# COMP 540: Assignment #2

Due on Monday, February 8, 2016

**Misiura Mikita & Jesse Hellemn**

# Problem 1

1) We have

$$\begin{aligned}\frac{dg(x)}{dx} &= \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)\\ &= -\frac{-e^{-x}}{(1+e^{-x})^2}\\ &= \frac{e^{-x}+1-1}{(1+e^{-x})^2}\\ &= \frac{(1+e^{-x})-1}{(1+e^{-x})^2}\\ &= \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2}\\ &= g(x)[1-g(x)]\end{aligned}$$

2) We have NLL($\theta$):

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

It's derivative wrt $\theta$:

$$\frac{\partial NLL(\theta)}{\partial\theta} = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\frac{\partial}{\partial\theta}\log h_\theta(x^{(i)}) + (1-y^{(i)})\frac{\partial}{\partial\theta}\log(1-h_\theta(x^{(i)}))\right] \tag{2}$$

Following the chain rule of calculus and the result above (keeping in mind that $h_\theta(x) = g(\theta^T x)$) we get:

$$\frac{\partial}{\partial\theta}\log h_\theta(x^{(i)}) = \frac{\frac{\partial}{\partial\theta}h_\theta(x^{(i)})}{h_\theta(x^{(i)})} = \frac{h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))\frac{\partial}{\partial\theta}\theta^T x^{(i)}}{h_\theta(x^{(i)})} = (1-h_\theta(x^{(i)}))x^{(i)} \tag{3}$$

$$\frac{\partial}{\partial\theta}\log(1-h_\theta(x^{(i)})) = \frac{\frac{\partial}{\partial\theta}(1-h_\theta(x^{(i)}))}{1-h_\theta(x^{(i)})} = -\frac{h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))\frac{\partial}{\partial\theta}\theta^T x^{(i)}}{1-h_\theta(x^{(i)})} = -h_\theta(x^{(i)})x^{(i)} \tag{4}$$

$$\begin{aligned}\frac{\partial NLL(\theta)}{\partial\theta} &= -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}x^{(i)}(1-h_\theta(x^{(i)})) - (1-y^{(i)})x^{(i)}h_\theta(x^{(i)})\right] &(5)\\ &= \frac{1}{m}\sum_{i=1}^{m}\left[x^{(i)}(h_\theta(x^{(i)}) - y^{(i)})\right] &(6)\end{aligned}$$

3) Let $S_{i,i}$ by the $i$th element of the diagonal of $S$. Since $S_{i,i} \geq 0$, then for any non-zero vector $z$ of dimension $n$, where $X$ has dimension $m$ by $n$, we have

$$z^T H z = z^T X^T S X z = (Xz)^T S (Xz) = \sum_{i=1}^{n} S_{i,i}(Xz)_i^2 > 0$$

# Problem 2

We have for the MLE estimate that

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

and thus that

$$\prod_{i=1}^{m} g(\theta_{MLE}^T x^{(i)}) \geq \prod_{i=1}^{m} g(\theta^T x^{(i)}) \qquad \text{for all } \theta \tag{7}$$

$$\prod_{i=1}^{m} \frac{g(\theta_{MLE}^T x^{(i)})}{g(\theta^T x^{(i)})} \geq 1 \qquad \text{for all } \theta \tag{8}$$

Likewise, for the MAP estimate we have

$$P(\theta_{MAP}) \prod_{i=1}^{m} g(\theta_{MAP}^T x^{(i)}) \geq P(\theta) \prod_{i=1}^{m} g(\theta^T x^{(i)}) \qquad \text{for all } \theta$$

Then, since $\theta_{MLE}$ is a valid $\theta$,

$$\begin{aligned} P(\theta_{MAP}) \prod_{i=1}^{m} g(\theta_{MAP}^T x^{(i)}) &\geq P(\theta_{MLE}) \prod_{i=1}^{m} g(\theta_{MLE}^T x^{(i)}) \\ \frac{P(\theta_{MAP})}{P(\theta_{MLE})} &\geq \prod_{i=1}^{m} \frac{g(\theta_{MLE}^T x^{(i)})}{g(\theta_{MAP}^T x^{(i)})} \\ \frac{P(\theta_{MAP})}{P(\theta_{MLE})} &\geq 1 \qquad \text{by equation 8} \end{aligned}$$

Thus $P(\theta_{MAP}) \geq P(\theta_{MLE})$. Now since both $P(\theta_{MAP})$ and $P(\theta_{MLE})$ are $N(0, \alpha^2 I)$,
Thus

$$\begin{aligned} P(\theta_{MAP}) &\geq P(\theta_{MLE}) \\ \frac{1}{\sqrt{(2\pi)}^{-k}\sqrt{|\alpha^2 I|}} exp\left(-\frac{\theta_{MAP}^T \theta_{MAP}}{2\alpha^2}\right) &\geq \frac{1}{\sqrt{(2\pi)}^{-k}\sqrt{|\alpha^2 I|}} exp\left(-\frac{\theta_{MLE}^T \theta_{MLE}}{2\alpha^2}\right) \\ -\theta_{MAP}^T \theta_{MAP} &\geq -\theta_{MLE}^T \theta_{MLE} \\ ||\theta_{MAP}||_2 &\leq ||\theta_{MLE}||_2 \end{aligned}$$

# Problem 3

### 3B3.

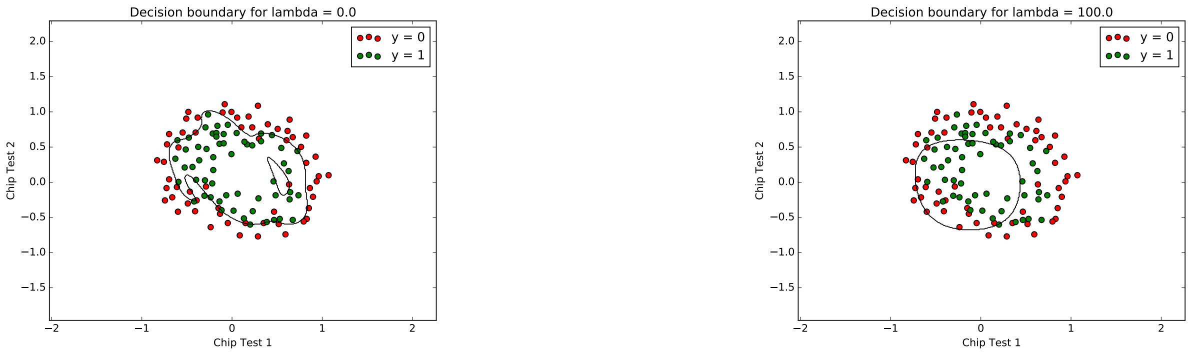
Decision boundaries for over-fitted ($\lambda = 0$) and under fitted ($\lambda = 100$) cases are shown in the figure 1.

### 3C.

For this particular data set we would recommend L2 regularization with log-transformation of the data. The reason of choosing L2 over L1 is that both methods give about the same accuracy, while L2 is computationally less expensive. Log-transformation, on the other hand, in both cases improves accuracy by about 2%, which is significant ( 94% vs 92%).

### 3D.

The Mel Cepstral representation did much better with every genre (we tried the FFT form with regularization levels of 0.1, 1, and 10, and found no significant improvent for any level of regularization), so we report its results here. The easiest genre to classify is Classical, followed by pop, then disco and metal. The hardest genre to classify is rock, followed by country and jazz.



(a) Decision boundary for $\lambda = 0$

(b) Decision boundary for $\lambda = 100$

Figure 1: Decision boundaries for over-fitted ($\lambda = 0$) and under-fitted ($\lambda = 100$) cases.