EE5907/EE5027 Week 4: Logistic Regression Solutions

Exercise 8.3

a. Given $\sigma(a) = \frac{1}{1+e^{-a}}$

$$\begin{aligned}
\frac{d\sigma(a)}{da} &= -\frac{1}{(1+e^{-a})^2}\frac{d(1+e^{-a})}{da} \\
&= \frac{1}{(1+e^{-a})^2}e^{-a} \\
&= \frac{1}{(1+e^{-a})^2}(1+e^{-a}-1) \\
&= \frac{1}{1+e^{-a}} - \frac{1}{(1+e^{-a})^2} \\
&= \sigma(a) - \sigma^2(a) \\
&= \sigma(a)(1-\sigma(a))
\end{aligned}$$

b. Given $NLL(w) = -\sum_{i=1}^{N}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)]$ where $\mu_i = \sigma(w^T x_i)$, we have

$$\begin{aligned}
g = \frac{d}{dw}NLL(w) &= \frac{d}{dw}\left(-\sum_{i=1}^{N}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)]\right) \\
&= \left(-\sum_{i=1}^{N}\frac{d}{d\mu_i}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)]\frac{d\mu_i}{d(w^T x_i)}\frac{dw^T x_i}{dw}\right) \\
&= -\sum_{i=1}^{N}\left[\left(\frac{y_i}{\mu_i} - \frac{1-y_i}{1-\mu_i}\right)\frac{d\mu_i}{d(w^T x_i)}\frac{dw^T x_i}{dw}\right] \\
&= -\sum_{i=1}^{N}\left[\frac{y_i - y_i\mu_i - \mu_i + \mu_i y_i}{\mu_i(1-\mu_i)}\mu_i(1-\mu_i)x_i\right] \\
&= \sum_i(\mu_i - y_i)x_i
\end{aligned}$$

c. The Hessian is positive definite if $z^T H z$ is positive for all $z \in \mathbb{R}^n$ and $z \neq \vec{0}$.

$$z^T H z = z^T X^T S X z = (Xz)^T S(Xz)$$

Since X is full rank and $z \neq \vec{0}$, then $Xz \neq \vec{0}$. Let the i-th entry of Xz be a_i, then

$$(Xz)^T S(Xz) = \sum_{i=1}^{n} a_i^2 \mu_i (1 - \mu_i)$$

Since $0 < \mu_i < 1$, therefore $a_i^2 \mu_i (1 - \mu_i) \geq 0$ for all i with at least one term greater than 0 because $Xz \neq \vec{0}$. Therefore $z^T H z$ is greater than 0, and H is positive definite.

Exercise 8.6

a. False. According to the proof in Exercise 8.3, the Hessian is positive definite and thus $NLL(w)$ is convex. $\lambda ||w||_2^2$ is also convex. Thus $J(w)$ is also convex because the sum of two convex functions is convex. Therefore there is only one local optimum and that local optimum is also a global optimum.

b. False. For L2 regularization, we tend to not get sparse estimates. The intuitive reason is that as a number x decreases from 10 to 5 to 0, x^2 changes from 100 to 25 to 0. Therefore decreasing x by a constant amount (10 to 5 to 0) yields diminishing returns (100 to 25 to 0). In contrast, $L1$ regularization tends to lead to sparse estimates because as x decreases from 10 to 5 to 0, $|x|$ changes from 10 to 5 to 0, thus there is no such diminishing returns.

c. True. If the training data is linearly separable, the MLE is obtained when $||w|| \to \infty$ without any regularization. Thus some weight ω_j might become infinite.

d. True. The log likelihood of the training set will decrease as λ increases because more weight is given to the regularization so w is not as "free" to fit the data. Therefore, the$NLL(w)$ of training set will increase as λ increases.

e. False. As we increase λ, we can potentially improve results on the test set and so the $NLL(w)$ of the test set can initially decrease. However, as we increase λ, $NLL(w)$ of both training and test sets can become bad (i.e., increase).

Exercise 8.7

a. The decision boundary will satisfy

$$p(y = 1|x) = \text{sigm}(\omega^T x) = 0.5 \implies \omega^T x = 0$$

Thus we have $\omega_0 + \omega_1 x_1 + \omega_2 x_2 = 0$, which implies the possible decision boundary should be a straight line. The rough decision boundary is displayed as the red line in Figure 1. There is zero classification error made on the training set. Note that although there are many possible decision boundaries that result in zero classification error, but there is a unique value of w (and hence unique decision boundary) that minimize $J(w)$.

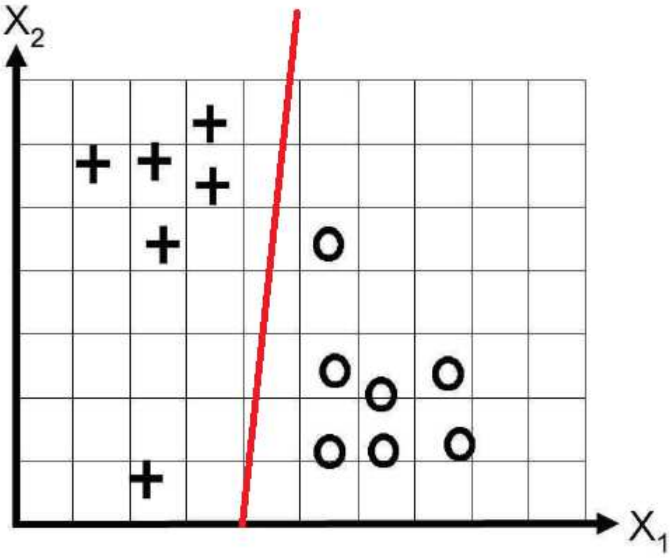

Figure 1: Decision boundary for part (a)

b. Suppose ω_0 is regularized all the way to zero. Then we have $x_2 = -\frac{\omega_1}{\omega_2}x_1$. Thus a decision boundary will definitely pass through the origin. The red line in Figure 2 shows the possible decision boundary. There is one classification error made on the training set. Again note that while there are many decision boundaries that result in one classification error, there is a unique decision boundary that minimize $J(w)$.

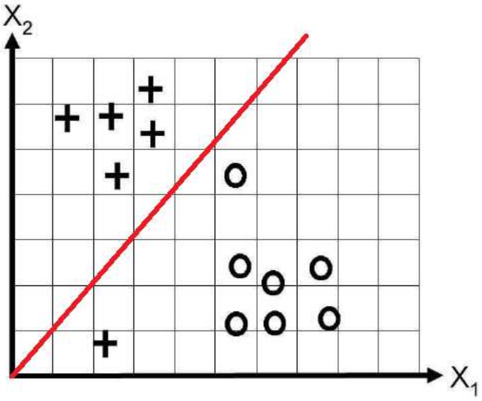

Figure 2: Decision boundary for part (b)

c. Suppose ω_1 is heavily regularized. Then we have $x_2 = -\frac{\omega_1}{\omega_2}x_1 - \frac{\omega_0}{\omega_2}$ where the coefficient of x_1 is approximately zero and $-\frac{\omega_0}{\omega_2}$ is a constant . Thus the decision boundary will be horizontal. The red line in Figure 3 shows a possible decision boundary. There are two classification errors made on the training set. Again note that while there are many decision boundaries that result in two classification errors, there is a unique decision boundary that minimize $J(w)$.

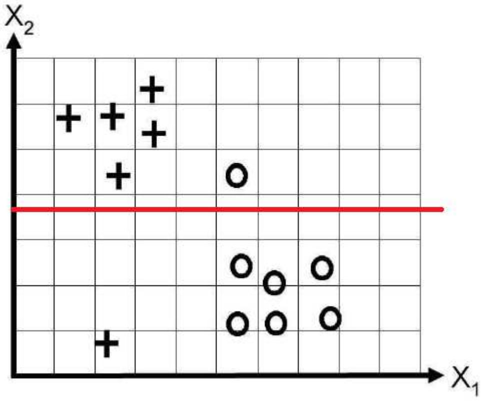

Figure 3: Decision boundary for part (c)

d. Suppose ω_2 is heavily regularized. Then we have $x_1 = -\frac{\omega_2}{\omega_1}x_2 - \frac{\omega_0}{\omega_1}$ where the coefficient of x_2 is approximately zero and $-\frac{\omega_0}{\omega_1}$ is a constant . Thus the decision boundary will be vertical. The red line in Figure 3 shows a possible decision boundary. There is zero classification error made on the training set. Note that although there are many possible decision boundaries that result in zero classification error, but there is a unique value of w (and hence unique decision boundary) that minimize $J(w)$.

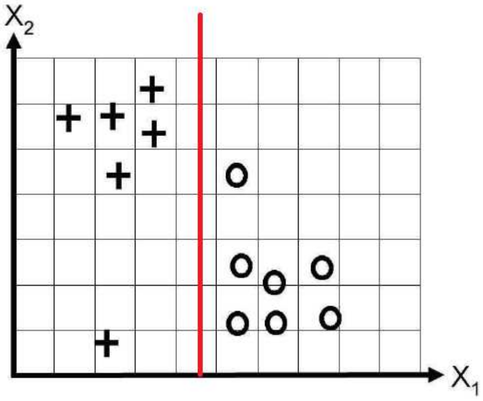

Figure 4: Decision boundary for part (d)