# HW3

*Rodrigo De Luna Lara*

*15 October 2016*

## Definition of the Gradient for Logistic Regression

Considering a dataset with a feature matrix $\langle x \rangle$, binary response vector $\langle y \rangle$, a logistic regression can be trained to define the values of a parameter vector $\langle \theta \rangle$ which correctly predicts the response. This can be done applying maximum likelihood estimation on the logistic regression probabily function.

The maximum likelihood estimator (MLE) of a probability function is defined as

$\hat{\theta}_{MLE} = \arg\max_{\theta} \log p_\theta(Y = y^{(1)}|X = x^{(1)}) + \ldots + \log p_\theta(Y = y^{(n)}|X = x^{(n)})$

$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} \log p_\theta(Y = y^{(i)}|X = x^{(i)})$

Given the probability function for the logistic regression $P(Y = y|X = x) = \dfrac{1}{1 + exp(y\langle\theta, x\rangle)}$

The MLE for logistic regression can be defined as $\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} \log \dfrac{1}{1 + exp(y^{(i)}\langle\theta, x^{(i)}\rangle)}$

Given that $\log A^{-1} = -\log A$ then $\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} -\log\left(1 + exp(y^{(i)}\langle\theta, x^{(i)}\rangle)\right)$

The maximization of a negative function is the minimization of the same function, so the function is redefined to $\hat{\theta}_{MLE} = \arg\min_{\theta} \sum_{i=1}^{n} \log\left(1 + exp(y^{(i)}\langle\theta, x^{(i)}\rangle)\right)$

The optimization problem becomes finding a value of $\theta$ that minimizes $\mathcal{L}(\theta) = \sum_{i=1}^{n} \log\left(1 + exp(y^{(i)}\langle\theta, x^{(i)}\rangle)\right)$

The gradient $\nabla\mathcal{L}$ is the direction of steepest descent in the function. To calculate the gradient the partial derivative of $\theta$ is applied to both sides of the function

$\dfrac{\partial}{\partial\theta^{(j)}}\mathcal{L}(\theta) = \dfrac{\partial}{\partial\theta^{(j)}} \sum_{i=1}^{n} \log\left(1 + exp(y^{(i)}\langle\theta, x^{(i)}\rangle)\right)$

The derivative of a sum is equal to the sum of the derivatives according to the sum rule in differentiation, which indicates that $\dfrac{d}{dx}\left(\sum_{i=1}^{n} f_i(x)\right) = \sum_{i=1}^{n}\left(\dfrac{d}{dx} f_i(x)\right)$

Therefore $\dfrac{\partial}{\partial\theta^{(j)}}\mathcal{L}(\theta) = \sum_{i=1}^{n} \dfrac{\partial}{\partial\theta^{(j)}} \log\left(1 + exp(y^{(i)}\langle\theta, x^{(i)}\rangle)\right)$

Then, given that $\dfrac{d}{dx}\log\left(f(x)\right) = \dfrac{f'(x)}{f(x)}$ the equation can be expressed as $\dfrac{\partial}{\partial\theta^{(j)}}\mathcal{L}(\theta) = \sum_{i=1}^{n} \dfrac{\frac{\partial}{\partial\theta^{(j)}}\left(1 + exp(y^{(i)}\langle\theta, x^{(i)}\rangle)\right)}{1 + exp(y^{(i)}\langle\theta, x^{(i)}\rangle)}$

Next, given that $\dfrac{d}{dx}\left(f(x) + g(x)\right) = \dfrac{d}{dx}f(x) + \dfrac{d}{dx}g(x)$ and that $\dfrac{d}{dx}c = 0$ then $\dfrac{\partial}{\partial\theta^{(j)}}\mathcal{L}(\theta) = \sum_{i=1}^{n} \dfrac{\frac{\partial}{\partial\theta^{(j)}}\left(exp(y^{(i)}\langle\theta, x^{(i)}\rangle)\right)}{1 + exp(y^{(i)}\langle\theta, x^{(i)}\rangle)}$

And if $\frac{d}{dx}exp(f(x)) = exp(f(x)) \cdot f'(x)$ then $\frac{\partial}{\partial\theta^{(j)}}\mathcal{L}(\theta) = \sum_{i=1}^{n} \frac{exp(y^{(i)}\langle\theta, x^{(i)}\rangle) \cdot \frac{\partial}{\partial\theta^{(j)}}(y^{(i)}\langle\theta, x^{(i)}\rangle)}{1 + exp(y^{(i)}\langle\theta, x^{(i)}\rangle)}$

Finally, since $exp(y^{(i)}\langle\theta, x^{(i)}\rangle) = exp\left(y^{(i)}\sum_{k=1}^{d}\theta_k x_k^{(i)}\right)$ the expression becomes

$$\frac{\partial}{\partial\theta^{(j)}}\mathcal{L}(\theta) = \sum_{i=1}^{n} \frac{exp\left(y^{(i)}\sum\limits_{k=1}^{d}\theta_k x_k^{(i)}\right) \cdot \frac{\partial}{\partial\theta^{(j)}}\left(y^{(i)}\sum\limits_{k=1}^{d}\theta_k x_k^{(i)}\right)}{1 + exp\left(y^{(i)}\sum\limits_{k=1}^{d}\theta_k x_k^{(i)}\right)}$$

The only $k$ that results in non-zero terms for the derivative part of the equation is $k = j$, so the expression becomes

$$\frac{\partial}{\partial\theta^{(j)}}\mathcal{L}(\theta) = \sum_{i=1}^{n} \frac{exp\left(y^{(i)}\sum\limits_{k=1}^{d}\theta_k x_k^{(i)}\right) \cdot \frac{\partial}{\partial\theta^{(j)}}\left(y^{(i)}\theta_j x_j^{(i)}\right)}{1 + exp\left(y^{(i)}\sum\limits_{k=1}^{d}\theta_k x_k^{(i)}\right)}$$

Since $\frac{exp(x)}{1+exp(x)} = \frac{1}{exp(-x)+1}$ and given that $\frac{\partial}{\partial\theta^{(j)}}\left(y^{(i)}\theta_j x_j^{(i)}\right) = y^{(i)}x_j^{(i)}$ the equation can be further simplified to $\frac{\partial}{\partial\theta^{(j)}}\mathcal{L}(\theta) = \sum_{i=1}^{n} \frac{y^{(i)}x_j^{(i)}}{1 + exp\left(-y^{(i)}\sum\limits_{k=1}^{d}\theta_k x_k^{(i)}\right)}$, which is the final form of the gradient function.

The gradient has $d$ components $\nabla\mathcal{L}(\theta) = \left(\frac{\partial\mathcal{L}}{\partial\theta_1}(\theta), \ldots, \frac{\partial\mathcal{L}}{\partial\theta_d}(\theta)\right)$

## Pseudo-code for Logistic Regression

Having defined the function for the gradient, the pseudo-code for the logistic regression is:

1. Define the vectors $\langle x\rangle$,$\langle y\rangle$ and $\langle\theta\rangle$
   (a) $\langle x\rangle$ is the vector of features
   (b) $\langle y\rangle$ is the response vector (-1 or +1)
   (c) $\langle\theta\rangle$ is the parameter vector of $\langle x\rangle$
2. Define training and validation data (usually in a 70-30 ratio)
3. Define number of maximum iterations, innovation threshold and step size $\alpha$
4. With the training data:
   (a) Initialize mangitudes of $\langle\theta\rangle$ to random values
   (b) For $j = 1, \ldots, d$ update $\theta_j \leftarrow \theta_j - \alpha \cdot \nabla\mathcal{L}$
   (c) Repeat (b) until the updates are smaller than the innovation threshold or the maximum number of iterations is reached
5. With the final value of $\theta$ calculate the response as $sign(\langle\theta, x\rangle)$
6. Determine the accurracy of the training data, and if acceptable validate the resulting $\theta$ on the validation set.