COMP 540 Homework 2

Peiguang Wang, Xinran Zhou

Due: February 2, 2018

Part 1: Gradient and Hessian of $NLL(\theta)$ for logistic regression

1. Let $g(z) = \frac{1}{1-e^{-z}}$. Show that $\frac{\partial g(z)}{\partial z} = g(z)(1-g(z))$.

   Proof.
   $$\frac{\partial g(z)}{\partial z} = -\frac{e^{-z}}{(1-e^{-z})} = \frac{1}{1+e^{-z}}(1-\frac{1}{1+e^{-z}}) = g(z)(1-g(z))$$

   □

2. Using the previous result and the chain rule of calculus, derive the following expression for the gradient of the negative log likelihood function $NLL(\theta)$ for logistic regression.
   $$\frac{\partial}{\partial\theta}NNL(\theta) = \sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$$

   Proof. Let h_θ denote $h_\theta(x^{(i)})$. Then the $NLL(\theta)$ is:
   $$NLL(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}log(h_\theta) + (1-y^{(i)})log(1-h_\theta)$$

   Then
   $$\frac{\partial}{\partial(\theta x)}NNL = -\sum_{i=1}^{m} y^{(i)}\frac{h_\theta(1-h_\theta)}{h_\theta} + (-1)(1-y^{(i)}\frac{h_\theta(1-h_\theta)}{1-h_\theta})$$

   Simplify the equation above
   $$\frac{\partial}{\partial(\theta x)}NNL = -\sum_{i=1}^{m} y^{(i)}(1-h_\theta) + (y^{(i)}-1)h_\theta = -\sum_{i=1}^{m} y^{(i)} - h_\theta$$

   Derive $\frac{\partial}{\partial(\theta)}NNL$ from $\frac{\partial}{\partial(\theta x)}NNL$:
   $$\frac{\partial NNL}{\partial\theta} = \frac{\partial NLL}{\partial(\theta x)}\frac{\partial(\theta x)}{\partial\theta} = \sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$$

   □

3. The Hessian or second derivative of the $NLL(\theta)$ can be written as $H = X^T S X$ where
   $$S = diag(h_\theta(x^{(1)})(1-h_\theta(x^{(1)})), ..., h_\theta(x^{(m)})(1-h_\theta(x^{(m)})))$$

   Show that H is positive definite. You may assume that $0 < h_\theta(x^{(i)}) < 1$, so the elements of S are strictly positive and that X is full rank.

   Proof. For any given vector u, compute $u^T H u$:
   $$u^T H u = u^T X^T S X u$$

   Let vector $v = (v_1, v_2, ...v_m)$ denote Xu. since X is full rank, then v is not a zero vector.
   $$u^T H u = v^T S v = \sum_{i=1} m v_i^2 S_{ii}$$

   Since the elements of S are strictly positive, then $u^T H u > 0$. So H is positive definite. □

Part 2: Properties of L2 regularized logistic regression

1. (True or False) $J(\theta)$ has multiple locally optimal solutions.

   Solution. This statement is **False**. Since $J(\theta)$ is a convex function, it only have one global optimal point.

2. (True or False) Let $\theta^\star = argmin_\theta J(\theta)$ be a global optimum. $\theta^\star$ is sparse (has many zero entries).

   Solution. This statement is **False**. Since in this regression problem, we use L2 regularization, it won't make $\theta^\star$ become sparse. L2 norm will make θ have small values. If L1 norm is used, then $\theta^\star$ will become sparse.

3. (True or False) If the training data is linearly separable, then some coefficients θ_j might become infinite if $\lambda = 0$.

   Solution. This statement is **True**. When the plain is

4. (True of False) The first term of $J(\theta)$ always increases as we increase λ.

   Solution. This statement is **True**. When adding λ, the cross-entropy loss

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

   will become larger to prevent overfitting.