Shanghai Jiao Tong University

Machine Learning PS: 1

Author: *LIUQiaoan 520030910220*

Course: *Autumn 2018, CS229: Machine Learning*
Date: *July 9, 2022*

Problem 1

(a) We have known from lecture that:

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}$$

Then

$$\begin{aligned}
\frac{\partial^2 J(\theta)}{\partial \theta_j \partial \theta_k} &= \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} \frac{\partial h_\theta(x^{(i)})}{\partial \theta_k} \\
&= \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} \frac{\partial g(\theta^T x^{(i)})}{\partial \theta_k} \\
&= \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} g(\theta^T x^{(i)}) \left[1 - g(\theta^T x^{(i)})\right] x_k^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)}) \left[1 - h_\theta(x^{(i)})\right] x_j^{(i)} x_k^{(i)}
\end{aligned}$$

Therefore, the Hessian $H = \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)}) \left[1 - h_\theta(x^{(i)})\right] x^{(i)} {x^{(i)}}^T$.

For any vector z,

$$\begin{aligned}
z^T H z &= z^T \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)}) \left[1 - h_\theta(x^{(i)})\right] x^{(i)} {x^{(i)}}^T z \\
&= \frac{1}{m}\sum_{i=1}^{m}\sum_{j}\sum_{k} h_\theta(x^{(i)}) \left[1 - h_\theta(x^{(i)})\right] z_j x_j^{(i)} x_k^{(i)} z_k \\
&= \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)}) \left[1 - h_\theta(x^{(i)})\right] ({x^{(i)}}^T z)^2 \\
&\geq 0
\end{aligned}$$

so it holds that $z^H z \geq 0$.

(c) Since

$$\begin{aligned} p(y=1|x;\phi,\mu_0,\mu_1,\Sigma) &= \frac{p(x|y=1)p(y=1)}{p(x)} \\ &= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)} \end{aligned}$$

then we can write the posterior distribution as:

$$\begin{aligned} p(y=1|x;\phi,\mu_0,\mu_1,\Sigma) &= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)} \\ &= \frac{\phi\exp\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\}}{\phi\exp\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\}+(1-\phi)\exp\left\{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right\}} \\ &= \frac{1}{1+\frac{1-\phi}{\phi}\exp\left\{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\}} \\ &= \frac{1}{1+\frac{1-\phi}{\phi}\exp\left\{\mu_0^T\Sigma^{-1}x-\mu_1^T\Sigma^{-1}x+\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1-\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0\right\}} \\ &= \frac{1}{1+\exp\left(-(\theta^Tx+\theta_0)\right)} \end{aligned}$$

with

$$\begin{cases} \theta^T = (\mu_1-\mu_0)^T\Sigma^{-1} \\ \theta_0 = \log\phi - \log(1-\phi) + \dfrac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \dfrac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 \end{cases}$$

(d)

$$\begin{aligned} l(\phi,\mu_0,\mu_1,\Sigma) &= \log\prod_{i=1}^{m} p(x^{(i)},y^{(i)};\phi,\mu_0,\mu_1,\Sigma) \\ &= \log\prod_{i=1}^{m} p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)p(y^{(i)};\phi) \\ &= \sum_{i=1}^{m}\left\{\log\left[(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)\right]+\log p(y^{(i)};\phi)\right\} \end{aligned}$$

First, we take partial derivative for ϕ:

$$\frac{\partial l}{\partial\phi} = \sum_{i=1}^{m}\left[\frac{y^{(i)}}{\phi}+\frac{y^{(i)}-1}{1-\phi}\right] = \sum_{i=1}^{m}\frac{y^{(i)}-\phi}{\phi(1-\phi)} = 0 \iff \phi = \frac{1}{m}\sum_{i=1}^{m}y^{(i)}$$

Second, we consider the partial derivative for μ_0 and μ_1. We first consider $p(x|y=i) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)\right)$ for $i=0,1$:

$$\begin{aligned} \frac{\partial\log p(x|y=i)}{\partial\mu_i} &= \frac{\partial(-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i))}{\partial\mu_i} \\ &= \Sigma^{-1}(x-\mu_i) \end{aligned}$$

Then we have:

$$\frac{\partial l}{\partial \mu_0} = 0 \iff \sum_{i=1}^{m} 1\left\{y^{(i)} = 0\right\} \Sigma^{-1}(x^{(i)} - \mu_0) = 0 \iff \mu_0 = \frac{\sum_{i=1}^{m} 1\left\{y^{(i)} = 0\right\} x^{(i)}}{\sum_{i=1}^{m} 1\left\{y^{(i)} = 0\right\}}$$

(g) On dataset 1, GDA seems to perform worse than logistic regression, because for dataset 1, $p(x|y)$ is not a Gaussian distribution.

Problem 2

(a) We use the knowledge of probability:

$$\begin{aligned} p(y^{(i)} = 1|x^{(i)}) &= p(t^{(i)} = 1|x^{(i)}) * p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)}) \\ &= p(t^{(i)} = 1|x^{(i)}) * p(y^{(i)} = 1|t^{(i)} = 1) \end{aligned}$$

So $\alpha = p(y^{(i)} = 1|t^{(i)} = 1) \in \mathbb{R}$.

(b) This subproblem uses result of subproblem (a), when $x^{(i)} \in V_+$, we have

$$\begin{aligned} h(x^{(i)}) &\approx p(y^{(i)} = 1|x^{(i)}) \\ &= p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)}) * p(t^{(i)} = 1|x^{(i)}) \\ &= p(y^{(i)} = 1|t^{(i)} = 1) * p(t^{(i)} = 1|x^{(i)}) \\ &\approx \alpha * 1 \\ &= \alpha \end{aligned}$$

(e) There is a update of θ after we get α,

$$\begin{aligned} \frac{1}{1 + e^{-\theta'^T x}} &= \frac{1}{\alpha} \frac{1}{1 + e^{-\theta^T x}} \\ &\geq 0.5 \end{aligned}$$

Then

$$\theta'^T x = \theta^T x + \log\left(\frac{2}{\alpha} - 1\right) \geq 0 \iff \theta_0' = \theta_0 + \log\left(\frac{2}{\alpha} - 1\right)$$

Problem 3

(a) We write Possion distribution as a function in exponential family:

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

Then

$$\begin{cases} T(y) = y \\ \eta = \log \lambda \\ a(\eta) = e^{\eta} \\ b(y) = \dfrac{1}{y!} \end{cases}$$

(b) In this problem, **canonical response function** is:

$$\begin{aligned} g(\eta) &= \mathbb{E}[y|x;\lambda] \\ &= \lambda \\ &= e^{\eta} \\ &= e^{\theta^T x} \end{aligned}$$

(c)

$$\begin{aligned} l(\theta) &= \log p(y^{(i)}|x^{(i)};\theta) \\ &= \log(\frac{1}{y!}\exp(y\log\lambda - \lambda)) \\ &= -\log y^{(i)} + \theta^T x^{(i)} y^{(i)} - e^{\theta^T x^{(i)}} \end{aligned}$$

Then we take derivative:

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

Then

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

Problem 4

(a) I get inspiration from the hint, first we take derivative of $\int p(y;\eta)dy$ w.r.t η:

$$\begin{aligned} \frac{\partial}{\partial \eta} \int p(y;\eta)dy &= \int \frac{\partial}{\partial \eta} p(y;\eta)dy \\ &= \int \left[y - \frac{\partial a(\eta)}{\partial \eta} \right] p(y;\eta)dy \\ &= 0 \qquad \text{(Since } \int p(y;\eta)dy \text{ is constant 1 no matter what } \eta \text{ is)} \end{aligned}$$

So we get:

$$\mathbb{E}\left[Y|X;\theta\right] = \int yp(y;\eta)dy = \frac{\partial a(\eta)}{\partial \eta} \cdot \int p(y;\eta)dy = \frac{\partial a(\eta)}{\partial \eta}$$

(b) Similar to sub-problem (a), we take derivative of $\int yp(y;\eta)dy$ w.r.t η:

$$\begin{aligned}
\frac{\partial}{\partial\eta}\int yp(y;\eta)dy &= \int \frac{\partial}{\partial\eta}yp(y;\eta)dy \\
&= \int \left[y - \frac{\partial a(\eta)}{\partial\eta}\right] yp(y;\eta)dy \\
&= \int y^2p(y;\eta)dy - \frac{\partial a(\eta)}{\partial\eta}\cdot\int yp(y;\eta)dy \\
&= \mathbb{E}\left[Y^2|X;\theta\right] - \{\mathbb{E}\left[Y|X;\theta\right]\}^2 \\
&= \frac{\partial}{\partial\eta}\mathbb{E}\left[Y|X;\theta\right] \\
&= \frac{\partial^2 a(\eta)}{\partial\eta^2}
\end{aligned}$$

Then $Var(Y|X;\theta) = \mathbb{E}\left[Y^2|X;\theta\right] - \{\mathbb{E}\left[Y|X;\theta\right]\}^2 = \frac{\partial^2 a(\eta)}{\partial\eta^2}$.

(c) We take second derivative of $l(\theta)$ w.r.t θ:

$$\begin{aligned}
l(\theta) &= -\sum_{i=1}^{m}\log p(y^{(i)}|x^{(i)};\theta) \\
&= -\sum_{i=1}^{m}\log\left[b(y^{(i)})\exp\left(\eta y^{(i)} - a(\eta)\right)\right] \\
&= -\sum_{i=1}^{m}\left[\log b(y^{(i)}) + \eta y^{(i)} - a(\eta)\right]
\end{aligned}$$

Then first derivative:

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial\theta_j} &= -\sum_{i=1}^{m}\left[y^{(i)}x_j^{(i)} - \frac{\partial a(\eta)}{\partial\eta}x_j^{(i)}\right] && \text{Use chain rule} \\
&= \sum_{i=1}^{m}\left[\frac{\partial a(\eta)}{\partial\eta} - y^{(i)}\right]x_j^{(i)}
\end{aligned}$$

Then second derivative:

$$\frac{\partial^2 l(\theta)}{\partial\theta_j\partial\theta_k} = \sum_{i=1}^{m}\frac{\partial^2 a(\eta)}{\partial\eta^2}x_j^{(i)}x_k^{(i)}$$

As a result, the Hessian matrix is:

$$H = Var(Y|X;\theta)\sum_{i=1}^{m}x^{(i)}{x^{(i)}}^T$$

For any $z \in \mathbb{R}^n$, we have

$$\begin{aligned} z^T H z &= Var(Y|X;\theta) \sum_{i=1}^{m} z^T x^{(i)} {x^{(i)}}^T z \\ &= Var(Y|X;\theta) \sum_{i=1}^{m} (z^T x^{(i)})^2 \\ &\geq 0 \end{aligned}$$

So the Hessian is always PSD.

Problem 5

(a) i.

$$X = \begin{bmatrix} -(x^{(1)})^T- \\ \vdots \\ -(x^{(m)})^T- \end{bmatrix} \in \mathbb{R}^{m\times n} \qquad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

So

$$X\theta - y = \begin{bmatrix} \theta^T x^{(1)} - y^{(1)} \\ \vdots \\ \theta^T x^{(m)} - y^{(m)} \end{bmatrix} \quad \Longrightarrow \quad W = \frac{1}{2}\text{diag}(w^{(i)}, \ldots, w^{(m)})$$

ii.

$$\begin{aligned} J(\theta) &= (\theta^T X^T - y^T) W (X\theta - y) \\ &= \theta^T X^T W X\theta - \theta^T X^T W y - y^T W X \theta + y^T W y \\ &= \theta^T X^T W X\theta - 2y^T W X\theta + y^T W y \end{aligned}$$

Take derivative w.r.t θ and set it to 0:

$$\nabla_\theta J(\theta) = 2X^T W X\theta - 2X^T W y = 0 \iff \theta = (X^T W X)^{-1} X^T W y$$

iii. The maximum likelihood $l(\theta)$:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta) \\ &= \sum_{i=1}^{m} \log \frac{1}{\sqrt{(2\pi)}\sigma^{(i)}} - \sum_{i=1}^{m} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2} \end{aligned}$$

Then maximizing $l(\theta)$ is equivalent to minimizing $\sum_{i=1}^{m} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2} = \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(y^{(i)} - \theta^T x^{(i)})^2$, so $w^{(i)} = \frac{1}{(\sigma^{(i)})^2}$.