Problem Set #1: Supervised Learning

Trung H. Nguyen

1. **Newton's method for computing least squares**

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2$$

(a) Taking the partial derivative of the cost function $J(\theta)$ w.r.t to each entry of θ, we have:

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j}\frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2 \\ &= \frac{1}{2}\sum_{i=1}^{m} 2(\theta^T x^{(i)} - y^{(i)})\frac{\partial}{\partial \theta_j}(\theta^T x^{(i)} - y^{(i)}) \\ &= \sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})x_j^{(i)}\end{aligned}$$

Then we can compute each entry of the Hessian as follow:

$$\begin{aligned}\frac{\partial^2 J(\theta)}{\partial_j \partial_k} &= \frac{\partial}{\partial \theta_k}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})x_j^{(i)} \\ &= \sum_{i=1}^{m} x_k^{(i)} x_j^{(i)}\end{aligned}$$

Therefore, the Hessian of the cost function $J(\theta)$ is $\nabla_\theta^2 J(\theta) = X^T X$

(b) For a given arbitrary $\theta^{(0)}$, following the update rule of Newton's method for the first iteration, we have:

$$\begin{aligned}\theta^{(1)} &:= \theta^{(0)} - (\nabla_\theta^2 J(\theta^{(0)}))^{-1}\nabla_\theta J(\theta^{(0)}) \\ &= \theta^{(0)} - (X^T X)^{-1}(X^T X\theta^{(0)} - X^T \vec{y}) \\ &= \theta^{(0)} - \theta^{(0)} + (X^T X)^{-1} X^T \vec{y} \\ &= (X^T X)^{-1} X^T \vec{y}\end{aligned}$$

2. **Locally-weighted logistic regression**

3. **Multivariate least squares**

(a) We have:

$$\begin{aligned}J(\Theta) &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p}\left((\Theta^T x^{(i)})_j - y_j^{(i)}\right)^2 \\ &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p}(X\Theta - Y)_{ij}^2 \\ &= \frac{1}{2}\sum_{k=1}^{m}(X\Theta - Y)_k{}^T(X\Theta - Y)_k \\ &= \frac{1}{2}\operatorname{Tr}\left((X\Theta - Y)(X\Theta - Y)^T\right)\end{aligned}$$

(b) Taking the gradient of $J(\Theta)$ w.r.t Θ, we have:

$$\begin{aligned}
\nabla_\Theta J(\Theta) &= \nabla_\Theta \left(\frac{1}{2} \operatorname{Tr}\left((X\Theta - Y)(X\Theta - Y)^T \right) \right) \\
&= \frac{1}{2} \nabla_\Theta \left(\operatorname{Tr}\left(X\Theta\Theta^T X^T - X\Theta Y^T - Y\Theta^T X^T + YY^T \right) \right) \\
&= \frac{1}{2} \nabla_\Theta \left(\operatorname{Tr}(X\Theta\Theta^T X^T) - 2\operatorname{Tr}(X\Theta Y^T) \right) \\
&= \frac{1}{2} (2X^T X\Theta - 2X^T Y) \\
&= X^T X\Theta - X^T Y
\end{aligned}$$

Setting the gradient to zero, we obtain the solution for Θ that minimizes $J(\Theta)$:

$$\Theta = (X^T X)^{-1} X^T Y$$

(c) We have θ_j is the least squares solution of the j^{th} linear model:

$$\theta_j = (X^T X)^{-1} X^T \overrightarrow{y_j}$$

Put θ_j's into the columns of a matrix, we have:

$$\begin{aligned}
\begin{bmatrix} \theta_1 & \cdots & \theta_p \end{bmatrix} &= \begin{bmatrix} (X^T X)^{-1} X^T \overrightarrow{y_1} & \cdots & (X^T X)^{-1} X^T \overrightarrow{y_p} \end{bmatrix} \\
&= (X^T X)^{-1} X^T \begin{bmatrix} \overrightarrow{y_1} & \cdots & \overrightarrow{y_p} \end{bmatrix} \\
&= (X^T X)^{-1} X^T Y \\
&= \Theta
\end{aligned}$$

Therefore, the parameters from these p independent least squares problems is the exact same as the multivariate solution.

4. **Naive Bayes**

(a) We have the joint likelihood function:

$$\begin{aligned}
l(\varphi) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \varphi) \\
&= \log \prod_{i=1}^{m} p(x^{(i)} | y^{(i)}; \varphi) p(y^{(i)}; \varphi) \\
&= \sum_{i=1}^{m} \left[\log \prod_{j=1}^{n} p(x^{(i)} | y^{(i)}; \varphi) + \log p(y^{(i)}; \varphi) \right] \\
&= \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{n} \log p(x^{(i)} | y^{(i)}; \varphi) \right) + y^{(i)} \log \phi_{y^{(i)}} + (1 - y^{(i)}) \log(1 - \phi_{y^{(i)}}) \right] \\
&= \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{n} x_j^{(i)} \log \phi_{j|y^{(i)}} + (1 - x_j^{(i)}) \log(1 - \phi_{j|y^{(i)}}) \right) + y^{(i)} \log \phi_{y^{(i)}} + (1 - y^{(i)}) \log(1 - \phi_{y^{(i)}}) \right]
\end{aligned}$$

(b) To get the maximum likelihood estimate, we set the gradient of the log-likelihood w.r.t to each parameter equal to zero.

- Taking the gradient of the log-likelihood w.r.t $\phi_{j|y=0}$, we have:

$$
\begin{aligned}
\nabla_{\phi_{j|y=0}} l(\varphi) &= \nabla_{\phi_{j|y=0}} \sum_{i=1}^{m} \left(\sum_{k=1}^{n} x_k^{(i)} \log \phi_{k|y^{(i)}} + (1 - x_j^{(i)}) \log(1 - \phi_{k|y^{(i)}}) \right) \\
&= \nabla_{\phi_{j|y=0}} \sum_{i=1}^{m} 1\{y^{(i)} = 0\} \left(x_j^{(i)} \log \phi_{j|y=0} + (1 - x_j^{(i)}) \log(1 - \phi_{j|y=0}) \right) \\
&= \sum_{i=1}^{m} 1\{y^{(i)} = 0\} \left(\frac{x_j^{(i)}}{\phi_{j|y=0}} - \frac{1 - x_j^{(i)}}{1 - \phi_{j|y=0}} \right) \\
&= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} \left(x_j^{(i)} - \phi_{j|y=0} \right)}{\phi_{j|y=0} \left(1 - \phi_{j|y=0} \right)}
\end{aligned}
$$

  Setting the gradient equal to zero, we have:

$$
\begin{aligned}
0 &= \sum_{i=1}^{m} 1\{y^{(i)} = 0\} \left(x_j^{(i)} - \phi_{j|y=0} \right) \\
\Leftrightarrow \phi_{j|y=0} &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} x_j^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} \\
&= \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}}
\end{aligned}
$$

- Similarly, we have:

$$
\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}
$$

- Taking the gradient of the log-likelihood w.r.t ϕ_y, we have:

$$
\begin{aligned}
\nabla_{\phi_y} l(\varphi) &= \sum_{i=1}^{m} y^{(i)} \log \phi_{y^{(i)}} + (1 - y^{(i)}) \log(1 - \phi_{y^{(i)}}) \\
&= \sum_{i=1}^{m} \frac{y^{(i)}}{\phi_y} - \frac{1 - y^{(i)}}{1 - \phi_y} \\
&= \frac{\sum_{i=1}^{m} \left(y^{(i)} - \phi_y \right)}{\phi_y (1 - \phi_y)}
\end{aligned}
$$

  Then setting the gradient equal to zero lets us obtain:

$$
\begin{aligned}
\sum_{i=1}^{m} \left(y^{(i)} - \phi_y \right) &= 0 \\
\sum_{i=1}^{m} 1\{y^{(i)} = 1\} - m\phi_y &= 0 \\
\Leftrightarrow \phi_y &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}
\end{aligned}
$$

(c) We have:

$$
\begin{aligned}
p(y = 0|x) &= \frac{p(x|y = 0)p(y = 0)}{p(x)} \\
p(y = 1|x) &= \frac{p(x|y = 1)p(y = 1)}{p(x)}
\end{aligned}
$$

thus,

$$
\begin{aligned}
& p(y=1|x) \geq p(y=0|x) \\
\Leftrightarrow & \frac{p(y=1|x)}{p(y=0|x)} \geq 1 \\
\Leftrightarrow & \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)} \geq 1 \\
\Leftrightarrow & \frac{\phi_y \prod_{j=1}^n \left(\phi_{j|y=1}\right)^{x_j} \left(1-\phi_{j|y=1}\right)^{1-x_j}}{(1-\phi_y) \prod_{j=1}^n \left(\phi_{j|y=0}\right)^{x_j} \left(1-\phi_{j|y=0}\right)^{1-x_j}} \geq 1 \\
\Leftrightarrow & \frac{\phi_y}{1-\phi_y} \prod_{j=1}^n \left(\frac{\phi_{j|y=0}}{\phi_{j|y=1}}\right)^{x_j} \left(\frac{1-\phi_{j|y=1}}{1-\phi_{j|y=0}}\right)^{1-x_j} \geq 1 \\
\Leftrightarrow & \log \frac{\phi_y}{1-\phi_y} + \sum_{j=1}^n x_j \log \frac{\phi_{j|y=1}}{\phi_{j|y=0}} + (1-x_j) \log \frac{1-\phi_{j|y=1}}{1-\phi_{j|y=0}} \geq 0 \\
\Leftrightarrow & \log \frac{\phi_y}{1-\phi_y} + \sum_{j=1}^n \log \frac{1-\phi_{j|y=1}}{1-\phi_{j|y=0}} + \sum_{j=1}^n x_j \left(\log \frac{\phi_{j|y=1}}{\phi_{j|y=0}} - \log \frac{1-\phi_{j|y=1}}{1-\phi_{j|y=0}}\right) \geq 0 \\
\Leftrightarrow & \theta^T \begin{bmatrix} 1 \\ x \end{bmatrix} \geq 0
\end{aligned}
$$

where in the last step, we choose:

$$
\begin{aligned}
\theta_0 &= \log \frac{\phi_y}{1-\phi_y} + \sum_{j=1}^n \log \frac{1-\phi_{j|y=1}}{1-\phi_{j|y=0}} \\
\theta_j &= \log \frac{\phi_{j|y=1}}{\phi_{j|y=0}} - \log \frac{1-\phi_{j|y=1}}{1-\phi_{j|y=0}} \quad \text{for } j = 1 \text{ to } n
\end{aligned}
$$

5. **Exponential family and the geometric distribution**

(a) We have the geometric distribution parameterized by φ:

$$
\begin{aligned}
p(y;\phi) &= (1-\phi)^{y-1}\phi \\
&= \exp\left((y-1)\log(1-\phi) + \log\phi\right) \\
&= \exp\left(y\log(1-\phi) - \log\frac{1-\phi}{\phi}\right)
\end{aligned}
$$

It can be seen that the Geometric distribution above is in the form of Exponential family with:

$$
\begin{aligned}
\eta &= \log(1-\phi) \\
T(y) &= y \\
b(y) &= 1 \\
a(\eta) &= \log\frac{1-\phi}{\phi} = \log\frac{e^\eta}{1-e^\eta}
\end{aligned}
$$

(b) We have the canonical response function:

$$
g(\eta) = E\left(T(y);\eta\right) = E\left(y;\eta\right) = \frac{1}{\phi} = \frac{1}{1-e^\eta}
$$

(c) Consider the derivative of the log-likelihood of a training example w.r.t θ_j, we have:

$$
\begin{aligned}
\frac{\partial l^{(i)}(\theta)}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j} \log p\left(y^{(i)}|x^{(i)};\theta\right) \\
&= \frac{\partial}{\partial \theta_j}\left(\eta \cdot y^{(i)} - a(\eta)\right) \\
&= \frac{\partial}{\partial \theta_j}\left(\eta \cdot y^{(i)} - \log \frac{e^{\eta}}{1-e^{\eta}}\right) \\
&= x_j^{(i)} y^{(i)} - \frac{1}{1-e^{\eta}} x_j^{(i)} \\
&= \left(y^{(i)} - \frac{1}{1-e^{\theta^T x^{(i)}}}\right) x_j^{(i)}
\end{aligned}
$$

which gives us the stochastic gradient ascent update rule for each θ_j:

$$
\begin{aligned}
\theta_j &:= \theta_j + \alpha \frac{\partial l^{(i)}(\theta)}{\partial \theta_j} \\
&= \theta_j + \alpha\left(y^{(i)} - \frac{1}{1-e^{\theta^T x^{(i)}}}\right) x_j^{(i)}
\end{aligned}
$$