[Summary]Logistic Regression 1

Don't be confused by the name "Logistic Regression"; it is named that way for historical reasons and is actually an approach to **classification problems, not regression problems**.

Binary Classification

Instead of our output vector y being a continuous range of values, it will only be 0 or 1.

$$y \in \{\, 0,\ 1 \,\}$$

Where 0 is usually taken as the "negative class" and 1 as the "positive class", but you are free to assign any representation to it.

We're only doing two classes for now, called a "Binary Classification Problem."

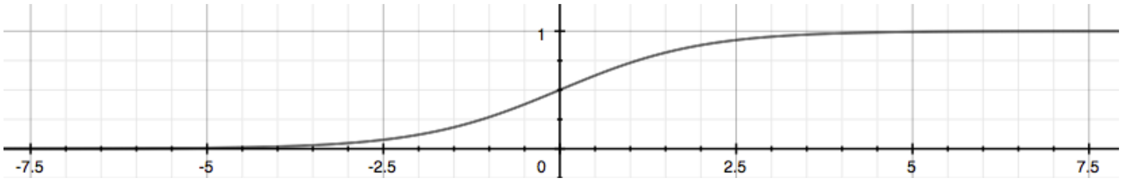
Hypothesis Representation

Our hypothesis should satisfy:

$$0 \leq \mathrm{h}_\theta(\mathrm{x}) \leq 1$$

Our new form uses the "Sigmoid Function," also called the "Logistic Function":

$$\mathrm{h}_\theta(x) = g(\theta^T x)$$
$$z = \ \theta^T x$$
$$g(z) = \frac{1}{1 + e^{-z}}$$



h_θ will give us the **probability** that our output is 1. For example, $\mathrm{h}_\theta(x) = 0.7$ gives us the probability of 70% that our output is 1.

$$\mathrm{h}_\theta(x) = P(y = 1|x;\theta) = 1 - P(y = 0|x;\theta)$$
$$P(y = 0|x;\theta) + P(y = 1|x;\theta) = 1$$

Decision Boundary

In order to get our discrete 0 or 1 classification, we can translate the output of the hypothesis function as follow:

$$\mathrm{h}_\theta(x) \geq 0.5 \ \rightarrow y = 1$$
$$\mathrm{h}_\theta(x) < 0.5 \ \rightarrow y = 0$$

The way our logistic function g behaves is that when its input is greater than or equal to zero, its output is greater than or equal to 0.5:

$$g(z) \geq 0.5$$
$$when\, z \geq 0$$

Remember:

$$z = 0,\ e^0 = 1 \Rightarrow g(z) = \frac{1}{2}$$
$$z \ \rightarrow \infty,\ \mathrm{e}^{-\infty} \rightarrow 0 \Rightarrow \mathrm{g(z)} = 1$$
$$z \ \rightarrow -\infty\,,\ e^{\infty} \rightarrow \infty \Rightarrow g(z) = 0$$

So if our input to g is $\theta^T X$, then that means:

$$\mathrm{h}_\theta(x) = g(\theta^T x) \geq 0.5$$
$$when\ \theta^T x \geq 0$$

From these statements we can now say:

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

The **decision boundary** is the line that separates the area where y = 0 and where y = 1. It is created by our hypothesis function.

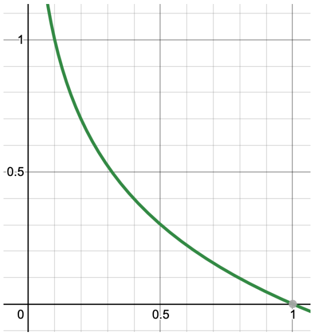
The input to the sigmoid function g(z)(e.g. $\theta^T X$) doesn't need to be linear, and could be a funcation that describes a circle (e.g. $z = \theta_0 + \theta_1 x_1^2 + \theta_2 x_2^2$) or any shape to fit our data.

Cost Function

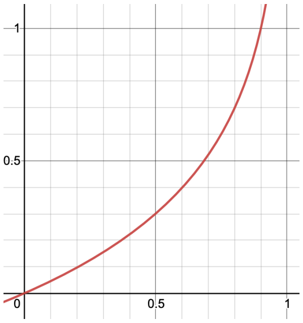
We cannot use the same cost function that we use for linear regression because the Logistic Function will cause the output to be wavy, causing many local optima. In other words, it will not be a convex function.

Instead, our cost function for logistic regression looks like:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(\mathrm{h}_\theta(x^i),\ y^i)$$
$$Cost(\mathrm{h}_\theta(x),\ y) = -\log\big(\mathrm{h}_\theta(x)\big)\ if\ y = 1$$
$$Cost(\mathrm{h}_\theta(x),\ y) = -\log\big(1 - \mathrm{h}_\theta(x)\big)\ if\ y = 0$$



If y =1



If y = 0

The more our hypothesis is off from y, the larger the cost function output. If our hypothesis is equal to y, then our cost is 0:

$$Cost(\mathrm{h}_\theta(x), y) = 0 \; if \; \mathrm{h}_\theta(x) = y$$
$$Cost(\mathrm{h}_\theta(x),\; y) \rightarrow \infty \; if \; y = 0 \; and \; \mathrm{h}_\theta(x) \rightarrow 1$$
$$Cost(\mathrm{h}_\theta(x),\; y) \rightarrow \infty \; if \; y = 1 \; and \; \mathrm{h}_\theta(x) \rightarrow 0$$

Simplified Cost Function and Gradient Descent

We can compress our cost function's two conditional cases into one case:

$$Cost(\mathrm{h}_\theta(x),\; y) = -y \log\big(\mathrm{h}_\theta(x)\big) - (1 \; - y)\log\big(1 - \mathrm{h}_\theta(x)\big)$$

We can fully write out our entire cost function as follows:

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

A vectorized implementation is:

$$\mathrm{h} = g(X\theta)$$
$$J(\theta) = \frac{1}{\mathrm{m}} \cdot \left(-\mathrm{y}^{\mathrm{T}} \log(\mathrm{h}) - (1-\mathrm{y})^{\mathrm{T}} \log(1-\mathrm{h})\right)$$

Gradient Descent

Remember that the general form of gradient descent is:

Repear {

$$\theta_j := \theta_j - \alpha \frac{\delta}{\delta\theta_j} J(\theta)$$

}

We can work out the derivative part using calculus to get:

Repeat {

$$\theta_j := \theta_j - \frac{\alpha}{m} \sum_{i=1}^{m} \left(\mathrm{h}_\theta\left(x^i\right) - y^i\right) x_j^i$$

}

Notice that this algorithm is identical to the one we used in linear regression. We still have to simultaneously update all values in theta.

Partial derivative of $\mathbf{J(\theta)}$

First calculate derivative of sigmoid function (it will be useful while finding partial derivative of J(θ)):

$$\sigma(x)' = \left(\frac{1}{1+e^{-x}}\right)'$$
$$= \frac{-(1+e^{-x})'}{(1+e^{-x})^2}$$
$$= \frac{-1' - (\mathrm{e}^{-\mathrm{x}})'}{(1+\mathrm{e}^{-\mathrm{x}})^2}$$
$$= \frac{0-(-x)'(e^{-x})}{(1+e^{-x})^2}$$
$$= \frac{-(-1)(e^{-x})}{(1+e^{-x})^2}$$
$$= \frac{e^{-x}}{(1+e^{-x})^2}$$
$$= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{e^{-x}}{1+e^{-x}}\right)$$
$$= \sigma(x)\left(\frac{+1-1+e^{-x}}{1+e^{-x}}\right)$$
$$= \sigma(x)\left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right)$$
$$= \sigma(x)\big(1-\sigma(x)\big)$$

Now we are ready to find out resulting partial derivative:

$$\frac{\delta}{\delta\theta_j}J(\theta) = \frac{\delta}{\delta\theta_j}\frac{-1}{m}\sum_{i=1}^{m}\left[y^i \log\left(\mathrm{h}_\theta(x^i)\right) + (1-y^i)\log\left(1-\mathrm{h}_\theta(x^i)\right)\right]$$
$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y^i\frac{\delta}{\delta\theta}\log\left(\mathrm{h}_\theta(x^i)\right) + (1-y^i)\frac{\delta}{\delta\theta_j}\log\left(1-\mathrm{h}_\theta(x^i)\right)\right]$$
$$= -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^i\frac{\delta}{\delta\theta_j}\mathrm{h}_\theta(x^i)}{\mathrm{h}_\theta(x^i)} + \frac{(1-y^i)\frac{\delta}{\delta\theta_j}\left(1-\mathrm{h}_\theta(x^i)\right)}{1-\mathrm{h}_\theta(x^i)}\right]$$
$$= -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^i\frac{\delta}{\delta\theta_j}\sigma(\theta^T x^i)}{\mathrm{h}_\theta(x^i)} + \frac{(1-y^i)\frac{\delta}{\delta\theta_j}\left(1-\sigma(\theta^T x^i)\right)}{1-\mathrm{h}_\theta(x^i)}\right]$$
$$= -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^i\sigma(\theta^T x^i)\left(1-\sigma(\theta^T x^i)\right)\frac{\delta}{\delta\theta_j}\theta^T x^i}{\mathrm{h}_\theta(x^i)} + \frac{-(1-y^i)\sigma(\theta^T x^i)\left(1-\sigma(\theta^T x^i)\right)\frac{\delta}{\delta\theta_j}\theta^{\mathrm{T}}\mathrm{x}^{\mathrm{i}}}{1-\mathrm{h}_\theta(x^i)}\right]$$
$$= -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^i\mathrm{h}_\theta(x^i)\left(1-\mathrm{h}_\theta(x^i)\right)\frac{\delta}{\delta\theta_j}\theta^T x^i}{\mathrm{h}_\theta(x^i)} - \frac{(1-y^i)\mathrm{h}_\theta(x^i)\left(1-\mathrm{h}_\theta(x^i)\right)\frac{\delta}{\delta\theta_j}\theta^T x^i}{1-\mathrm{h}_\theta(x^i)}\right]$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y^i\left(1-\mathrm{h}_\theta\left(x^i\right)\right)x_j^i - (1-y^i)\mathrm{h}_\theta\left(x^i\right)x_j^i\right]$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y^i\left(1-\mathrm{h}_\theta\left(x^i\right)\right) - (1-y^i)\mathrm{h}_\theta\left(x^i\right)\right]x_j^i$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y^i - y^i\mathrm{h}_\theta\left(x^i\right) - \mathrm{h}_\theta\left(x^i\right) + y^i\mathrm{h}_\theta\left(x^i\right)\right]x_j^i$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y^i - \mathrm{h}_\theta\left(x^i\right)\right]x_j^i$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left[\mathrm{h}_\theta\left(x^i\right) - y^i\right]x_j^i$$