

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

**Lecture - 08**  
**Train error vs Test error (Recap)**

(Refer Slide Time: 00:12)

• Let us consider the problem of fitting a curve through a given set of points

• We consider two models :

$$\begin{array}{l} \text{Simple} \\ \text{(degree:1)} \end{array} \quad y = \hat{f}(x) = w_1x + w_0$$
$$\begin{array}{l} \text{Complex} \\ \text{(degree:25)} \end{array} \quad y = \hat{f}(x) = \sum_{i=1}^{25} w_i x^i + w_0$$

• Note that in both cases we are making an assumption about how  $y$  is related to  $x$ . We have no idea about the true relation  $f(x)$

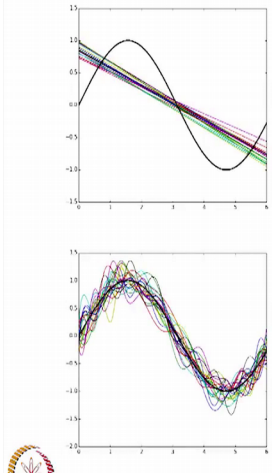
• The training data consists of 100 points

The points were drawn from a sinusoidal function (the true  $f(x)$ )

NPTEL Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8 5/84

So, we spoke about bias and variance and we saw that simple models have a high bias, but low variance and complex models have a low bias high variance and so on. And we saw it some illustrative examples that what that is what that means.

(Refer Slide Time: 00:18)

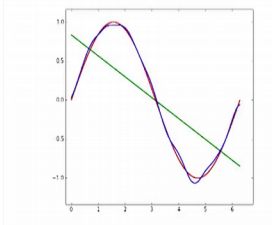


- Simple models trained on different samples of the data do not differ much from each other
- However they are very far from the true sinusoidal curve (under fitting)
- On the other hand, complex models trained on different samples of the data are very different from each other (high variance)

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

Then, the important thing to note was these two formal definitions of bias.

(Refer Slide Time: 00:26)



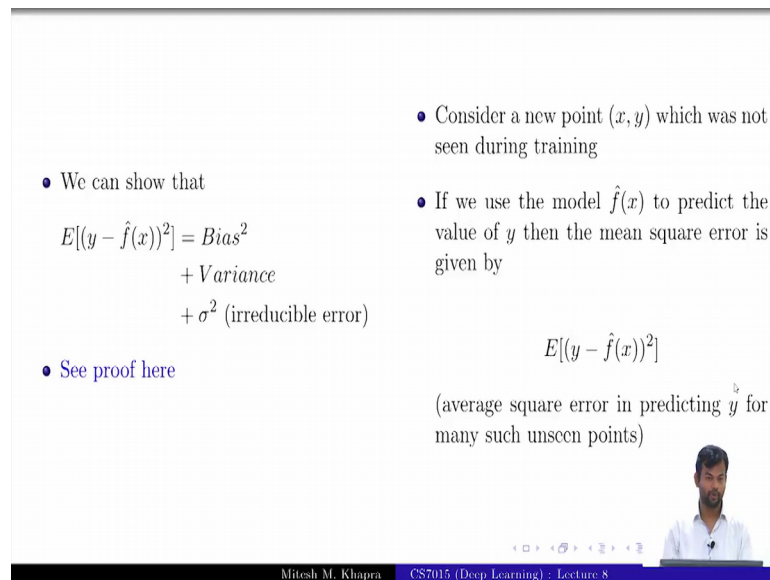
Green Line: Average value of  $\hat{f}(x)$  for the simple model  
Blue Curve: Average value of  $\hat{f}(x)$  for the complex model  
Red Curve: True model ( $f(x)$ )

- Let  $f(x)$  be the true model (sinusoidal in this case) and  $\hat{f}(x)$  be our estimate of the model (simple or complex, in this case) then,  
$$\text{Bias}(\hat{f}(x)) = E[\hat{f}(x)] - f(x)$$
- $E[\hat{f}(x)]$  is the average (or expected) value of the model
- We can see that for the simple model the average value (green line) is very far from the true value  $f(x)$  (sinusoidal function)
- Mathematically, this means that the simple model has a high bias

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

The formal definition of variance which you all know anyways, and then the important concept that we spoke about was the strain error versus test error.

(Refer Slide Time: 00:36)



• We can show that

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

• See proof here

• 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

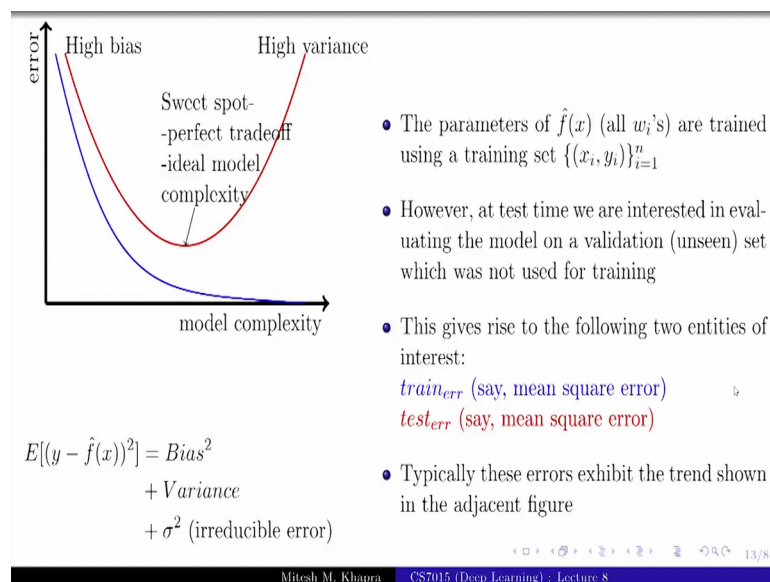
$$E[(y - \hat{f}(x))^2]$$

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

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

So, this was the curve that we were interested in and one corner of this curve was related to high bias low variance. And the other corner was related to low bias high variance, right.

(Refer Slide Time: 00:39)



High bias High variance

Sweet spot - perfect tradeoff - ideal model complexity

error

model complexity

• 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)

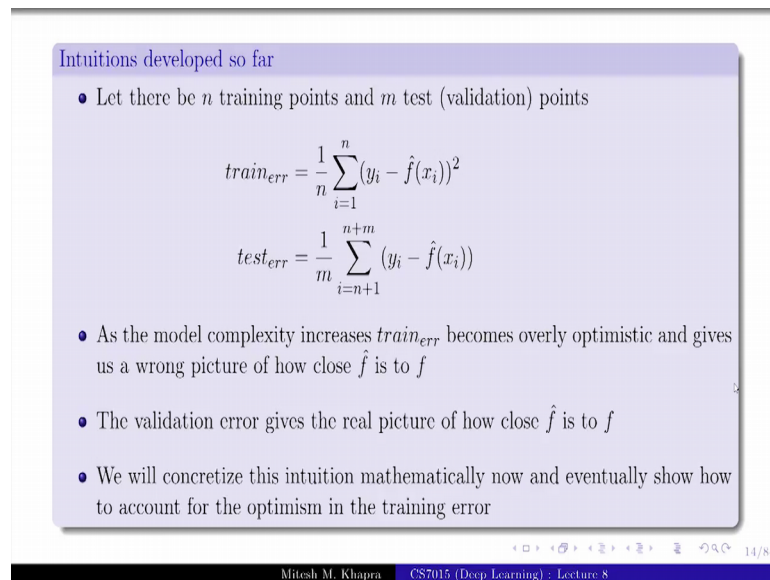
• Typically these errors exhibit the trend shown in the adjacent figure

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

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

So, we were looking for something in the middle. That is what our quest is in this lecture, right and we want to find ways of falling somewhere in middle.

(Refer Slide Time: 01:06)



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

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8 14/84

This led to the definition of two quantities of interest or training error and test errors. So, training error is computed from the training points. These are the points that you actually look at while you are solving this optimization problem. So, the training always involves solving an optimization problem which is the objective that you want to optimize or maximize and the test error is something that you want to use it for at eventually.

So, you all have these two quantities of interest that we design and we realize that the training error is more optimistic whether the test errors actually gives us the real picture of what we do and we tied those back to things that you have done previously in the machine learning; or other courses that we always split the data into training valid and test train it on the training data, do some validations on the validation data, but never look at the test data. That is for the final evaluation.

So, this is this intuition which I have been trying to build with these two curves is the explanation for why we do things that we.

(Refer Slide Time: 02:03)

• 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  $\varepsilon$  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  $T \subset D$  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  $\hat{y}_i$

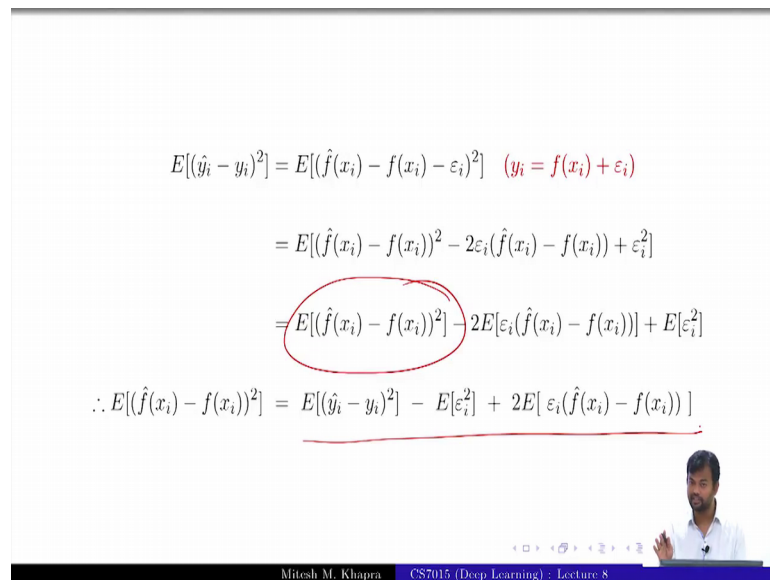
15/84

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

Now, we are interested in doing a more mathematical, mathematically rigorous analysis of this intuition, right. So, that is where we left off. So, what we are interested in? So, now I will just start from this point is that we are given some data which is  $m$   $n$   $m$  training points. And end testing points and we know that there is a true function between the outputs and the inputs, and we are also expecting or accepting some noise in this relation just as in any other relation which means that  $y$  is related to  $x$   $i$ , but by some true function. But there is also this noise and for simplicity we assumed as this noise comes from a normal distribution with zero mean and some small variance and as usual we never know  $f$ . But we are trying to approximate this  $\hat{f}$ , and we come up with some parametric form for  $\hat{f}$  and then, try to learn the parameters of  $\hat{f}$  from the training subset of the data that is given to us.

So, this is what we always do and we have already seen different variations of  $\hat{f}$ . One of them being the deep neural network and what we are actually interested in is this quantity, the expected difference or square difference between the predictions made by our model and the true value of the output with respect to the true function weight. Then, we asked I asked you whether we can actually estimate this quantity and all of you said no. Why? It is because you do not know what  $f$  of  $x$  is, right. So, we will see how to estimate this empirically.

(Refer Slide Time: 03:44)

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


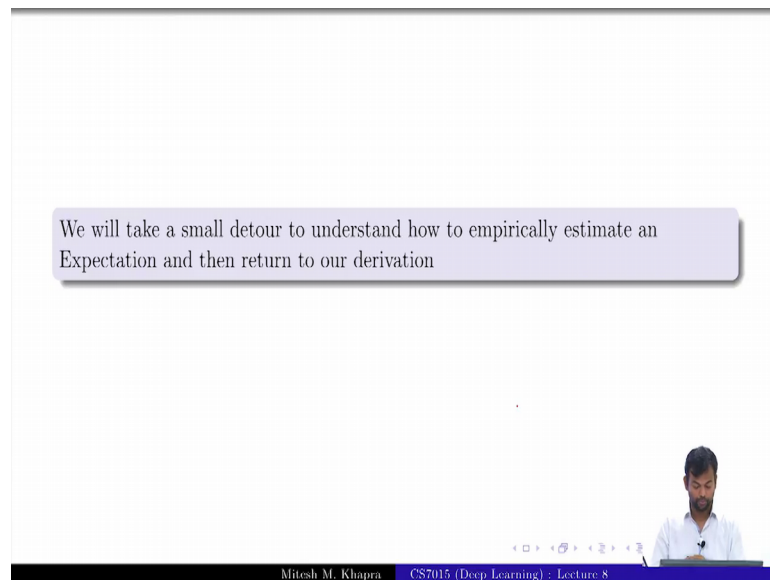
So, then we started off with this information that we have we know what  $y_i$  hat is because that is the prediction that we make and we know  $y_i$ . What  $y_i$  is, we do not know the function, but we see the output of the function in the form of the training data points given to us or any data points given to us.

So, we wrote this by making this particular substitution where we notice that  $y_i$  that we see is actually the true function plus some noise and then, we did some trickery and try to simplify this. And then, we just realize that this is the term that we are interested in. So, we moved it to the other side of the equation and came up with this neat left hand side or neat right hand side that we need to analyze now. So far everything is clear.

This is where we ended the last class, right. You just went to it very quickly, but I assume everything is clear at this point, fine. So, we are left with a bunch of expectations, right and I am assuming we have no clue how to estimate this. Remember that when you are dealing with expectations as always this true expectation and then, there is this empirical estimation. So, what we are going to move towards? So, these all equations when I write  $E$  here, capital  $E$  here, I am talking about the true expectation.

Now, we will see how to approximate the true expectation with an empirical expectation and then, based on that we will make some observations.

(Refer Slide Time: 05:12)

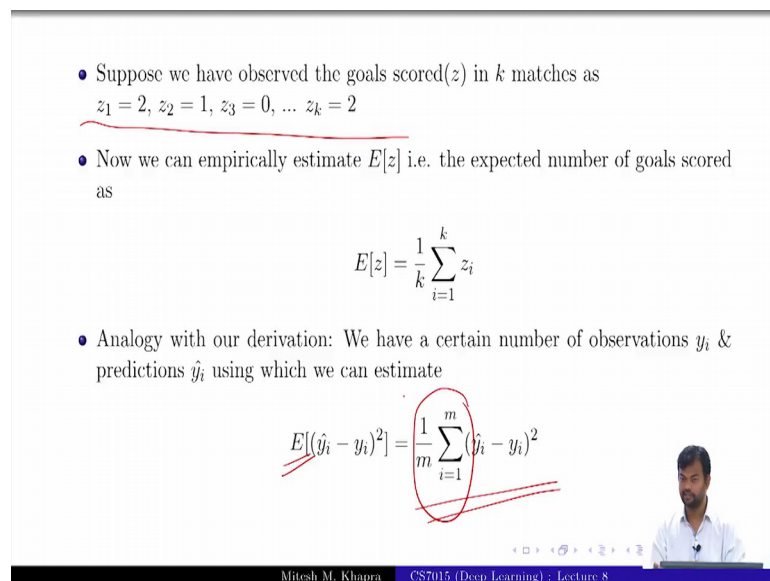


We will take a small detour to understand how to empirically estimate an Expectation and then return to our derivation

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

So, that is what we will do?

(Refer Slide Time: 05:14)



- Suppose we have observed the goals scored ( $z$ ) in  $k$  matches as  $z_1 = 2, z_2 = 1, z_3 = 0, \dots, z_k = 2$
- Now we can empirically estimate  $E[z]$  i.e. the expected number of goals scored as

$$E[z] = \frac{1}{k} \sum_{i=1}^k z_i$$

- Analogy with our derivation: We have a certain number of observations  $y_i$  & predictions  $\hat{y}_i$  using which we can estimate

$$E[(\hat{y}_i - y_i)^2] = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2$$

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

Now, so we will just take a small d 2 and I will just tell you what expectations are or what empirically expectation is, how to compute them. So, suppose we have observed the goals code in  $k$  matches. There is some  $k$  football matches that we have seen and we have seen that the goals code were the following.

Now, if I asked  $q$  what is the expected value of the goal? Now, the number of goals for what will you do; take the average of this. This is what you will do. So, what is it that

you are doing here? You are taking a dash estimate of the expectation, empirical estimate. You are making some observations. These are the observations given to you, these are the  $k$  matches. Watch as much, as many football matches as you want after the semester ends and then, notice the number of goals that were scored in them and then, you can compute this expectation, and this is how you do empirically. So, there is something that we do on a regular basis, but I just want you to realize that what you are doing is actually an implicit estimate of the true expectation. Is that fine?

Now, can you relate this to the quantity that we are interested in? We are interested in computing a certain expectation which is this. Can you take an analogy and tell me how you would do this? The hint is we have done this a million times in the course already, fine. So, this is how we will do it and have actually done this a million times in the course. So, when you compute this, we are actually doing an empirical estimate of the data.

So, let us just take a minute to understand this. We are given some data, we are interested in this to expectation which we cannot compute. So, we will take this data, we will assume there is enough of this. We are given  $m$  samples which are enough and from that we will make an empirical estimate and just as in the case of these old score, right. As you see more and more matches, you will have a better understanding of how many goals can be scored when two particular teams are playing. In the same analogy holzer as you see more and more data, your estimate would become better, but that is how you will do the estimation, ok.

So, now we will come back to, so now do not get surprised when I am going to replace all these  $e_s$  by this all the  $e_s$  that we had in our original equation. I am going to replace them by these summations, fine.




(Refer Slide Time: 07:22)

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

- We can empirically evaluate R.H.S using training observations or test observations

Case 1: Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{\substack{= \text{covariance } (\varepsilon_i, \hat{f}(x_i) - f(x_i)) \\ \times \quad \checkmark}}$$



So, this was our original equation that we had derived and we were interested in this left hand side quantity which is a sum of some terms on the right hand side. So, now this expectation I told you that we can estimate it from data, but which data? Training data or test data, both. So, we will try to estimate it from both and see if there is any difference which arises when you estimate it from one data and the other data.

So, the first thing that I am going to do is, I am going to use test observations to estimate this. So, can you tell me what are my summations going to look like? It is summation over n plus 1 to n plus m. We assume that the first endpoints are training points and the remaining points are test points.

So, the quantity on the left hand side is true error. Remember that because that has f x which we do not know quantity on the right side, the first thing is empirical estimation of the error, ok. The second thing is a small constant. However, the epsilon i square and we assume that comes from a normal distribution with a small variance. What is the third quantity? Actually I have given you the answer already, but I want you to think about it. I am saying it is the co variance between two things.

When I say it is the co variance between two things, what is the first thing that I need to prove is that the two things are dash random variables. I mean first thing we need to see is that the two things are random variables epsilon is clear. It is a random variable.

What about this other thing or rather epsilon is a random variable what about the other thing and depending on the training instance that you have sampled, this ongoing difference is going to differ. You are having your training or test instance whatever is this  $x_i$  this is going to differ because these  $x$ s are different. They are all random variables. So, there is difference between these two quantities also going to be a random variable. Is that fine, but still is this true.

So, then I have told you this is  $x$  and this is  $y$  and what I am saying is that the co variance between  $x$  and  $y$  is just  $e$  of  $x$  into  $y$ . Is that correct?

(Refer Slide Time: 09:34)

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

- We can empirically evaluate R.H.S using training observations or test observations

Case 1: Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{\text{covariance } (\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

$\therefore \text{covariance}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$   
 $= E[(X)(Y - \mu_Y)]$  (if  $\mu_X = E[X] = 0$ )  $\varepsilon$   
 $= E[XY] - E[X\mu_Y] = E[XY] - \mu_Y E[X]$

20/84

That is how you define co variance. What is the definition of co variance; if you have bothered to look at the prerequisites, no expectation in the form of  $e$ . So, co variance is  $e$  of  $x$  minus  $\mu$  of  $x$  into  $y$  minus  $\mu$  of  $y$ , what is our  $x$  epsilon and what is our  $y$ . What is  $\mu$  of  $x$  0?

So, I will just simplify this a bit. I will open up the product. What is  $\mu$  of  $y$  into  $e$  of  $x$ ? What is  $e$  of  $x$ ? What is the expected value of the noise 0?

(Refer Slide Time: 10:01)

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

- We can empirically evaluate R.H.S using training observations or test observations

Case 1: Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{\text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

∴ covariance(X, Y) = E[(X - μ<sub>X</sub>)(Y - μ<sub>Y</sub>)]  
 = E[(X)(Y - μ<sub>Y</sub>)] (if μ<sub>X</sub> = E[X] = 0)  
 = E[XY] - E[Xμ<sub>Y</sub>] = E[XY] - μ<sub>Y</sub>E[X] = E[XY]

20/84

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

So, then this turns out to be as that. Is that fine? That is why we are writing the covariance is just the product of the two things. Is it fine?

So, let us just take a minute to again understand this. The true error is the empirical estimation of the error plus I mean plus or minus a small constant, and then, this nasty quantity that we do not know what to do with it. So, let us look at this quantity and see what we can say about it.

(Refer Slide Time: 10:30)

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{\text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

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

20/84

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

Now, what is the co variance between these two? I am trying to compute this expectation from the test data. Just remember that. So, each  $i$  here is a test instance are these two random variables dependent or independent is the question that I am trying to ask. It is independent. So, let us look at it piece wise. So, remember that we had said that  $y$  is equal to  $f$  of  $x_i$  plus  $\epsilon_i$ . This  $\epsilon_i$  I had no relation to  $f$  of  $x_i$ . I mean I could choose any  $x_i$ , but this noise is going to be random. So, there is no relation between these two.

Now, is there a relation between  $\hat{f}$  of  $x_i$  and  $\epsilon_i$ ? We are doing tests. So, how did we come up with  $\hat{f}$  of  $x_i$ ? How did when I say how did we come up with  $\hat{f}$  is  $i$  mean? How did we learn the parameters of a fact using the training data and what are we computing expectation with respect to now tested on these? Did these  $\epsilon_i$  improve, influence the parameters that we had learned further from the training data? No, since there is no dependence between these two guys.

So, that is why  $\epsilon_i$  is independent of the other random variable that you see in this expectation. Is that clear? Do you get the intuition  $\hat{f}$  of  $x_i$  further? No, but this is the mean. This noise is what is present in the test data and you have not seen this add training time. When you are training the parameters, you did not look at this noise. You are looking at the noise in the training data.

So, this is not participated in the estimation of the parameters of  $\hat{f}$ , but that was for the training data right, but this now I am doing the expectation from a test data.

(Refer Slide Time: 12:10)

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{\text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

- None of the test observations participated in the estimation of  $\hat{f}(x)$  [the parameters of  $\hat{f}(x)$  were estimated only using training data]  
 $\therefore \varepsilon \perp (\hat{f}(x_i) - f(x_i))$   
 $\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] = E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i)] = 0 \cdot E[\hat{f}(x_i) - f(x_i)] = 0$   
 $\therefore \text{true error} = \text{empirical test error} + \text{small constant}$
- Hence, we should always use a validation set (independent of the training set) to estimate the error

21/84

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

So, these two random variables are independent. That means, I can write this as is this fine. What will happen to this 0? So, what did we eventually conclude that the true error is equal to empirical test error plus a small constant?

So, what does this tell you? Now, tell me forget the math. Tell me in English, right. What does this take? What does this mean? Can you relate it to; now why you do this training error, validation error, test error? So, what does this tell me? This tells me that if I have trained a model and now if I take an estimate of the error on some data which I had not used for the training, then that error which I see is actually very close to the true error. It only differs by this small constant.

How many forget that? That is why when I look at the validation error, it is not being overly optimistic. It is giving me a true picture of what the actual error is, right. So, there are two things that you need to understand here. One, this is the quantity that we are interested in which we cannot estimate. We are trying to estimate it by using this; we are trying to make an approximation. So, we are trying to see how good this approximation is. What this derivation is telling us is that if you are approximated it using the test error or the test data, then this approximation is actually very close to the true error and how close it is actually? It just differs by this small constant.

So, you get the importance of what we are seeing here right. Now, to truly appreciate this I need to tell you what would have happened if you had used the training data for this

estimation, right. It is largely dependent, but that is again a normal assumption that you make. So, this is ok. Good that you asked at this point. I will be doing a couple of things today where we will be deriving some things. We will try to prove some things mathematically, but all of these would have underlying some assumptions.

So, if you remember the atom derivation with this we did there, also we had made this funny assumption that the gradients are actually coming from a stationary distribution which will not happen in practice. So, this reminds me of this joke from Big Bang theory. If it says that I have a solution, but it only works for squared eggs in a vacuum, right. So, it is basically all these things always have some assumptions underlying them. But the idea is to kind of ignore those assumptions and see what happens in a neat setting and at least see whether in a neat setting everything works fine or not.

So, that is what is happening here. So, is a valid point that you are assuming that the noise comes from a zero mean distribution. Now, if the noise did not come from a zero mean distribution, then this would have not gone down to zero and the mean would have been higher than this is no longer a small constant and so on. So, those things are there. So, this is going to happen in some of the other derivations that I do. Today it is not that I am teaching you something wrong. It is just that you have to take it with a pinch of salt in the sense that these assumptions are there and the original derivations these are not my assumptions. And they work only under those assumptions.

So, you have to be careful about that, but the idea is that still with these assumptions, can we at least make something meaningful out of it. Is that fine with everyone? Can we all work with that basic premise? So, what I have done so far is told you that if you are estimating the errors from the validation data, you are doing a good job.

(Refer Slide Time: 15:37)



Case 2: Using training observations

$$\begin{aligned}
 & \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\
 &= \underbrace{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2}_{\text{small constant}} + \underbrace{2 E[\varepsilon_i (\hat{f}(x_i) - f(x_i))]}_{\text{covariance } (\varepsilon_i, \hat{f}(x_i) - f(x_i))}
 \end{aligned}$$

Now,  $\varepsilon \not\perp \hat{f}(x)$  because  $\varepsilon$  was used for estimating the parameters of  $\hat{f}(x)$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] \neq E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i)] \neq 0$$

Hence, the empirical train error is smaller than the true error and does not give a true picture of the error

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

Now, let us see if I would estimate the error from the training data. Take a guess what would happen? What would my argument for this be? Now, this will not disappear, right because these two are not independent. Now, I cannot write it as a product of two expectations. That means, it will not go down to 0, fine. So, that is the argument which I am going to make.

So, hence actually the true error if you see, it is equal to the empirical estimation plus some quantity. That means the true error is dash as compared to the empirical. That means the empirical error that we see is pessimistic or optimistic? Optimistic: that is what I started with that. You gave a very optimistic estimation of your error if you are looking at this empirical estimation from the training data because you have ignored this quantity. Is it fine? So, what is missing in the story?

Let us see now what was this quantity. So far all our discussions LT term right, but now suddenly I have realized that my true error is actually L theta plus something else. You see where I am headed with this. So, that is what we need to see now. Now think it would be we should, but I am pretty sure it is positive. I cannot work it out.

Now, I am pretty sure it is positive and you can see and if you find it is not then let me know.

(Refer Slide Time: 17:09)

Case 2: Using training observations



$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

Now,  $\varepsilon \perp \hat{f}(x)$  because  $\varepsilon$  was used for estimating the parameters of  $\hat{f}(x)$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] \neq E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i)] \neq 0$$

Hence, the empirical train error is smaller than the true error and does not give a true picture of the error

But how is this related to model complexity? Let us see



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

So, how is all this related to model complexity? We started off with this idea that model complexity tells you how much is the bias, how much is the variance and because of that you get these two curves that you are not happy with. One curve being very optimistic and the other curve being a bit pessimistic. Now, how does this discussion tie up to model complexity?