Course Name: Machine Learning and Deep learning - Fundamentals and Applications

Professor Name: Prof. M. K. Bhuyan

Department Name: Electronics and Electrical Engineering

Institute Name: Indian Institute of Technology, Guwahati

Week-1

Lecture-3

Welcome to NPTEL online course on machine learning and deep learning fundamentals and applications. In my first class, I introduced the concept of bias and the variance. So if I consider the problem of classification or the problem of regression, if I consider high bias, the problem is the error is significant in case of the training and also it is significant in case of the testing. That means for the unseen data, the error is significant.

That means high bias means I am considering a very simple model for classification or maybe for regression. And another case is the high variance.

That means I can consider a very complex model for classification or maybe for regression. So if I consider high variance, the problem is during the testing, the error will be significant. But during the training, the error is minimum because I am considering a complex model.

So during the training, the error will be minimum,

but for the unseen test data, the error will be significant.

So that is the high variance. So that is why we should compromise between bias and the variance. Bias should not be too high and also the variance should not be too high. That means the model should not be too simple and the model should not be too complex for the problem of classification and the regression.

So there should be a tradeoff between bias and the variance.

So today I am going to discuss the concept of bias and the variance and the tradeoff between bias and the variance. So let us discuss this concept, the concept of bias and the variance and how to consider the tradeoff issue of bias and the variance. So in this figure you can see I have shown that is the concept of the bias and the concept of the variance.

In the x-axis you can see I have shown model complexity and in the y-axis I have shown the error. So if I consider a very simple model, that means I can consider it is a very simple model here and this side I can consider a very complex model.

So let us see what will happen in these two cases. So if I consider a very simple model, you can see here during the testing or first let us consider during the training the error is significant. So here you can see this is the training, this dotted curve is for training. So error is significant because we are considering a very simple model and you can see the testing also during the testing also error is significant. So this green curve it shows the testing that is for the unseen data and in case of a high variance that means if I consider a complex model

then you can see the training error that is the training error

that is shown by the black color dotted line the curve.

So this is not significant it is very small in case of the training, but during the testing that is shown by the green colored curve. So you can see here during the testing the error is significant and that is the overfitting that is the concept of the overfitting. Overfitting is because of the high variance and underfitting is because of the high bias. So one is the underfitting and another one is the overfitting.

So this green curve shows the testing and this black dotted curve shows the training error and this blue curve the dotted blue curve shows the variance.

If you see if I consider a very complex model that means in this case the variance is very high and if I consider a very simple model the variance is small. So this is about the variance and you can see I have shown also the bias that is the bias squared. So this is the bias squared. So for a simple model this bias is very high and for a complex model the bias is less.

So you can see the difference between these two one is the underfitting another one is the overfitting.

So that means underfitting is not good and the overfitting is also not good. So there should be some compromise between the high bias and the high variance. So if I consider the model complexity at this point that means not a very simple model and not a very complex model then corresponding to this case you can see during the testing also the error is not so high and during the training also the error is also not high in this case. So that means there should be some trade-offs between bias and the variance.

So there should be a trade-off between bias and the variance.

So bias should not be too high and variance should not be too high. So the generalization test error that is the error in unseen data can be decomposed into bias error that means I can write the test error that is the error in unseen data and that can be decomposed into the bias error that is because of the high bias that is the error from wrong model assumptions plus another error I can consider that is the error due to high variance. So that is the error from sensitivity to small fluctuations in training data plus I can consider another error that is the irreducible error. So in the figure also I have shown the irreducible error that is because of inherent noise in the problem itself.

This irreducible error is because of inherent noise in the problem itself.

So this irreducible error is irrelevant of the underlying model and it is because of the inherent noise in the problem. So maybe the noise coming from the data quality or maybe inaccuracies in collecting the data. So for these reasons I have the irreducible errors and high bias already I told you because we are considering a very simple model and high variance because of the over complex assumptions. So this is the concept of the bias variance and you can see the test error I can decompose into bias error the error due to high variance and also the irreducible error.

So corresponding to the high bias and it is actually the under fitting and corresponding to the high variance that is nothing but the over fitting.

So our objective is to build a model that achieves a balance between bias and the variance. So that the combined error of these competing forces is minimum and that is the objective of the bias variance tradeoff. So move to the next figure here I have shown a classification problem. So two class problems. So in the first case what I am considering I am considering a very simple model and this simple model is not good for classification because I am getting a decision boundary between the classes and you can see there are many misclassifications and in this case I am considering a very simple model.

In the third case I am considering a very complex model and corresponding to this during the training I am getting this decision boundary between the classes and during the testing the error will be high because for the unseen test data there will be misclassifications. So that is the case of over fitting. The first case is the case of under fitting and there should be compromise between these over fitting and under fitting. So in the middle I have shown the good compromise and I am getting a decision boundaries between these classes.

So this concept is also true for regression.

So here in this in the curve I have shown corresponding to the under fitting that is the

high bias you can see the error is significant both for training and for the testing. So training is shown by the green curve and the testing is shown by the red curve and for the over fitting that is the high variance for the training the error is not significant it is minimum but during the testing that means for the unseen data the error is significant. So that is why we have to go for good compromise that is the tradeoff between high bias and the high variance. So if you consider this point here then you can see during the training the error is not so significant and during the testing also the error is not significant. So that means we have to consider the tradeoff between bias and the variance.

So mathematically how to consider this problem. So let us move to the next slide. So the problem definition suppose we have one independent variable X independent variable is suppose X and one is a dependent variable. So one dependent variable is Y so Y depends on X so Y value I can write like this $Y = FX + \varepsilon$ that means Y value depends on X and also Y value can also be affected by noise. This noise cannot be modeled explicitly.

So $Y = FX + \varepsilon$ that means the Y depends on X and also the Y value can be affected by noise. So that is why I am writing $Y = FX + \varepsilon$. So noise is modeled by a random variable epsilon. So this is the noise so that is modeled by a random variable epsilon with 0 mean and the variance sigma epsilon square.

So this noise is the 0 mean and the variance of the noise is sigma epsilon square that is the variance.

So this magnitude of variance represents the level of uncertainty about the underlying phenomenon. So I am repeating this I am considering the random variable epsilon actually the noise is modeled by random variable epsilon with 0 mean and the variance sigma epsilon square and the magnitude of the variance represents the level of uncertainty. Since we are considering 0 mean random variable epsilon so that means the expected value of epsilon is equal to 0 and the variance of epsilon that is nothing but the expected value of epsilon square and that is nothing but sigma epsilon square.

So that is the variance. Now I want to find a function f hat.

So suppose I am considering I want to find this one to determine. So I have to determine the function f hat such that it is as close to the true function f the true function is f and I want to find a function the function is f hat such that it is as close to the true function f and this f hat that can be learned from the training data. So you can see this I want to write to find a function f hat such that it is as close to the true function the true function is f. So I have to find a function f hat such that it is as close as to the true function f and it can be learned from the training data. So this function f hat is learned by minimizing a loss function.

So this f hat is learned by minimizing the loss function by considering the training data. So what is the goal? Goal is to bring predictions. So predictions actually from the training data prediction of training data goal is to bring prediction of training data as close to as possible to their observed value.

That means the mathematically I can show this y should be approximately equal to the predicted value is of $\hat{f}x$.

So that is the goal. So goal is to bring predictions of training data as close as possible to the observed value. So observed value is y and the predicted value is f hat x. So for this I am considering one loss function and that loss function is MSE that is the mean squared error. So mean squared error we are considering that is a loss function. So the MSE is defined like this MSE is equal to expected value $y - \hat{f}x^2$.

So this is the expression for the MSE that is the average squared difference of a prediction. The prediction is f hat x from its true value y. So that means what is the meaning of the MSE? The average squared difference of a prediction the prediction is f hat x from the true value y. So that is the meaning of the MSE the mean squared error. So now how to define the bias? So bias is mathematically defined like this.

So bias f hat x is equal to expected value $\hat{f}x - fx$. So mathematically the bias is defined like this. It is the difference of the average value of prediction over different realization of training data to the true underlying function f x for a given unseen point x. So that is the meaning of the bias. So we are considering over different realization of training data that means we are considering different training data sets.

So I can write it is the difference of the average value of prediction. This prediction is actually the over different realization of training data to the true underlying function f x for a given unseen point unseen I can say unseen test point test point x. So that is the meaning of the bias. So now let us define the variance.

So move to the next slide. So now how to define the variance? So variance I can write f hat x that is nothing but the $E\hat{f}x - E\hat{f}x^2$. So this is the definition of the variance. So what is the actual variance? It is actually the mean square deviation of f hat x from its expected value. So what is the expected value? Expected value f hat x over different realization of training data that means we are considering different training data sets. So this I can write like this variance is the mean square deviation of $\hat{f}x$.

So mean square deviation we are calculating. So mean square deviation of $\hat{f}x$ from its

expected value expected value what is the expected value? Expected value of f hat x that is the expected value over different realization of training data. So that is the meaning of the variance. So I want to determine a formula that connect the MSE to bias, variance and the irreducible error. So that means I want to decompose the MSE the mean square error into bias, variance and the irreducible error.

So the expression for this expected value expected value $y - \hat{f}x$ is equal to expected value bias f dash x square plus expected value variance f hat x plus sigma epsilon square. So this is a very important relationship. So what I am considering? I am decomposing MSE the mean squared error into bias, variance and the irreducible error.

So in this expression you can see the first term is the bias,

the second term is the variance and this is the irreducible error.

So this is a very important relationship. I am decomposing the MSE into bias, variance and the irreducible error. So this term is nothing but the MSE the mean squared error. So in this expression if you see the first expectation we are considering the first expectation in the term expected value expected value y minus f hat x whole square is over the distribution of unseen point x. So I am repeating this the first if you see here I am showing actually two expectations. The first expectation in the term expected value expected value y minus f hat x whole square is over the distribution of unseen test point x.

The second expectation over the distribution of training data and the random variable epsilon. So that is the interpretation of this and this is a very important point. So I am repeating this the first expectation in this term is over the distribution of unseen test point x while the second over the distribution of the training data and the random variable epsilon. So this is the meaning of the these two expectations in the first term. Here you can see the mean squared error can be decomposed into three terms.

One is the bias, one is the variance and finally the irreducible error. Now what is the proof of this how to prove this equation. So let us consider the proof of this bias variance decomposition. So move to the next slide.

That is the proof of bias variance decomposition. So that is actually the proof of the equations that already I have explained how to decompose the mean squared error into bias variance and the irreducible error. So MSE the mean squared error is nothing but expected value y minus f hat x square that is the mean squared error. So I can write like this expected value f x plus epsilon minus f hat x whole square . So I can write this in this way.

So suppose this is equation number one. What actually we are doing here you can see here we have this information $Y = FX + \varepsilon$. So just I am putting the value of y the value of y is nothing but $FX + \varepsilon$. After this let us expand this equation. So that is equal to expected value f x minus f hat x square. This is simple expansion expected value of epsilon square plus twice expected value f x minus f hat x epsilon.

So this is a simple expansion. So in this case one important point is when two random variables are independent the expectation of their product is equal to the product of their expectations. So this is the fundamental concept of the random variable. I am repeating this when two random variables are independent the expectation of their product is equal to the product of their expectations. So again I am expanding this one expected value f x minus f hat x whole square plus this expected value of epsilon square expected value of f x minus f hat x expected value of epsilon.

So I am getting this expression. So in this expression if you consider this term that is nothing but sigma epsilon square that is the variance. And since we are considering the zero mean noise so this term will be equal to zero. So this term will be zero. So finally this MSE the mean square error I can write like this $y - \hat{f}x$ is equal to expected value f x minus f hat x square plus sigma epsilon square.

So suppose this is my equation number 2 and this is my equation number 3.

So finally I am getting this expression. So we can see that this MSE the mean squared error decomposed to the irreducible error and the expected value of f x minus f hat x whole square. So that means I am getting these two terms if I decompose the mean squared error. So first term is this and another one is that is the irreducible error.

So I am getting the equation number 3.

So now I am considering this part. So how to expand this part? So how I have to expand this part to expand I have to expand this part. So let us move to the next slide how to expand this part of this equation. So we have the expected value from the previous slide f x f hat x square. So we have this expression and how to decompose this expression into bias and the variance.

So that is equal to expected value f x minus expected value f hat x minus f hat x minus expected value of $\hat{f}x$.

So I can write this like this I am getting equation number 4. So how to actually get this one? This actually I am getting by subtracting that is a subtract and add expected value of f hat x. So I am getting this like this. So subtract and add expected value of f hat x.

This is very simple. After this move to the second step. So we are now expanding this one this expected value of f x minus f hat x square. So now we are expanding this one. So it is expected value expected value of f hat x minus f x square plus expected value f hat x minus expected value f hat x minus 2 expected value of f x minus expected value of f hat x f hat x minus expected value of f hat x. So this is the equation number 5.

So that means we are expanding the term inside the square. So I am getting the equation number 5. After this you can see the expected value f hat x minus f x square plus expected value f hat x minus expected value f hat x. Minus twice f x minus expected value f hat x. Expected value of f hat x minus expected value of f hat x.

So this is the equation number 6. So in this expression if you see here this is the actually the bias that is the bias of f hat x that is the bias. And if you see this term this term is nothing but the variance of f hat x. So bias is the expected value of f hat x minus f x whole square that is the bias. It is a constant since we subtract f x from the expected value of f hat x. So here f x is a constant and the expected value of f hat x is also a constant.

So that is why the bias is a constant. So just at this point I can write here this bias is a constant bias is a constant since we subtract f x f x is a constant from the expected value of f hat x which is also a constant. So the bias is a constant since we subtract f x is a constant from the expected value of f hat x that is also a constant. So therefore applying expectation to squared bias does not have any effect. So that is the meaning of this. I am repeating this so applying expectation to squared bias does not have any effect.

So the meaning is expected value we are taking the expected value of f hat x minus f x square is equal to expected value of f hat x minus f x square. So this is the interpretation of this. In equation number 6 we are able to pull f x minus expected value of f hat x out of the expectation because it is a constant. So that principle we are applying in equation number 6.

So after this again we are going to expand this one. So move to the next slide. So from the previous slide we are considering this term. So from equation number 6 I can write bias f hat x square that is a squared bias plus variance f hat x minus twice f x minus expected value of f hat x expected value of f hat x minus expected value of f hat x. So this is the equation number 7. So equation number 7 actually we are considering because of the linearity of expectation that principle we are applying the linearity of expectation.

So we are getting the equation number 7. So finally I am getting the expected value of f x minus f hat x whole square that is equal to bias of f x hat that is the squared bias plus

variance f hat x. So we are getting this expression. So this is the equation number 8. So we see in the equation number 8 that is the expected value of f x minus f hat x whole square is the sum of squared bias and the variance.

So I am getting the equation number 8. Now we can combine equations 3 and 8. So if you see the equation number 3 in my previous slide and equation 8 already I have derived. So from these two equations I can write expected value of y minus f hat x is equal to bias that is the squared bias plus variance plus irreducible error. So this expression is only for the given test point x but usually we have a set of test points. So if I consider a set of test point then this expression I can write like this. So expected value y minus f hat x is equal to just I am giving that expected value because I am considering a set of test point.

So that is why I am taking the expectation bias plus expected value variance plus irreducible error. So I am getting this final equation. So this is the proof of bias variance decomposition.

So this is nothing but the mean square error but we are considering a set of test points and this is the bias term.

This is the variance term and this is nothing but the irreducible error. So you can see here that MSE can be decomposed into bias variance and the irreducible error. So that is the proof of bias variance decomposition. So this is the proof of bias variance decomposition. So in this class I explain the concept of bias and the variance and also one important thing is the bias variance decomposition.

The mean square error can be decomposed into bias, variance and the irreducible error.

And already I told you the high bias and the high variance it is not good. High bias means we are considering a very simple model and high variance means we are considering a very complex model. So high variance means overfitting and the high bias means underfitting. So we should compromise between high variance and the high bias and that is the bias variance trade-off. And finally I have shown already the mathematical equations for bias variance decompositions. So let me stop here today. Thank you.