know this? What is the difference between the two? Most of you should, but it is not confident about it. So, we do not know what f x i is the true thing, but what do we know, we are given some training data.

We know these y is for was training leaders and we know these why I had for those training dinner, fine.

(Refer Slide Time: 09:11)



$$E[(\hat{y}_i - y_i)^2] = E[(\hat{f}(x_i) - f(x_i) - \varepsilon_i)^2] \quad (y_i = f(x_i) + \varepsilon_i)$$

$$= E[(\hat{f}(x_i) - f(x_i))^2 - 2\varepsilon_i(\hat{f}(x_i) - f(x_i)) + \varepsilon_i^2]$$

$$= E[(\hat{f}(x_i) - f(x_i))^2] - 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))] + E[\varepsilon_i^2]$$

$$\therefore E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\,\varepsilon_i(\hat{f}(x_i) - f(x_i))\,].$$

NPTEL

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

So, this is something that we can estimate. Yes or no, this is given to us. So, this expectation is going to be an empirical estimate, because we are going to look at some

1000 10000 20000 training points and estimate this. It is an empirical estimate. How many of you get that? Now, I am just going to rewrite some of this. So, what I have done is, I just defined at y is equal to f x i plus epsilon i. So, I have just replaced y i by that, ok. Is that fine? Now, this is of the form a minus b the whole square. So, I am going to treat it as that and just open up the bracket. So, I will have a square minus 2 a b plus b square and now, this is a sum or difference of expectation. So, I can push the expectation inside. So, this is what I get this fine.

Now, I am just going to rearrange the term. So, remember this was the quantity that we were actually interested in, but this is the quantity that we had a handle over because these were the data points given to us. So, I will just rearrange the terms and I can write this which was my quantity of interest as this. Can you estimate everything on LHS, on RHS? What is this variance? Sigma square; we assumed it came from 0, sigma square distribution and this can estimate the answer is no for the same reason. We do not know what f of x is.