Logistic Regression

What is?

* Used for classification
* Logistic regression is a statistical and machine learning technique for classifying records of a dataset based on the values of the input fields.
* Only predicts binary categorical (True False) or multiclass. (and probability?)
* Data should be continuous. If not it needs to be dummy or indicator coded:

Transformed to some continuous value.

When to use?

* If data is binary
* If probabilistic results are required
* When you need a linear decision boundary
* If you know polynomials, boundaries can be complex polynomials
* Need to understand the impact of the feature, these can be selected based on coefficients or parameters.

Building a model

So

X is a dataset in the space of real numbers of m by n (m features, n records) that we want to predict, which can be ether 0 or 1.

**Logistic vs. Linear Regression**

Why can linear regression not be used for some binary classification problems?

* 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.

In a linear regression the line can be described as

Or

Where represents the intersect and slope

The line has two parameters shown with ‘vector theta’

So

Transpose simply means to convert theta from a column vector to a row vector so its dimensions properly align with X for mathematical operation which in this case represents the feature vector. So represents the ‘dot product’ or multiplication of the coefficient vector ‘theta transpose’ and the feature vector ‘X’. Theta 0 in this case will always be represented as alone as the first term of X is always 1 so Theta 0 \* X 0 will always just equal theta 0.

Sigmoid function

Also called the logistic function