Chapter 7
Logistic Regression

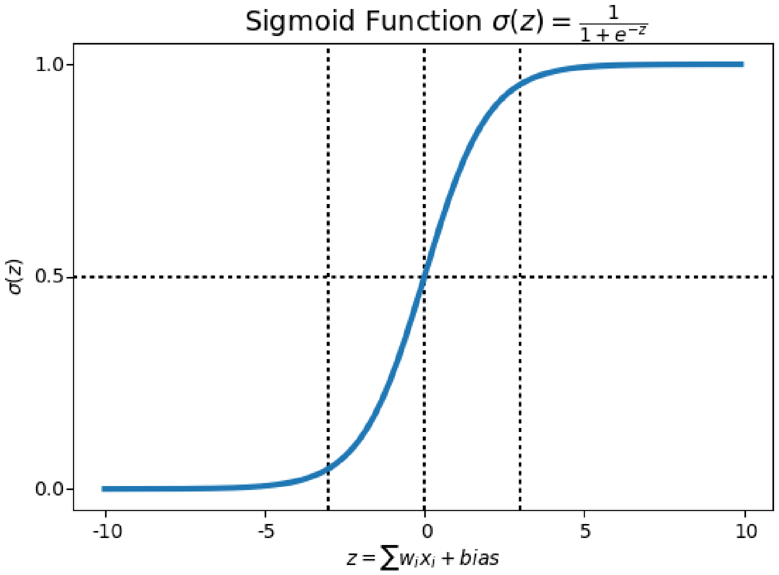
Logistic Regression is a Machine Learning algorithm which is used for the classification problems. it is a predictive analysis algorithm and based on the concept of probability. Logistic regression is probably the most widely used general-purpose classifier. It is very scalable and can be very fast to train. It's used for

- Spam filtering
- News message classification
- Web site classification
- Product classification
- Most classification problems with large, sparse feature sets

7.1 Introduction

Logistic regression extends the ideas of linear regression to the situation where the dependent variable, Y, is categorical. Logistic regression is designed as a binary classifier (output say {0,1}) but actually outputs the probability that the input instance is in the "1" class. Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the dependent variable is a linear one. Nor does it assume that the dependent variable or the error terms are distributed normally.

Logistic Regression uses a more complex cost function; this cost function can be defined as the '*Sigmoid function*' or also known as the 'logistic function' instead of a linear function. The *Sigmoid function* maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.



The formula of a sigmoid function:

$$f(x) = \frac{1}{1+\mathrm{e}^{-(x)}}$$

7.2 Hypothesis Representation

When using linear regression, we used a formula of the hypothesis i.e.

$$h_\theta(x) = \theta_0 + \theta_1 x$$
$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ . \\ . \\ x_n \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ . \\ . \\ \theta_n \end{bmatrix}$$

$$h_\theta(x) = \theta^T x$$

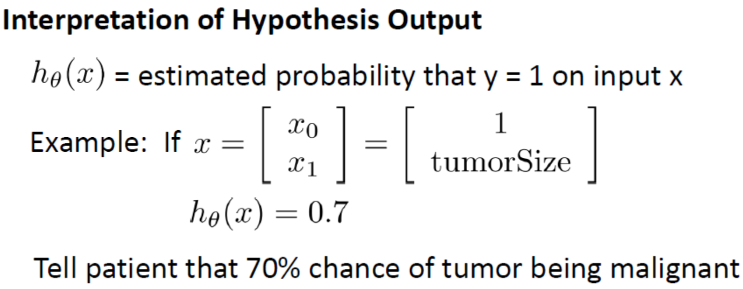
For logistic regression we are going to modify it a little bit i.e.

$$h_\theta(\mathbf{x}) = g(\theta^\top \mathbf{x})$$

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

Where

$$0 \le h_\theta(x) \le 1$$



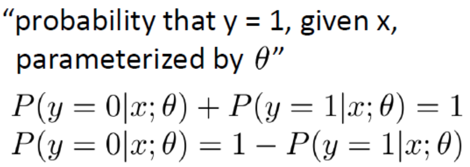
Interpretation of Hypothesis Output

$h_\theta(x)$ = estimated probability that y = 1 on input x

Example: If $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

$h_\theta(x) = 0.7$

Tell patient that 70% chance of tumor being malignant



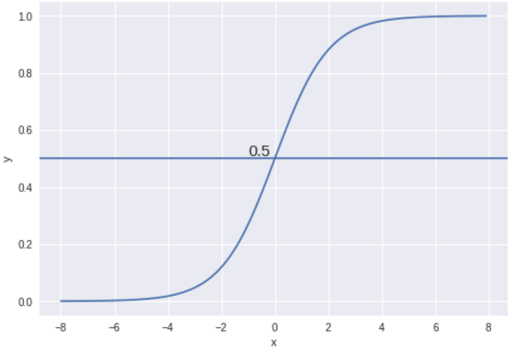
"probability that y = 1, given x, parameterized by θ"

$$P(y=0|x;\theta) + P(y=1|x;\theta) = 1$$
$$P(y=0|x;\theta) = 1 - P(y=1|x;\theta)$$

7.3 Decision Boundary

We expect our classifier to give us a set of outputs or classes based on probability when we pass the inputs through a prediction function and returns a probability score between 0 and 1. For Example, We have 2 classes, let's take them like cats and dogs (1 — dog , 0 — cats). We basically decide with a threshold value above which we classify values into Class 1 and of the value goes below the threshold then we classify it in Class 2.

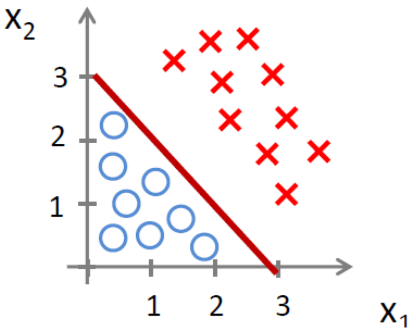
As shown in the below graph we have chosen the threshold as 0.5, if the prediction function returned a value of 0.7 then we would classify this observation as Class 1(DOG). If our prediction returned a value of 0.2 then we would classify the observation as Class 2(CAT).



For Example: Suppose

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

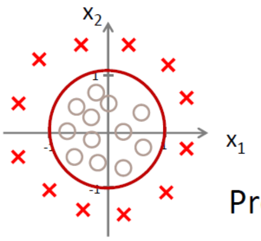


$$\theta^T x \geq 0 \Rightarrow y=1$$

Predict y=1, it $-3+x_1+x_2 \geq 0$

$$x_1 + x_2 = 3$$

Non-linear decision boundaries



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2) \quad \theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$$x_1^2 + x_2^2 = 1$$

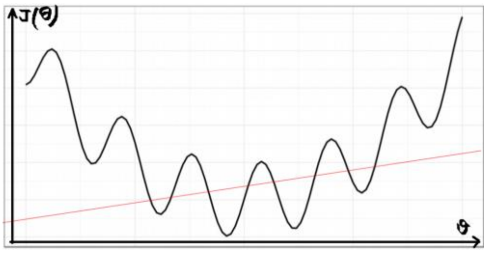
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \dots)$$

7.4 Cost function

The cost function represents optimization objective. The cost function J(θ) in the Linear regression is:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2.$$

With the modified hypothesis function, taking a square error function won't work as it no longer convex in nature and tedious to minimize. We take up a new form of cost function.

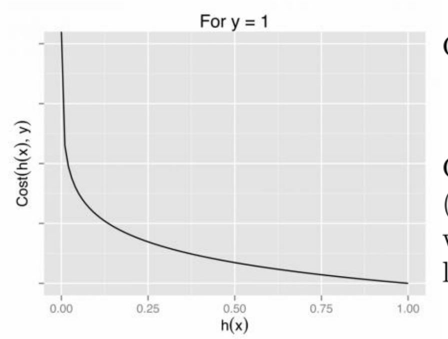


For logistic regression, the Cost function is defined as:

$$Cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & \text{if } y = 1 \\ -log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

The above two functions can be compressed into a single function i.e.

$$J(\theta) = -\frac{1}{m}\sum\left[y^{(i)}\log(h\theta(x(i))) + \left(1 - y^{(i)}\right)\log(1 - h\theta(x(i)))\right]$$

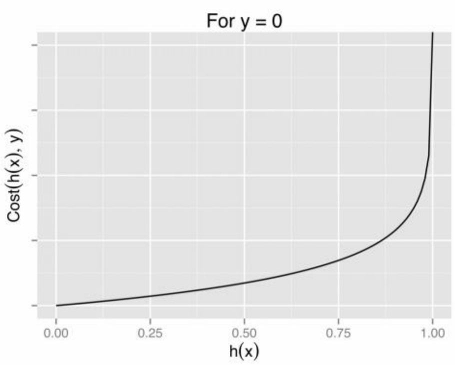


Cost $= 0$ if $y = 1, h_\theta(x) = 1$

But as $h_\theta(x) \to 0$

$Cost \to \infty$

Captures intuition that if $h_\theta(x) = 0$, (predict $P(y = 1|x;\theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.



- **If** y = 0 **and** h(x) = 0**,** Cost = 0
- **But for** $h(x) \to 1$

  $Cost \to \infty$

7.5 Cost function optimization using Gradient Descent

The cost function measures how well our parameters θ s are doing on the training data set. So, it seems natural to minimize the cost function for minimal error across the training data set to find value θ s. We would achieve the value of the parameters using gradient descent technique.

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

We want to minimize $J(\theta)$

Repeat

$\{$

$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$ OR $\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$

$\}$

simultaneously update all θ s

7.6 Linear Regression Vs Logistic Regression

Linear Regression	Logistic Regression
Linear Regression is used for solving Regression problem.	Logistic regression is used for solving Classification problems.
In Linear regression, we predict the value of continuous variables.	In logistic Regression, we predict the values of categorical variables.
In linear regression, we find the best fit line, by which we can easily predict the output.	In Logistic Regression, we find the S-curve by which we can classify the samples.
Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.
In Linear regression, it is required that relationship between dependent variable and independent variable must be linear.	In Logistic regression, it is not required to have the linear relationship between the dependent and independent variable.