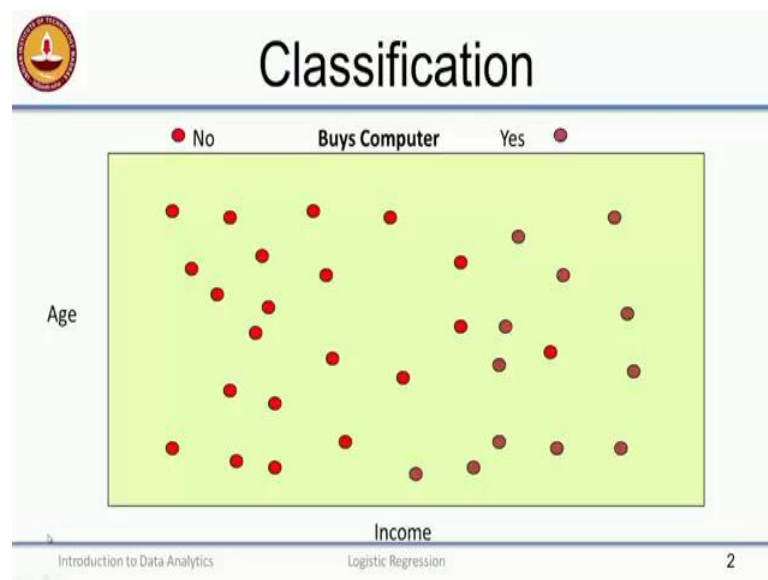**Introduction to Data Analytics**
Prof. Nandan Sudarsanam and
Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 05
**Lecture - 23**
Logistic Regression

Hello and welcome to this module on Logistic Regression.
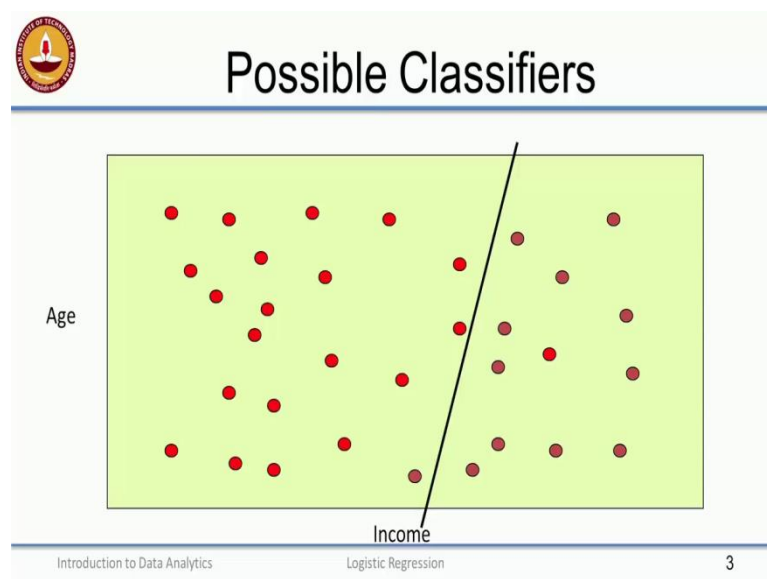
(Refer Slide Time: 00:15)

So, we have looked at the problem of classification earlier and here is an example from one of the earlier modules. So, the users not in brown here are those who bought a computer and those marked in red are people, who did not buy computer.
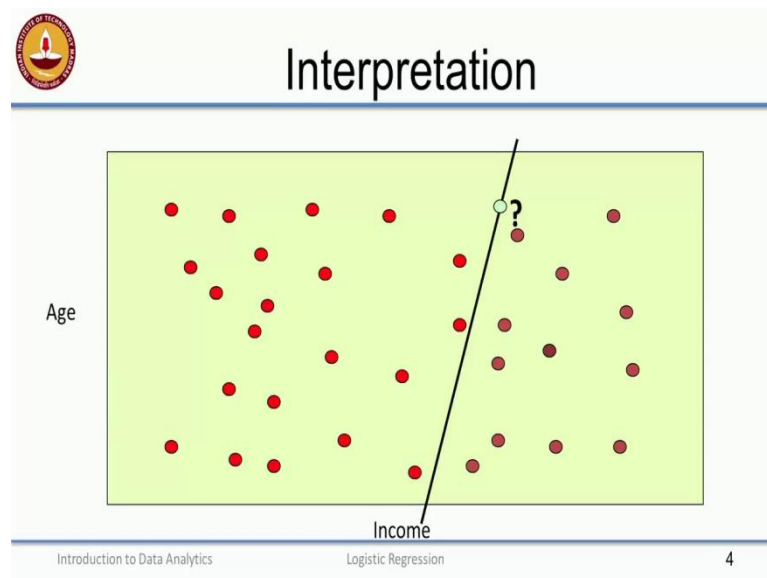
(Refer Slide Time: 00:29)
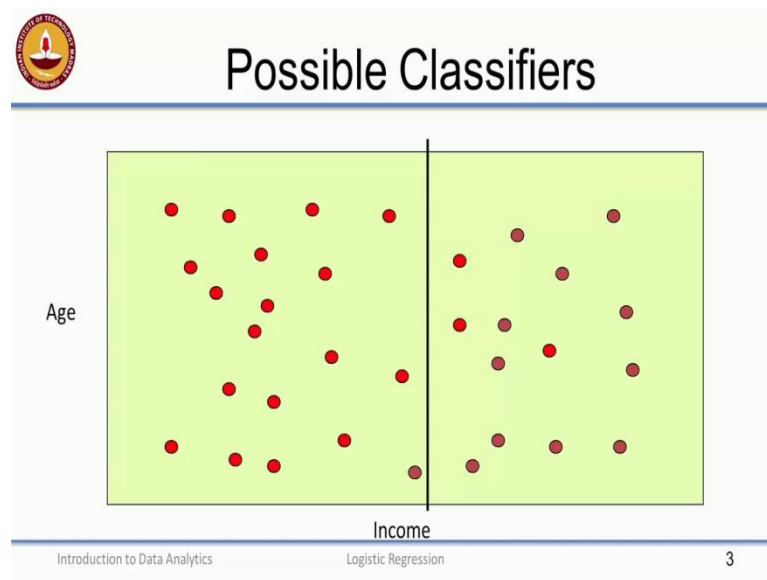
And the goal of classification we said earlier is to find a decision surface that would help us separate people who buy computers from those who do not buy computers. There are different ways in which you could have these decision surfaces and we looked at a few.
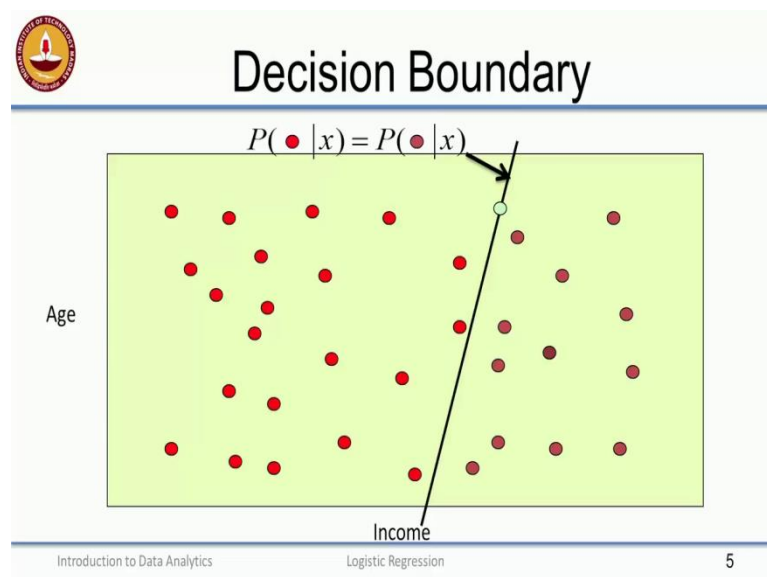
(Refer Slide Time: 00:41)

Now, let us step back and ask the question, what exactly does this decision surface mean.

Specifically let me ask the question, what is the data point that lie on a decision surface belong to, is it buy computers or does not buy computers.

So, one way of thinking about it is to say that this decision surface denotes all the data points for which the probability of it being red is equal to the probability of it being brown. This essentially means that for the points on the boundary the decision boundary you are not able to make a decision as to whether you will buy computer or does not buy a computer.

(Refer Slide Time: 01:29)
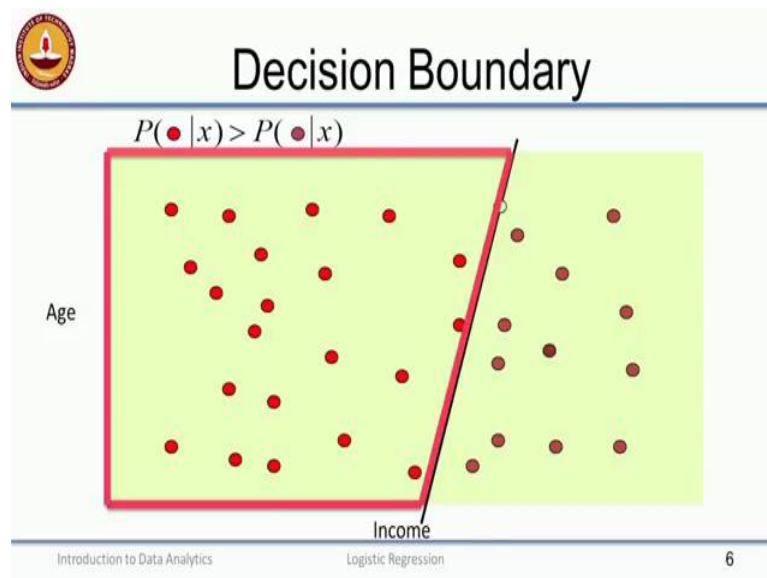
So, what is it tell us about the points that lie to one side of the boundary? So, the points that lie to one side of the boundary are those, where the probability that the person will not buy a computer in this case is higher than the probability that he will buy a computer. So, the one way of thinking about the decision boundary is that it models all the points, where both the classes are equally likely or equally probable to occur.

(Refer Slide Time: 01:57)

So, if you want to go beyond classification, so you might be interested in knowing what is the actual probability of a specific class given a data point. Not just in finding the right

classification, you really like to know what is the probability that the person buys a computer given the age and income of the person versus the probability that the person will not buy a computer given the age and income of the person. So, why would you want to know this kind of probability or the class label?

So, one example is you could think of in medical domain. Suppose I say that, you have a specific disease or the patient walks into the hospital and the doctor says that the patient has a specific disease and you would like to know if the, how confident is the Doctor of the prediction. So, the Doctor says I am 95 percent sure that this patient has the disease, then you certainly would go into the treatment. So, like wise when you have a classifier that is going to give you a class label you would like to know, how sure the classifier is of the class label and that is one application, where you would like to see these kinds of probabilities.

So, one way to approach predicting probabilities instead of just the class labels could be to treat it as a regression problem. So, let us stop and think about how you would treat classification as a regression problem.

(Refer Slide Time: 03:25)

Regression for classification

- Use an indicator variable for class
  - 1 for *buys* and 0 for *does not buy*

$X_1 = \langle 30000, 25 \rangle, Y_1 = DoesnotBuyComputer$
$X_2 = \langle 80000, 45 \rangle, Y_2 = BuysComputer$
$\vdots$

$\longrightarrow$

$X_1 = \langle 0.15, 0.25 \rangle, Y_1 = 0$
$X_2 = \langle 0.4, 0.45 \rangle, Y_2 = +1$
$\vdots$

  - Use linear regression!
  - $f(x)$ can be interpreted as $p(y=1|x)$

Introduction to Data Analytics   Logistic Regression   8

So, normally in classification, so you have labels, who does not buy a computer or buys a computer. So, instead of using these labels you could use an indicator variable for the class. So, if the user is or the customer is going to buy a computer I would say the output is 1, if the customer is not going to buy a computer I would say the output is 0. Now,

your data gets transformed into a regression problem now instead of a classification problem, where you have 0's and 1's as your response variables and the actual attributes of the data has the predicted variables for the regression problem.

And you could use linear regression here, we all know about linear regression now; you could use linear regression here. And the finally, their function that if it f of x can be interpreted as the probability that the output y will be 1 given the data x, that seems like a reasonable way of doing classification. So, whenever the probability is greater than 0.5, you would say that x belongs to class 1, the probability is less than 0.5 you will say x belongs to class 0, that it is actually a valid way of doing a classification using linear regression, but there are some problems with that. So, what are the problems?

(Refer Slide Time: 04:50)

So, linear regression is not really limited in range, the output can go from -infinity to +infinity. So, typically this output cannot be interpreted as a probability, when you troublesome it is the fact that the output can be negative and therefore, this certainly cannot be interpreted as a probability even if you think of doing some kind of normalization. Having said that I should say, it actually works in practice, if you do not really want to treat it as probability, but just as a classifier, you know if it is greater than 0.5 it take it as 1 and lesser than 0.5 take it as 0 it works well, it works in practice, but not that well and there is way of doing better than just using simple linear regression.

(Refer Slide Time: 05:33)

So, I want to use linear regression still, but I am going to do that on a transformed function. If the transformation that we are going to talk about here is called the logistic function or the logit function, so let us have some notation here. Let p(x) denotes the probability that the output y is 1 given x, and then the logit transformation is given by the $\log\left(\frac{p(x)}{1-p(x)}\right)$.
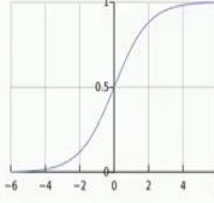
So, if you think about the binary problem, so p(x) is the probability of the output being 1 and 1 - p(x) is a probability of the output being 0. So, essentially you are taking this ratio of the probability of success to the probability of failure. So, this is known as odds and so this sometimes known as the log odds function.

So, now, what are we going to do in logistic regression is essentially try to fit a linear regression model to this logistic function as the output. So, essentially we end up saying that your $\log\left(\frac{p(x)}{1-p(x)}\right)$ can be modeled as some linear function, which is $\beta_0 + x * \beta_1$. So, if you think about it you can solve for p(x) from this kind of an expression and then you end up having p(x) looking like a sigmoid function. So, $\frac{e^{\beta_0 + x\beta}}{1 + e^{\beta_0 + x\beta}}$ and you can simplify that and the functional form that you are going to get is something like this.

So, you can see that the it behaves like a probability function. So, it transfer only from 0 to 1 and by varying the value of β what you are going to do is your going to vary the slope and by varying the value of β_0, you are going to vary where the function is going to rise. So, this gives us a very valid way of fitting probabilities, there is no problem with interpreting p(x) fitted in this fashion as a probability. So, earlier we trying to interpret f(x) in a linear regression model as a probability had problems, so we could not do that, because it could be a negative as we saw earlier.

But, in this case since p(x) is going to be limited between 0 and 1 he might as well interpreted as a probability is it that right model for doing it that is an open question, it depends on the domain that your working in, but it is fairly widely used and it is very power in resolves in a very powerful classifier and which you can use in variety of different settings, whether this assumption is actually supported by the data or not it seems to work well in practice.

So, that fig did with the linear regression case we will predict the classes 1 and the probability of x is greater than 0.5 and 0 otherwise and this essentially you can show if this minimizes the misclassification rate given the form of the predictor that we had on the previous line one thing to note. So, even though p(x) is given by this exponential function, the actual classification boundary…

So, what is the decision boundary? Decision boundary is the point, where the probability of class 1 is equal to the probability of class 2 or class 1 and class 0 probabilities are equal. So, you with the little bit of thought you can see that the decision boundary is still given by a line which essentially $\beta_0 + x\beta_1 = 0$. So, that gives you the decision boundary of the logistic regression classified as well and hence this is also a linear classifier and I mentioned earlier it is pretty powerful and works well in practice.
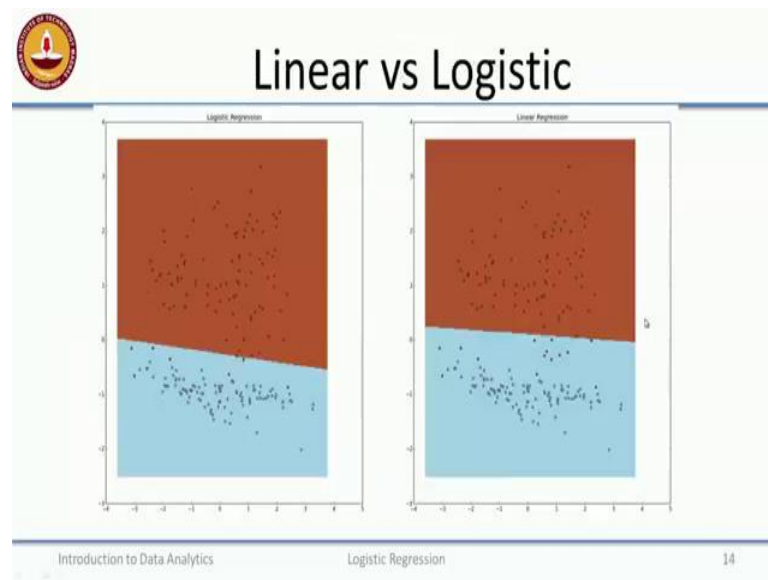
So, let us look at an example of what happens when we fit data using logistic regression versus linear regression. So, here is a two class problem, so the data points or either in blue or in red and the shading in the region indicates, what is the class label that would be predicted by the classifier in those regions. So, on the right hand side you have slides I mean you have the prediction made by fitting a linear regression to the indicator variable on the left hand side you have the output given by logistic regression.

So, you can see that linear regression actually makes a certain errors closer to the boundary that is because linear regression is essentially limited at the rate at which the curves can climb and when closer to the boundary when there are points that are bunch together from one class, but little further away from the rest of the class linear regression is not able to model those successfully, while logistic regression by virtue of the fact that you could have a steep climb from 0 to 1 is able to capture those data points. So, this is essentially the difference between linear and logistic regression. So, far I have been talking about binary classification problems, because they are easier to illustrate and kind of understand the basics behind.

(Refer Slide Time: 11:04)

Multiple Classes

- Suppose there are k classes
- Each class gets a set of parameters

$$p(Y = c \mid x) = \frac{e^{\beta_0^{(c)} + x \cdot \beta^{(c)}}}{\sum_{l} e^{\beta_0^{(l)} + x \cdot \beta^{(l)}}}$$

  - Traditionally the parameters of the first or the last class in some arbitrary order is set to 0

But, then logistic regression can be extended to multiple classes as well, suppose there are k classes then I would say that each class gets the different set of parameters β_0 and β for that specific class. So, in that case what happens is your probability of... So, the probability that particular class is the right class for a data point is given by $e^{\beta 0(c) + x\beta(c)}$ it is our essentially the parameters specific to the class and divided by the total the normalizing factor, which is essentially the numerators sums for all the data points.

To make the problem somewhat easier traditionally the parameters of one of the classes, it could be either the first class by numbering from 0 to k or it could be the last class which is k is set to 0 and you can think about it, it really does not affect what the classifier the decision boundary that you are going to learn it will change the parameters that you are learning, but the decision boundary that you learn will not be affected.

So, in a sense you will be left with fewer parameters that you have to estimate that is because you are talking about probability distributions here and we know that as soon as you fix n outcomes in a discrete probability distribution of the n + 1 the outcome is automatically fixed in total of n + 1 outcome. So, far we have been looking at the basic model and logistic regression and I will end this module here and for the next module we will look at how will actually learn the parameters of this logistic regressions classifier.