Tutorial Dataset

In this tutorial we will use a contrived dataset.

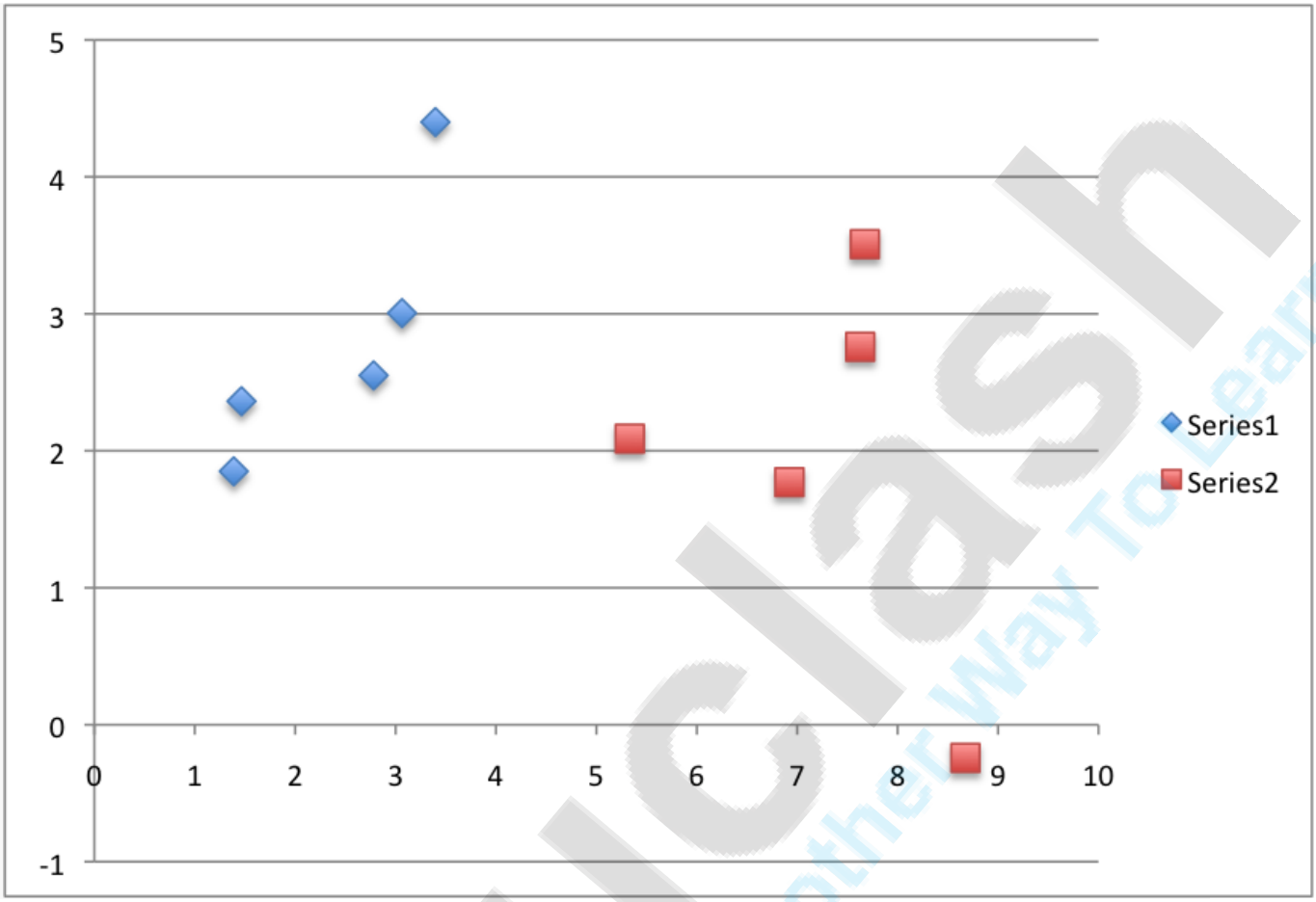
This dataset has two input variables (X1 and X2) and one output variable (Y). In input variables are real-valued random numbers drawn from a Gaussian distribution. The output variable has two values, making the problem a binary classification problem.

The raw data is listed below.

	X1	X2	Y
2	2.7810836	2.550537003	0
3	1.465489372	2.362125076	0
4	3.396561688	4.400293529	0
5	1.38807019	1.850220317	0
6	3.06407232	3.005305973	0
7	7.627531214	2.759262235	1
8	5.332441248	2.088626775	1
9	6.922596716	1.77106367	1
10	8.675418651	-0.2420686549	1
11	7.673756466	3.508563011	1

Below is a plot of the dataset. You can see that it is completely contrived and that we can easily draw a line to separate the classes.

This is exactly what we are going to do with the logistic regression model.



Logistic Regression Tutorial Dataset

Logistic Function

Before we dive into logistic regression, let's take a look at the logistic function, the heart of the logistic regression technique.

The logistic function is defined as:

Logistic Function (Sigmoid Function)

transformed = 1 / (1 + e^-x)

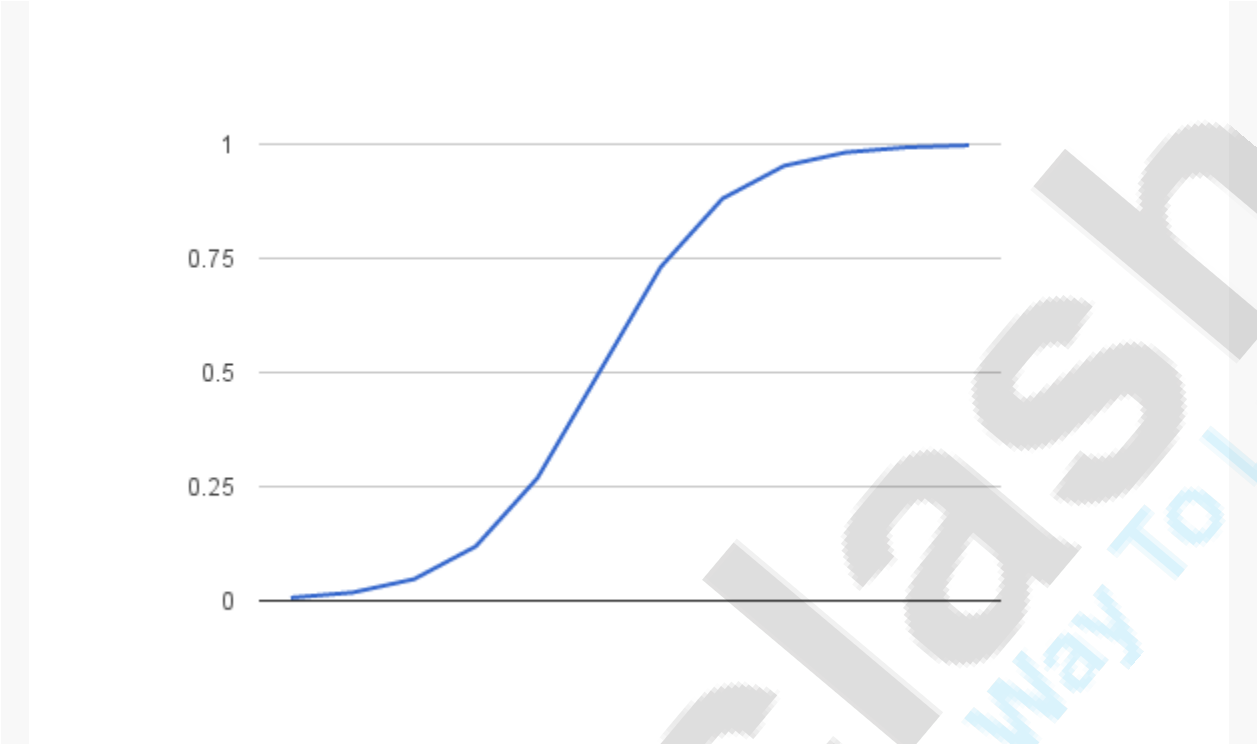
Where e is the numerical constant Euler's number and x is a input we plug into the function.

Let's plug in a series of numbers from -5 to +5 and see how the logistic function transforms them:

	X	Transformed
1	X	Transformed
2	-5	0.006692850924
3	-4	0.01798620996
4	-3	0.04742587318
5	-2	0.119202922
6	-1	0.2689414214
7	0	0.5
8	1	0.7310585786
9	2	0.880797078
10	3	0.9525741268
11	4	0.98201379
12	5	0.9933071491

You can see that all of the inputs have been transformed into the range [0, 1] and that the smallest negative numbers resulted in values close to zero and the larger positive numbers resulted in values close to one. You can also see that 0 transformed to 0.5 or the midpoint of the new range.

From this we can see that as long as our mean value is zero, we can plug in positive and negative values into the function and always get out a consistent transform into the new range.



Logistic Function

Logistic Regression Model

The logistic regression model takes real-valued inputs and makes a prediction as to the probability of the input belonging to the default class (class 0).

If the probability is > 0.5 we can take the output as a prediction for the default class (class 0), otherwise the prediction is for the other class (class 1).

For this dataset, the logistic regression has three coefficients just like linear regression, for example:

output = b0 + b1*x1 + b2*x2

The job of the learning algorithm will be to discover the best values for the coefficients (b0, b1 and b2) based on the training data.

Unlike linear regression, the output is transformed into a probability using the logistic function:

$$p(class=0) = 1 / (1 + e^{(-output)})$$

In your spreadsheet this would be written as:

$$p(class=0) = 1 / (1 + EXP(-output))$$

Calculate Prediction

Let's start off by assigning 0.0 to each coefficient and calculating the probability of the first training instance that belongs to class 0.

$$B0 = 0.0$$

$$B1 = 0.0$$

$$B2 = 0.0$$

The first training instance is: x1=2.7810836, x2=2.550537003, Y=0

Using the above equation we can plug in all of these numbers and calculate a prediction:

$$prediction = 1 / (1 + e^{(-(b0 + b1*x1 + b2*x2))})$$

$$prediction = 1 / (1 + e^{(-(0.0 + 0.0*2.7810836 + 0.0*2.550537003))})$$

$$prediction = 0.5$$