Getting Started with Machine Learning

Raunak J

So far now we've covered regression in detail.

Now let's move on to Classification.

In this tutorial we'll cover Logistic Regression.

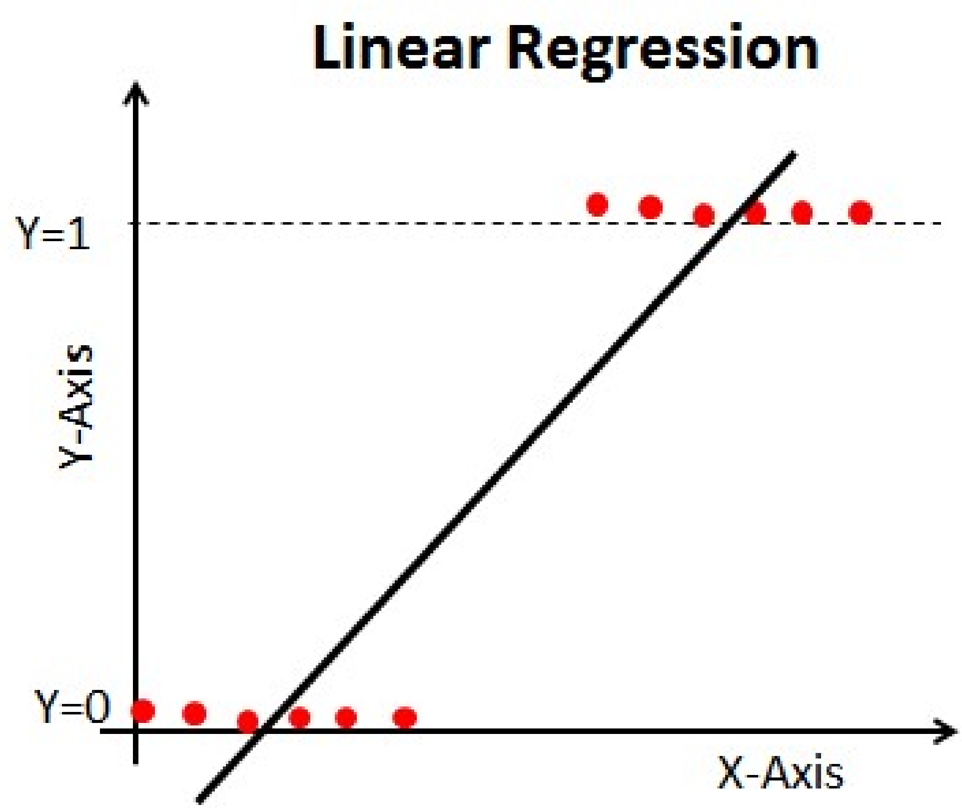
Well, isn't it Logistic REGRESSION ?

Even though it is Logistic Regression, it is a binary classification problem.

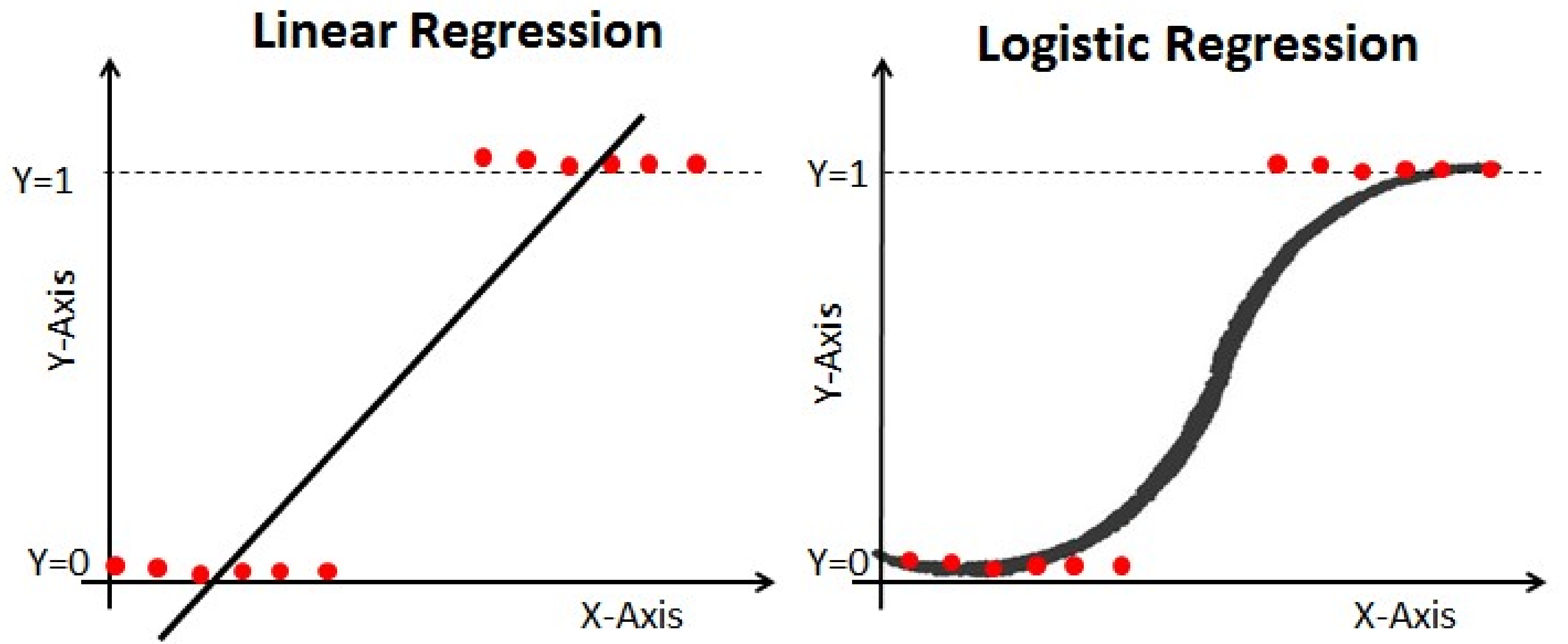
Formal Definition of Logistic Regression :

In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc.

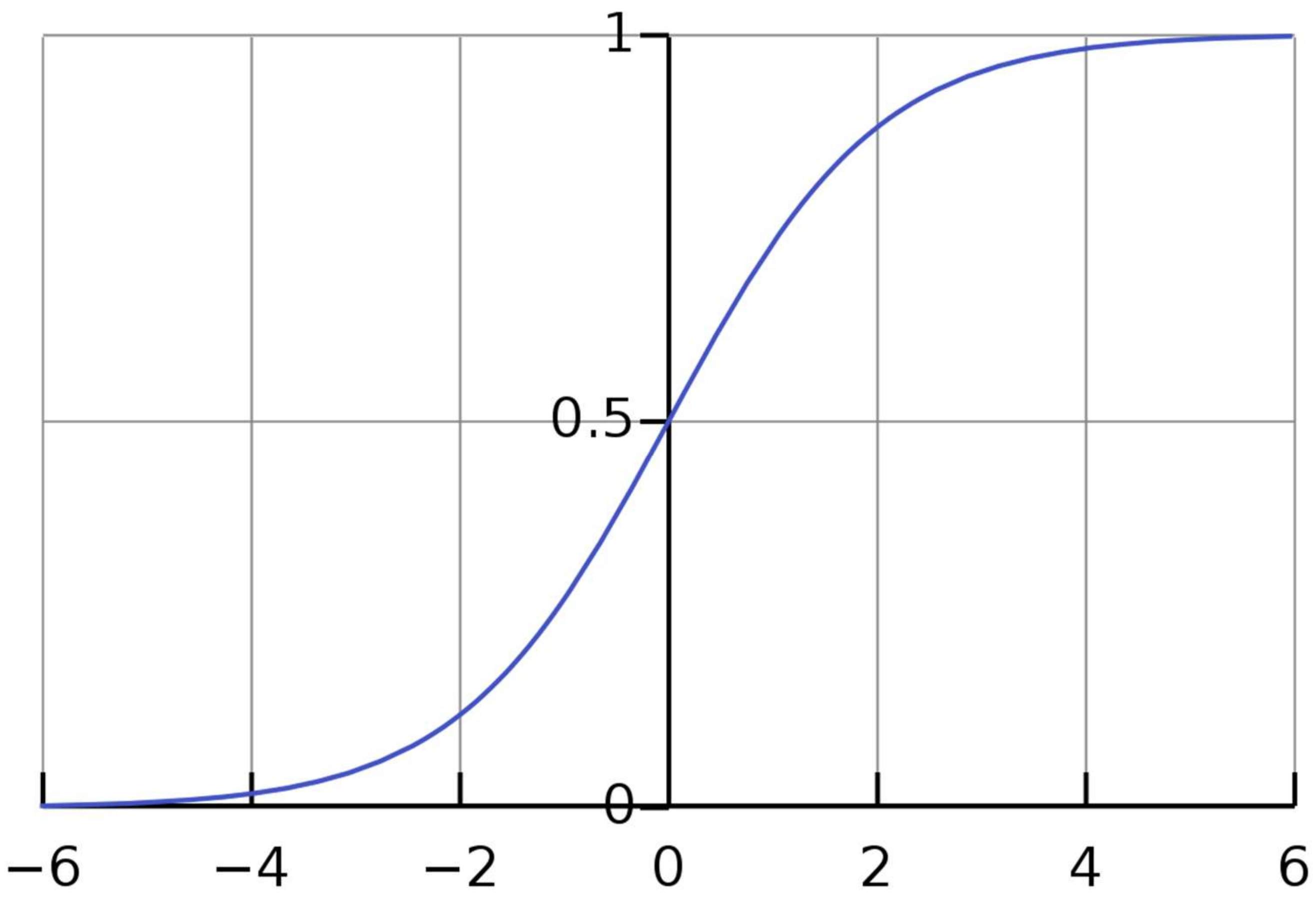
Source :
https://en.wikipedia.org/wiki/Logistic_regression



Source : https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python



Source : https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python

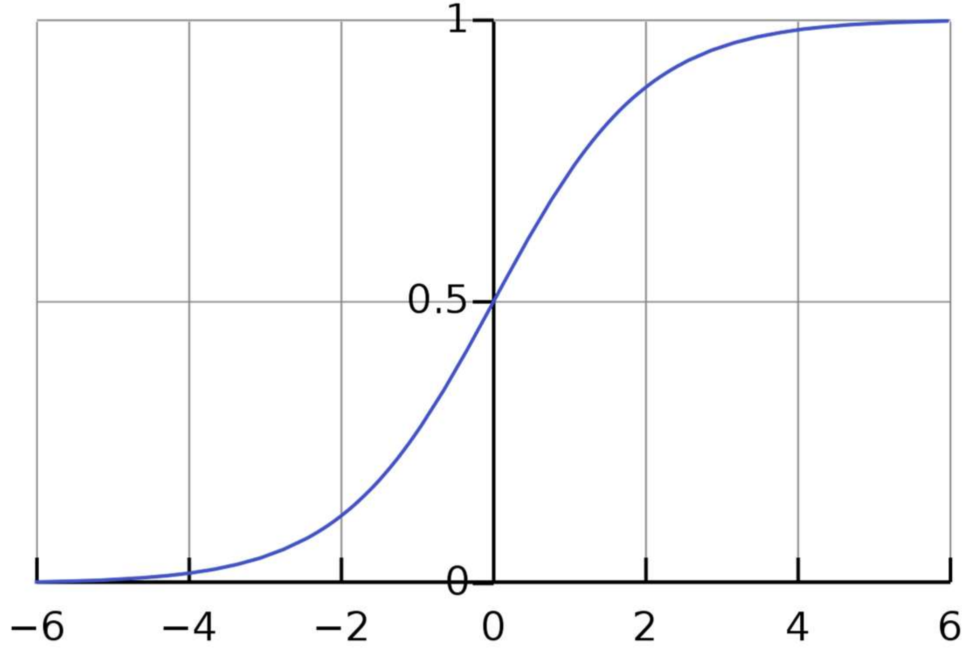


Source : https://en.wikipedia.org/wiki/Logistic_function

Logistic Function.

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

$$x_0 = x\ value\ of\ midpoint$$
$$L = Max\ Value$$
$$k = growth\ rate$$



Logistic Function is also known as Sigmoid Function.

$$\sigma(Z) = \frac{1}{1 + e^{-Z}}$$

Where Z is an equation.

In our case we'll use function of Linear Regression and convert into Logistic Regression using Sigmoid.

Logistic Regression is given by :

$$Z = \beta_0 + (\beta_1 * x_i)$$

$$\sigma(Z) = \frac{1}{1 + e^{-Z}}$$

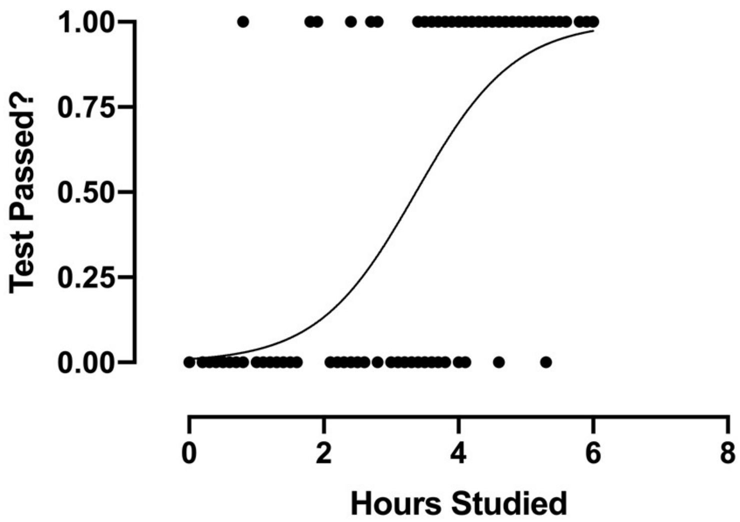
$$\sigma(Z) = \frac{1}{1 + e^{-\{\beta_0 + (\beta_1 * x_i)\}}}$$

$$\hat{y} = \sigma(Z)$$

So $\hat{y} = \sigma(Z)$ yields Probability.

Probabilities lie between 0-1 i.e 0-100%

So we now set a threshold of .5 i.e 50%



Source : https://www.graphpad.com/guides/prism/8/curve-fitting/reg_simple_logistic_and_linear_difference.htm

$So\ \hat{y}\ basically\ yeilds\ probability.$
$So\ probability\ of\ being\ 1\ \&\ probaility\ of\ being\ 0\ should\ be\ calculated.$

$For\ that\ we\ set\ the\ threshold. Usually\ threshold\ is\ set\ to\ 0.5$

$$\hat{y} \geq 0.5 \mid prediction\ y = 1$$

$$\hat{y} < 0.5 \mid prediction\ y = 0$$

Understanding algorithm in terms of Probability :

$$\sigma(Z) = P(\hat{y} = 1|Z; \{\beta_0, \beta_1\})$$

$$\sigma(Z) = P(\hat{y} = 0|Z; \{\beta_0, \beta_1\})$$

$$P(\hat{y} = 0|Z; \{\beta_0, \beta_1\}) = 1 - P(\hat{y} = 1|Z; \{\beta_0, \beta_1\})$$

Understanding Logistic Regression with an example :

Considering the prediction of classifying whether the student is going to pass/fail the examination based on number of hours they've devoted to studying.

Hours of Studying (X)	Examination Outcome (y)
0.50	0
1.50	0
2.00	0
4.25	1
3.25	1
5.50	1

Hours of Studying (X)	Examination Outcome (y)	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x}).(y_i - \overline{y})$
0.50	0	-2.33	-.5	5.4289	.25	1.165
1.50	0	-1.33	-.5	1.7689	.25	.665
2.00	0	-.83	-.5	.6889	.25	.415
4.25	1	1.42	.5	2.0164	.25	.71
3.25	1	.42	.5	.1764	.25	.21
5.50	1	2.67	.5	7.1289	.25	1.335
Sum = 17	Sum = 3	Sum = .02	Sum = 0	Sum = 17.2084	Sum = 1.5	Sum = 4.5

Hours of Studying (X)	Examination Outcome (y)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}).(y_i - \bar{y})$
0.50	0	-2.33	-.5	5.4289	.25	1.165
1.50	0	-1.33	-.5	1.7689	.25	.665
2.00	0	-.83	-.5	.6889	.25	.415
4.25	1	1.42	.5	2.0164	.25	.71
3.25	1	.42	.5	.1764	.25	.21
5.50	1	2.67	.5	7.1289	.25	1.335
Sum = 17	Sum = 3	Sum = .02	Sum = 0	Sum = 17.2084	Sum = 1.5	Sum = 4.5

$$Mean\ of\ x:\ \bar{x} = \frac{\sum x_i}{N} = \frac{17}{6} = 2.83$$

$$Mean\ of\ y:\ \bar{y} = \frac{\sum y_i}{N} = \frac{3}{6} = .5$$

Hours of Studying (X)	Examination Outcome (y)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}).(y_i - \bar{y})$
0.50	0	-2.33	-.5	5.4289	.25	1.165
1.50	0	-1.33	-.5	1.7689	.25	.665
2.00	0	-.83	-.5	.6889	.25	.415
4.25	1	1.42	.5	2.0164	.25	.71
3.25	1	.42	.5	.1764	.25	.21
5.50	1	2.67	.5	7.1289	.25	1.335
Sum = 17	Sum = 3	Sum = .02	Sum = 0	Sum = 17.2084	Sum = 1.5	Sum = 4.5

$$\beta_1 = \frac{(x_i - \bar{x}).(y_i - \bar{y})}{(x_i - \bar{x})^2} = \frac{4.5}{17.2084} = 0.2615$$

$$\beta_0 = \bar{y} - (\beta_1 * x_i) = .5 - (.2615 * 2.83) = -.240045$$

Hours of Studying (X)	Examination Outcome (y)	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x}).(y_i - \overline{y})$
0.50	0	-2.33	-.5	5.4289	.25	1.165
1.50	0	-1.33	-.5	1.7689	.25	.665
2.00	0	-.83	-.5	.6889	.25	.415
4.25	1	1.42	.5	2.0164	.25	.71
3.25	1	.42	.5	.1764	.25	.21
5.50	1	2.67	.5	7.1289	.25	1.335
Sum = 17	Sum = 3	Sum = .02	Sum = 0	Sum = 17.2084	Sum = 1.5	Sum = 4.5

$$Z = \beta_0 + (\beta_1 * x_i) = -.240045 + (.2615 * x_i)$$

$$\sigma(Z) = \frac{1}{1 + e^{-Z}} = \frac{1}{1 + e^{-\beta_0 + (\beta_1 * x_i)}} = \frac{1}{1 + e^{-\{-.240045 + (.2615 * x_i)\}}}$$

$$\hat{y} = \sigma(Z)$$

Hours of Studying (X)	Examination Outcome (y)	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x}).(y_i - \overline{y})$
0.50	0	-2.33	-.5	5.4289	.25	1.165
1.50	0	-1.33	-.5	1.7689	.25	.665
2.00	0	-.83	-.5	.6889	.25	.415
4.25	1	1.42	.5	2.0164	.25	.71
3.25	1	.42	.5	.1764	.25	.21
5.50	1	2.67	.5	7.1289	.25	1.335
Sum = 17	Sum = 3	Sum = .02	Sum = 0	Sum = 17.2084	Sum = 1.5	Sum = 4.5

$Predict\ the\ status\ of\ examination\ outcome\ where\ the\ student\ was\ studying\ for\ 2.7\ hours\ on\ daily\ basis.$

$$\hat{y} = \sigma(Z) = \frac{1}{1+e^{-Z}} = \frac{1}{1+e^{-\beta_0+(\beta_1 * x_i)}} = \frac{1}{1+e^{-\{-.24004 \quad (.2615*2.7)\}}}$$

$$\hat{y} = .6144 \approx 61.44\%$$

This exceeds the threshold of 50% which indicates the outcome is 1, i.e The student is going to pass based on 2.7 hours of preparation on daily basis.

Hours of Studying (X)	Examination Outcome (y)	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x}).(y_i - \overline{y})$
0.50	0	-2.33	-.5	5.4289	.25	1.165
1.50	0	-1.33	-.5	1.7689	.25	.665
2.00	0	-.83	-.5	.6889	.25	.415
4.25	1	1.42	.5	2.0164	.25	.71
3.25	1	.42	.5	.1764	.25	.21
5.50	1	2.67	.5	7.1289	.25	1.335
Sum = 17	Sum = 3	Sum = .02	Sum = 0	Sum = 17.2084	Sum = 1.5	Sum = 4.5

$$Predict\ the\ status\ of\ examination\ outcome\ where\ the\ student\ was\ studying\ for\ 0.75\ hours\ on\ daily\ basis.$$

$$\hat{y} = \sigma(Z) = \frac{1}{1+e^{-Z}} = \frac{1}{1+e^{-\beta_0+(\beta_1 * x_i)}} = \frac{1}{1+e^{-\{-.240045+(.2615*.75)\}}}$$

$$\hat{y} = .4890 \approx 48.90\%$$

This does not exceed the threshold of 50% which indicates the outcome is 0, i.e The student is going to fail based on .75 hours of preparation on daily basis.

Thank You