NPTEL

NPTEL ONLINE CERIFICATION COURSE

Introduction to Machine Learning

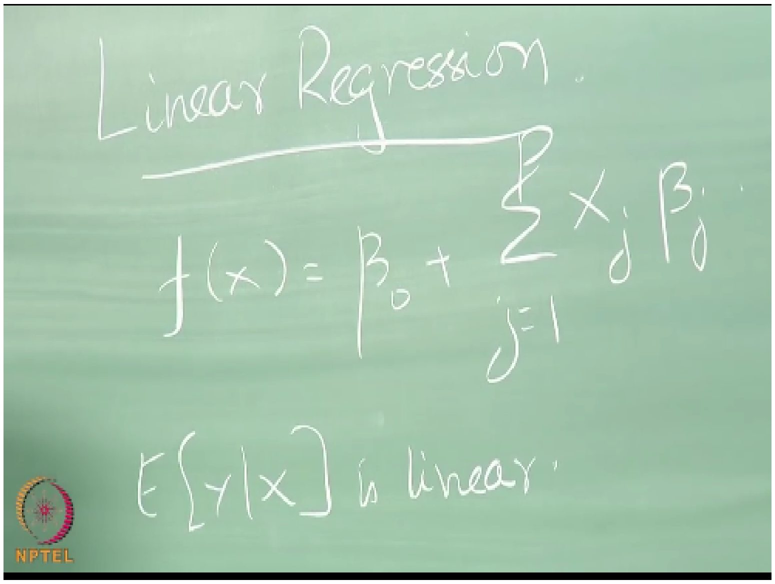
Lecture 8

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

Linear Regression

So there is a basic assumption that we had earlier.

(Refer Slide Time: 00:55)



Linear Regression.

$$f(x) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$$

$E[y|x]$ is linear.

So we are going to assume that the expected value of y given x is linear right. So that is essentially what, this is telling us right. So f(x) is so here is the expected value of y. there is some kind of a noise corrupted the training data that is given to you; the expected value of y given x is linear right. And so we had this discussion earlier, so the nice thing is linear regression is not as weak as you think right.

So x can be a variety of different things right, x need not just be real valued inputs. They are assume to be drawn from Rp right x are assumed to be drawn from some T dimensional real valued space. But they did not just be real valid inputs they could be any kind of encoding right.

We talked about basis expansions like we talked about basis expansion, which essentially is blowing up your input space by some kind of transformation of the input variables right.

So if my original data is x1 to xp. I can think of right that this as my input so that is basically basis expansion right I could also think of interaction terms right.

(Refer Slide Time: 2:26)

I have to think of more complex transformation, x could be qualitative inputs as well. What I mean by that hot cold right, tall short medium height how I would handle that. Weights has levels in the input right or it could be just red blue green mean it does not really correspond to any level. I mean young and old we can think of saying okay. Young is 1 and old is 2 and middle age is 1.5 or whatever right but what about red blue and green things like that.

Encode each color right, encoding how do you do the encoding. Yeah! you are right you have to do some kind of encoding how do you do the encoding. As soon as I said identify yourself and talk nobody talks. I could do some kind of binary encoding right, so I can think of saying that okay I have four colors right. So I will have two bits to encode the four colors right two bit 4 gets translated into two bits.

It turns out that that is usually too much of a compression in the encoding right and if you have four possible values this thing can take it is better to sometimes use four bits right. So it is sometimes called one of n encoding right so only one of those four bits will be one for any input right. So red means the first bit will be one blue means the second bit will be one and so on and so for right or sometimes it is called one hot encoding.

So this one of the inputs will be hot the others will all be cold right. So sometimes called one of N or one hot encoding, so you could take care of qualitative inputs like that categorical inputs also you could do that. And whatever you do right mean however you are expanded your basis or however you are encoded the thing. Finally the model you fit will be linear, it except that if your original dimension was 1 in this case right I had a one color input.

It could take four values now my input dimensions become 4. Similarly I had P input earlier now input dimensions has become this case depends. Depends on whether I am feeding in x1 to xp also right, if I only feed in the second order terms it is still P but the the class of functions I can model is restricted. And if I feed in x1 to xp as well as the squares then it is 2p and the class of functions I can model is become larger.

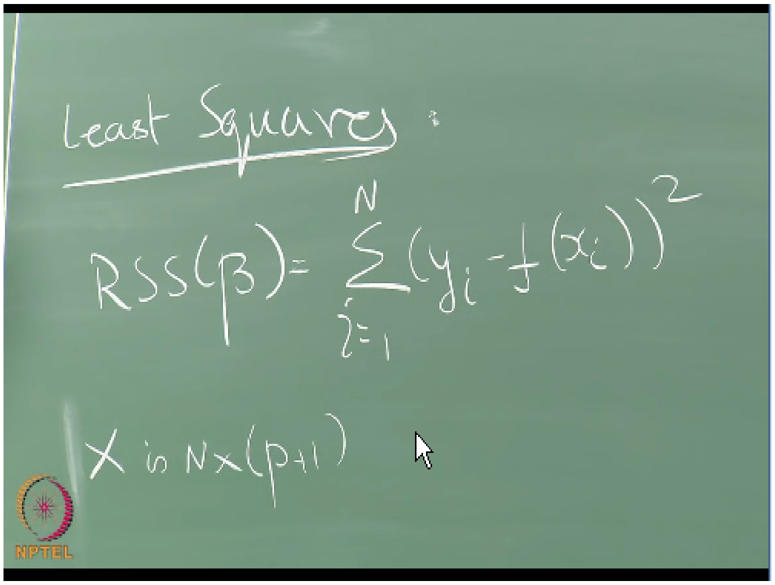
So that is basically the underlying set up right the model is still linear. person why is it two-bit including each to the degree 4 bit inputting is better than 2-bit encodings it okay. The point is so when I have two bits encoding so there will be the same input variable that gets activated for two different colors right suppose I am using red this zero one okay and blue is 11 okay. so that 11 will get activated for both red then blue.

And likewise when there is 1 0and 11 right. So so the same bit gets activated for multiple inputs okay. And that gives you some amount of interference in the training right. We can still train it with two bits you probably need a lot more training data to take care of the interference from one to the other right. When you have these kinds of 4 bits essentially you have independent weights modeling the influence of each of the levels.

So red that is one weight by weight I mean 1β here okay for red there will be 1β that will be modeling the effect of red. For blue there will be another beta modeling the effect of blue right. That way there will not be much interference between the variables right. So technically you can model it with two bits and get away with it is that you will probably need more data for the estimation. That is why I say in practice four bits is better.

Let us go and continue looking at this, so I am my training data so fonts are find so the training data is I am going to assume this of the form and that each xi. I am going to I am going to assume that I have n data points each of the form x1 y 1 to x n y n right.

(Refer Slide Time: 08:09)



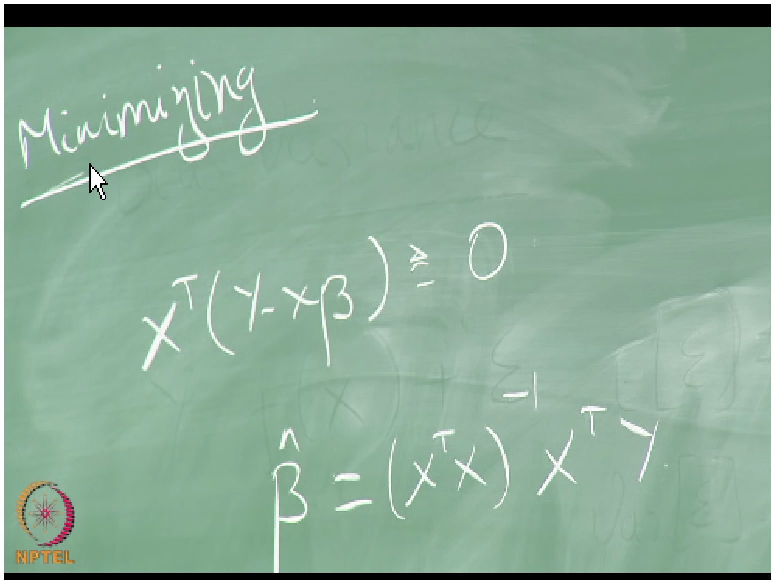
And the way we are going to fix this is using least squares. So we are going to translate this into matrix notation to show you some things right and in matrix notation when I write x at least for today it is N x (P +1) matrix, where the first column is all ones. So we have seen this already so it is N x P matrix where the first column is all ones okay. So I can write it like this in matrix form. Let me think about it that square thingy there becomes this.

Because f (x) now becomes just xβ. Would that be this all done the linear algebra tutorial you should be able to tell me what there is, so I am going to let you if you cannot see that immediately. I am going to let you work it out yourself okay and yourself okay. We should get really familiar with doing this kind of derivatives of matrices. Because we will be using this quite often whenever we write this kinds of error functions in terms of matrices ready to use this.

Intuitively see it is right but you just need to work out the math here. And at least this seems easy enough I am taking a derivative of this with respect to β and the only term where β P is $X^T X$ okay right. So if X has full column rank okay $X^T X$ will be positive definite. So it low it will certainly be invertible okay. So it will be and no it is not just invertible it will be positive definite and therefore we can assume that it is here maxima or minima.

Now anyways so if I equate this to 0 I will get an extremism point right I will get either a maximum or a minimum okay. So what would it be anyway think about it I am going to anyway minimize the error that that should give you a clue right okay. So I am essentially I have to set this to 0 if I want to find the minima of the error and this is going to give me right.

(Refer Slide Time: 13:26)



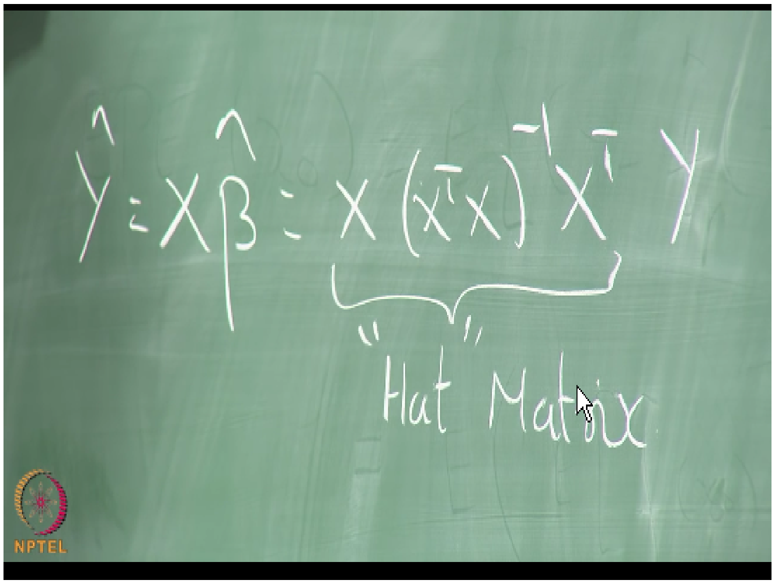
Minimizing

$$X^T(Y - X\beta) \hat{=} 0$$

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

This is all standards if you already know what the solution of linear regression is we saw that in the the last class and you should have revised things by now.

(Refer Slide Time: 14:16)



$$\hat{Y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{\text{"Hat" Matrix}} Y$$

They tell you that if you read in the previous whatever we have covered till the previous class and come the next class will be easier right. So we already seen the solution right and so if I put this together I basically get $Y^{\wedge=}$ X times X^T X inverse X^T Y. So this expression is sometimes called the Hat matrix you know why it takes Y and puts a hat on it right. So it is called the Hat matrix so hat essentially means what this is shorter and for estimates okay.

Hat so hats off shortens for estimates denotes that it is not the true quantity so why is the the true random variable and y hat is an estimate of the value of y okay. So this is essentially the estimator matrix right. So in that sense you can think of it as a hat matrix, so another way of thinking about it is the following. So what do we what can we say about Y right the vector Y not the output random variable Y.

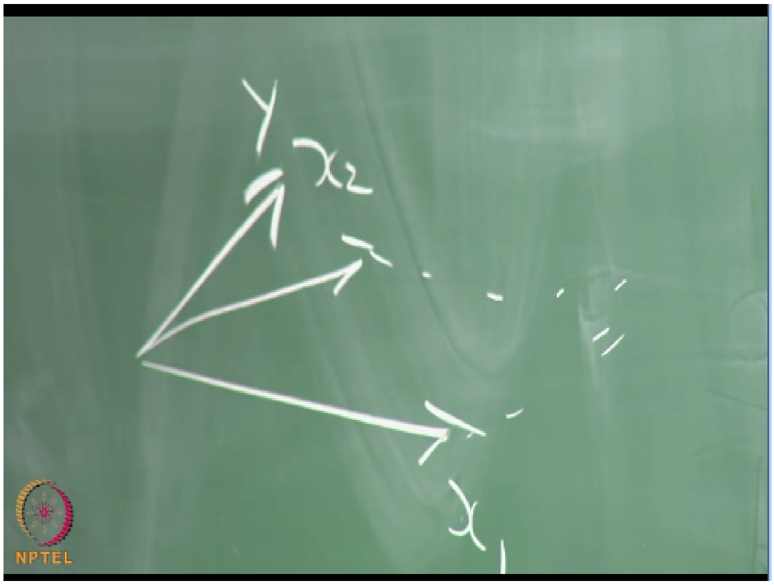
I am talking about the vector Y right, I should say that. So X is n x p +1what about Y N x 1 right. So Y is actually a point in right it is N x 1, so you can take the P + 1 columns of X right so X is going to have P + 1 columns right. You can take the P +1 columns of X as set of basis vectors right. So what is the dimensionality of each column n okay. So each column is a RN is a vector in RN right and I have P + 1 such vector in RN okay.

Now I can think of these vectors as a set of basis function basis vectors right. So it ideally I would like them to span a P +1 dimensional subspace of RN. It is where all the linear algebra tutorial supposed to help. So so you have a P + 1 subspace of RN, your X beta is what we will be a point in that P + 1 subspace, dimensional subspace right. Because X or my basis vectors right and I am combining the basis vectors by some set of scalars β right β 1 β 2 like β1 β2 right.

That although scalars just am getting just getting a linear combination of my basis vector so it is going to give me some point in the p + 1 dimensional space right. In fact if I am doing linear regression all I can do is express a point in that P +1 dimensional space. If I take the columns of my x matrix any output that I can learn will be a point in that p +1 dimensional space. Makes sense right so what is the best possible point in that p +1 dimensional space that I can predict.

So let us say I have two vectors x 1 and x 2 these are not the data points okay. These are the the column vectors okay so since there are two dimension 2 vectors here so X 1 and X 2 right. And this is the space and let us supposes I have a vector Y which is in the N dimensional space okay. I have a vector Y in the n-dimensional space what is the best prediction I can make.

(Refer Slide Time: 18:45)



So Y is in r3 I mean if you mean if you can buy into my drawing skills okay so X 1 and X 2span that two dimensional subspace of r3 right. And Y is a point in r3 right so that is what Y is what is the best prediction that I can make that fits into the X 1 X 2 space. The projection of Y right that is what I had should be and will I be able to make that prediction. am I making the prediction? yes, because if you look at the error.

So Y minus y hat is essentially orthogonal to the space spanned by X so that is what our minimization condition is telling us right, X^T Y X β is 0 right. So essentially it is telling I said okay ,this vector this is y minus y hat this vector is orthogonal to the plane spanned by X that is essentially what the minimizing condition is telling is okay. So this is the best possible estimate that you can make for why given that you are restricted to the plane spanned by the columns of X okay.

That make sense so this is a geometric interpretation of what linear regression is doing. It is also lets us think about some other things right so what happens if X is not full rank that would mean that some of the columns are dependent on each other right or linearly dependent on each other. That essentially means that it is not really spanning a P +1 dimensional space its spanning a smaller subspace right.

It is planning a smaller dimensional subspace therefore your approximation is going to be worse. That is one part of it, then anything else the formula would not be valid. So we have to think of different ways of doing it right, so that is the next thing but still regardless of that the best fit that you can get will still be the projection of your Y onto the space spanned by the XS. You have to have to come up with different ways offending it.

But it will still be the projection right, so that is the thing. So one of the easiest ways of doing it is what? Now we know exactly now we know that, it is just in the space that is spanned by these vectors that is important right and we are supposed to find the projection onto the space. And if there are redundant vectors that will not help us define the space. We can throw them out but even though I have all this P+1 dimension right whatever is redundant that is not helping me define the subspace I can throw them on.

So there are some very simple checks that you can do right in fact if you use some standard tools like R and you are trying to do linear regression unless you explicitly tell it not to it will automatically do the check for you it will automatically do the check and throw out the independent curves it will pick some subset of independent bases and then use that to figure out what the projection should be okay.

Great so what about the case with the number of dimensions is much larger than the number of data points? Do you think that will happen yes, no possible how many of you here work with images or have done any work with image data. So more often than not that is the case right so because image data is very high dimensional right and unless you are able to generate huge volumes of such data and more often than not P will be greater than n.

So you have to think of some kind of regularizing the fit so that you get actually a valid answer right. If P is larger essentially what it means that you have a much larger space and Y actually exists in a smaller space than what is given to you. So so we have to figure out a way of regularizing the problems so that adding additional constraints on what kind of projections you are looking for because otherwise it does not make sense to talk about the projection of Y on this P plus 1 dimensional space okay.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Higher Education
Government of India

www.nptel.ac.in