

Solution to Homework 2

Shoeb Mohammed and Zhuo Chen

February 7, 2016

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

1.1

Given

$$g(z) = \frac{1}{1+e^{-z}} \quad (1)$$

Proof.

$$\frac{\partial g(z)}{\partial z} = \frac{(1+e^{-z}) \cdot 0 + e^{-z}}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{g(z)}{(1-g(z))} \quad (2)$$

□

1.2

For logistic regression, negative log likelihood is

$$NLL(\theta) = -\sum_{i=1}^m (y_i \log(h_\theta(x_i)) + (1-y_i) \log(1-h_\theta(x_i))) \text{ where } h_\theta(x_i) = g(\theta^T x^i) \quad (3)$$

Proof. Using equation 2 and chain rule for differentiation we have

$$\begin{aligned} NLL(\theta) &= -\sum_{i=1}^m \left(\frac{y_i}{h_\theta(x_i)} h_\theta(x_i)(1-h_\theta(x_i)) \frac{\partial \theta^T x^i}{\partial \theta} - \frac{(1-y_i)}{(1-h_\theta(x_i))} h_\theta(x_i)(1-h_\theta(x_i)) \frac{\partial \theta^T x^i}{\partial \theta} \right) \\ &= -\sum_{i=1}^m (y_i(1-h_\theta(x_i)) - (1-y_i)h_\theta(x_i)) x_i \text{ because } \frac{\partial}{\partial \theta} \theta^T x^i = x^i \\ &= \sum_{i=1}^m (h_\theta(x_i) - y_i) \end{aligned} \quad (4)$$

□

1.3

Given

$$\begin{aligned} H &= X^T S X \text{ where } S = \text{diag}(\mu_1 \dots \mu_m) \\ \mu_i &= h_\theta(x_i)(1-h_\theta(x_i)) \text{ for } i=1 \dots m \\ \text{and } 0 < \mu_i < 1 \text{ for } i=1 \dots m \end{aligned} \quad (5)$$

Proof. For any vector $u \neq 0$ we have,

$$\begin{aligned} u^T H u &= u^T (X^T S X) u \\ &= (Xu)^T S (Xu) \\ &= v^T S v \text{ where } v = [v_1 \dots v_m]^T = Xu \neq 0 \text{ since } X \text{ is full rank} \\ &= \sum_{i=1}^m v_i^2 \mu_i \\ &> 0 \text{ since } \mu_i \text{ is positive and } v_i \neq 0 \end{aligned} \quad (6)$$

Thus, H is positive definite.

□

2 Regularizing logistic regression

Proof. The maximal likelihood and MAP estimates for θ are

$$\begin{aligned}\theta_{MLE} &= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^m P(y^{(i)}|x^{(i)}; \theta) \\ \theta_{MAP} &= \underset{\theta}{\operatorname{argmax}} P(\theta) \prod_{i=1}^m P(y^{(i)}|x^{(i)}; \theta) \text{ where } P(\theta) \sim N(0, \alpha^2 I)\end{aligned}\tag{7}$$

Equation 7 can be rewritten using log likelihood $LL(\theta)$:

$$\begin{aligned}\theta_{MLE} &= \underset{\theta}{\operatorname{argmax}} LL(\theta) \text{ where } LL(\theta) = \sum_{i=1}^m \log(P(y^{(i)}|x^{(i)}; \theta)) \\ \theta_{MAP} &= \underset{\theta}{\operatorname{argmax}} \log(P(\theta)) + LL(\theta) \\ &= \underset{\theta}{\operatorname{argmax}} K - \frac{d}{2\alpha^2} \theta^T \theta + LL(\theta) \text{ where } K \text{ is constant. This follows from } P(\theta) \sim N(0, \alpha^2 I)\end{aligned}\tag{8}$$

$$= \underset{\theta}{\operatorname{argmax}} LL(\theta) - \frac{d}{2\alpha^2} \|\theta\|_2^2$$

Now,

$$\begin{aligned}LL(\theta_{MAP}) - \frac{d}{2\alpha^2} \|\theta_{MAP}\|_2^2 &\geq LL(\theta_{MLE}) - \frac{d}{2\alpha^2} \|\theta_{MLE}\|_2^2 && \text{from definition for } \theta_{MAP} \\ &\geq LL(\theta_{MAP}) - \frac{d}{2\alpha^2} \|\theta_{MLE}\|_2^2 && \text{from definition for } \theta_{MLE} \\ \implies \frac{d}{2\alpha^2} \|\theta_{MAP}\|_2^2 &\leq \frac{d}{2\alpha^2} \|\theta_{MLE}\|_2^2 \\ \implies \|\theta_{MAP}\|_2^2 &\leq \|\theta_{MLE}\|_2^2 \\ \implies \|\theta_{MAP}\|_2 &\leq \|\theta_{MLE}\|_2\end{aligned}\tag{9}$$

□