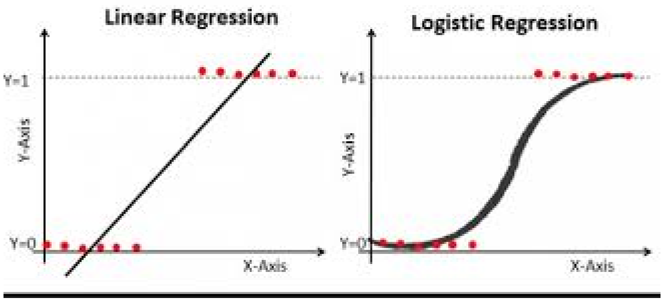
Logistic Regression



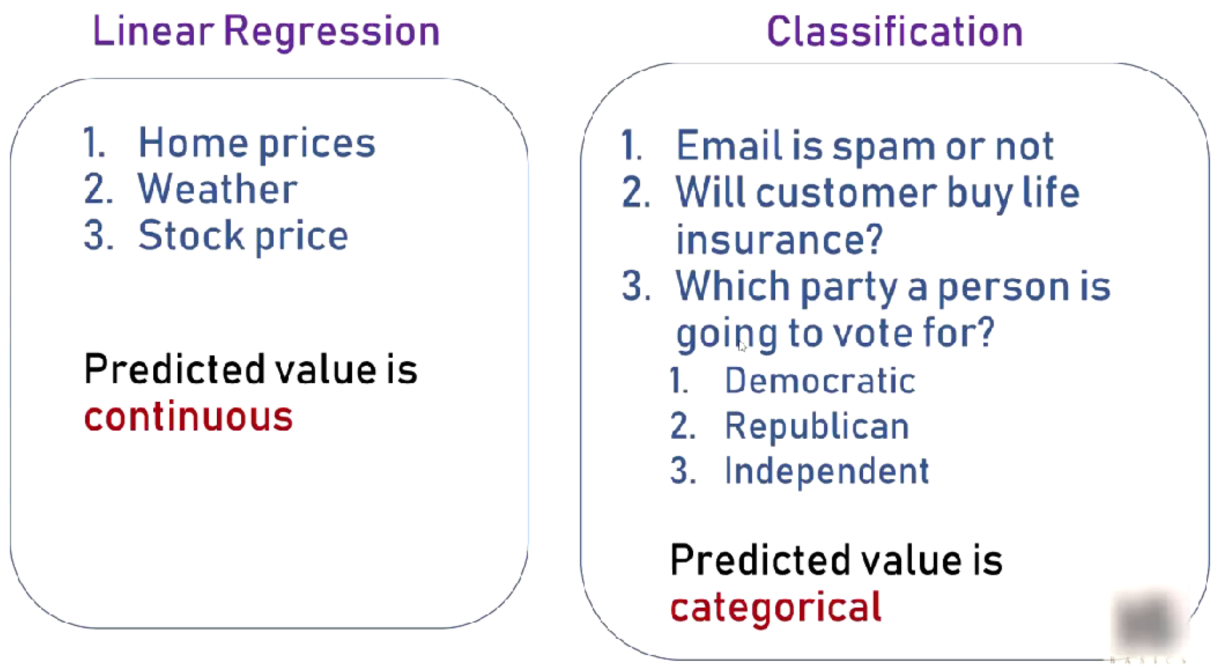
What is logistic regression?

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

The goal of this tutorial is to solve a simple classification problem using Logistic Regression.

Linear Regression versus Logistic Regression

Linear Regression can be used to predict something like homeprices or weather and in all this examples the prediction value is **continues**. There are other type of problems such as Email spam, election. Inn this case the prediction value is **categorical**, because the answer in Email spam is yes/no or in election the answer is one of the person. Hence, the second type of problems is called classification problem.



Linear Regression

1. Home prices
2. Weather
3. Stock price

Predicted value is **continuous**

Classification

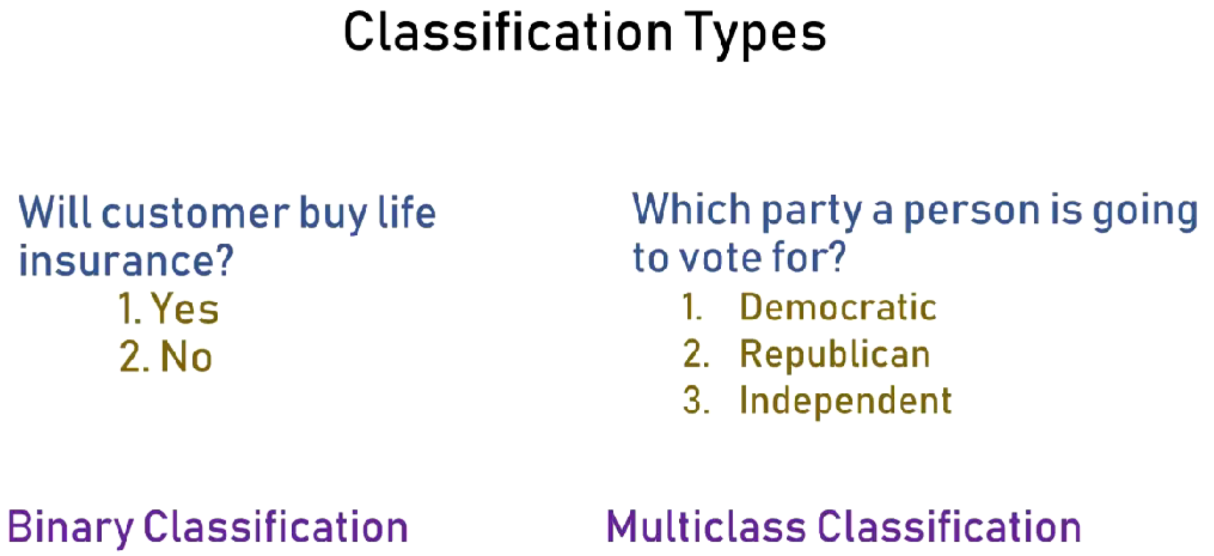
1. Email is spam or not
2. Will customer buy life insurance?
3. Which party a person is going to vote for?
   1. Democratic
   2. Republican
   3. Independent

Predicted value is **categorical**

Logistic regression is one of the techniques used for classification.

There are two types of classification problems:

1. Binary classification (Yes/No)
2. Multiclass Classification (You have more than two categories)



Classification Types

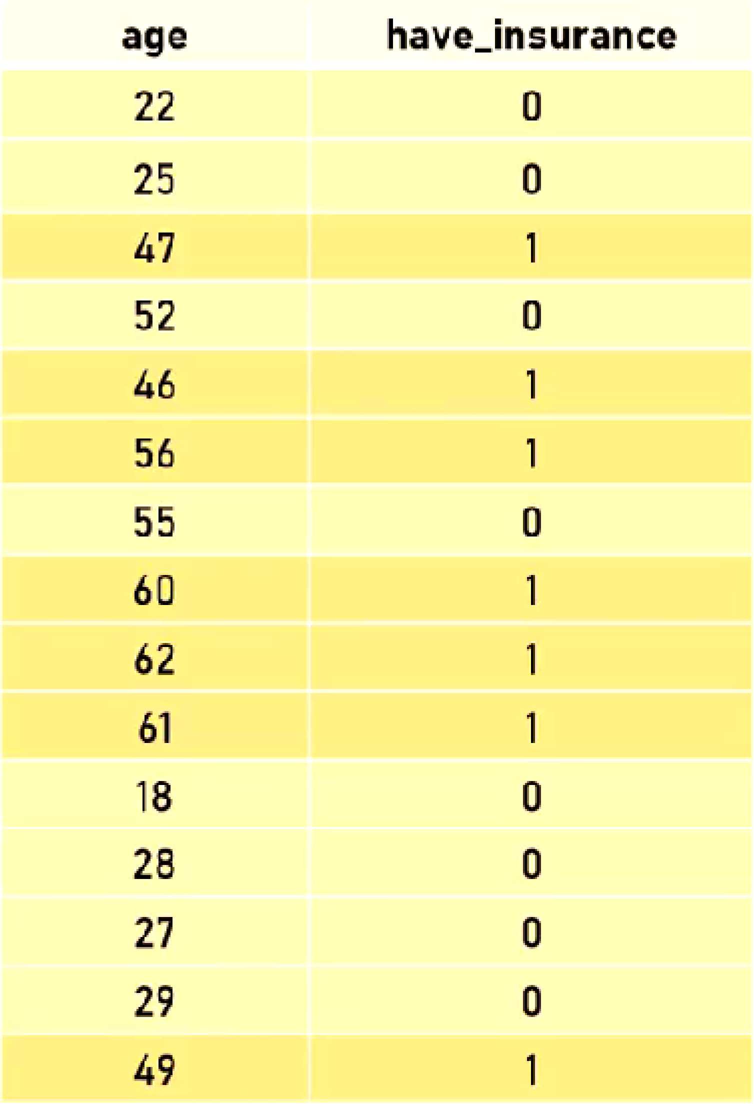
Will customer buy life insurance?
1. Yes
2. No

Binary Classification

Which party a person is going to vote for?
1. Democratic
2. Republican
3. Independent

Multiclass Classification

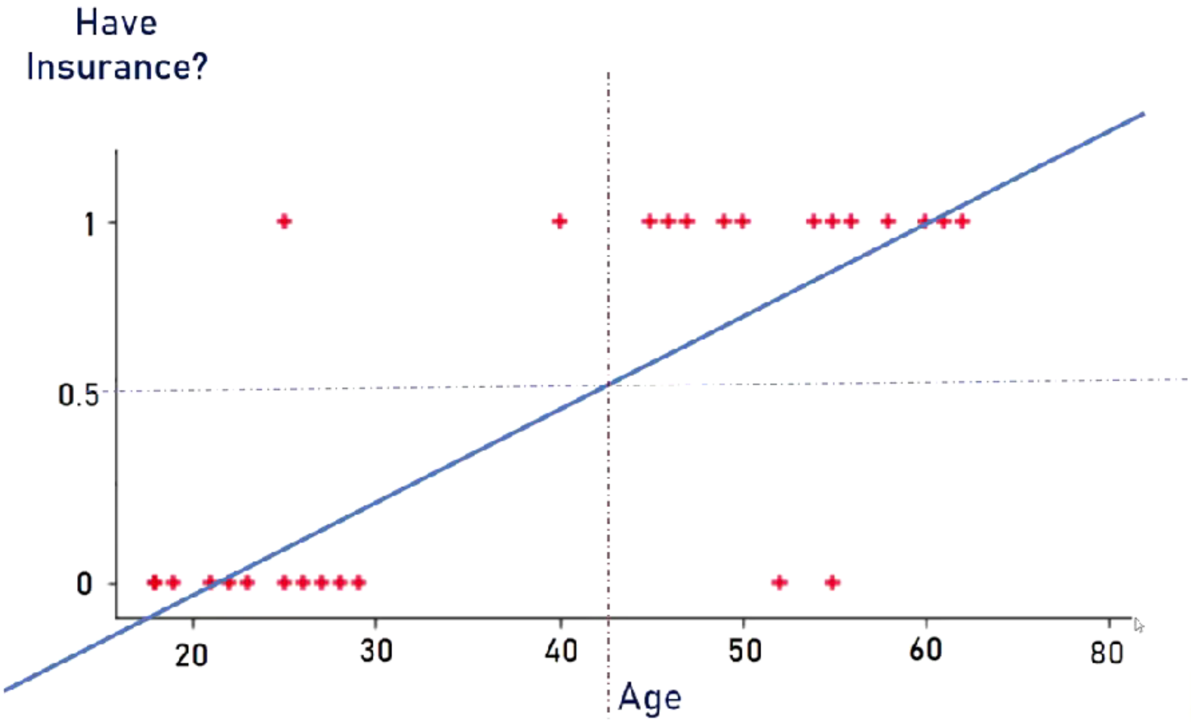
Predicting if a person would buy life insurnace based on his age using logistic regression



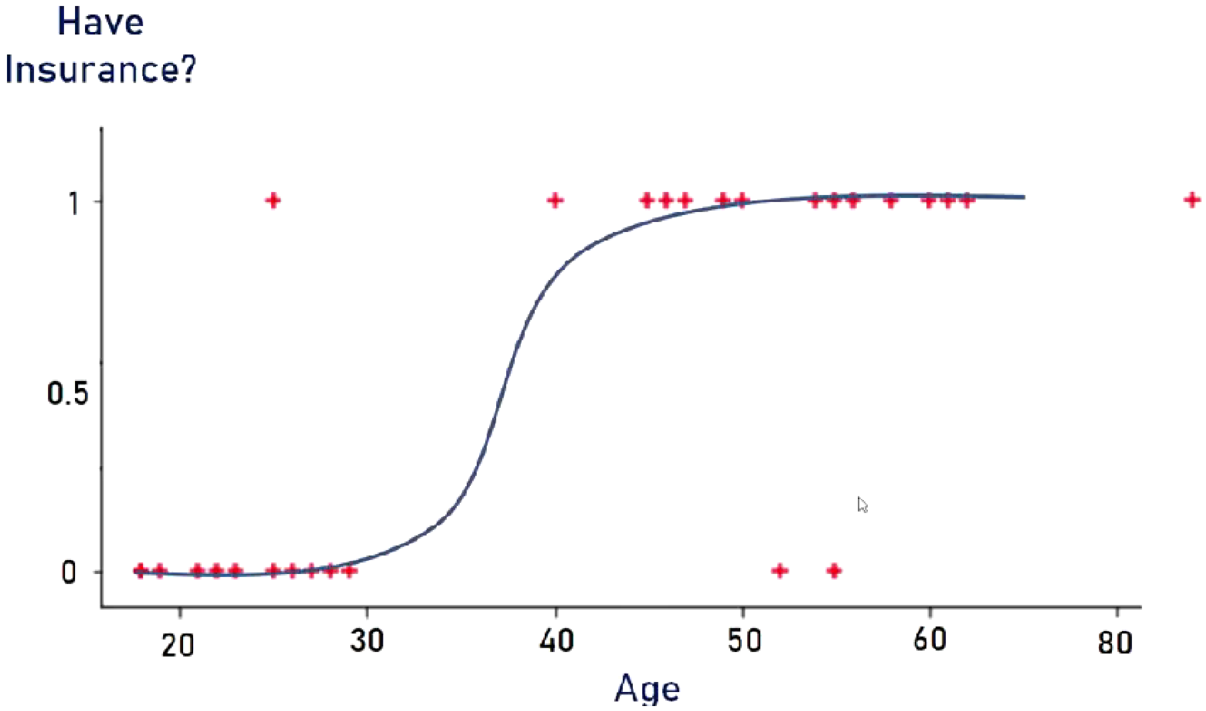
age	have_insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1
18	0
28	0
27	0
29	0
49	1

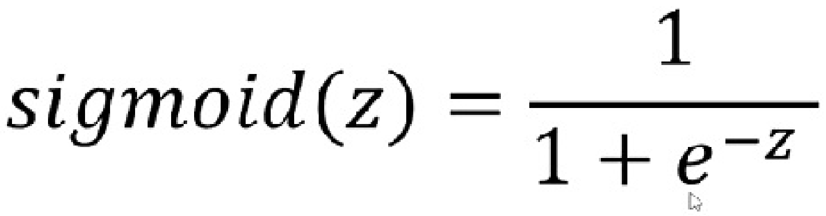
Above is a binary logistic regression problem as there are only two possible outcomes (i.e. if person buys insurance or he/she doesn't).

Predict with Linear Regression

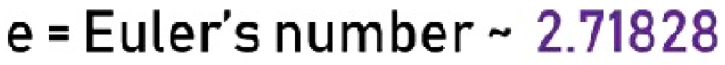


Predict with Logistic Regression

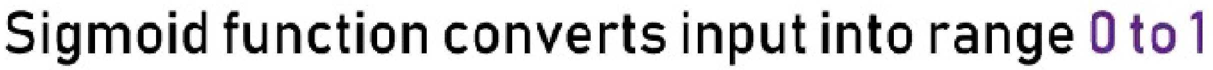




$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$

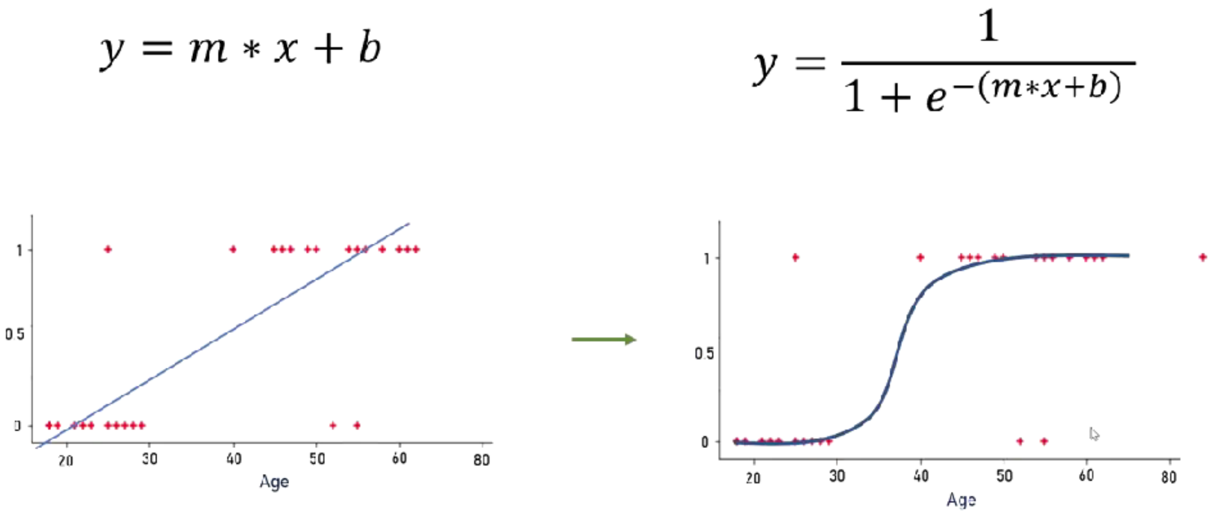


e = Euler's number ~ 2.71828



Sigmoid function converts input into range 0 to 1

z = m∗x + b



$$y = m * x + b$$

$$y = \frac{1}{1 + e^{-(m*x+b)}}$$

In [1]:

```
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline
```

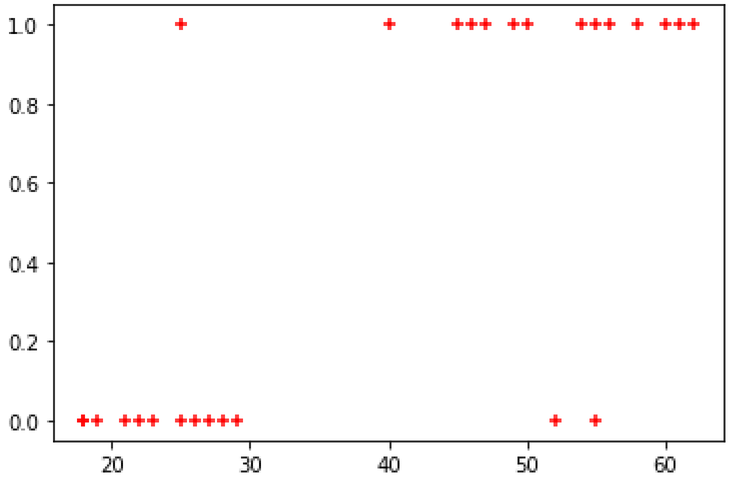
In [2]:
```
df = pd.read_csv("D:/Data_Science/My Github/Machine-Learning-with-Python/7. logist
df.head()
```

Out[2]:

	age	bought_insurance
0	22	0
1	25	0
2	47	1
3	52	0
4	46	1

In [3]:
```
plt.scatter(df.age,df.bought_insurance,marker='+',color='red')
```

Out[3]:
```
<matplotlib.collections.PathCollection at 0x9952e50>
```



In [5]:
```
df.shape
```

Out[5]:
```
(27, 2)
```

In [6]:
```
from sklearn.model_selection import train_test_split
```

In [9]:
```
# Use 90% for train_size
X_train, X_test, y_train, y_test = train_test_split(df[['age']],df.bought_insuranc
```

In [10]:
```
X_test
```

Out[10]:

	age
9	61
24	50
12	27

In [11]:
```
X_train
```

Out[11]:

	age
18	19
21	26
5	56
4	46
22	40
23	45
10	18
15	55
8	62
1	25
19	18
17	58
2	47
20	21
16	25
6	55
26	23
13	29
3	52
25	54
14	49
0	22
11	28
7	60

In [12]:
```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
```

In [13]:
```
model.fit(X_train, y_train)
```

Out[13]:
```
LogisticRegression()
```

In [14]:

```
X_test
```

Out[14]:

	age
9	61
24	50
12	27

In [15]:

```
model.predict(X_test)
```

Out[15]:

```
array([1, 1, 0], dtype=int64)
```

[1,1,0] for age [61,50,27] seems true the old people buy the insurance(i.e. 1) & vice versa

In [16]:

```
# measure the accuracy of your model
model.score(X_test,y_test)
```

Out[16]:

```
1.0
```

which means our model is perfect and this is because out data size is small (only 27 samples)

In [17]:

```
# Predict a probability
model.predict_proba(X_test)
```

Out[17]:

```
array([[0.05938775, 0.94061225],
       [0.20064434, 0.79935566],
       [0.81810182, 0.18189818]])
```

It shows the probability of X_test being in one class versus the other. The first class if the customer will not buy the insurance.

Predict the probebility of age 16?

In [19]:

```
model.predict([[16]])
```

Out[19]:

```
array([0], dtype=int64)
```

So 16 not buy the insurance

Exercise

Download employee retention dataset from here: https://www.kaggle.com/giripujar/hr-analytics (https://www.kaggle.com/giripujar/hr-analytics).

1. Now do some exploratory data analysis to figure out which variables have direct and clear impact on employee retention (i.e. whether they leave the company or continue to work)
2. Plot bar charts showing impact of employee salaries on retention
3. Plot bar charts showing corelation between department and employee retention
4. Now build logistic regression model using variables that were narrowed down in step 1

5. Measure the accuracy of the model

Date	Author
2021-09-02	**Ehsan Zia**