Intro to Logistic Regression

In this video, we'll learn a machine learning method called Logistic Regression which is used for classification.

In examining this method, we'll specifically answer these three questions.

- ✓ What is logistic regression?
- ✓ What kind of problems can be solved by logistic regression?
- ✓ In which situations do we use logistic regression?

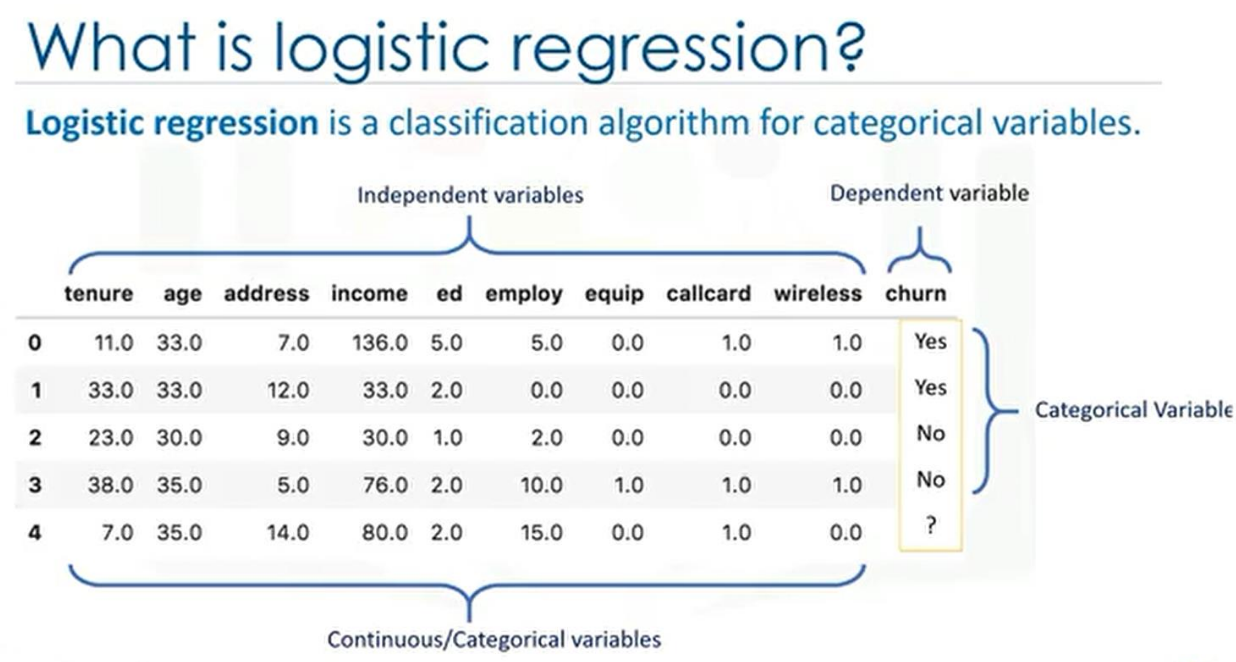
So, let's get started.

Logistic regression: Logistic regression is a statistical and machine learning technique for classifying records of a dataset based on the values of the input fields.

Let's say we have a telecommunication dataset that we'd like to analyze in order to understand which customers might leave us next month. This is historical customer data where each row represents one customer.

Imagine that you're an analyst at this company and you have to find out who is leaving and why?

You'll use the dataset to build a model based on historical records and use it to predict the future churn within the customer group.



	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

The dataset includes information about services that each customer has signed up for, customer account information, demographic information about customers like gender and age range and also customers who've left the company within the last month. The column is called churn.

We can use logistic regression to build a model for predicting customer churn using the given features. In logistic regression, we use one or more independent variables such as tenure, age, and income to predict an outcome, such as churn, which we call the dependent variable representing whether or not customers will stop using the service.

Logistic regression: Logistic regression is analogous to linear regression but tries to predict a categorical or discrete target field instead of a numeric one.

Note: In linear regression, we might try to predict a continuous value of variables such as the

- price of a house,
- blood pressure of a patient, or
- fuel consumption of a car.

Note: in logistic regression, we predict a variable which is binary such as

- yes/no,
- true/false,
- successful or not successful,
- pregnant/not pregnant, and so on,

all of which can be coded as zero or one.

In logistic regression independent variables should be continuous.

If categorical, they should be dummy or indicator coded. This means we have to transform them to some continuous value.

Note: Please note that logistic regression can be used for both

- ✓ binary classification and
- ✓ multi-class classification.

But for simplicity in this video, we'll focus on binary classification. Let's examine some applications of logistic regression before we explain how they work. As mentioned, logistic regression is a type of classification algorithm, so it can be used in different situations.

For example,

- to predict the probability of a person having a heart attack within a specified time period, based on our knowledge of the person's age, sex, and body mass index. Or
- to predict the chance of mortality in an injured patient or to predict whether a patient has a given disease such as diabetes based on observed characteristics of that patient such as weight, height, blood pressure, and results of various blood tests and so on.
- In a marketing context, we can use it to predict the likelihood of a customer purchasing a product or halting a subscription as we've done in our churn example.
- We can also use logistic regression to predict the probability of failure of a given process, system or product.
- We can even use it to predict the likelihood of a homeowner defaulting on a mortgage.

These are all good examples of problems that can be solved using logistic regression.

Note: Notice that in all these examples not only do we predict the class of each case, we also measure the probability of a case belonging to a specific class.

There are different machine algorithms which can classify or estimate a variable.

Q: The question is, when should we use logistic regression?

Here are four situations in which logistic regression is a good candidate.

First, when the target field in your data is categorical or specifically is binary. Such as zero/one, yes/no, churn or no churn, positive/negative and so on.

Second, you need the probability of your prediction. **For example**, if you want to know what the probability is of a customer buying a product. Logistic regression returns a probability score between zero and one for a given sample of data. In fact, logistic regression predicts the probability of that sample and we map the cases to a discrete class based on that probability.

Third, if your data is linearly separable. The decision boundary of logistic regression is a line or a plane or a hyper plane.

A classifier will classify all the points on

- ✓ one side of the decision boundary as belonging to one class and
- ✓ all those on the other side as belonging to the other class.

For example, if we have just two features and are not applying any polynomial processing, we can obtain an inequality like $\theta_0 + \theta_1 x_1 + \theta_2 x_2 > 0$, which is a half-plane easily plottable.

Note: Please note that in using logistic regression, we can also achieve a complex decision boundary using polynomial processing as well, which is out of scope here. You'll get more insight from decision boundaries when you understand how logistic regression works.

Fourth, you need to understand the impact of a feature. You can select the best features based on the statistical significance of the logistic regression model coefficients or parameters. That is, after finding the optimum parameters, a feature X with the weight Theta one close to zero has a smaller effect on the prediction than features with large absolute values of Theta one.

Note: Indeed, it allows us to understand the impact an independent variable has on the dependent variable while controlling other independent variables.

Let's look at our dataset again. We defined the independent variables as X and dependent variable as Y.

Notice, that for the sake of simplicity we can code the target or dependent values to zero or one.

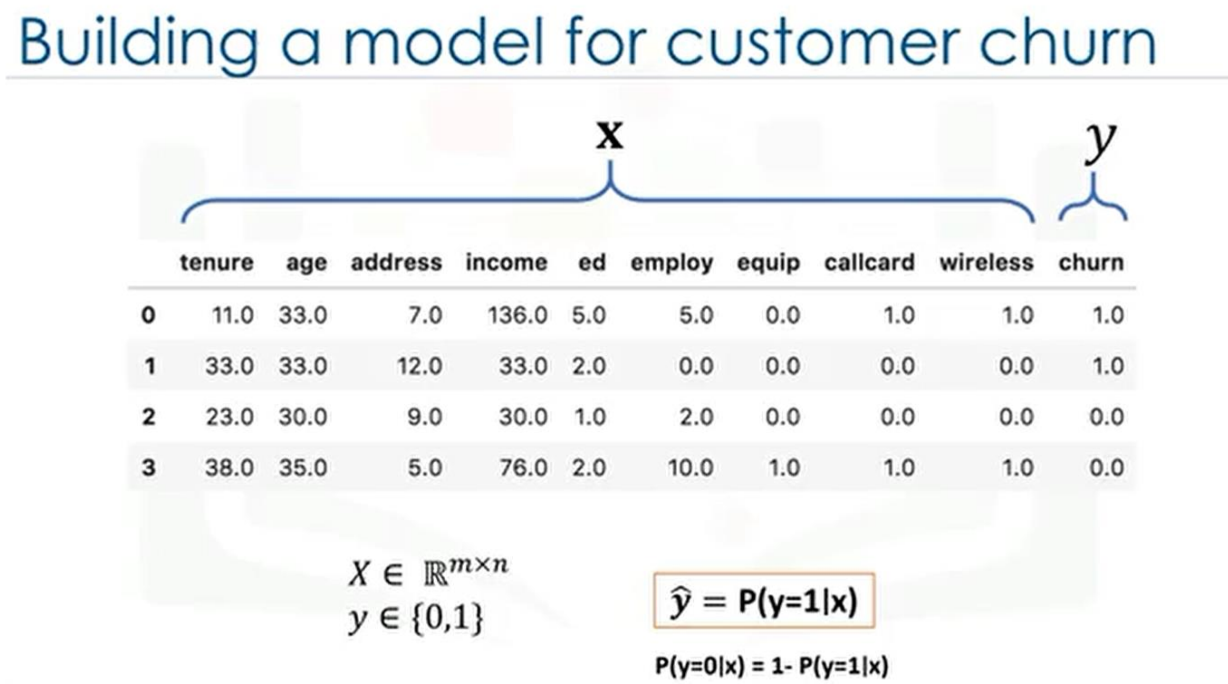
Goal: The goal of logistic regression is to build a model to predict the class of each sample which in this case is a customer, as well as the probability of each sample belonging to a class.

Given that, let's start to formalize the problem.

X is our dataset in the space of real numbers of m by n. That is, of m dimensions or features and n records, and

Y is the class that we want to predict, which can be either zero or one. Ideally, a logistic regression model, so-called Y hat, can predict that the class of the customer is one, given its features X.

It can also be shown quite easily that the probability of a customer being in class zero can be calculated as one minus the probability that the class of the customer is one.



	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1.0
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1.0
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0.0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0.0

$X \in \mathbb{R}^{m \times n}$

$y \in \{0,1\}$

$\boxed{\hat{y} = P(y=1|x)}$

$P(y=0|x) = 1 - P(y=1|x)$

Logistic regression vs Linear regression

In this video, we will learn the difference between linear regression and logistic regression.

We go over linear regression and see why it cannot be used properly for some binary classification problems. We also look at the sigmoid function, which is the main part of logistic regression.

Let's start. Let's look at the telecommunication dataset again.

The goal of logistic regression is to

- ✓ build a model to predict the class of each customer and also
- ✓ the probability of each sample belonging to a class.

Ideally, we want to build a model, y hat,

- that can estimate that the class of a customer is one given its feature is x.
- I want to emphasize that y is the label's vector, also called actual values, that we would like to predict, and
- y hat is the vector of the predicted values by our model.

Model of customer churn data

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
	X									y
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0

$$\boxed{\hat{y} = P(y=1|x)}$$

Mapping the class labels to integer numbers, can we use linear regression to solve this problem?

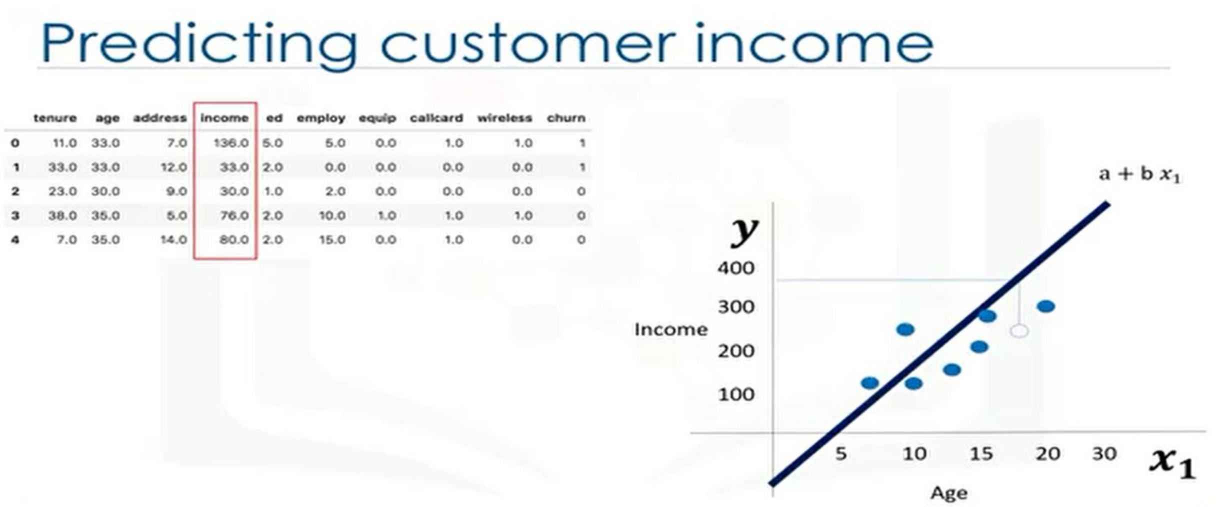
First, let's recall how linear regression works to better understand logistic regression. Forget about the churn prediction for a minute and assume our goal is to predict the income of customers in the dataset. This means that instead of predicting churn, which is a categorical value, let's predict income, which is a continuous value. So, how can we do this?

Let's select an independent variable such as customer age and predict the dependent variable such as income. Of course, we can have more features but for the sake of simplicity, let's just take one feature here. We can plot it and show

- ✓ age as an independent variable and
- ✓ income as the target value we would like to predict.

With linear regression, you can fit a line or polynomial through the data

We can find this line through training our model or calculating it mathematically based on the sample sets. We'll say, this is a straight line through the sample set. This line has an equation shown as $a+bx_1$. Now, use this line to predict the continuous value, y.



	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0

That is, use this line to predict the income of an unknown customer based on his or her age, and it is done.

Q: What if we want to predict churn?

Q: Can we use the same technique to predict a categorical field such as churn?

Okay, let's see. Say, we're given data on customer churn and our goal this time is to predict the churn of customers based on their age.

We have

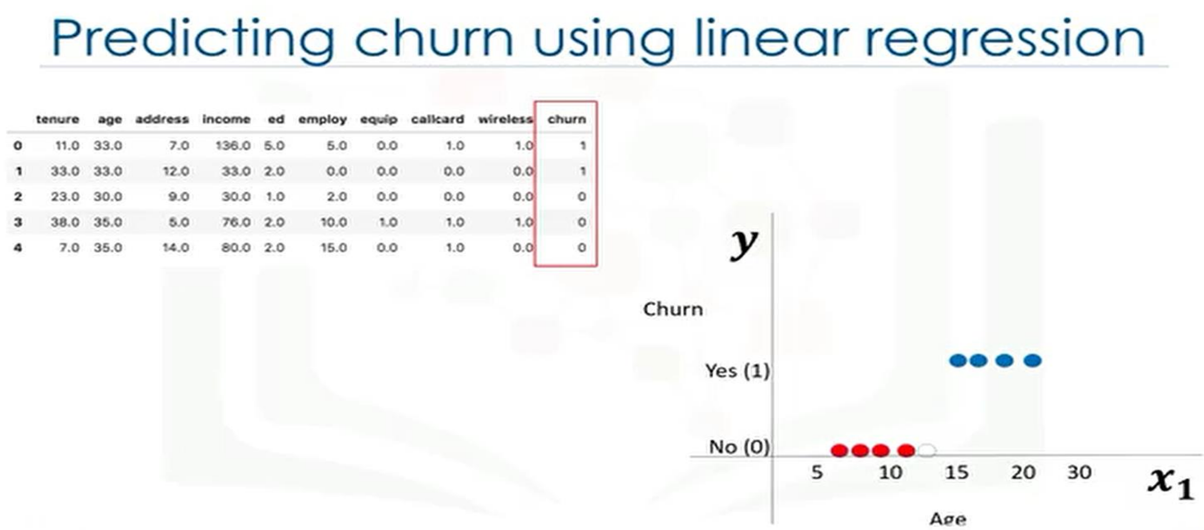
- ✓ a feature, age denoted as x_1, and
- ✓ a categorical feature, churn, with two classes, churn is yes and churn is no.

As mentioned, we can map yes and no to integer values zero and one.

Q: How can we model it now?

Well, graphically, we could represent our data with a scatterplot, but this time, we have only two values for the y-axis. In this plot,

- class zero is denoted in red, and
- class one is denoted in blue.



	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0

Note: Our goal here is to make a model based on existing data to predict if a new customer is red or blue.

Let's do the same technique that we used for linear regression here to see if we can solve the problem for a categorical attribute such as churn. With linear regression, you again can fit a polynomial through the data, which is shown traditionally as $a+bx_1$. This polynomial can also be shown traditionally as $\theta_0 + \theta_1 x_1$.

This line has two parameters which are shown with vector Theta where the values of the vector are θ_0 and θ_1 ($\theta^T = [\theta_0, \theta_1]$). We can also show the equation of this line formally as $\theta^T X = \theta_0 + \theta_1 x_1$. Generally, we can show the equation for a multidimensional space as $\theta^T X = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$,

where

- Theta is the parameters of the line in two-dimensional space or
- parameters of a plane in three-dimensional space, and so on.

Note: As Theta is a vector of parameters and is supposed to be multiplied by x, it is shown conventionally as transpose Theta. Theta is also called the weights factor or confidences of the equation, with both these terms used interchangeably, and X is the feature set which represents a customer.