Binary Cross Entropy Loss.

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

Derivation:

$$P(y | x^{(i)}, \theta) = h(x^{(i)}, \theta)^{y^{(i)}} * (1 - h(x^{(i)}, \theta))^{(1-y^{(i)})}$$

when $y=1$ we get $h(x^{(i)}, \theta)$

$y=0$ we get $(1 - h(x^{(i)}, \theta))$

We want to maximize the function $h(x^{(i)}, \theta)$ by making it as close to 1 as possible.

The likelihood is defined as follows

$$L(\theta) = \prod_{i=1}^{M} P(y | x^{(i)}, \theta) = \prod_{i=1}^{M} h(x^{(i)}, \theta)^{y^{(i)}} * (1 - h(x^{(i)}, \theta))^{(1-y^{(i)})}$$

Since we are trying to maximize ~~the~~ $h(x^{(i)}, \theta)$ in $L(\theta)$, we can introduce the log and just maximize the log of the function

$$\max_{h(x^{(i)}, \theta)} \log L(\theta) = \log \prod_{i=1}^{m} h(x^{(i)}, \theta)^{y^{(i)}} (1 - h(x^{(i)}, \theta))^{(1-y^{(i)})}$$

$$= \sum_{i=1}^{m} \log h(x^{(i)}, \theta)^{y^{(i)}} (1 - h(x^{(i)}, \theta))^{(1-y^{(i)})}$$

$$= \sum_{i=1}^{m} \log h(x^{(i)}, \theta)^{y^{(i)}} + \log (1 - h(x^{(i)}, \theta))^{(1-y^{(i)})}$$

$$= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}, \theta) + (1-y^{(i)}) \log (1 - h(x^{(i)}, \theta))$$

Dividing by m, because we want average cost

$$= \frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}, \theta) + (1-y^{(i)}) \log (1 - h(x^{(i)}, \theta))$$

Since we are ~~min~~ maximizing $h(\theta, x^{(i)})$, we can minimize the negative of above equation.

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

Gradient descent:

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

Here, $h(\cdot)$ = sigmoid function $\quad h'(\cdot) = h(\cdot)(1-h(\cdot))$

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

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

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

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

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

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

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

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

Vectorized version:

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