NPTEL

NPTEL ONLINE CERTIFICATION COURSE

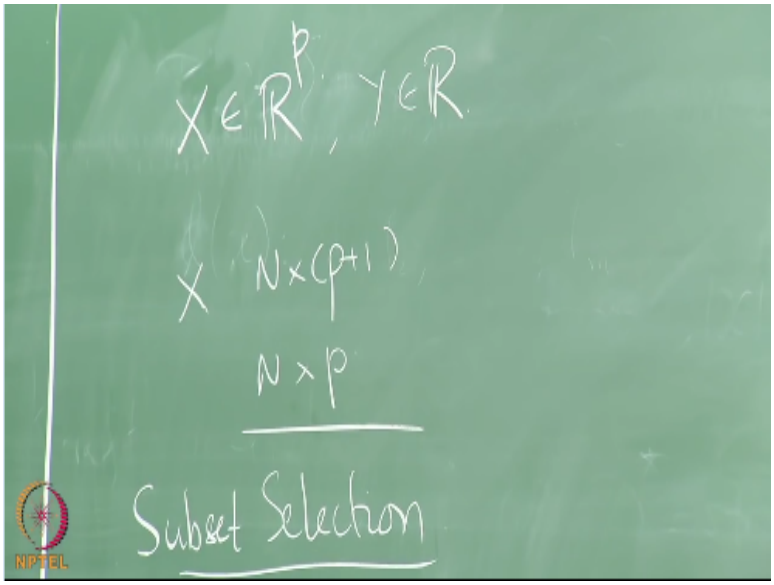
Introduction to Machine Learning

Lecture 14

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

Subset Selections 1

(Refer Slide Time: 00:24)



Right so we were looking at linear regression right, so we are assuming that the X is coming from R^P but I told you that it is not necessarily that it has to come from the real numbers and we talked about various ways in which you can encode the data and so on so forth and then if we assume that the Y comes from R and I told you depending on the circumstances so we will talk about the input matrix X okay which might be of N x P + 1 or N x P right so it will be N x P + 1 when we actually have an explicit intercept right.

So a β0 so term for the β0 term will have a column of once added to the data right the input instead of thinking of it as a P-dimensional vector will think of it as P + 1 dimensional vector and

when I do not have the intercept okay it just becomes a P-dimensional input right so that is that is the basic setup that we have and we are essentially looking at minimizing some squared error right so we looked at the simple linear regression we looked at the case where there were multiple inputs.

And we looked at how you can interpret that in terms of single variable regression univariate regression is essentially what we looked at in the last class. And so this class we will go on to look at a little more you know complex things that you can do with linear regression, so linear regression is great because it is so easy to solve right and it is very efficient runs very quickly and all that but there are a couple of drawbacks to linear regression so the first one is that if you remember I was mentioning in last class also that Lanier by doing this least squares fit.

You are actually getting a fit that has very low very low variance but it also has how much price okay did I say that in fact I think people are hallucinating so what I did say is that if you do least squares fit and if the if linear happens to be the right choice then this squares which gives you the 0 bias solution okay right so the least squares fit gives you 0 bias solution but the problem is the variance can be relatively high and it turns out that by not just doing the straightforward least squares fit by doing more tricks with their data with the with the models that we have we can trade off a little bit bias okay.

I am going to get a biased estimator for the fit for the line okay I am still fitting straight lines okay I have not done anything different I am still fitting straight lines but the fit I am getting will no longer be a bias-free fit but then I can reduce the variance a lot more by adding certain constraints additional constraints to the problem okay so essentially what I what I would really like to do is reduce the number of variables which I am trying to fit instead of trying to fit P + 1 if you can somehow reduce the number of variables what does is equal to it is equivalent to setting some of the β to 0.

So if I can somehow set the β to some of the β to 0 then essentially what it means I have lot fewer parameters that I am estimating right so the fewer the parameter set I am estimating the lower the variance still right but because I am now restricting the class of models that I am going to be looking at because I have to set some numbers to 0 so my bias will go up slightly right this is assuming that I am fitting straight lines you know the straight lines are the right things to do

but still my bias will increase a little bit right in this case of course that is always this residual bias because the language of straight lines is not powerful enough.

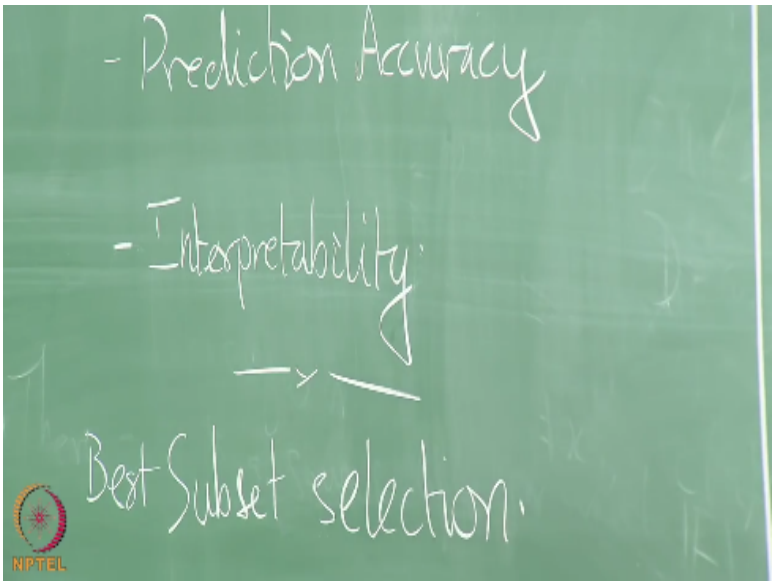
So that bias is still there we are not talking about that I am saying even assuming straight lines are the right thing to do if I am going to force certain coefficients to be 0 I am increasing the bias in the estimator right but the variance will go down because I have a lot fewer parameters that I am going to estimate okay, so that is essentially what we are going to look at and so what I mean by subset selection here is that I am going to select a subset of the input variables to use for fitting the line okay right.

So one thing is we can reduce the variance significantly and that is where we can improve the prediction accuracy of the model okay that is one of the reasons we would like to do subset selection there any other reason you can think of for wanting to work with the few smaller subset less computation yeah but it is a question of there you have to do more computation to figure out what the subset is and so on so forth yeah but less computation is one answer but there is another one I think I heard somebody say this somebody someone or rather speaks with you synchronously always see a trend you know no there is yet another major advantage I mean.

Yeah so related to this right, so there are variables which could potentially have high noise and so it will end up with a small coefficient right but then if I if I tell you okay here is this model m and it has like 135 coefficients and it becomes hard for you to even figure out what this means that instead of that if it is okay here is this model you gave me a 135 variables but these are the 4 important variables that I need for doing a linear fit right.

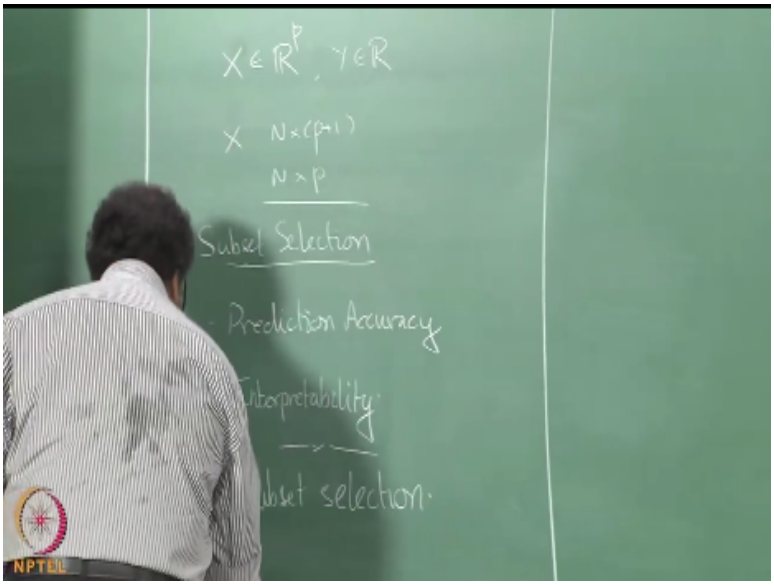
Now it is easy much easier for you to interpret what is going on right so interpretation is a very big component of any kind of data analytics that you want to do like ultimately what you are doing with machine learning is trying to understand the data right, so one of the things you would like to have this interpretability right.

(Refer Slide Time: 07:00)



So, the first one is to okay this is interpretability and that is prediction accuracy right so there are many ways in doing this the simplest kind of approach is essentially to take this literally right and try to select from subsets of features right, so why is that a simplistic approach so why is that a simplistic approach oh come on easy it is combinatorial now exactly.

(Refer Slide Time: 08:03)



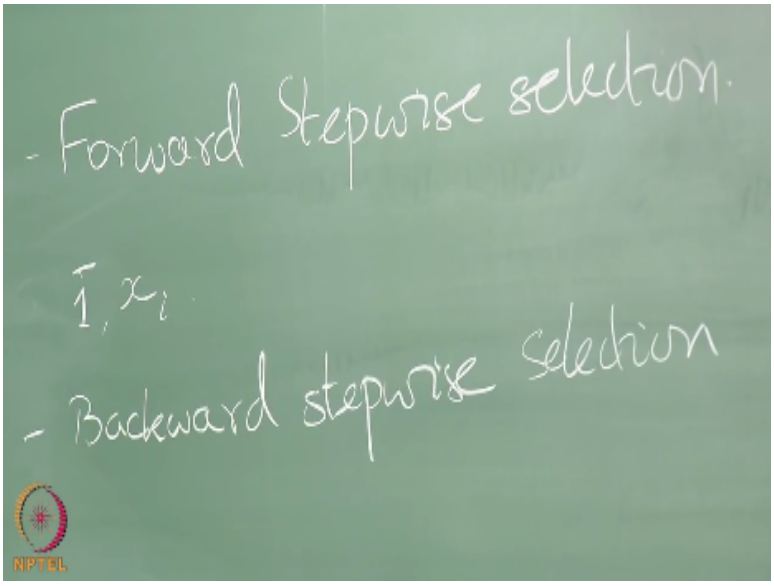
So this will just do subsets in a bit subset selection essentially I would say that okay here first pic subsets of size 1subsets of size two subsets of size 3 and so on so forth right, and it turns out that you can see yourself if you start playing around with some linear regression tools that what is the variables that go into the best subset of size one right basically the 11 best variable need not figure in the best subset of size 2 so it does not have a nice inclusion property.

So for it to have a nice inclusion property you need to have certain very nice conditions on the data set so in general it does not have this inclusion property right, so you have to just redo the whole thing again so it is not enough if you just do one say okay you cannot be greedy basically I cannot say I will do the first subset and then I just add the one best variable to it and then I will find the next one, so essentially you have to do a combinatorial selection so people have come up with a very clever ways of organizing this right.

And my in fact of using QR decomposition to do things more rapidly I am not going to go into details of that but then up to like 30 or 40 variables you can scale well right, you mean can hope to get answers in your lifetime kind of scaling right but if you go for much larger variables right much larger number of variables like many problems of interests like in text or image domains and things like that then there is no hope to do an exhaustive search but that is a base length which you can do people.

Actually come up with algorithms which do this kind of a subset selection okay, so there is one very interestingly named algorithm called leaps and bounds which does pretty efficient subset selection but just a more of a informational thing for you right.

(Refer Slide Time: 10:22)



This thing is forward step by selection so when he gets is what that is exactly it being greedy right it is just trying to do best subset selection by being greedy, so you start off with so what is the feature you for sure want to have all once okay you need the intercept right otherwise your line has to pass through the origin okay, so you need the intercepts you start with the intercept okay so what should be the coefficient for the intercept we already looked at it before now the mean of the Y's right.

So that will be the coefficient so we already have fitted that so now what you do this you start off with that variable okay, now then you add the next one okay let some xi add that as the next variable such that it gives you the best fit model oh the set they are already taken so you are not going to disturb all the stages the variables you have taken to in step k now we will add a new variable such that among all the variables I could add at this k plus 1stage this one gives me the maximum of improvement in the performance.

So how will I measure performance some kind of residual error on the test data right so that is how I measure performance I keep doing this, until a point where the error does not change much is there any other thing I can say so the residual is orthogonal to any of the other directions I

could add right there is one thing which we know of right, so we know that at the end of the right fit you get the residual will be orthogonal to the space spanned by the excess weight that means individually if I take any of the excess right the CD will be orthogonal to that individual direction as well, so when you find that none of the directions that are left have any kind of component along the direction of the residual then I can stop.

Or I mean that may take a long time to do because that might happen only when I have the full least squares fit so I can stop and the residual drops below a certain threshold that you can say okay I am happy with the prediction accuracy or getting l stop here. So there in many ways of doing that right so the other way is to do what will you do in this case I will start with the fit that has all the variables okay and then I will keep dropping one by one right so one thing to note is that you can do this.

If the number of data points is greater than the number of dimensions so then you can actually find the fit if p is > than n now as people pointed out last time the formula itself that, we are using is no longer valid right so because if the matrix will not be invertible so we will have to think of other tricks for doing the fit and so on so forth but forward stepwise selection you can do even if p is greater than n because I am anyway fitting one direction at a time right so it is fine I am adding one direction at a time I can keep going until I reach n directions at which point I should have a full least squares.

Fit okay yeah so when people actually even come up with all kinds of variants on this so one thing which you could do is think of some kind of hybrid approach where I keep adding and deleting directions right, so if I if you remember we talked about how greedy is not always the best way to grow things right, so you can grow up to a certain level again maybe then dropping one of the earlier dimensions will actually give you a slightly better performance could be right so that is one possibility and right they could do some kind of a hybrid version of this so one thing I should point out here, so you may think that forward stepwise selection because it is greedy is going to be much worse than best subset selection right.

So it turns out that in many real cases many real data sets that you work with right the greedy selection is actually not a bad thing to do right in fact, so much so many statistics packages like are but a little bit to do this right so you would not find this in many of the machine learning tools like so they would not have this kind of a forward feature selection and things like that they

have other ways of doing feature selection which we will talk about later right but then statistical packages actually have this stage wise edition of witches because they seem to work well on a variety of data sets okay.

IIT Madras Production

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

www.nptel.ac.in