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

Part 1

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

$$\begin{aligned} g(z) &= (1+e^{-z})^{-1} \\ \frac{dg(z)}{dz} &= \frac{-(-e^{-z})}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})(1+e^{-z})} = \left(\frac{1}{1+e^{-z}}\right)\left(\frac{e^{-z}}{1+e^{-z}}\right) \\ &= \left(\frac{1}{1+e^{-z}}\right)\left(\frac{(1+e^{-z})-1}{1+e^{-z}}\right) = \left(\frac{1}{1+e^{-z}}\right)\left(1-\frac{1}{1+e^{-z}}\right) \\ &= g(z)(l-g(z)) \end{aligned}$$

Part 2

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

Let $x = x^{(i)}, y = y^{(I)}$. Since $h_\theta = g(\theta^T x)$, $\frac{h_\theta(x)}{d\theta} = h_\theta(x)(1-h_\theta(x))x$

$$\begin{aligned} \frac{d}{d\theta}NLL(\theta) &= -\frac{1}{m}\sum_{i=1}^{m}\frac{y}{h_\theta(x)}(h_\theta(x))(1-h_\theta(x))x + \frac{1-y}{1-h_\theta(x)} \times -h_\theta(x)(1-h_\theta(x))x \\ &= -\frac{1}{m}\sum_{i=1}^{m}h_\theta(x)(1-h_\theta(x))x\left(\frac{y}{h_\theta(x)} + \frac{y-1}{1-h_\theta(x)}\right) \\ &= -\frac{1}{m}\sum_{i=1}^{m}h_\theta(x)(1-h_\theta(x))x\left(\frac{y(1-h_\theta(x)) + (y-1)(h_\theta(x))}{h_\theta(x)(1-h_\theta(x))}\right) \\ &= -\frac{1}{m}\sum_{i=1}^{m}x(y-h_\theta(x)) \\ &= -\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})x^{(i)} \end{aligned}$$

Part 3

A matrix is positive definite if $U^T AU > 0$ for all non-zero vector x.

$$\begin{aligned} U^T AU &= U^T SXU \\ if Xu &= y, \\ U^T AU &= y^T Sy \\ &= \sum_{i=1}^{m} y^{(i)2}h_\theta(x^{(i)})(1-h_\theta(x^{(i)})) > 0 \end{aligned}$$

2 Regularizing logistic regression

$$\theta_{MLE} = argmax_{\theta} \prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)}; \theta)$$
$$\theta_{MAP} = argmax_{\theta} P(\theta) \prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)}; \theta)$$

Let $\prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)}; \theta) = F(\theta)$.

$F(\theta_{MLE}) \geq F(\theta)$,

$\therefore F(\theta_{MLE}) \geq F(\theta_{MAP})$. (1)

$P(\theta_{MAP})F(\theta_{MAP}) \geq P(\theta)F(\theta)$,

$\therefore P(\theta_{MAP})F(\theta_{MAP}) \geq P(\theta_{MLE})F(\theta_{MLE})$. (2)

Combining (1) and (2), we get $P(\theta_{MAP})F(\theta_{MLE}) \geq P(\theta_{MLE})F(\theta_{MLE})$.

Eliminating $F(\theta_{MLE})$ on both sides of the equation leaves us with $P(\theta_{MAP}) \geq P(\theta_{MLE})$.

Since both are Gaussian distributions, $\theta_{MLE} \geq \theta_{MAP}$.

$\therefore ||\theta_{MLE}||_2 \geq ||\theta_{MAP}||_2$