NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction of Machine Learning

Lecture 9

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

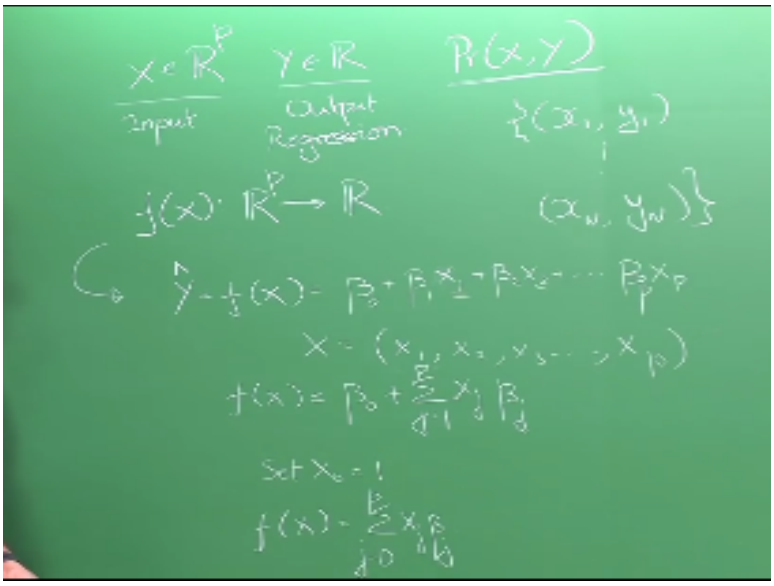
Statistical Decision Theory –
Regression

Hello and welcome to this module on statistical decision theory, so the goal here is try to give you a framework that we will keep using for the rest of the course right for at least for the majority of the rest of the course and it introduce you to some of the basic notations and also to talk about some kind of a unifying idea behind what we will look at in different classification algorithms and regression algorithms right.

To set the tone let's consider the inputs which will denote by X as being gone drawn from some P dimensional space right so which we will call our P right. So if you think about what we did in the previous modules we talked about input that had age and income as the attributes, so that would mean that P was two dimensions right so one of the dimensions represented H other dimension represented income.

So what we are doing here now is trying to move to a more general setting where I am talking about any kind of a p dimensional space right and what p could be much larger than two and the output that we are going to be looking at least in the initial regression case that we will see I will assume that.

(Refer Slide Time: 01:37)



The output is drawn from the real numbers again so this will be like the temperature that we saw in this second example in the previous modules right. So the input X is drawn from a p dimensional real space and the output Y is drawn from again a real from the real numbers and in the case of regression right, so the case of classification will see the little bit later the output will come from a discrete space alright and we will also make an assumption that the data comes to you from some kind of a problem joint probability distribution right.

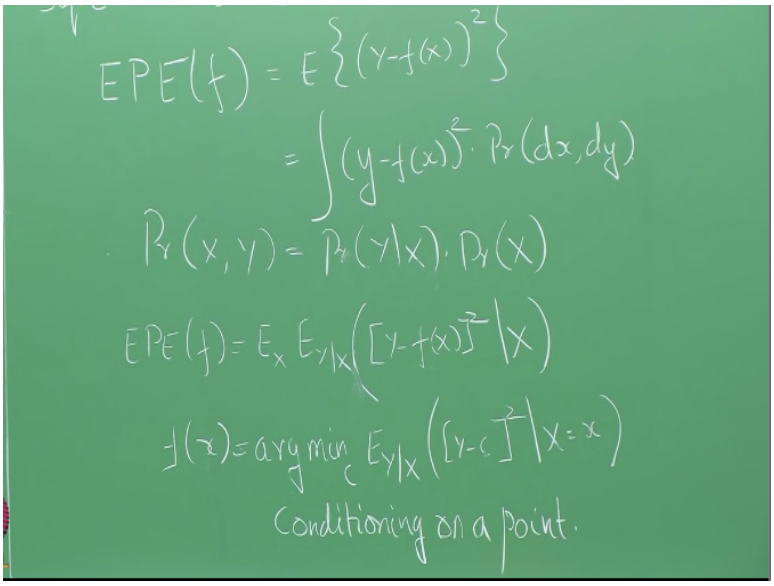
So you do not know this joint distribution a priori right so nobody tells you what is the distribution from which the data is coming but the assumption that we are going to make is that there is an underlying data distribution right like a joint distribution over the inputs and the outputs right and that it is fixed right and you are going to be given samples and drawn from and you are going to give in a set of samples that is drawn from this probability distribution over X Y so this will be your training data right which you will use for both training and possibly for validation if required right.

So you are going to get an X_1 with a corresponding Y_1 X_2 with the corresponding Y_2 and so on so forth, so the goal is given such a set of training data so learn a function f of X that goes from a p-dimensional space to the real line right there so the p dimensional space sassily corresponds to a point in the input space and the real line corresponds to the output space so the function f is going to take any input that is given to it and produce a number right so the F could take different forms right so we looked at F being a straight line right in the example that we saw earlier.

So in such cases right so my one example of F would be here saying that I am going to predict Yˆ = f (X) at given an X I am going to predict a label Yˆ hat right and that is going to be given by some H is going to be given by $\beta_0 + \beta_1 X_1 \beta_2 X_2$ all the way up to $\beta_P X_P$ so one thing which I want you to note here is that this $X_1 X_2$ so on so forth are essentially the coordinates of X. So when I say X here X essentially comprises of so this could be age this could be income and so on so forth each one of these corresponds to a different attribute that describes the describes the data right.

So I can look at this and then I can write f of X is essentially compactly I can write it as right is an alternate way of writing this is to declare that set right so I will set $X \neq 1$, and then I can just remove this special treatment of β_0 and I can just write the summation J = 0 to P X J βj right so is that clear right. So this is essentially what do you do in when you are doing linear regression right so another example of doing this is a very popular and other classifier which we will call the nearest neighbor classifier.

(Refer Slide Time: 06:52)



So where my Yˆ (f) is given by one by some number K summing over all X_I that belong to some neighborhood of X right and sum up all the Y_I is corresponding to that X_I, so let us assume that my training data looks something like this right so there is a X here as an X here there is an X here there is an X here and there is an X here and an X here right and let us say my K is 3 right, so if I get a query point say somewhere here I get a query so this is my X so this is my X_1 X_2 X_3 so these are my training data the X_1 to X_6 are my training data and X is the point for which I want to predict the output right.

So these are the places which I have already measured it this is a new point and I want to produce the output here so in this case what do I do is I pick the three nearest neighbors because K is 3 I pick the 3 points that are closest to this data point right find the corresponding Y's so in this case I will pick y_2 y_3 and y_4 and I will take the average of these three points and I will report the value of the function at X is the average of this three point right so this is called the K nearest neighbor regressor right.

So I will just take the average of the outputs of Y- y_3 and y_4 and report that as the value of the X right so depending on where X is I will be picking different three neighbors and reporting their values okay this is the K nearest neighbor so there are different ways in which you can define this function f right but remember that we had this discussion in the last set of modules that unless you make an assumption about the form of f you really cannot do any generalization we needed to talk about lines in the previous class but now we are talking about different

assumptions for the function f need not necessarily be lines okay in this case it is a straight line right but in this case it is a average it is a local average and that gives me the function that I want to learn okay.

So how do you choose this function right so that could be many different ways in which you can define the β as right, so given that I have chosen that this is the way to model the function how do I pick the β right so how or how do I even choose this form for my for my predictor so how do I know that this is a valid waveform to choose so we have to look at some performance measure. So which we will consider in this case is the loss function right which will compare the true output Y right with the predicted output f (X) right.

So I have the true output Y and I have a particular put f (X), so I will have some loss function that compares f of X with y and my goal is to find an f of X such that this loss function is minimized one of the most popular loss functions that people use in the literature is known as the square error so basically look at right and the performance measure that I am interested in is the expected prediction error of the function f right so that is equal to the expected value of $y - f(F)^2$ right in the case of squared error so the expected prediction error is the expected value of $y - f(F)^2$ right.

So what is the distribution with respect to which you are taking this expectation so whenever we talk about the expectation of a random variable so we want to talk about the underlying distribution right so what is the distribution with respect to which you are taking this expectation right exactly the Joint Distribution of X and Y whether it is the Joint Distribution between x and y so that is a distribution with respect to which you are taking this expectation therefore I can write this.

So I can do a little bit more sleight of hand here right and talk about the conditional distribution right so if you remember the probability of X, Y can be written as probability of Y given X into probability of X so it can written as probability of Y given X into probability of X this is just the product rule in probability so I get this, so what does this tell me that okay there is some chance with which I can choose a data point X right having chosen a data point X so what is the probability of seeing a particular output value right.

So Y is y why are we looking at probabilities here again? So this helps us to kind of you know model a variety of different scenarios so the first one is if there is noise in the measurement no so I am talking about what is the probability of Y given X suppose I am telling you that I am measuring the temperature at 3 o'clock every day right so there will be some kind of a natural variation in the temperature is measured at 3 in the morning right.

So that is modeled by this probability right that will be some set of temperatures that are very power probable and some set of temperatures that are not right so for example if I am measuring temperature at 3 a.m. right so 40^0 is not a probable value right, so those will have lower probabilities and then say something in the 20s will have a higher probability so I am talking about Chennai if people are wondering how you are getting 20^0 early in the morning right.

The second factor that this allows us to look at is our ignorance about the whole system you know so I might have just chosen the time of day maybe there are other factors I should have taken into consideration while I am forming my data, so these factors about which I do not know anything will appear as noise you know so it is not important whether I take the temperature at 3:00a.m.

Maybe it is important where in the building I do the measurements maybe I am measuring it next to the kitchen where things will be warmer or maybe I am measuring it next to an air-conditioner where things would be actually warmer if I am measuring it on the external of the building right and or it could be measuring it on the inside of an air-conditioned room the temperatures could be lower so even though I say I measure it at 3 a.m. there could be many such factors for natural variations which I have not modeled okay.

So this is beyond the natural variations in the system right, so one way of arguing about it could be to say that hey the natural variations are due to factors that you do not know anything about right so that is a valid argument so it could very well be that. So it is really there is nothing like a natural variation there is no real noise so everything arises all the uncertainty in the data arises from my lack of knowledge but that is a philosophical question.

So there is things that are measurable which we do not measure right and that I would call as lack of knowledge and things which are immeasurable which I would call as noise right there could be both of these sources which introduced the probability interval system right, so it is not

just a mathematical whimsy that we model this as a joint distribution but there is an actual practical reason for talking about probability distribution share right. So now I can go back and write my expected prediction error as an expectation over X of so an expectation over X of the expectation over Y given X because I have written this probabilities out like this.

So I can write the expectations out also like this of this quantity then it is the same quantity earlier the only difference is now I am conditioning it on the value of x right so what this expression says is a I will tell you what X is okay I will tell you what X is then you tell me what the error will be right, so the uncertainty here is over the value of y right. So I will give you X I will fix X you tell me what Y is so I am going to look at the error just on conditioning after conditioning on X this will only look at the variation on Y.

And the outer expectation gives me the variation of X the outer expectation takes care of the variation on X, so now I can try to find the minimum of this prediction error where I can try to find the minimum of this prediction error by conditioning on a specific value for X right so I will not look at this expectation right I am not making any assumptions about F and I am just assuming that F can be anything like any function in the world F can be any function of the work so what I want to do now is I want to look at each and every value of X and I want to say that I will pick an F such that for every value of x right it makes the best possible prediction right I will pick an F such that for every value of x it will make the best possible prediction alright so what does that mean so it will produce the prediction so f (f) for a specific value of X f of X will give the output such that this inner expectation is minimized right.

So I am going to write it down like this so for a specific value of x right this is f (f) for a specific X so for I was writing capital X which is a random variable right but Here I am using a specific X right, so given an X right f of X has to be a specific number right so let us say somebody with an age of 25 an income of 15,000 rupees walks into my shop I can only give one output it is going to buy a computer or does not buy a computer or I am going to say I am measuring the temperature at 3 a.m.

What will be the value and I can give you only one number since I have already fixed the input description so I can we are going to give you one output corresponding to that input description so that output let me call as C right so that is C, C is the output that I am going to give for fx let

see is the output I am going to give for f of x so what is the value that I should choose for C okay.

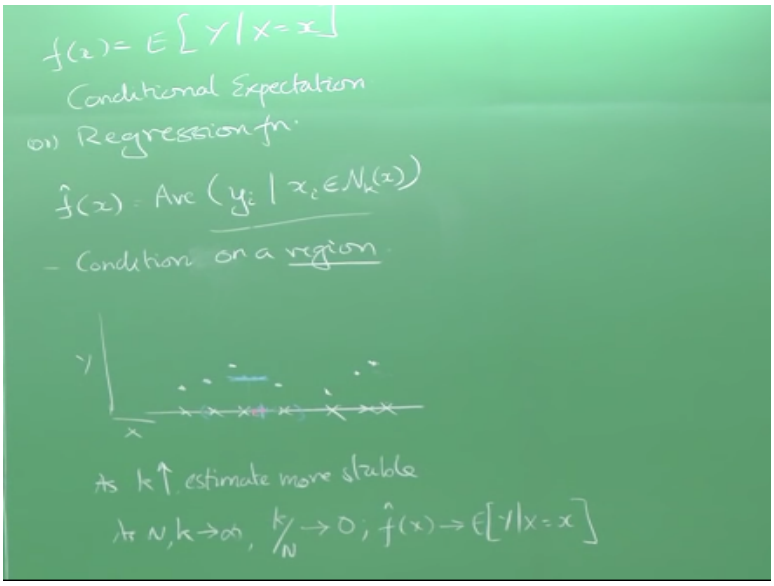
So it should be such that the error which is $Y-C^2$ is as small as possible right so I am minimizing over the different possible values of C that I could assign for f of x I am trying to pick that C which gives me the smallest error right so Arg min means first minimize with respect to C and take that argument but basically take that value of C which achieve this minimum if there are multiple values that gives you the minimum.

I can pick any one right so this is essentially called conditioning on appoint so instead of conditioning on the random variable X right so conditioned on a specific point where X equal to small x right and then I can find this so now what happens for every possible input X that I could have small X that I could have I will find the corresponding C and I will say f of X equal to that C right so the thing to note here is I have not made any functional assumptions about what this what should F look like right.

So F could be something really, really jagged I do not care right so this is a recipe for disaster as we saw earlier that you might end up over fitting the data but just work with me here because they are just trying to build some general principles right so we can do little they can go little further right now that we have decided to say that the, the minimizer is the one that the value that you are signed to f of X so what is the value of C that will minimize this expression right so I have to look at $Y-C^2$ so I have to assign a single value for C.

So circuses imagine what does it mean right suppose I give the input as x, I make a measurement let us call it say y1 and I give it same input X again I make another measurement say y2 I give the same input X again I make another measurement y3 so I have three measurements y1 y2 y3 for the same input X right now I am asking you to give me a prediction for what will be the output given X right.

(Refer Slide Time: 23:38)



So what will what should your prediction be it will be the average of these three wise right so y1+y2+y3 divided by three so that it should be the prediction so why is that the case because we are talking about squared error right the quantity that will minimize the squared error is essentially the average right so I can I will end up writing that f of x equal to the expected value of why given X equal to X right so this is essentially what my prediction would be so this is the known as the conditional expectation right or so sometimes called the regression function the conditional expectation or the regression function.

So there are a few problems with this right so what are the problems a I do not know this distribution right I do not know the distribution with respect to which I am taking the expectation so what is the distribution with respect to which I am taking the expectation that is essentially probability of Y given X right so that is the distribution with respect to which this expectation is being taken and I do not know that if I know that my problem my life is a lot simpler right.

So I actually have to estimate it from the data so what is the data that is given to me I have this pairs of x1 y1 to x and y n so that is the data that has been given to me and I have to do this estimate of this expectation from that data right so how will you do that so it is very simple you know that you can always estimate the expectation by taking averages so what you would do is from your data you pick all the training data points that have this value of x right find the corresponding Y take an average.

And you are done right so one simple way of thinking about it is to say that okay I cannot find the true f so I am going to find an estimate of that which is called as f hat so is equal to average of all the Y's such that x I equal to X right there is a problem here why so how many samples do you think you are going to get of the same input X right first the second thing is you are not going to be able to make a prediction for any data point which is not there in the input right so we are trying to make an estimate of the expectation by using the averages but if you have known time enough measurements than your average is going to be bad right and second thing is you are making an average at that point and if the point does not exist in your training data.

You are not going to be able to return an estimate for it right so we need to address this somehow right so what we will do here is we will relax the conditioning right instead of conditioning on a point we will conditioning will condition we will condition on the region right so what does it mean so I am taking the average here of all those data points for which Xi equal to X now that is not going to work right because there are too few data points what I am going to do is I am going to take this as the average of all the data points which belong to some region around X which is essentially the, the neighborhood right.

That we are talking about right so that circle there would correspond to the neighborhood around X so I am going to be conditioning on this region which is given by this neighborhood around X so there is it makes sense right so we are not going to condition on the point we are conditioning around the region so the one assumption that we are making an implicit assumption that we are making here so why are we conditioning around a region so that instead of taking an average of one data point.

I have at least K data points of which I will be taking the average right so that gives me a better estimate of the expectation right that is that is the reason we are doing this conditioning or a region but more importantly we are also making an implicit assumption you know we needed to make assumptions if you remember our inductive bias said that we needed to make assumptions the assumption that we are making is that the output of the function over this region is going to be a constant right.

We are going to be making the assumption that the output of the function over this region will be a constant right so let us let us try and do a little example so that becomes a thing clear to people right let me go back to my one-dimensional example so it makes it easier for me to draw things

so I have let us say I have multiple data points like this alright so I have a query point and then the corresponding outputs.

So these are the Y's right this is X and that is Y so these are my xx and Y is okay this is the training data I have and now given a query point let us say I am given a query point here right I want to know what is the output value for this X so let us say I pick my three nearest neighbors which would be these three data points right and then I will try to take the average of this, this and this which will be somewhere here.

I will say that is my output right exit so like this I am going to make cases for variety of data points but one thing which I want to point out here is so I assume that my data point lies here so what if I had assumed that my data point was here if my query point was here so what would have been the output so my neighbors remain the same let me three neighbors do not change these are my three nearest neighbors whether the query was here right whether the query was here or the query was here my nearest neighbors do not change right.

So what will be the output for this input point as well so I will be taking the average of these three points so the output will also again be that right so in fact for certain some region around here where these three are the nearest points the output will be a constant right I said output will always be a constant so this is what I mean by saying that we make the assumption that the output is constant in a region right so this region so for all those data points for which these three are the nearest neighbors.

So the output is going to be here so this is essentially the assumption we are making that the output is going to be consistent over a region so that I can write an expectation over the region as my substitute okay so if you think about it so what have we come up with here this is essentially your nearest neighbor classifier right so you take the generic idea of minimizing the expected prediction error right and then add certain conditions to it so what are the conditions you add into it so you are well you cannot do the expectation.

So you are going to take averages and you cannot do an average on the training data and therefore you are going to do an average over a region assuming that the output is constant over a region so conditioning on the data point wise and then conditioning on the region relaxing that the conditioning on the region gives us nearest neighbor classifiers so in some sense you can

argue that one way of minimizing the expected prediction error yields a nearest neighbor classifier okay.

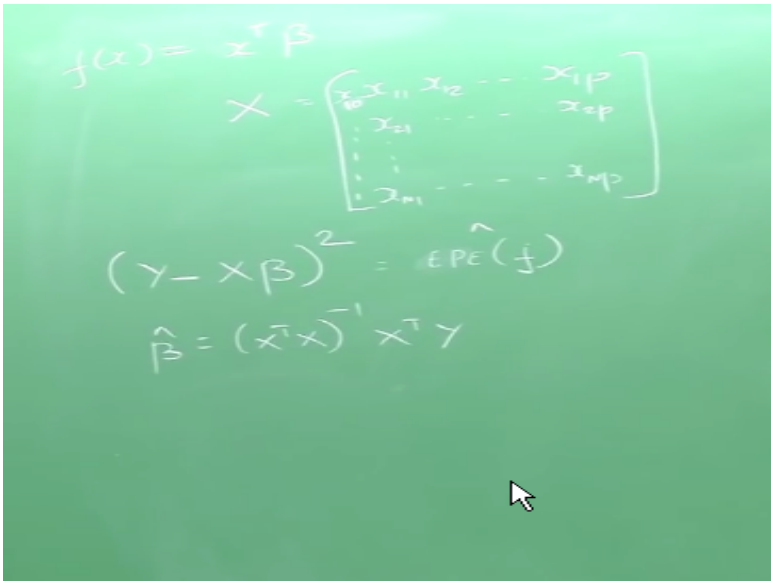
So in fact it is a very powerful classifier and you can show that as SK increases right so the estimate becomes more and more stable you know so for small changes in the input data then the, the classifier does not change tremendously right and so as and as N and K 10 to infinity or they become large right so your ratio at K by N would go to zero right in such a case your F hat of X will go to right as K increases the estimate becomes more stable in particular as K and N becomes large that is and my number of data points is very, very large.

And so the number of points I can look in the neighborhood also becomes larger and larger so such that of course the data points have to grow at a faster rate than the size of the neighborhood that is what K by N means so in which case I can show that my actual prediction I make using this average actually approaches the true prediction that I am interested in right so there are a few caveats here that I need to point out so I assuming I am saying that N goes to infinity I mean that is a pretty place statement to make because N rarely goes to infinity right in fact never so coming up with large data sets is hard except for very rare cases right therefore you cannot really have a classifier that gives you the right output right.

And the other problem is as P becomes larger weight as the dimensionality becomes larger right generally the data tends to becomes sparse you know so if I am looking at K neighbors in like a thousand dimensional space the area or the volume covered by these K neighbors would be very large because they are very sparse space and it is usually not a good assumption to make that the input is actually a constant over this I am sorry the output is a constant over this large area right.

So one thing is if P is large right then if the dimension of the input is large if you have like 10,000 dimension vector as your input then using K nearest neighbors is not really a good idea right and alternatively you should also remember that in some cases having a little bit of a bias is actually not a bad thing right and therefore we have to look at an appropriate way of representing the function f remember we did not make any assumptions about the function f right so the function f could change as drastically as we as we want and so that means that we are trying to keep the bias as minimal as possible.

(Refer Slide Time: 37:03)



And so we would like to remove that assumption right so moving on let us look at the linear regression case where we actually made a significant assumption about the form of the function f right expresses specifically we assume that F is going to be linear in the input parameter so there can be written as β1 X1+β2 X2and so on so forth right so essentially your f of X is you need to be in vector notation we assume it to be some x transpose beta right assume it be X transpose β and so if you look at it from the training data point of view so I can think of having a vector notation for this right.

So I can think of a matrix X right so a matrix X in which each row corresponds to an input X Y x1 right so X will be something like right X will be matrix that looks like that and so my β would be a vector of the coefficient so β^0 2βP right and of course my X is going to have a right so the zero dimension is going to be 1 right and they are going to have β so I can look at the output as essentially being Xβ so my overall error is going to be Y - right so that is going to be and that is going to be a the estimate of the expected prediction error based on the training data.

That is being given to me and I can minimize this I can take the derivative right and then I can equate it to zero and I can do some minimization to get the value of β and that is essentially going to be I take the differential of this equate it to zero simplify for β right I am going to get β at equal to this is primarily because of the square here so you are going to get this is your

simplified expression so this is essentially my β, β at vector and so remember that the X that we put in here is essentially a matrix where we the columns the rows are the data points right the columns or the feature.

So this will be the like the age of every customer that comes in this will be the income of every customer and so on so forth and each row is a complete data point right so what we have done here is make the assumption that my function is globally linear and then I have tried to solve for it to give you the parameters β at right and in the nearest neighbor case we made the assumption that my function is locally constant right.

So we start off with the same formulation we wanted to minimize the expected prediction error and we make different assumptions one assumption leads us to linear regression the assumption that we made was the data is going to be globally linear and another assumption that we made where the data is going to be global locally constant illness K nearest neighbor okay.

IIT Madras Production

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

www.nptel.ac.in