

NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

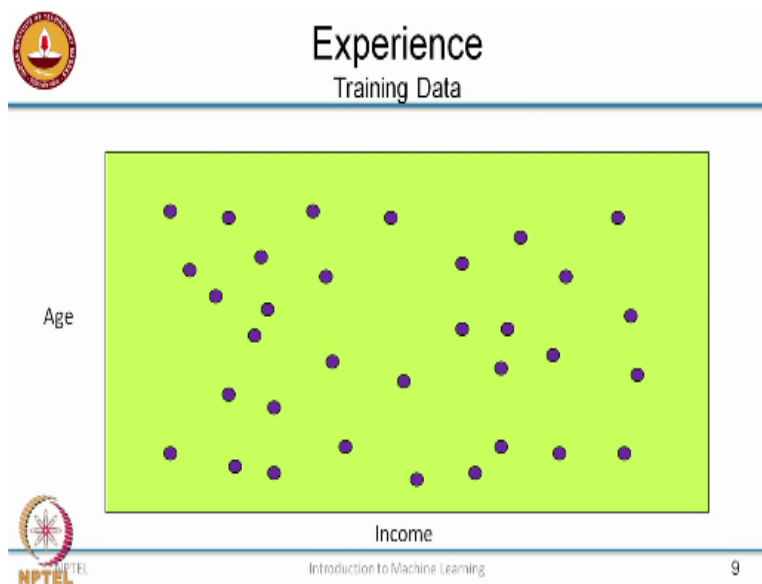
Lecture 2

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

Supervised Learning

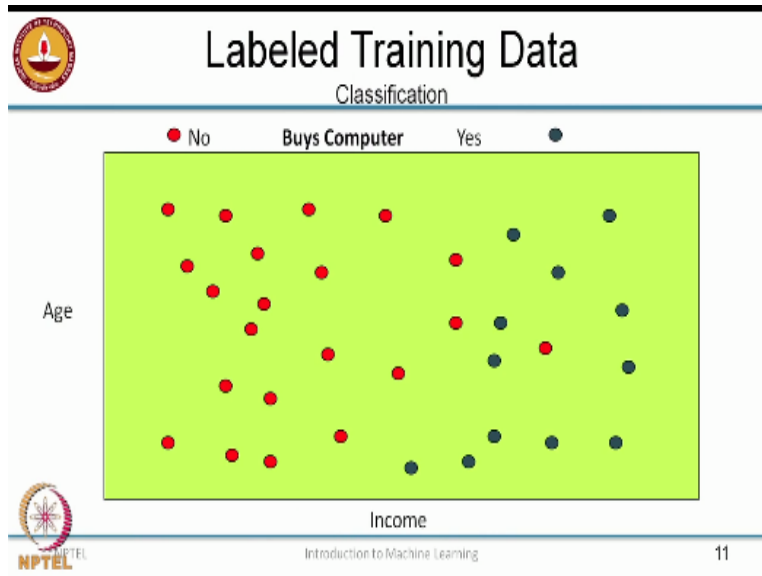
So in this module we will look at supervised learning right.

(Refer Slide Time: 00:21)



If you remember in supervised learning we talked about experience right where you have some kind of a description of the data. So in this case let us assume that I have a customer database and I am describing that by two attributes here, age and income. So I have each customer that comes to my shop I know the age of the customer and the income level of the customers right.

(Refer Slide Time: 00:48)



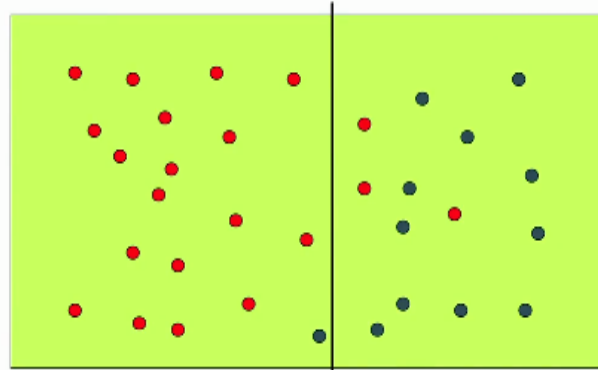
And my goal is to predict whether the customer will buy a computer or not buy a computer right. So I have this kind of labeled data that is given to me for building a classifier right, remember we talked about classification where the output is a discrete value in this case it is yes or no, yes this is the person will buy a computer, no the person will not buy a computer. And the way I describe the input is through a set of attributes in this case we are looking at age and income as the attributes that describe the customer right.

And so now the goal is to come up with a function right, come up with a mapping that will take the age and income as the input and it will give you an output that says the person will buy the computer or not buy the computer. So there are many different ways in which you can create this function and given that we are actually looking at a geometric interpretation of the data, I am looking at data as points in space.

(Refer Slide Time: 01:57)



Possible Classifiers



The one of the most natural ways of thinking about defining this function is by drawing lines or curves on the input space right. So here is one possible example, so here I have drawn a line and everything to the left of the line right. So these are points that are red right, so everything to the left of the line would be classified as will not buy a computer, everything to the right of the line where the predominantly the data points are blue will be classified as will buy a computer.

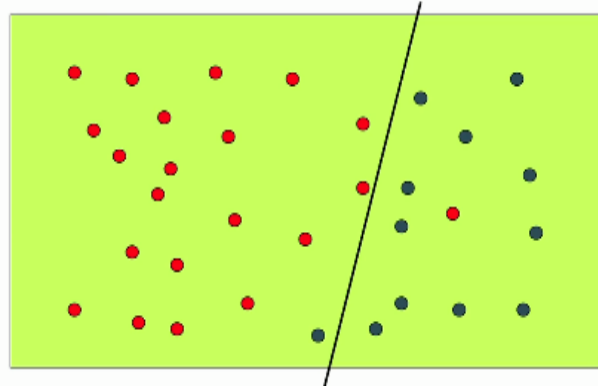
So how would the function look like, it will look like something like if the income of a person remember that the x-axis is income and the y-axis is age. So in this case it basically says that if the income of the person is less than some value right, less than some X then the person will not buy a computer. If the income is greater than X the person will buy your computer. So that is the kind of a simple function that we will define.

It will just notice that way we completely ignore one of the variables here which is the age. So we are just going by income, if the income is less than some X then the person will not buy a computer, if the income is greater than X the person will buy a computer. So is this a good rule more or less I mean we get most of the points correct right except a few right. So it looks like yeah, we can we can survive with this rule right. So this is not too bad right, but then you can do slightly better.

(Refer Slide Time: 03:29)



Possible Classifiers



All right, so now we got those two red points that those just keep that points are on the wrong side of the line earlier. Now seem to be on the right side right, so everything to the left of this line will not buy a computer, everything to the right will buy a computer right, everyone moves to the right will buy a computer. So if you think about what has happened here, so we have improved our performance measure right.

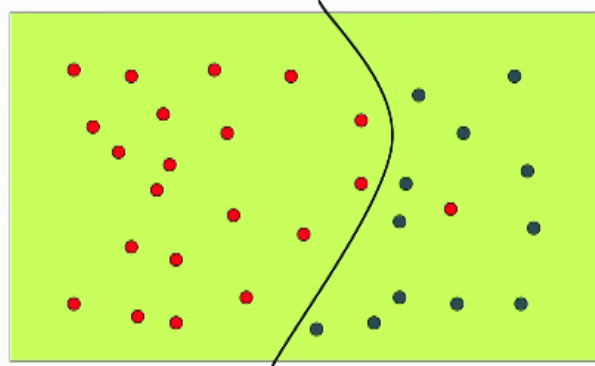
So the cost of something, so what is the cost here. So earlier we are only paying attention to the income right, but now we have to pay attention to the age as well right. So the older you are right, so the income threshold at which we will buy a computer is higher right. So the younger you are, younger means lower on the y axis, so the younger you are the income threshold at which you will buy a computer is lower right.

So is that clear, so the older you are right, so the income threshold is shifted to the right here right so the older you are, so you need to have a higher income before you buy a computer and the anger you are your income threshold is lower, so you do not mind buying a computer even if your income is slightly lesser right. So now we have to start paying attention to the age right, but then the advantage is you get much better performance right can you do better than this yes okay.

(Refer Slide Time: 04:54)



Possible Classifiers



Now almost everything is correct except that one pesky red point, but everything else is correct. And so what has happened here we get much better performance, but at the cost of having a more complex classifier right.

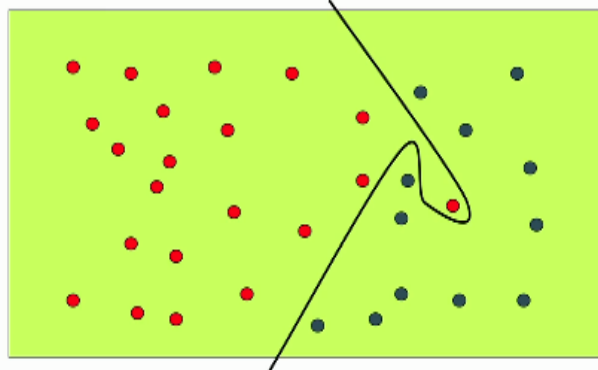
So earlier if you thought about it in geometric terms, so first you had a line that was parallel to the y-axis therefore, I just needed to define a intercept on the x-axis right. So if X is less than some value then it was one class was greater than some value was another class. Then the second function it was actually a slighting line like that, so I needed to define both the intercept and the slope right.

And now here it is now a quadratic so I have to define three parameters right. So I have to define something like $ax^2+ bx+c$, so I have defined the ABC the three parameters in order to find the quadratic, and I am getting better performance. So can you do better than this.

(Refer Slide Time: 05:57)



Possible Classifiers



Okay the sum for does not seem right correct seems to be too complex a function just to be getting this one point there right. And I am not sure I am not even sure how many parameters you need for drawing that because Microsoft use some kind of spline PowerPoint use some kind of spline interpolation to draw this curve I am pretty sure that it is lot, lot more parameters than it is worth another thing to note here is that that particular red point that you see is actually surrounded by a sea of blue right.

So it is quite likely that there was some glitch there either the person actually bought a computer and we never we have not recorded it has been having what computer or there are some extremist reason the person comes into the shop sure that is going to buy a computer but then gets a phone call saying that some emergency please come out immediately and therefore he left without buying a computer right there could be variety of reasons for why that noise occurred and this will probably be the more appropriate classifier right.

So these are the kinds of issues I would like to think about what is the complexity of the classifier that I would like to have right and versus the accuracy of the classifier, so how good is the classifier in actually recovering the right input output map and or their noise data in the in the input in the experience that I am getting is it clean or is there noise on it and if so how do I handle that noise these are the kinds of issues that we have to look at okay.

(Refer Slide Time: 07:31)



Inductive Bias

- Need to generalize \longrightarrow Assumptions about lines!
- In general, **Inductive bias**
 - Language bias
 - Search bias



So these kinds of lines that we drew right kind of hiding one assumption that we are making so the thing is the data that comes to me comes as discrete points in the space right and from these discrete points in the space I need to generalize and be able to say something about the entire state space right so I do not care where the data point is on the x and y-axis right I should be able to give a label to that right.

If I do not have some kind of assumption about these lines right and if you do not have some kind of assumptions about these lines the only thing I can do is if the same customer comes again hey or somebody who has exact same age and income as that cause customer comes again I can tell you whether the person is going to buy a computer or not buy a computer but I will not be able to tell you about anything else outside of the experience right.

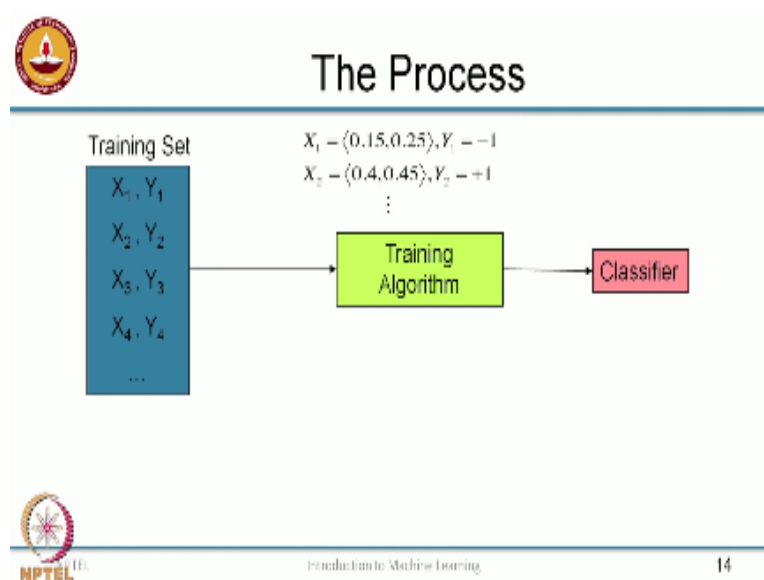
So the assumption we made is everything to the left of a line is going to do one thing or the other right so everything to the left of the line will not buy the computer everything to the right or everyone to the right will buy a computer this is an assumption I made the assumption was the Lions are able to segregate people who buy from who do not buy the lines or the curves were able to segregate people who will buy from who will not buy so that is a kind of an assumption I made about the distribution of the input data and the class labels.

So this kind of assumptions that we make about these lines are known as inductive biases in general inductive bias has like two different categories one is called language bias which is essentially the type of lines that I am going to draw my gonna draw straight lines or am I going

to draw curves and what order polynomials am I going to look at and so on so forth these for my language bias and such bias is the other form of inductive bias that tells me how in what order am I going to examine all these possible lines right.

So that gives me the gives me a search bias right, so putting these two these things together we are able to generalize from a few training points to the entire space of inputs right I will make this more formal as we go on and then in the next night set of modules right.

(Refer Slide Time: 10:01)



And so here is one way of looking at the whole process so I am going to be giving you a set of data which we will call the training set so the training set will be will consists of say as an input which we'll call as X and an output which we call as Y right, so I am going to have a set of inputs I have X_1, X_2, X_3, X_4 likewise I will have Y_1, Y_2, Y_3, Y_4 and this data is fed into a training this

data is fed into a training algorithm right and so the data is going to look like this in our case right.

So remember our X's are the input variable success all the inputs so in this case that should have the income and the age, so x_1 is like 30,000 and 25 and x_2 is like 80,000 and 45 and so on so forth and the Y's or the labels they correspond to the colors in the previous picture right so y_1 does not buy a computer Y_2 buys a computer and so on so forth so this essentially gives me the color coding so y_1 is essentially red and y_2 is blue right and I really if I am going to use something numeric this is what we will be doing later on I really cannot be using these values first of all wise or not numeric and the X is varied too much right.

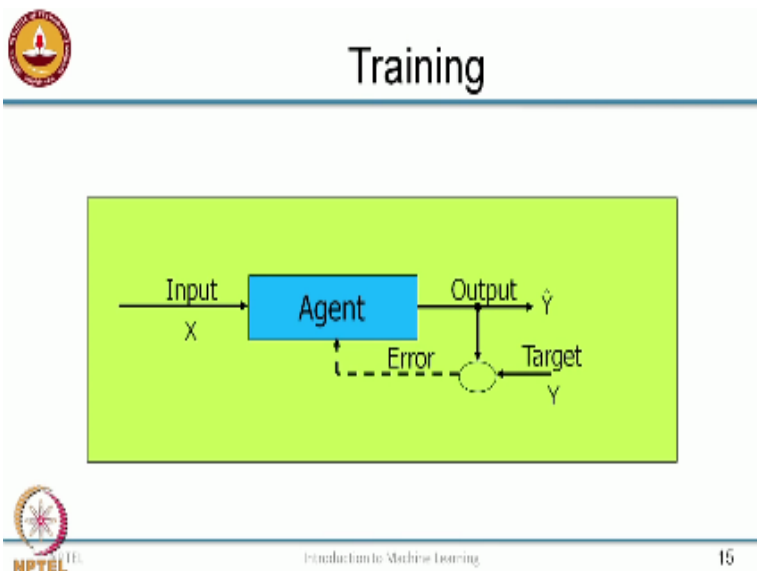
So the first coordinate in the X is like 30,000 and 80,000 and so on so forth and the second coordinate is like 25 and 45 so that is a lot a lot smaller in magnitude so this will lead to some kind of numerical instabilities, so what will typically end up doing is normalizing these so that they form appropriate approximately in the same range so you can see that I have try to normalize these X values between 0 and 1 right.

So have chosen an income level of say 2 lakhs it is the maximum and age of 100 and you can see the normalized values and likewise for buys and not buy I have taken not by as - 1 and by as computer is + 1 these are arbitrary choices, now but later on you will see that there are specific reasons for wanting to choose this encoding in this way alright and then the training algorithm chugs over this data right and it will produce a classifier so now this classifier I do not know I do not know whether it is good or bad right so we had a straight line in the first case right an axis parallel line if we did not know the good or bad and we needed to have some mechanism by which we evaluate this right.

So how do we do the evaluation typically is that you have what is called a test set or a validation set right so this is another set of x and y paths like we had in the training set, so again in the test set we know what the labels are it is just that we are not showing it to the training algorithm we know what the labels are because we need to use the correct labels to evaluate whether your trading algorithm is doing good or bad right so, so this process by which this evaluation happens is called validation later then of the validation.

If you are happy with the quality of the classifier we can keep it if you are not happy they go back to the training algorithm and say hey I am not happy with what you produced give me something different right, so we have to either iterate over the algorithm again we will go over the data again and try to refine the parameter estimation or we could even think of changing some parameter values and then trying to redo the training algorithm all over again but this is the general process and we will see that many of the different algorithms that we look, look at in the course of fitting the course of these lectures actually follow this kind of a process okay so what happens inside that green box.

(Refer Slide Time: 13:48)

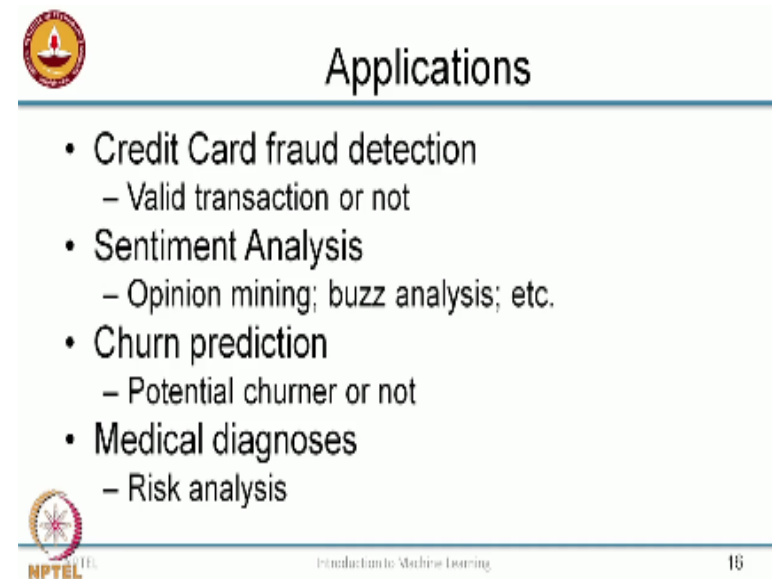


So inside the training algorithm is that there will be this learning agent right which will take an input and it will produce an output which it thinks is the correct output right but it will compare it against the actual target which it was given for the in the training right, so in the training you actually have a target which so it will compare it against a target which right and then figure out what the error is and use the error to change the agent right so then it can produce the right output next time around this is essentially an iterative process so you see that input okay produce an output \hat{Y} and then you take the target Y .

You can compare it to the \hat{Y} figure out what is the error and use the error to change the agent again right and this is by and large the way most of the learning algorithms will operate most

of the classification algorithms or even regression algorithms will open it and we will see how each of this works as, we go on right there are many, many applications.

(Refer Slide Time: 14:46)



The slide features a title 'Applications' in a large, bold, black font, centered at the top. To the left of the title is a circular logo with a red and gold border, containing a stylized figure. Below the title is a horizontal blue line. Underneath the line is a bulleted list of four items, each with a sub-point. The first item is 'Credit Card fraud detection' with a sub-point '- Valid transaction or not'. The second is 'Sentiment Analysis' with a sub-point '- Opinion mining; buzz analysis; etc.'. The third is 'Churn prediction' with a sub-point '- Potential churning or not'. The fourth is 'Medical diagnoses' with a sub-point '- Risk analysis'. At the bottom left of the slide is the NPTEL logo, and at the bottom right is the page number '16'. The text 'Introduction to Machine Learning' is faintly visible at the bottom center.

- Credit Card fraud detection
 - Valid transaction or not
- Sentiment Analysis
 - Opinion mining; buzz analysis; etc.
- Churn prediction
 - Potential churning or not
- Medical diagnoses
 - Risk analysis

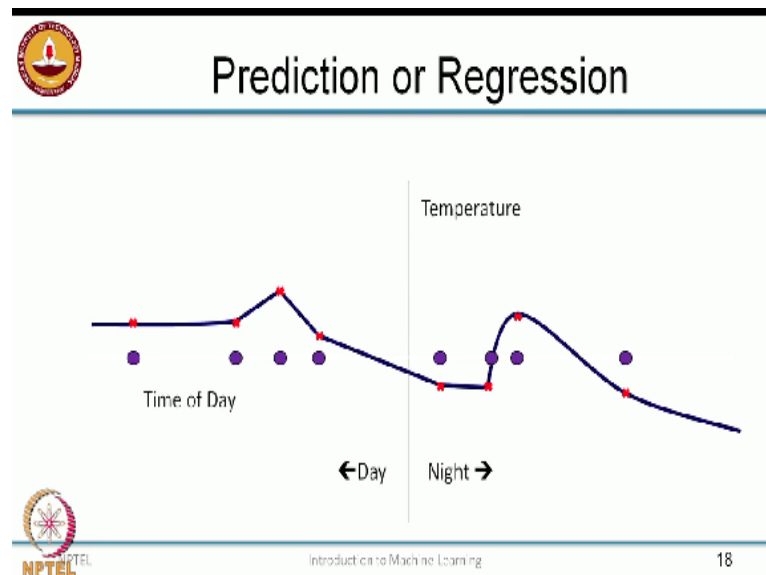
I mean this is too numerous to list here are a few examples you could look at say a fraud detection right, so we have some data where the input is a set of transactions made by a user and then you can flag each transaction as a valid transaction or not you could look at sentiment analysis you know varied Lee called opinion mining or buzz analysis etc. Where I give you a piece of text or a review written about and a product or a movie and then you tell me whether the movies whether the review is positive or whether is negative and what are the negative points that people are mentioning about and so on so forth and.

This again a classification task or you could use it for doing churn prediction where you are going to say whether a customer who is in the system is likely to leave your system is going to continue using your product or using your service for a longer period of time, so this is essentially churn so when a person leaves your services you call the person earner and you can label what the person is Churn or not and I have been giving you examples form medical diagnosis all through apart from actually diagnosing whether a person has the disease or not you

could also use it for risk analysis in the slightly indirect way I talked about that when we when we do the algorithms for classification.

So we talked about how we are interested in learning different lines or curves that can separate different classes in supervised learning and, so this curves can be represented using different structures and throughout the course we will be looking at different kinds of learning mechanisms like artificial neural networks support vector machines decision trees nearest neighbors and Bayesian networks and these are some of the popular ones and we look at these in more detail as the course progresses so another supervised learning problem is the one of prediction.

(Refer Slide Time: 16:45)



Or regression where the output that you are going to predict is no longer a discrete value it is not like we will buy a computer whereas not buy a computer it is more of a continuous value so here is an example, where at different times of day you have recorded the temperature so the input to the system is going to be the time of day and the output from the system is going to be the temperature that was measured at a particular point at the time right so you are going to get your experience or your training data is going to take this form so the blue points would be your input and the red points would be the outputs that you are expected to predict.

So note here that the outputs are continuous or real value right and so you could think of this in this toy example as points to the left being day and the points to the right being night right and just as in the previous case of classification, so we could try to do these simple as possible fit in this case which would be to draw a straight line that is as close as possible to these points now you do see that like in the classification case when it choose a simple solution there are certain points at which we are making large errors right so we could try to fix that.

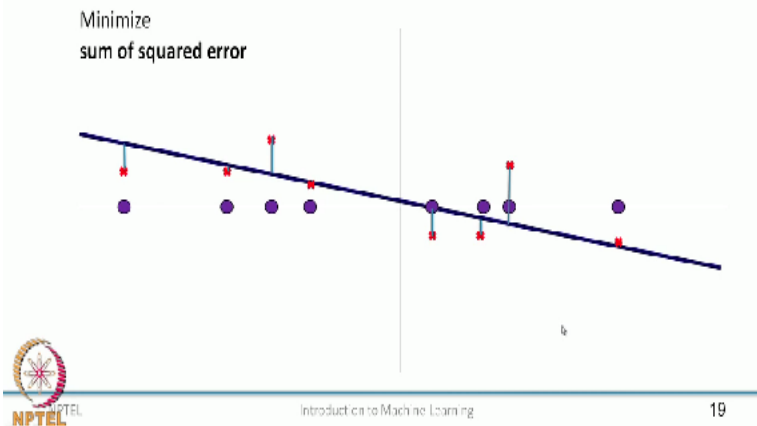
And try to do something more fancy but you can see that while the daytime temperatures are more or less fine with the night times we seem to be doing something really off right because we are going off too much to thee the right-hand side all right how are you could do something more complex just like in the classification case where we wanted to get that one point right so we could try and fit all these temperatures that were given to us by looking at a sufficiently complex curve.

And again this as we discussed earlier is probably not the right answer and you are probably in this case surprisingly or better off fitting the straight line right and so these kinds of solutions where we trying to fit the noise in the data we are trying to make the solution predict the noise in the training data correctly are known as over fitting over fit solutions and one of the things that we look to avoid in, in machine learning is to over fit to the training data.

(Refer Slide Time: 19:21)



Linear Regression



So we will talk about this again and then new course right and so what we do is typically we would like to do what is called linear regression some of you might have come across this and of different circumstances and the typical aim in linear regression is to say take the error that your line is making so if you take an example point let us say I take any let us say I take an example point somewhere here right.

So this is the actual training data that is given to you and this is the prediction that your line is making at this point so this quantity is essentially the, the prediction error that this line is making and so what you do is you try to find that line that has the least prediction error right so you take the square of the errors that your prediction is making and then you try to minimize the, the sum of the squares of the errors why do we take the squares.

(Refer Slide Time: 20:31)



Linear Regression

- Minimize sum squared error
- With sufficient data simple enough
- With many dimensions, challenge is to avoid over fitting
 - Regularization
- Higher order functions?
 - Basis transformations
 - Ex: $(x_1, x_2) \rightarrow (x_1^2, x_2^2, x_1x_2, x_1, x_2)$



Because errors could be both positive or negative and we want to make sure that you are minimizing that regardless of the sign of the error okay and so with sufficient data right so a linear regression is simple enough you could just already using matrix inversions as we will see later but with many dimensions like the challenge is to avoid over fitting like we talked about earlier and then there are many ways of avoiding this.

And so I will again talk about this in detail when we look at linear regression right so one point that I want to make is that linear regression is not as simple as it sounds right so here is an example so I have two input variables x_1 and x_2 right and if I try to fit a straight line with x_1 and x_2 I will probably end up with something like $a_1 x_1$ plus $a_2 x_2$ right and that looks like, like a plane in two dimensions right.

But then if I just take these two dimensions and then transform them transform the input so instead of saying just the x_1 and x_2 if I say my input is going to look like x_1 square x_2 squared $x_1 x_2$ and then the x_1 and x_2 s it was in the beginning so instead of looking at a two-dimensional input if I am going to look at a 5 dimensional input right.

So that will and out now I am going to fit a line or a linear plane in this 5 dimensional input so that will be like $a_1 x_1$ squared plus $a_2 x_2$ square plus $a_3 x_1 x_2$ plus $a_4 x_1$ plus $a_5 x_2$ now that is no longer the equation of a line in two dimensions right so that is the equation of a second-order polynomial in two dimensions but I can still think of this as doing linear regression because I am

only fitting a function that is going to be linear in the input variables right so by choosing an appropriate transformation of the inputs.

(Refer Slide Time: 22:38)



Applications

- Time series predictions
 - Rainfall in a certain region
 - Spend on voice calls
- Classification!
- Data reduction
- Trend analysis
 - Linear or exponential
- Risk factor analysis
 - Factors contributing most to output



I can fit any higher-order function so I could solve very complex problems using linear regression and so it is not really a weak method as you would think at first, first glance again we will look at this in slightly more detail in the later lectures right and regression our prediction can be applied in a variety of places one popular places in time series prediction you could think about predicting rainfall in a certain region or how much you are going to spend on your telephone calls you could think of doing even classification using this.

If you think of you remember our encoding of plus 1 and minus 1 for the class labels so you could think of plus 1 and minus 1 as the outputs right and then you can fit a regression line regression curve to that and if the output is greater than 0 you would say this class is plus 1 its output is less than 0 you see the class is minus 1 so it could use the regression ideas to fitness will solve the classification problem and you could also do data addition. So I really do not want to you know give you all the millions of data points that I have in my data set but what I would do is essentially fit the curve to that and then give you just the coefficients of the curve right.

And more often than not that is sufficient for us to get a sense of the data and that brings us to the next application I have listed their which is trend analysis so I am not really interested in quite

many times. I am not interested in the actual values of the data but more in the, the trends so for example I have a solution that I am trying to measure the running times off and I am not really interested in the actual running time because with 37seconds to 38 seconds is not going to tell me much.

But I would really like to know if the running time scales linearly or exponentially with the size of the important all right so those kinds of analysis again can be done using regression and in the last one here is again risk factor analysis like we had in classification and you can look at which are the factors that contribute most to the output so that brings us to the end of this module on supervised learning.

IIT Madras Production

**Funded by
Department of Higher Education
Ministry of Human Resource Development**

Government of India

www.nptel.ac.in

Copyrights Reserved