NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

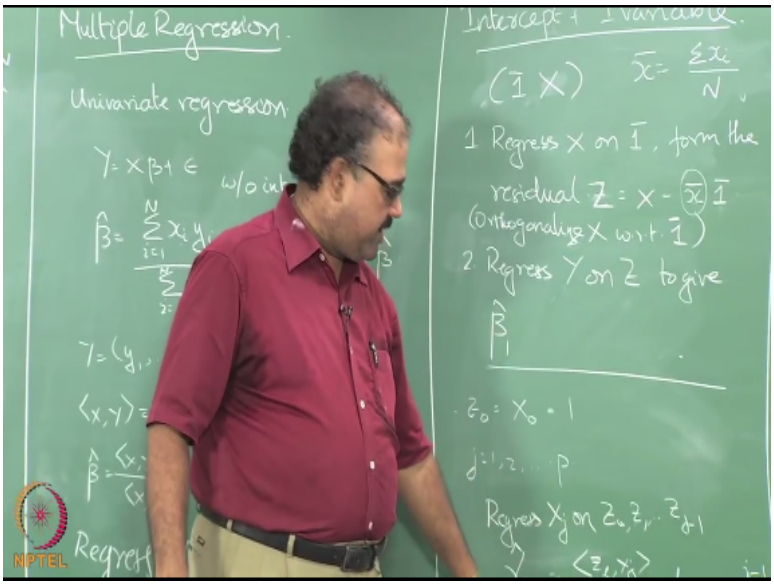
Lecture 13

Prof. Balaraman Ravibdran
Computer Science and Engineering
Indian institute of technology

Multivariate Regression

So far I have really not worried about the fact that we have multiple dimensions in the input space, that we just had this way of handling it but then if you actually look at how statisticians typically present linear regression they will start off with a univariate regression they start off with one input variable and one output variable right, so one independent variable and one dependent variable, so the independent variable is the input variable a.

(Refer Slide Time: 00:54)



So whatever we have looked at so far is usually called multiple regression, we will still typically start off with univariate regression people usually start off with univariate regression because it is easier to analyze you can derive a lot of intuition into what exactly is happening with the regression right, in fact if you think about it this picture I drew for you is with you it is univariate

regression with an intercept right, so that there is a column of once and then there is one other variable that is all right.

This is essentially univariate in the regression with the bias term right, so this kind of mean you can very easily develop all kinds of intuitions and also analysis very clean and more importantly you can understand multivariate regression okay by a series of univariate regressions, so let us look at it very quickly and then we will see what happens all right. So this is the basic model that we have but here we are going to assume that X is a number right X is a single number so it is a single vector now right, so my data will be of the form some X Y that is not a that is not a vector is just a simple X.

So without intersection, so why is it called the intercept, so the constant value you add is what the value of where it will cut the y-axis okay that is why it is intercept, so I have no intercept that means there is no β naught here okay so now this 1β is given by right essentially our original this case for a univariate case I am going to denote by r_i the residual error that I am making, so I made the prediction right. So x_i β hat is the prediction I am making, so y_i is the actual output. I saw in the training data, so $y_i - x_i$ β is the residual error okay that is before I am going to define okay people.

I hope are familiar with the inner product notation of this form essentially okay, now can you tell me what β height should be the inner product notation somebody said something I had a small squeak somewhere, this is fairly simple, so one thing I just point out in passing here right I am not going to cover it you guys are happy to I mean I we are free to read up in has T tips Ronnie Friedman later chapter.

Which we will not get to but the fact that I am using inner products here okay should tell you that I can apply the ideas of linear regression on any inner product space okay not just in real number space okay, so yeah I will leave it at that gives you a good generalization. So what we are doing here I will call this as, so regressing Y on X okay and we get the coefficient β hat so this is remember what we are talking about so far is a univariate regression at no intercept nothing.

Suppose that your columns are all orthogonal not only are they independent they are all orthogonal okay you can little bit of thought you can convince yourself that each β, so β 1 β 2 β 3 and so on so forth or just given by regressing Y on X 1, X 2, X 3 and so on so forth so β 1 this regression of way on X 1 β 2 is regression of Y on X 2 right why is that the case, so now my X 1

and X 2 are orthogonal they are actually the orthogonal basis an orthogonal basis for the P dimensional space the p + 1dimensional space I am talking about and each coefficient that I am going to get essentially would mean will be the intercept on each of the individual dimensions.

The projection on each of the individual dimensions because they are orthogonal in the lowered lower space right, so that is easy to convince yourself, what is interesting is what happens if the XS are not orthogonal, they are independent let me say they are still spanning a P plus 1 dimensional space right but they are not orthogonal, so what do the coefficients represent in that case so that is essentially what we are going to look at okay. So we will start off by taking one step at a time look at the intercept plus one variable.

So far I said that is one variable without intercept okay now I am adding the intercept, so what will be what does it essentially mean for us, my x becomes 1comma X right, so here I had just a single value x right I could just write it like this, now my vector x becomes so I am going to consider is a column of once and my original vector x okay this is my new vector that I am going to consider. So what I am going to do is the first step I am going to do is tell you about that, so this upper case exists the actual column vector excess of consists of x okay this is the actual input I am going to look at so let me define $\bar{x}$ as the average of the all the inputs I have seen all the inputs I have received is my training data.

So a regress x on I write and form the form the residual, so what will the residual be but in this case what would it be I am saying because I am regressing on one all once right, so if all one system only input variable I have write what should be the best possible prediction I can give $\bar{x}$ right so $\bar{x}$ is the only output I can give that will be the one that minimizes the prediction error right because I am looking at squared error the output should be $\bar{x}$. So my β hat will be $\bar{x}$ in this case right.

So the residual which I will denote by is by Z will be okay so this $\bar{1}$ is just to indicate that it is a vector of once okay, so this x is the x is a vector so this $\bar{x}$ is a scalar value which is the average of all the inputs and $\bar{1}$ is the vector of one so that this gives you the residual okay does it make sense, this is the vector of residuals, so I usually put the bar on the x and the z then the middle to differentiate it from two right but sometimes it looks like lower case then, so is it fine let us adopt the convention that even if, I put the bar there it is still uppercase dead okay because either way I will have to be very careful about distinguishing 2 and Z.

And things like that so this is an upper case and if I really want you to look at lower case it and write that okay, so the second thing I will do is now regress like, so this is one univariate regression this is another univariate regression, so essentially this tells me okay. I have taken the average value out of the input variables because the average value can be used to predict the average output if, I have taken out the average value so whatever is left okay is the individual variations on the data point okay and use that to predict my output value away.

So this essentially means that so given that there are two dimensions 1 and x okay so the β 1 had tells me what is the contribution of x okay after I have used 1to explain the output okay, so given I have taken care of one already what is x, so if you think about what I have done here this is essentially making it orthogonal to the 1 vector right, the Z vector is essentially the x vector right the component of x that is orthogonal to the 1 vector going back to how we did the univariate regression? That is what we have done here so this is remained new people of anything already looked at gram smith, the people have come across gram-smith orthogonalization right, this is essentially something very similar to that.

So I start off with $\bar{1}$ and the x_s the 2 vectors that span some space now I am orthogonalize I am essentially giving you an orthogonal basis now one is one vector and z this the other vector but together they span the same space that was spanned originally by 1 index it except that they are orthogonal and people agreed with me earlier when I said that if the columns of x are orthogonal then they can independently do regression on each of those columns that is essentially what we are doing yet so I have done a regression on list to get me β 1.

So going back to our picture here, so essentially I had some x_1 I had some x_2 right, so what I did was I first rigorous x_2 on this and so essentially I am getting so that is my z right I am getting that orthogonal component to that, so now I have x_1 I have z and they are spanning the same well yeah they are spanning the that is in loop correct ratio metrically there you go that is 90^0, so they are essentially spanning the same space that z is a projection of x_2 imagine the plane is going into the board right so it does not look right to you but the plane is going into the board.

So z is actually perpendicular to x_1 and it is formed by projecting but by regressing x_2 on x_1 okay that is direction z and my virginal y which was going out of the plane, now I essentially project it on z to get the coefficient what right it does not matter see this is still going to project here okay

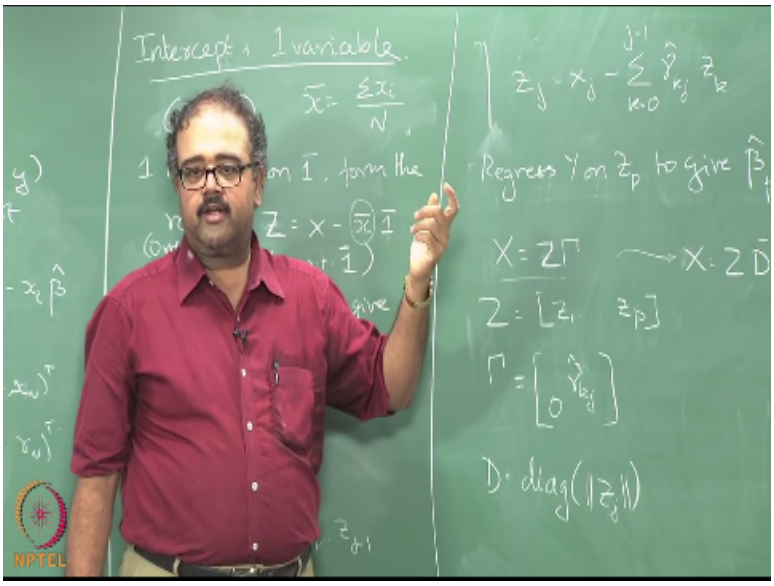
so earlier when I wanted the coefficient for x_1 and x_2 right I would have looked at this these points right here now I will basically look at these points that is essentially what I have done there is no change in the actual space.

So the same y hat is what I will get okay but the coefficients I will be using for representing the way that will be different, we can generalize this to P dimensions, so what will you get so j runs from 1 to P so I will regress x_j on all the previous z directions that I have determined right, so how would I written jet not will jet not I start off with z_1 would have been obtained by regressing x_1^0 is it not and then finding the residual, so that gives me so that is what we do here right, so I take $\bar{x}$ right I basically I regressed x on 1 bar and then take the residual and make that as z so likewise I will regress x_1 on z^0 take the residual and use that as z_1 ok then I will regress x_2 on is it not and z_1 ok then take the residual, so that is the γ coefficient which will so in this case it was $\bar{x}$ okay.

So if you think about it right that will be $\bar{x}$, so z δ inner product with z δ when z is all once is n right and the del inner product with x_1 when z it is all once is just x_1, so this essentially will be the summation of x_1 this will be just $\bar{x}$ the first case weight that is this is a comma it make sense or was it too quick yes no, so I am saying this $\bar{x}$ was delayed by just using the same formula right this was 1/2 this happens when I regress on the first variable right, so start off with z^0 is 1 I am regressing x_1 on z^0 when I regress x_1 on z^0 what do I get so z^0 inner product z^0 is n well just once right and they sum up all the ones.

So that is where dimension is y n that will be N and z^0 inner product x_1 will be summation x_1 right, so summation x_1 / n is essentially the average, so that is essentially what we had here okay so that is the same formula now I am generalizing to other dimensions so I am still continuing the loop here okay, so that loop that runs for J = 1 to P.

(Refer Slide Time: 21:14)



So for every j and that is how I derive my z_j okay, so I take the current coordinate under consideration x_j subtract all the previous dimensions I have basically looked at. So what I am left with what am I left with the orthogonal component of x_j okay orthogonal with respect to the dimensions I have already looked at, so in some what I am considering, so once I have done this for in some more I am considering it in some order right when they come to the pH dimension, so what do I get, so what is β P, had P now it is the actual coordinate sorry the actual coefficient that I will find for the variable if I had done the β the regression as we talked about earlier right, if I had done that if you estimated my β like this okay.

This is essentially what I will end up with okay but because of the process we have generated it we can interpret it in a slightly different way which is essentially β hat P tells you how much the pH variable contributes to the output given, that we have adjusted for all the other input variables given that we have adjusted for all the other input variables, how much does the pH variable contribute to the output, now we can go back and think about non independent vectors, if any of the variables is not independent right, so what happen in this case the receiver will be 0 and it essentially will be trying to find, how much would 0 contribute to the output okay that is not going to be a lot okay.

But it becomes even more interesting if my vectors are merely dependent but not exactly, so what will happen is if I subtract out everything else from that vector right, so think of it like this right this is x_1 that is x_2 okay this is the 2nd plane it is not like this is the plane right, so x_1 and x_2

are very close to each other there, so if I subtract out x_1 from x_2 I am going to get a small component that is orthogonal to this right, I am going to get something like this all right. Now if I take the inner product of that, so that will be a small number here, so this can become very large right.

So if my vectors are nearly dependent but not exactly so that the residual is NaN not zero exactly but close to zero then the whole thing can become very unstable the estimate whole estimation process can become unstable, so that is essentially what will happen if even if you eliminate perfectly dependent columns right there could still be possibility of getting numerical instability so to avoid all of these things people have come up with various techniques, that of course one of them is to essentially get rid of all the correlated or the nearly correlated columns right, but there are there are other ways of actually trying to get this to be stable okay.

So just an assay, so let z be the matrix that we create by taking z_1 to z_p columns okay, so I have done this set 1 to Z P in this elimination processor it is um in some order right, so I will take this z_1 to z_p columns okay and γ is the matrix where I store all my γ hat k_j there is an upper triangular matrix right, so for every combination k_j, I will have 1 γ hat value I will just put it in the upper triangular part and the lower triangular said I will just keep it as zero. So an upper triangular matrix there and you can think about it you can write the x_s z x comma x can be written as z times γ right.

So essentially the I am just stacking all of these things you have done together and we are writing it as is that times γ and so D is a diagonal matrix where the diagonal entries are the norm of the inner product of z_j with itself right, so the j entry or the j_{j8} entry in the D matrix would be the inner product of z_j with itself that is the norm of z_j, so I can write it like this, so this is called the Q_R decomposition of x right, so the thing about Q is it is orthogonal right. Q is orthogonal and R is upper triangular okay.

So this kind of a representation for the data matrix, so this kind of a QR representation of the data matrix essentially gives you some kind of ortho normal basis but Q is not just orthogonal is orthonormal way because I am dividing by the norm here okay, so it is so the product will be ones or zeros because they are orthogonal to begin with anyway okay I made them orthonormal so Q gives me an orthonormal basis and R is said upper triangular matrix that lets me reconstruct

the inputs x ok and this kind of composition is very convenient and it is used widely in other kinds of representation or transformation of the data and so on so forth.

IIT Madras Production

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

www.nptel.ac.in