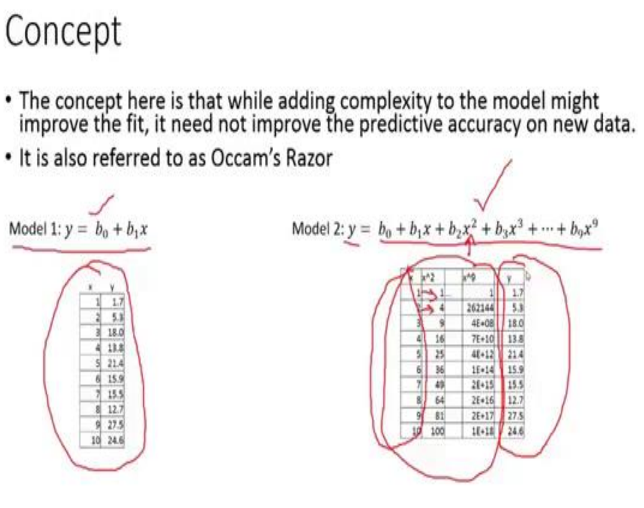
Introduction to Data Analytics

Prof. Nandan Sudarsanam and
Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 05
Lecture - 27
Bias-Variance Dichotomy

Hello and welcome to our lecture on Bias-Variance Dichotomy, this is a conceptual lecture. So, you must be right now going through lectures, where you are learning many of the machine learning tools and techniques and this is not one of them. So, we not using a new technique in this lecture, but we are introducing a very important concept that is machine learning and therefore, it applies to all the techniques I would say that you are learning and I would go in so for saying this is probably one of the most important concepts that you should understand in machine learning.

(Refer Slide Time: 00:54)



So, let us derive into what the core concept, the concept can be said in very simple sentence. The idea here is that while adding complexity to a model, while if you want to keep adding like say complexity to a model, you might improve the fit of the model. So, you might find out that you are better describing the data, but that need not improve the predictive accuracy of this model when you compare it to new data that you get and this concept is to for whatever type of model that you look at.

So, you can take an example on linear regression and you can add complexity to it by either adding more input variables or even with one input variable, you can add more complex transformations. So, for instance one way of just adding complexity to a simple problem, where you have one input variable, one output variable is that saying I am not only interested in looking it through the standard input variable, but I would like to look it as a polynomial. So, what is the model when you have x, x is the input variable, but you can also take x^2, x^3 and so on.

So, you can have a more complicated fit between y and x, because in simple regression you always see y = m x + c, but what we have $y = m 1 x + m 2 x^2 + c$. So, you can add complexity that way, you can add complexity by adding more variables. Now, this is again not confined to the regression any more. Almost any method that you take, you will typically find that there is some way of getting more and more complex.

So, for instance you might have already covered trees, classification regression trees. You can add more complexity by, you know creating more and more and more branches to the point where you have such a complicated tree, you have such a large tree where each terminal node or leaf is a single data point that we are using in your training set. So, you have really the more complexity, but k need choices k could be assigned complexity. With neural networks something that you would learn the future, the number of layers in the neural networks can make the complexity.

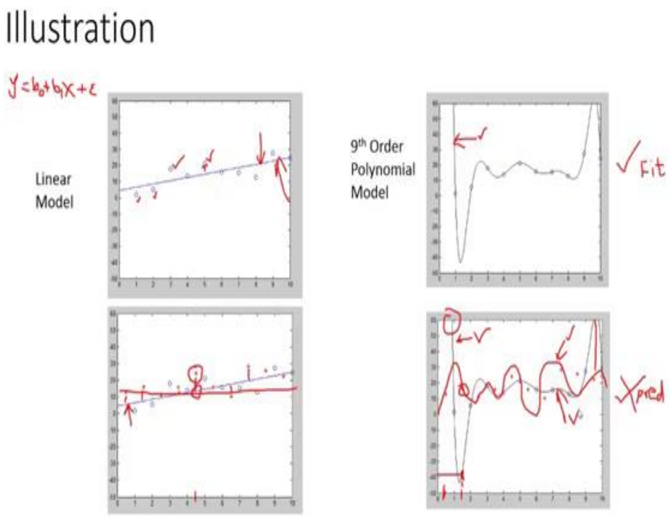
And this idea of complexity would become more clear as we, you know talk through some example. But, the idea is that you can make the model more and more and more complex, the model that you are going to use to create this relationship between input variables and the output variable, nothing become more and more and more complex and the more and more complex it becomes you will do a better job of fitting the data that you have.

But, that does not mean you are creating a better model and the answer is might not, because while in might fit a data better that you have, tomorrow we need to predict using this model, you might not do a better job of predicting and we are going to see how that can happen that can possibly happen. Now, this core concept in machine learning is also sometimes referred to as Occam's razor that is more of mathematical concept and it is definitely used permanently in machine learning, And the idea there is not which is that if there are two models with equal predictive accuracy, then you prefer the model that is simpler that is less complex.

So, that is not go back, but you see how these two highly related concepts, but the concept of bias variance dichotomy, you are essentially questions saying that I can add more complexity to the model and it will look like it is doing a better job of fitting a data, but am I getting better predictive. So, let us actually you know understand this through an example in this particular model in the two tables that have shown you here is model one which is your good old linear regression that you know and here is the data set.

So, use the same data set which is the same x on the right hand side and the same y, but I say that I do not necessary believe this is the right model that is model one is the right model. What are the relationship between x and y were more complex. So, I added an x^2 term and x^2 is nothing but it is really simple, it just I take x and i square it and I created new column and like that I keep on adding columns still x^9 and finally, I have all these as potential input variables as described by this model. So, y is some function of all these parameters and I do a multiple regression on that. So, how does it work up?

(Refer Slide Time: 06:14)



So, here is my linear model, here are the data points that you saw before and here is the that line that you see is the fitted line that goes to these data standard and so this is what you get, how this what are you... So, the polynomial the one that I created before is actually an ninth order polynomial. So, how does that will work?

Well it turns out that it does a really good job of fitting all the data points as we can see this ninth order polynomial described by this black curve goes through every single data point and that should not be surprised, you have ten data points and you using ninth order polynomial fit it there is enough flexibility in the model, in the coefficient is enough complexity in the model where you do not you can actual go through each data point, here in the first model it is a straight line, even if it wanted to it cannot even if I had the flexibility to put this line wherever you had I can move this line up and down I can rotate this line, but the best job that I can windup doing is actually the line that you see on this screen.

So, I did a simple linear regression which does try to you know fit as many data points as possible and this is as best as it could do now with the ninth order polynomial, where it tries to do that it does really good job it fits all the data points.

See you sitting there knowing like walk, so may be my system is ninth order polynomial, but it is not I let you learn a secret, this actual data between x and y was created by an actually a linear system with some amount of noise. So, you truly equation the true relationship which because I am the God out here, I am the one whose actually creating these data points. So, I have this oracle I am letting you on the secret that I actually created that this data points through some $y = (\beta_0 \text{ or } b_0) + b_1 x$ and not x^2 and x^3, but then was some error, some noise like any regular system.

So, given this let us it looks like this the ninth order polynomial still developed better job of fitting, the fit is great. But, let us look at how these two models compare will they have to predict that another occurrence, here is the fitted model, the same graph has above in the linear model. But, I created a new data set I created new data set and I see how well my line does as a job predicting what is going to happen next's and the answer is it does not do too bad. So, this is the predicted line, the blue line is predicted line that I got from my training data.

Now, I am go get new data it looks like I miss targeted couple of times and I probably should expect that given what I know right now that there is some amount of noise in the system that is irreducible. But, both this basically means is that tomorrow when you come to me with saying that hey 4.5 what you predict I say I am going predict this value and in reality I windup seeing this value that the line in red is what I windup seeing in the field when I makeup predicts in the prediction make is where I windup, what I windup using as the blue line.

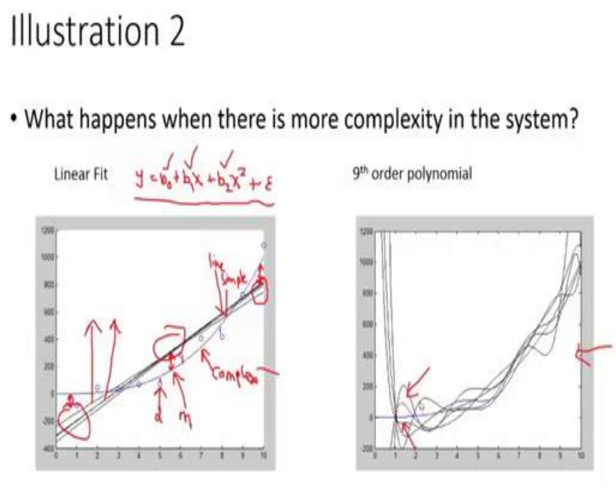
Now, what do you see in terms of predicting with the ninth order polynomial, you actual do quite terribly bad I mean look at this data point. So, this data point where it is 1.5, my prediction at 1.5 would be something close to – 40 or - 39 or whatever. So, add 1.5 if I were to use this ninth order polynomial as might fit it model I would be predicting minus 40, but look it what I actually got, I got + 50. So, I was almost of by x 55 units, whereas in that much with the linear model and you are going to see the same kind of we are practitioner. In fact, while at something like 0.5 this model just goes through the roof I cannot even fit it inside the graphs.

So, clearly the ninth order polynomial while is doing a great job of fitting does it very bad job of predicted and if my goal more often than I should say bad, but I did go more often than now is to really come up with good predicted accuracy. So, the most machinery from judges not trying to fit a model to the data that does not buy you the much, but you want to pay to model that can be generalized to other situations.

So, tomorrow when you get a data set, because what you going to do with the model here either going to predict or either you going interpret the model in either case you need to acknowledge that what you have the data that you have is nothing but, a sample and if you take another sample I if it turns out that you would have told a completely different story they may be the way doing things is not really correct, my whole points is that for instance, if you had seen the reds star data in the linear regression model you might have created a slightly different line may be the line would have looked little bit like this.

But, think about what you might done with the complex polynomial you might have created a you know completely different polynomial function that look than again will go through all the red data point, but if for one sample you create one story and for another sample you create completely different story can may be the way doing thing is not a really accurate. So, here is what we shown you the system where it was truly a linear system and clearly how using a linear model made more sense than having a more complex model.

(Refer Slide Time: 13:02)



Now, what happens when there is more complexity to this system, let us for instance say that my true model was quadratic and that is what I going show here, I going take up quadratic model and then see what happen when I try a linear fit. So, here is quadratic model this is the truth that the machine learning algorithm does not know that in the world I will never know that true system is quadratic. The only thing that I have is data show you the data I am just setting you in a secret that for today's exercise I created this data. So, this is the data, this is the model that I created this data using this model and adding some amount of noise or uncertainty.

So, the model that is creating is data has actually you know is more $+\beta_1 x_1$ that it because it only 1 x and $\beta_2 x^2$. So, I am using some model like this which some b_0, b_1, b_2 + some amount of noise. Obviously, if you knew the this was the model then this is the model you going to try a fit, I mean if you knew this is the model then this is the model you should use with that truly known b_1, b_2. So, you do not even need to do any kind of machine learning statistics excise.

But, sadly you are only given the data and you not told which model it is, now let us see what happens if you hand this data and then you try to fit a line this is one fit. But, the kind of give you a feel for what happens when you do this many times I generated another set of data points using this model. So, I have shown you only one set of blue dots, blue small mini circles, but effectively it will another set of blue mini circles and then fit a line, I fit another line.

But, one thing you should note this is look all these lines are in general more or less they trying to the same job and in general they wind up feeling chronically in certain cases they always windup underestimating in this region, because their always under the truth, they always windup over estimating. So, each time I do this exercise it looks like chronically of in certain areas, but I am fairly consistent each time when we do this exercise I windup kind of creating the same line.

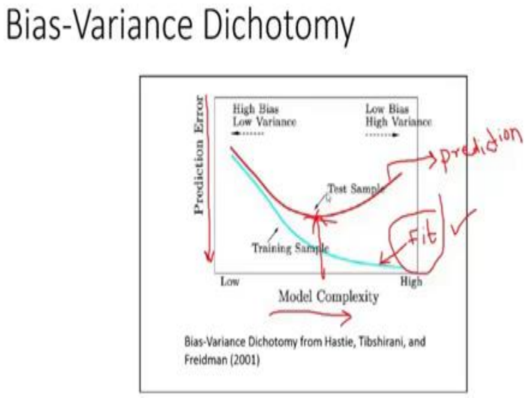
Now, what happens when you have a ninth order problem, what happens is you still not doing to great and the reason you not doing to great is because this is a quadratic system and when you are trying to fit something so complex. So, you still over shooting a lot, but there are couple of things note this one in general as expected as shown in the previous slide you not always telling the same story, one time you telling one story the next time you know you predicting vastly different this kind of extreme variance from one kind of prediction to the other is not seen in the linear fit.

But, take another look you are not; obviously, chronically off in certain areas, yes this is only five such fits, but imagine if you had five thousand such fits. Even if you had five thousand fits in the linear model, you will always be overestimating some regions, you always be underestimated some regions in that is called bias, whereas in ninth order polynomial the idea is that yes there so much variability, but if you were to do many, many, many fits that on average you might not be off from the blue line and that useful in concept it is not useful in practice, because in reality you are going to get only one data set and you are going to try one fit.

So, if it is off, it is off whether it is because of whatever reason, but it really helps to understand why it is off, here in the linear fit it is off because you might be chronically always going to be off. Because, you are trying to fit very rudimentary model, very simplistic model for something that little more in reality more complex, the model the reality is more complex, so this is more complex because of it is quadratic and the model you are trying fit is too simplistic, because it can only here model can many be a line. So, it is a line which is you know simple whereas...

So, here what you can to witness is a lot of a some region you are going to be always off whereas out here in the ninth order polynomial this is so much variability. Because, you are just getting fooled by the pure noise, the same thing that you saw in the previous equation, nothing is really in the previous line nothing is really change, you still trying to fit something that is a model that is overly complex to a system that is not that complex we went from a linear data source to quadratic data source, but that is still does not in ninth order polynomial, so this is a lot of variability.

(Refer Slide Time: 18:34)



And this is the point that is get captured in what is offend, what is really described as a bias variance dichotomy. This graph is taken from the ESL book Hastie and Tibshirani, and it captured what we kind of try to illustrate in the last two slides, which is as model complexity goes increases it looks like you are doing a very good job of fitting the data this is nothing but, the fit. So, you take the data and look here error keeps you want very low error.

So, as you keep on making a more and more and more and more complex model, you are going to go through more and more data points. But, at some point your ability to predict this is prediction on new data, the red line is prediction on new data, the blue line is the fit on the data that was given to you for training. So, it is call the training sample and the prediction is done on the test sample, you prediction keeps on getting low to a sweet spot for model complexity and then after that it goes up and what is the sweet spot, you that sweet spot would be at a, if you are able to match the exact complexity of the system.

So, if you had a quadratic system for instance and you use a quadratic model that could might be an sweet spot. And so if have a complexity that is perfect you know and what we are going to do is, in the real world you don't know what the true model is. So, how do you figure out what this complexity should be and that is going to be covered in or lectures on validation. How do you validate a model, how do you fine tune some parameters of a particular model, again it can be K-nearest neighbors, it can be trees, it can be neural networks, it can be support vector machine, it can be a simple regression.

But, if there is a some kind of tuning parameters that there can increase or decrease complexity. How do you go about increasing and decreasing complexity is seen what works best and then choosing the appropriate one that we captured in validation, the lectures and invalidation when this lecture we want to create an appreciation that as model complexity increases, the fit becomes better, the prediction you need to find the sweet spot of model complexity and that is the core idea.

The other idea here is that yes when model complexity is low, you do not do too well in terms of here prediction, but the reason for that is because there is high bias meaning in that linear regression if you remember, you always of when you trying to fit that linear regression to the quadratic function, you will always often certain regions. So, you had a high bias in a low variance, what we mean by that I go back to the slide is you are bias in certain regions. So, these are regions were you have bias, you are always going to be off, because of the nature of you try to fit a line through a curve, but you have a very low variance, if the variability between many such fits is low.

So, on any given day you get any given data set it is not like you are going to come up with the completely new equation. So, that is what we call us high bias and low variance. Now, you step over to the ninth order polynomial here the bias is not that high, it is not like that can tell you that you are going to chronically be under predicting or over predicting in some regions. So, this has low bias, so I cannot tell you upfront that you going to be always off in one direction, but it is got high variance.

What you mean by that is on any given day if I take any given data set I take the sample and I try to fit this polynomial I do not know which line I am going to get, I mean this line predicts some astronomical high value out here, whereas this curve predicts some astronomical low value out here. So, this like such high variance on a given data set I do not know how going to be predicting and therefore, if I can have such high variance, it just means that I am probably not going to do very good job of predicting. I do not even know what I am going to be predicting.

So, that is the concept between the bias variance dichotomy, which is that when you go for lower model complexity you get high bias and low variance and we go for a higher model complexity you will get low bias and high variance and this is the very important concept to be internal analyst with respect to Machine learning and it this going to be extremely important even in terms of applying more advance techniques. So, I hope the bias variance dichotomy is clear.

Thank you.