Optional Logistic Regression: Gradient

This is an optional reading where I explain gradient descent in more detail. Remember, previously I gave you the gradient update step, but did not explicitly explain what is going on behind the scenes.

The general form of gradient descent is defined as:

$$\textit{Repeat} \{$$
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$
$$\}$$

For all j. We can work out the derivative part using calculus to get:

$$\textit{Repeat} \{$$
$$\theta_j := \theta_j - \frac{\alpha}{m} \sum_{i=1}^{m} (h(x^{(i)}, \theta) - y^{(i)}) x_j^{(i)}$$
$$\}$$

A vectorized implementation is:

$$\theta := \theta - \frac{\alpha}{m} X^T (H(X, \theta) - Y)$$

Partial derivative of J(θ)

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

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

Note that we computed the partial derivative of the sigmoid function. If we were to derive $h(x^{(i)}, \theta)$ with respect to θ_j, you would get $h(x^{(i)}, \theta)(1 - h(x^{(i)}, \theta)) x_j^{(i)}$. Note that we used the chain rule there, because we multiply by the derivative of $\theta^T x^{(i)}$ with respect to θ_j. Now we are ready to find out resulting partial derivative:

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{\partial}{\partial \theta_j} \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} log(h(x^{(i)}, \theta)) + (1 - y^{(i)}) log(1 - h(x^{(i)}, \theta)) \right]$$
$$= -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \frac{\partial}{\partial \theta_j} log(h(x^{(i)}, \theta)) + (1 - y^{(i)}) \frac{\partial}{\partial \theta_j} log(1 - h(x^{(i)}, \theta)) \right]$$
$$= -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{y^{(i)} \frac{\partial}{\partial \theta_j} h(x^{(i)}, \theta)}{h(x^{(i)}, \theta)} + \frac{(1 - y^{(i)}) \frac{\partial}{\partial \theta_j} (1 - h(x^{(i)}, \theta))}{1 - h(x^{(i)}, \theta)} \right]$$
$$= -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{y^{(i)} \frac{\partial}{\partial \theta_j} h(x^{(i)}, \theta)}{h(x^{(i)}, \theta)} + \frac{(1 - y^{(i)}) \frac{\partial}{\partial \theta_j} (1 - h(x^{(i)}, \theta))}{1 - h(x^{(i)}, \theta)} \right]$$
$$= -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{y^{(i)} h(x^{(i)}, \theta)(1 - h(x^{(i)}, \theta)) \frac{\partial}{\partial \theta_j} \theta^T x^{(i)}}{h(x^{(i)}, \theta)} + \frac{-(1 - y^{(i)}) h(x^{(i)}, \theta)(1 - h(x^{(i)}, \theta)) \frac{\partial}{\partial \theta_j} \theta^T x^{(i)}}{1 - h(x^{(i)}, \theta)} \right]$$
$$= -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{y^{(i)} h(x^{(i)}, \theta)(1 - h(x^{(i)}, \theta)) \frac{\partial}{\partial \theta_j} \theta^T x^{(i)}}{h(x^{(i)}, \theta)} - \frac{(1 - y^{(i)}) h(x^{(i)}, \theta)(1 - h(x^{(i)}, \theta)) \frac{\partial}{\partial \theta_j} \theta^T x^{(i)}}{1 - h(x^{(i)}, \theta))} \right]$$
$$= -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} (1 - h(x^{(i)}, \theta)) x_j^{(i)} - (1 - y^{(i)}) h(x^{(i)}, \theta) x_j^{(i)} \right]$$
$$= -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} (1 - h(x^{(i)}, \theta)) - (1 - y^{(i)}) h(x^{(i)}, \theta) \right] x_j^{(i)}$$
$$= -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} - y^{(i)} h(x^{(i)}, \theta) - h(x^{(i)}, \theta) + y^{(i)} h(x^{(i)}, \theta) \right] x_j^{(i)}$$
$$= -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} - h(x^{(i)}, \theta) \right] x_j^{(i)}$$
$$= \frac{1}{m} \sum_{i=1}^{m} \left[h(x^{(i)}, \theta) - y^{(i)} \right] x_j^{(i)}$$

The vectorized version:

$$\nabla J(\theta) = \frac{1}{m} \cdot X^T \cdot (H(X, \theta) - Y)$$

Congratulations, you now know the full derivation of logistic regression :) !

Mark as completed