NPTEL

NPTEL ONLINE CERTIFICATION COURSE

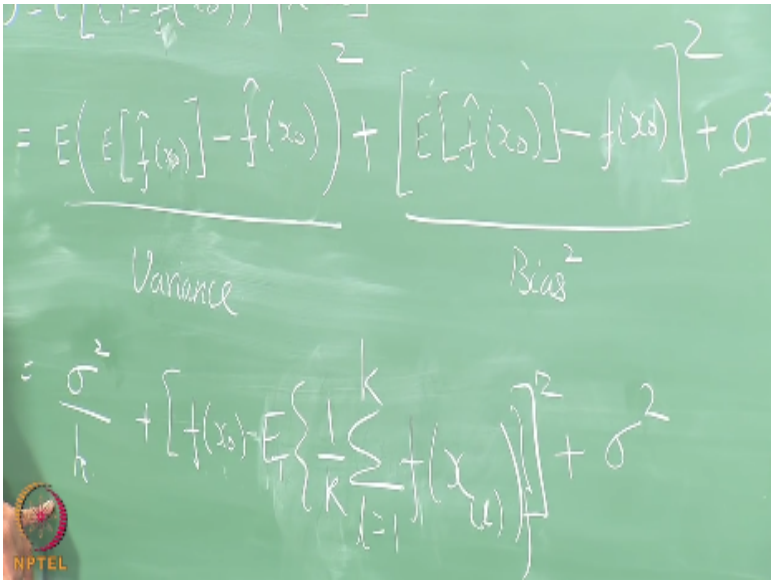
Introduction to Machine Learning

Lecture 11

Prof. Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

Bias-Variance

(Refer Slide Time: 00:15)



Bias-variance so I will explain what that is this is just a very preliminary introduction in this class and later on, as we progress we will come back to this at other points okay so let us start off with the assumption that and in many cases I will be looking at regression because it is easier to write but you can similar concepts you can also have for classification okay, so I am going to assume that your actual data is being generated by a system of this form right so there is a function f right which is what you are trying to learn about but the data that is given to you the Yz that are given to you are actually corrupted by some kind of noise okay.

Right so if you remember last class at least the last class I was teaching we were talking about a Joint Distribution over Y and X correct I said there is some Joint Distribution over Y and X you

do not know what the Joint Distribution is you are only given samples drawn from that distribution okay, here we are making a specific assumption about the form of the Joint Distribution I am assuming that there is some kind of an underlying deterministic function f here which is operating on my input X.

But then it is corrupted by some stochastic noise which we will call epsilon, and that gives me the Joint Distribution over X and Y when this process together gives me the Joint Distribution of X and Y and we are going to assume that expected value of epsilon is 0 variance of epsilon is some σ^2 so the expected prediction error at some point X. right is essentially the so we knew what the expected prediction error was right. So that was Y - f of X^2 but at the point X_0 and this conditioning on X_0 and.

This is my prediction error right, so it turns out that I can rewrite this expectation as a sum of three terms so what are the three terms the first term what is fine okay, so the first term is essentially the error that I am going to see by looking at the estimate that I will get from specific data instance from the estimate, that I will get as an expectation over the entire training the sample from which the training data is being drawn okay. So if I build a classifier multiple times if I build a classifier F hat multiple times okay.

So this is the expected output that I am going to get for F_0 X_0 right and this is the output I am getting for this specific instance of data that I have so that is one component of the error the other component is okay, look at the expected prediction I will make for X_0 in taken over multiple training instances, what is the expected prediction I will make for X_0 and what is the expected output I am going to get through output I am going to get it what is expected through output I will get is expectation of Y and what will be the expectation of Y.

In this case of X_0 and then there is a underlying error Sigma squared that just comes from the fact that I have a variance of Sigma squared in mind I am going to make any single prediction even if I am going to give you the output as f of X even if my output is f of X that will be an expected error or σ^2 because my why has that noisily right does it make sense so this term is typically called the okay, so this is this term is typically called the variance of the estimator f hat okay this term is called the bias this term is called the bias of the estimator effect.

So one way to think about it is the following right so f is my true function, so regardless of how much ever data I am getting whatever data I get right, regardless of whatever data I get I expect to make at least this much error from the true function f okay. So that is the bias and the variance is essentially given a specific instance of the training data okay, so what is the expected error I am going to make it so this is the bias this is the variance and that what about that part that is hopeless okay.

I mean regardless of how powerful your classifier is you cannot get rid of that σ^2 because that is inherent noise in the data okay, so that, that is nothing you can do about it okay so now by choosing your classifier appropriately you can trade-off between the bias and the variance so I will just for simple simplicity sake I will take the example of our K nearest neighbor classifier so all of you know about KNS right, so very easy to talk about bias and variance in Canaan's right.

So let us look at this so what do you think this variance term will be for the KNN case since I am looking at a prediction I am making over many, many instances right. And the specific prediction I make for one training set right what would that be if you know if you think about it the prediction I am making is essentially just the mean of K numbers right, so what will be the variance of that prediction from many, many different samples wrong it will be the base variance divided by the number of samples should have seen that in probability theory course if you have not okay I have a later I will be doing a session on statistics, okay and a little bit on hypothesis testing and so on so further at that point we will go back and look at it but just a this is the basic setup right.

So I have some distribution right I take samples from it as some distribution P, I draw samples from that distribution P and I try to estimate the mean and variance of that of that distribution through those samples okay, so now the variance of the estimates of the mean made from this samples is essentially given by the variance of the underlying distribution divided by the number of samples which you are drawing every time okay this assumes that you have drawn the K samples many times and they have made an estimate of what the mean will be right.

And this is the specific estimate of the mean this essentially the variance of that estimate right, so that is the σ^2 / K note about this, so this term right so this is essentially my expected prediction this there should be an expectation here over the training data so this is the expected prediction I am going to make right, so I am going to take the all the K nearest neighbors of a data point then

take the average of that that will be the prediction I am making so this is the prediction that I am going to make right.

And this is the expected value of Y that we have plugged in here is f of X_0 right, now let us try and look at this what happens when I change my K right if I increase my K what will happen to the variance it will decrease or increase decrease what will happen to the bias is an interesting question if increase my K what will happen to the bias increase why sorry it is a subtraction louder please, it is not just the subtractive part right so if I increase K essentially what is going to happen is I am going to start pulling in data points that further and further from X_0 right I will be pulling in data points further and further from X_0.

Therefore my estimate of f of X_0 is going to be an average of a lot of dissimilar data points, so the error is going to be higher so for a fixed dimension increasing K right is going to essentially pull in data points that are more and more dissimilar than the the query point right the query point was X_0 so I am going to go further out and therefore this will essentially become larger okay so as K becomes larger my bias increases in KNN and my variance decreases variance decreases just because I am taking an average of more data points right.

There is nothing to tell you that the average is correct digits that I am telling you the average will look the same even, if I change the training data right because I am averaging so many data points correct and this part of course we cannot do anything about this is the irreducible bias so what does this tell you. So last class we had this discussion about increasing K what does it do right what did we say when K becomes larger I did not say anything about it becoming more correct or not I said it will look more stable.

Why does it look more stable because my variance goes down right, so when I say that classifier is more stable estimator is more stable because the variance has gone down and also if you look at the classification surface that you will get right the separation surface that you will get it, will be a lot smoother if K is very large but I told you when K is 1 you are going to get lot of isolated islands of different classes and so on so forth and K for small values of K you will find that the classification surface is very complex like it is not like a linear thing or you can predict very complex functions also.

Easy to think of the complexity terms of the classification surface but function wise also you can think of very complex functions if K = 1 right, if K is larger and larger the function has to be smoother and smoother it cannot have rapid variations in the function that essentially means that when K becomes larger your function class becomes simpler right the kind of looks counterintuitive I am giving you a lot of K but then your function class typically becomes simpler because it has to have all the smoothness constraints on it right and as K is smaller then your function class can be larger.

So your regress or your classifier is more complex okay if K is smaller and it is less complex if K is larger and in general that is the case that if your classifier is more complex your variance will be higher that bias will be lower okay, if your classifier is less complex their bias will be higher and their variance will be lower okay. So this is usually the case and this also lets us understand why k-means does not perform that well in high dimensions why is that once more sorry one parameter gets changed a lot no see this is even simpler than that right, so take the bias term right so even for low values of K right.

So if you take a very high dimensional data you can with a little bit of analysis show that the very high probability, the nearest neighbors will be far away from the any query point it can take any query point right and draw a ball around it the ball is more likely to be empty then filled okay, so for the radius of the ball depends on the dimension it becomes larger and larger as the P becomes larger okay and so essentially it means that even for small values the bias will be high even for small K the bias will be high because the expected distance to the nearest point will be larger in the high dimensional space.

So not only will the variance be high because you have small k when the bias will be high now basically increasing K is not helping you in that case okay, so this disturb pretty rudimentary discussion on bias and variance at the tradeoff but I just wanted to give you a feel of that, and you have to keep this in mind later on as we are looking at every classifier that we will see right and specifically now we are going to go into linear regression right. So what about bias and variance in linear regression this linear regression have any bias must be right seems to be a very simple classifier.

Okay I will talk about that later but the point is yes so you have to be any classifier that you are going to be building or thinking about in the future right you will have to start thinking okay

what is the bias what is the variance okay is it the is it appropriate to use this classifier in this setting right and things like that.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in