**NPTEL**

**NPTEL ONLINE CERTIFICATION COURSE**

**Introduction of Machine Learning**

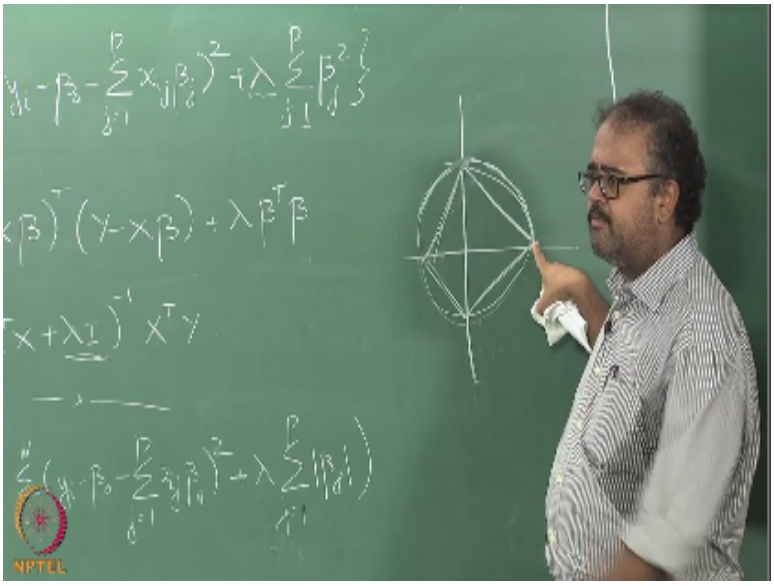
**Lecture 16**

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

**Shrinkage Methods**

So what are the shrinkage methods you think we can come up with, each of β is closed right so we imposed a $L_2$ norm on the β $L_2$ constraint on the β you can impose any other norm constraint on the β right I can I can impose an l4 norm or more commonly I can impose a l1 now it is called lasso right.

(Refer Slide Time: 01:00)



So lasso is essentially let us just ignore the absolute value of the β you sum up those and you want to keep them so you can write the same we can write the constraint formulation where I can say sum of β has to be less than some t write sum of mod β has to be less than 70 or I could just do this kind of a formulation right. And so to impose a constraint on each individual β would require you to know something about the variables themselves beforehand right otherwise if you

constrain and very important variable to have a small coefficient then it becomes a problem so you need to know something about the variables and you can say okay.

I know that these variables are very important make sure that the other variables don't I want to more than 0.5 times the coefficient of these variables, so something like you can think of all kinds of complex constraints once you have knowledge about the system but typically you do not write in such cases you will have to have some kind of uniform constraints like this and so this is a very popular constraint it actually makes life harder for us right so it does not have such a nice closed form solution any more why I mean this is no longer differentiable right.

So I can't I cannot write your nice closed form solution like this in fact I have to work very hard to solve this there are packet mean you don't have to work very hard I mean in distilling typically so you can just run it on R or β something you can always run lasso on it will give you the nice fit. So what is the nice thing about lasso I will try to give you an intuition about it so think of it this way so up suppose I have a non important coefficient okay, so if I can reduce it from say at 1002.3 okay oh let us not even look at it that way.

So I can reduce some coefficient from say 1000 to 999 okay right and there is another coefficient which you can reduce from let us say so there are many variables in my foot there is one variable whose coefficient I can reduce from thousand to 999 there is another variable whose coefficient is one I can reduce it from one to zero okay. And both of them cost the same change in my squared error both of them contribute equally to the squared error and making this change will make the same change to the squared error.

So which one would lasso which one would reach regression prefer to reduce 1000 to 999 because that causes a much larger reduction in this quiet penalty right which one would lasso prefer to reduce mother in matter either one but then I can make this thing slightly more contract right now you go which one would lasso prefer right so even though this is an absolute values this is a larger reduction Bridge would prefer still preferred 1000 to 999 right.

Because the fall is $1.1^2$ to $0^2$ which says thousands quite to 999 squared right still that is a larger reduction in error right. So what is a take-home message here la saw is more likely to drive coefficients to zero than ridge so rich would happily leave the coefficient at one point one right or even more dramatically it will happily leave coefficients at 0.3 0.2 0.8 so it will leave it at

small values it will not drive it all the way to zero okay well lasso given an opportunity right we will drive the coefficients to zero we need not do it zero the driver to zero at the cost of minimizing the error right it will still try to minimize theta right but given the chance it will more likely to drive coefficients to zero okay.

So in sometimes lasso is also called sparse regression right because this l1 norm constraint is also called a scarcity constraint because it makes your β vector lie more will have to have more zeros right so if you know what a sparse matrix is right so you have a matrix with a lot of zero entries in it and only few nonzero entries you call the matrix a sparse matrix right that people work with sparse matrices right some of you have all play sparse matrices.

So you really don't want to have an array representing your sparse matrix because most of the entries are 0so typically what you do in a sparse matrix representation is you store the index of the nonzero entry and the nonzero entry okay that actually takes a lot less memory than actually having a large M by n array with lots of 0 so this Way's called sparse things. So here the l1 regression has a tendency to make the β sparse okay to have a lot of 0 so sometimes called the scarcity constraint coefficients less than 1 all the coefficients of less than 1.

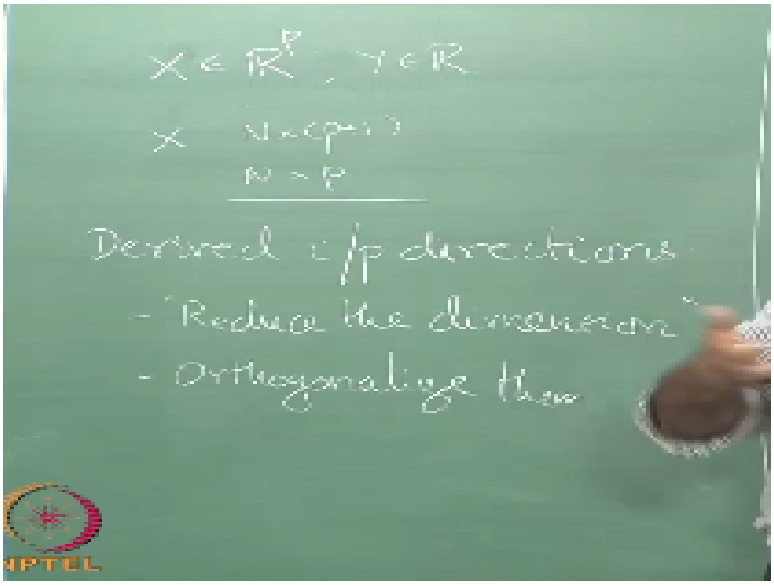
Yeah no see face if I take point 0 1 0.1and square it okay and the difference between that squared + $0^2$ is lesser than point 1 and 0 no, no, no, no, so but the drop in the value will be bigger in the lasso than in Ridge ok now it depends on what other compositing elements that you have no lasso this typically drives the coefficients to zero well rich does not this is that I was giving you an intuition as to why that is the case right it is not mathematically a sound argument but you can give a mathematically sound argument also that last saw is more likely to find spots fix then regression okay.

So I am being very careful I will be careful so I mean I can also think of a geometric intuition for it so if you think about the right the lasso constraints it will be something like this all right let me think about rich constraints that suppose to be a circle okay so the rich constraint will be something like this right so here if you know where the sum has to be a constant right so the sum has to be a constant in the sum of squares has to be a constant, so one will be a circle other one will be anything.

So when you're looking at the error surface corresponding to this right, so essentially you will have to find solutions that lie on this or live within this four Ridge I mean four lasso and live within this four Ridge okay and it turns out that you are more likely to hit a corner off you can show that more formally that you are more likely to hit a corner of the in the in the latter case and in the rich case you are likely to hit I mean so the probability of hitting something because this is the whole thing is convex the probability of hitting that side it is higher right.

So you can this is just the jump in a very rough intuition I do not want to get into showing things formally but you can show that the probability that lasso will give you these kinds of corners in the fitter corner obviously you can see that has one of the coefficients a 0 right so that you will get a corner as if it is much higher then you will get one of these axis points in the ridge. So in fact you can think of having higher-order penalties also okay like I said you can think of an L for norm penalty right and you can even think of so far we looked at two methods for variance reduction so one was subset selection okay further on what shrinkage based methods okay now there is another third class of methods which people use for getting better fits with possibly fewer variables or fewer parameters this is based on based on.

(Refer Slide Time: 11:29)



Derived input directions right so we talked about reducing the number of variables so far right but in both of the I mean at least in the subset selection part we retain some of the variables and then we ignore some of the other variables right likewise whether we are doing implicit subset selection by doing lasso or written regression we are reducing the coefficients of some variables and retaining some other variables right but at all points we were operating with the original set of basis vectors that were given to you right so what were the basis vectors we are talking about here what are the basis vectors you are talking about here the columns of the X matrix like the columns of the X matrix are the basis vector.

So we are working with the original basis vectors we are working with the same columns that were given to us right we in one case we picked some columns and threw out some of the other columns in the other case we tried to continuously adjust the weights of the column so that some of them were given more weight and some were given less weight. So when we talk about derived input directions now I am not going to stick with the original columns okay I am going to find a new set of columns and I am going to find a new set of features new set of directions right which I will then use for doing my regression okay.

So we actually talked a little bit about it in the when we do look at the orthogonalization right I told you and you two orthogonalization essentially what you are doing is you are finding an orthogonal basis for the into input space and then you are trying to find the coefficients there so likewise what we will do here is right, so we will reduce the dimensions okay as I put it in quotes I will explain why and then we will also orthogonalize the dimensions okay so what is the advantage of orthogonal icing the dimension well we could do unit variant regression on each dimension separately okay and then that will give us the coefficients okay.

So we do not have to actually do a multivariate regression okay we can do unit variant regression on each dimension because once I orthogonalize the directions they do not interfere with one another right, so I can do unit variants regression. So typically when I try to do these derived input directions I try to orthogonalize the directions okay and I also tried to find a reduced set of dimensions that will give me the original fit or as close to the original fit as possible.

**IIT Madras Production**

Funded by
Department of Higher Education

Ministry of Human Resource Development
Government of India

www.nptel.ac.in