Part 2 - Gradient of L2 penalized binary logistic regression

Using the previous result and the chain rule of calculus, derive the expression for the gradient of the L2 penalized cost function $J(\theta)$ (shown below) for logistic regression. $\lambda > 0$ is the regularization parameter.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)}log(h_\theta(x^{(i)})) + (1 - y^{(i)}log(1 - h_\theta(x^{(i)})))) + \frac{\lambda}{2m}\sum_{j=1}^{d}\theta_j^2$$

Solution

The L2 penalized cost function is defined as:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\log(h_\theta(x^{(i)})) + (1 - y^{(i)})\log(1 - h_\theta(x^{(i)}))\right) + \frac{\lambda}{2m}\sum_{j=1}^{d}\theta_j^2$$

where:

- m is the number of training examples.

- d is the number of features.
- $y^{(i)}$ is the target value for the i-th example.
- $x^{(i)}$ is the feature vector for the i-th example.
- $h_\theta(x^{(i)})$ is the sigmoid function, which is denoted as $g(z)$, where $z = \theta^T x^{(i)}$.
- $\lambda > 0$ is the regularization parameter.

First, let us recompute Part 1 for $h_\theta(x^{(i)}) = g(\theta^T x^{(i)})$, where $z = \theta^T x^{(i)}$:

$$\begin{aligned}\frac{\partial}{\partial \theta_j} h_\theta(x^{(i)}) &= \frac{\partial}{\partial \theta_j} g(\theta^T x^{(i)}) \\ &= g(\theta^T x^{(i)}) \cdot (1 - g(\theta^T x^{(i)})) \cdot \frac{\partial}{\partial \theta_j}(\theta^T x^{(i)}) \\ &= g(\theta^T x^{(i)}) \cdot (1 - g(\theta^T x^{(i)})) x_j^{(i)}\end{aligned}$$

Deriving the gradient of $J(\theta)$ with respect to θ by the chain rule and by Part 1 on non-regularized term first

$$\begin{aligned}\nabla J(\theta) &= \frac{\partial}{\partial \theta_j}\left(-\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right)\right) \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\frac{1}{h_\theta(x^{(i)})}\frac{\partial}{\partial \theta_j}h_\theta(x^{(i)}) + (1-y^{(i)})\frac{1}{1-h_\theta(x^{(i)})}\frac{\partial}{\partial \theta_j}(1-h_\theta(x^{(i)}))\right) \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\frac{1}{h_\theta(x^{(i)})}h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))x_j^{(i)} - (1-y^{(i)})\frac{1}{1-h_\theta(x^{(i)})}(-h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))x_j^{(i)})\right) \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}(1-h_\theta(x^{(i)})) - (1-y^{(i)})h_\theta(x^{(i)})\right)x_j^{(i)} \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - y^{(i)}h_\theta(x^{(i)}) - h_\theta(x^{(i)}) + y^{(i)}h_\theta(x^{(i)})\right)x_j^{(i)} \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}\end{aligned}$$

Computing the gradient of the regularization term:

$$\frac{\partial}{\partial \theta_j}\left(\frac{\lambda}{2m}\sum_{j=1}^{d}\theta_j^2\right) = \frac{\lambda}{2m}\frac{d}{d\theta_j}\left(\sum_{j=1}^{d}\theta_j^2\right) = \frac{\lambda}{2m}\cdot 2\theta_j = \frac{\lambda}{m}\theta_j$$

Combining both parts, we get:

$$\nabla J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} + \frac{\lambda}{m}\theta_j$$