

Today's Topics



Linear Regression, Logistic Regression and K - Nearest Neighbours!



Supervised Learning

All these algorithms are supervised learning algorithms.


What are supervised learning algorithms?

Supervised learning algorithms try to model relationships and dependencies between the target prediction output and the input features such that we can predict the output values for new data based on those relationships which it learned from the previous data sets.



Linear Regression

It's a method to predict a **target variable** by fitting the *best linear relationship* between the dependent and independent variable.

$$Y = X_1 + X_2 + X_3$$


Dependent Variable

Independent Variable

Outcome Variable

Predictor Variable

Response Variable

Explanatory Variable

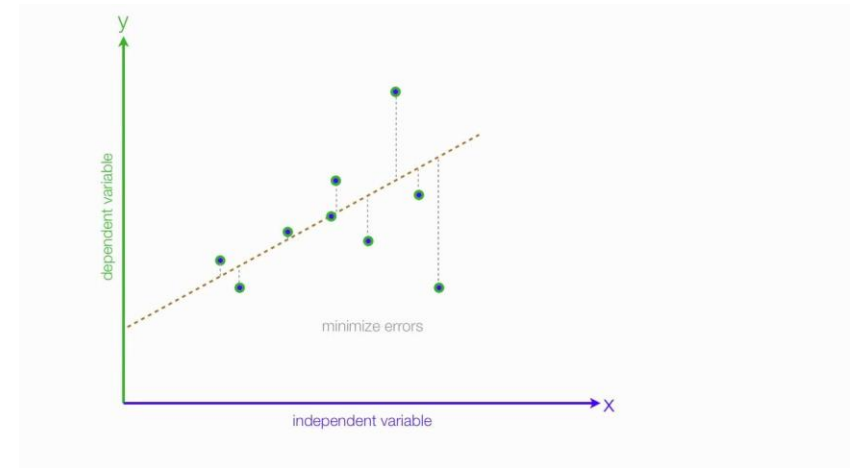
Parts of the equation

It can be of any shape depending on the number of independent variables (a point on the axis, a line in two dimensions, a plane in three dimensions, or a hyperplane in higher dimensions)

$$y = k + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Diagram illustrating the parts of the equation:

- y : Dependent Variable
- k : Intercept
- $\beta_1, \beta_2, \dots, \beta_n$: Coefficient
- x_1, x_2, \dots, x_n : Predictors



Best Fit!

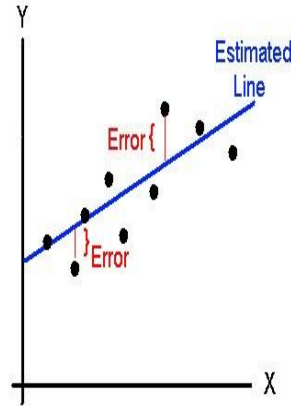
Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$



So what should be the best fit?

We want the line to minimize the error, and the way we want to do it is such that we can minimise the deviations between the two!



Mean Squared Error

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$



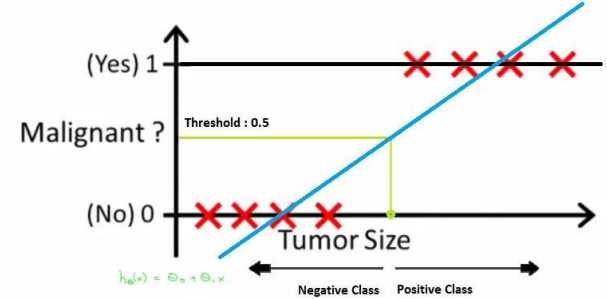
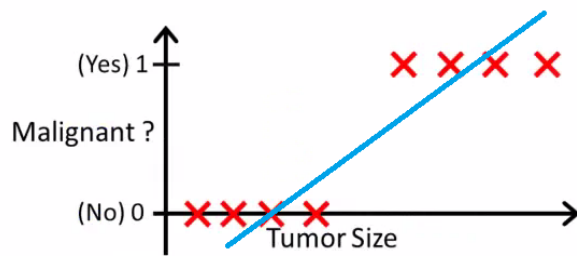
Logistic Regression

We identify problem as classification problem when independent variables are continuous in nature or real valued and dependent variable is in categorical form i.e. in classes like positive class and negative class. The real life example of classification example would be, to categorize the mail as spam or not spam, to categorize the tumor as malignant or benign and to categorize the transaction as fraudulent or genuine.

Logistic Regression actually gives us the probability of an instance belonging to the positive or negative class, i.e. Binary Classification

Two Class Classification		
$y \in \{0, 1\}$	1 or Positive Class	0 or Negative Class
Email	Spam	Not Spam
Tumor	Malignant	Benign
Transaction	Fraudulent	Not Fraudulent

So how are we going to do this? Linear Regression?



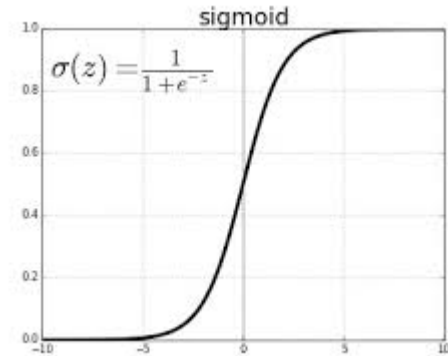
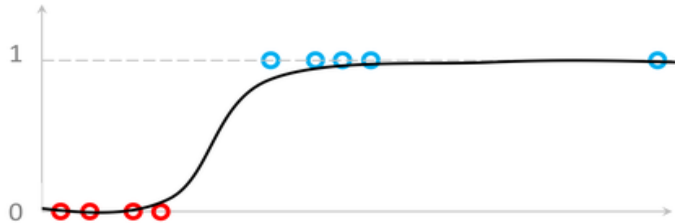
Logistic Regression uses Sigmoid Function

Let 'x' be the size of the tumor cell and we want classify benign or malignant

$$z = b + w x$$

$$y' = 1 / (1 + e ^{-z})$$

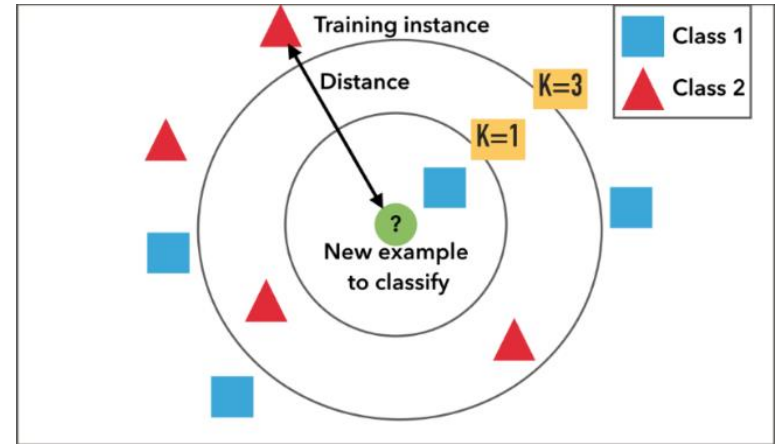
Where z is the equation of the line and y will belong to (0,1), hence giving us a probability



K- Nearest Neighbours

Choose your Nearest Neighbour!

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.





Distance Metrics and Correct K!

Now we need a distance metric between any 2 data points (training examples) to find out the near neighbours

1. Minkowski Distance
2. Manhattan Distance
3. Cosine Distance
4. Euclidean Distance

And choosing the correct K and value !

1. Mean
2. Mode

K= 1, K = 3, etc.

1
$$D(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^d |x_{il} - x_{jl}|^{1/p} \right)^p.$$

Here \mathbf{x}_i and \mathbf{x}_j are 2 training examples

2
$$|A - B| = \sum_{i=1}^d |a_i - b_i|$$

Here A and B are 2 training examples

3
$$\cos \theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$$

Here a and b are 2 training examples

4
$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

Here p and q are 2 training examples