Notation: let $x_p^{(i)} \in \mathbb{R}^d$, where $i = 1, \cdots, n$. $p = 1, 2, \cdots, d$

n is the sample size

d is the dimension of features

$$X = \begin{bmatrix} x^{T(1)} \\ x^{T(2)} \\ \vdots \\ x^{T(n)} \end{bmatrix}$$ is the total sample data

(1.a.) Let $z^{(i)} = \theta^T x^{(i)}$ where $\theta \in \mathbb{R}^d$

$$f_1(z^{(i)}) = -\log\left(\frac{1}{1+e^{-z^{(i)}}}\right) \quad ①$$

$$f_2(z^{(i)}) = -\log\left(1 - \frac{1}{1+e^{-z^{(i)}}}\right) = z^{(i)} + f_1(z^{(i)}) \quad ②$$

we can rewrite $J(\theta)$ as

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)} f_1(z^{(i)}) + (1-y^{(i)}) f_2(z^{(i)})\right] \quad ③$$

since $\frac{\partial f_1(z^{(i)})}{\partial \theta_p} = -\frac{\partial}{\partial \theta_p}\log\left(\frac{1}{1+e^{-z^{(i)}}}\right) = -\frac{\partial}{\partial z^{(i)}}\log\left(\frac{1}{1+e^{-z^{(i)}}}\right)\cdot\frac{\partial z^{(i)}}{\partial \theta_p}$

$$= [-1 + g(z^{(i)})]\, x_p^{(i)}$$

$$\frac{\partial f_1(z^{(i)})}{\partial \theta_p \partial \theta_q} = \frac{\partial}{\partial \theta_q}\left([-1+g(z^{(i)})]\cdot x_p^{(i)}\right) = \frac{\partial}{\partial z^{(i)}}\left([-1+g(z^{(i)})]\cdot x_p^{(i)}\right)\cdot\frac{\partial z^{(i)}}{\partial \theta_q}$$

$$= g(z^{(i)})[1-g(z^{(i)})]\, x_p^{(i)} \cdot x_q^{(i)} \quad ④$$

② $\Rightarrow$ $\frac{\partial f_2(z^{(i)})}{\partial \theta_p} = \frac{\partial}{\partial \theta_p}[z^{(i)} + f_1(z^{(i)})] = x_p^{(i)} + \frac{\partial}{\partial \theta_p} f_1(z^{(i)})$

$$\frac{\partial f_2(z^{(i)})}{\partial \theta_p \partial \theta_q} = \frac{\partial}{\partial \theta_q}\left[x_p^{(i)} + \frac{\partial}{\partial \theta_p} f_1(z^{(i)})\right] = \frac{\partial}{\partial \theta_p \partial \theta_q} f_1(z^{(i)}) \quad ⑤$$

④ ⑤ $\Rightarrow$ $H_{pq} = \frac{\partial^2 J(\theta)}{\partial \theta_p \partial \theta_q} = \frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\frac{\partial}{\partial \theta_p \partial \theta_q} f_1(z^{(i)}) + (1-y^{(i)})\frac{\partial}{\partial \theta_p \partial \theta_q} f_1(z^{(i)})\right]$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \theta_p \partial \theta_q} f_1(z^{(i)}) = \frac{1}{n}\sum_{i=1}^{n} g(z^{(i)})[1-g(z^{(i)})]\, x_p^{(i)} x_q^{(i)} \quad ⑥$$

⑥ $H = \frac{1}{n} X^T \begin{bmatrix} g'(z^{(1)}) & & & \\ & g'(z^{(2)}) & & \\ & & \ddots & \\ & & & g'(z^{(n)}) \end{bmatrix} X = \frac{1}{n} X^T \underset{i=1,\cdots,n}{\operatorname{diag}}\{g'(z^{(i)})\} X$

Therefore, for any vector z, $z^T H z = \frac{1}{2}(Xz)^T \underset{i=1,\cdots,n}{\operatorname{diag}}\{g'(z^{(i)})\}(Xz)$

since $g'(z^{(i)}) = g(z^{(i)})(1-g(z^{(i)})) \geq 0$, for $i = 1, \cdots, n$, we can conclude that $z^T H z \geq 0$

(1.C) Define $A = -\frac{1}{2}(X-\mu_1)^T \Sigma^{-1}(X-\mu_1)$ ①

$B = -\frac{1}{2}(X-\mu_0)^T \Sigma^{-1}(X-\mu_0)$ ②

$C = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}}$ ③

from Bayes Rule, we can get

$$p(y=1|X) = \frac{p(X|y=1)\,p(y=1)}{p(X|y=1)\,p(y=1) + p(X|y=0)\,p(y=0)} = \frac{Ce^A\phi}{Ce^A\phi + Ce^B(1-\phi)}$$

$$= \frac{1}{1+e^{-(A-B+\ln\phi-\ln(1-\phi))}} \quad ④$$

from ① ② $\Rightarrow$ $A-B = -\frac{1}{2}(X-\mu_1)^T\Sigma^{-1}(X-\mu_1) + \frac{1}{2}(X-\mu_0)^T\Sigma^{-1}(X-\mu_0)$

$= (\mu_1-\mu_0)^T\Sigma^{-1}X + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1$ ⑤

from ⑤ $\Rightarrow$ $A-B+\ln\phi-\ln(1-\phi) = (\mu_1-\mu_0)^T\Sigma^{-1}X + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \ln\phi - \ln(1-\phi)$

with $\theta = (\Sigma^{-1})^T(\mu_1-\mu_0)$, $\theta_0 = \ln\phi - \ln(1-\phi) + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1$

we can get

$A-B+\ln\phi = \theta^TX + \theta_0$

So we can rewrite ④ as

$$p(y=1|X) = \frac{1}{1+\exp(-(\theta^TX+\theta_0))}$$

(1.d) The log likelihood of the training dataset is given by

$$\ell(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)$$

$$= \sum_{i=1}^{n} \log p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{n} \log p(y^{(i)}; \phi)$$

$$= \sum_{i=1}^{n} \log \frac{1}{(2\pi)^{d/2}} + \frac{1}{2} \sum_{i=1}^{n} \log \frac{1}{|\Sigma|} - \frac{1}{2} \sum_{i=1}^{n} 1\{y^{(i)}=0\} (x^{(i)} - \mu_0)^T \Sigma^{-1} (x^{(i)} - \mu_0)$$

$$- \frac{1}{2} \sum_{i=1}^{n} 1\{y^{(i)}=1\} (x^{(i)} - \mu_1)^T \Sigma^{-1} (x^{(i)} - \mu_1)$$

Now, the likelihood is maximized by setting the derivative with respect to each of the parameters to zero:

For ϕ: $\frac{\partial \ell}{\partial \phi} = \frac{1}{\phi} \sum_{i=1}^{n} 1\{y^{(i)}=1\} - \frac{1}{1-\phi} \sum_{i=1}^{n} 1\{y^{(i)}=0\}$

setting $\frac{\partial \ell}{\partial \phi} = 0$ and solving for $\phi \Rightarrow \phi = \frac{1}{n} \sum_{i=1}^{n} 1\{y^{(i)}=1\}$

For μ_0: $\frac{\partial \ell}{\partial \mu_0} = -\frac{1}{2} \sum_{i=1}^{n} 1\{y^{(i)}=0\} \frac{\partial}{\partial \mu_0} (x^{(i)} - \mu_0)^T \Sigma^{-1} (x^{(i)} - \mu_0)$

$$= -\frac{1}{2} \sum_{i=1}^{n} 1\{y^{(i)}=0\} \frac{\partial}{\partial \mu_0} (x^{T(i)} \Sigma^{-1} x^{(i)} - x^{T(i)} \Sigma^{-1} \mu_0 - \mu_0^T \Sigma^{-1} x^{(i)} + \mu_0^T \Sigma^{-1} \mu_0)$$

$$= \frac{1}{2} \sum_{i=1}^{n} 1\{y^{(i)}=0\} \left[\frac{\partial}{\partial \mu_0} (x^{T(i)} \Sigma^{-1} \mu_0) + \frac{\partial}{\partial \mu_0} (\mu_0^T \Sigma^{-1} x^{(i)}) - \frac{\partial}{\partial \mu_0} (\mu_0^T \Sigma^{-1} \mu_0) \right]$$

$$= \frac{1}{2} \sum_{i=1}^{n} 1\{y^{(i)}=0\} \left[(\Sigma^{-1})^T x^{(i)} + \Sigma^{-1} x^{(i)} - \left(\Sigma^{-1} + (\Sigma^{-1})^T \right) \mu_0 \right]$$

since Σ^{-1} is symmetric, $(\Sigma^{-1})^T = \Sigma^{-1}$, setting $\frac{\partial \ell}{\partial \mu_0} = 0$ and solving for $\mu_0 \Rightarrow$

$$\mu_0 = \frac{\sum_{i=1}^{n} 1\{y^{(i)}=0\} x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)}=0\}}$$

similarly, for μ_1:

$$\frac{\partial \ell}{\partial \mu_1} = \frac{1}{2} \sum_{i=1}^{n} 1\{y^{(i)}=1\} \left[(\Sigma^{-1})^T x^{(i)} + \Sigma^{-1} x^{(i)} - \left(\Sigma^{-1} + (\Sigma^{-1})^T \right) \mu_1 \right]$$

since Σ^{-1} is symmetric, $(\Sigma^{-1})^T = \Sigma^{-1}$, setting $\frac{\partial \ell}{\partial \mu_1} = 0$ and solving for $\mu_1 \Rightarrow$

$$\mu_1 = \frac{\sum_{i=1}^{n} 1\{y^{(i)}=1\} x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)}=1\}}$$

1.d for Σ:

$$\frac{\partial \ell}{\partial \Sigma} = \frac{\partial}{\partial \Sigma} \frac{1}{2} \sum_{i=1}^{n} \left[\log\left(\frac{1}{|\Sigma|}\right) - \frac{1}{2} \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \right]$$

with $\frac{\partial}{\partial \Sigma} \log\left(\frac{1}{|\Sigma|}\right) = -(\Sigma^{-1})^T = -\Sigma^{-1}$ (since Σ is symmetric)

$\frac{\partial}{\partial \Sigma} \Sigma^{-1} = -\Sigma^{-1}\Sigma^{-1}$ (derived by $\frac{\partial I}{\partial \Sigma} = \frac{\partial}{\partial \Sigma}(\Sigma^{-1}\Sigma) = 0$)

we can further get

$$\frac{\partial \ell}{\partial \Sigma} = -\frac{1}{2} n \Sigma^{-1} + \frac{1}{2} \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T (\Sigma^{-1})(\Sigma^{-1})$$

setting $\frac{\partial \ell}{\partial \Sigma} = 0$ and solving for $\Sigma \Rightarrow$

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

1.f Observations: GDA seems to be affected more towards the outliers in the validation dataset 1. This indicates the GDA may have a higher bias compared to logistic regression.

1.g GDA seems to perform worse than logistic regression for Dataset 1; GDA makes more specific assumptions about the dataset then logistic regression, and GDA will perform well if those assumptions are true, so the difference might be caused by the difference of distribution between Dataset 1 and Dataset 2.

1.h The z-transformation given by

$$z^{(i)} = \frac{x^{(i)} - \mu}{s}, \quad i = 1, \ldots, n$$

where μ is the mean of the dataset and s is the standard deviation of the dataset might help improve the GDA's performance with Dataset 1.

(2.a) Since the exponential family is represented by

$$p(y|x;\eta) = b(y)e^{\eta^T T(y) - a(\eta)}$$

and since the poisson distribution can be re-written as

$$p(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

$$= \frac{e^{-\lambda}e^{\ln\lambda^y}}{y!} = \frac{1}{y!}\exp(y\ln\lambda - \lambda)$$

with $b(y) = \frac{1}{y!}$, $\eta = \ln\lambda$, $T(y) = y$, and $a(\eta) = \lambda$, we can conclude that poisson distribution is in the exponential family

(2.b) Canonical response function is given by

$$g(\eta) = E[T(y);\eta] = E[y;\eta] = \lambda = \exp(\eta)$$

(2.c)

The log likelihood of an example $(x^{(i)}, y^{(i)})$ is defined by $\ell(\theta) = \log p(y^{(i)}|x^{(i)};\theta)$ and from 2.a, we can further derive

$$\ell(\theta) = \log p(y^{(i)}|x^{(i)};\theta) = \log[b(y^{(i)})e^{\eta^T T(y^{(i)}) - a(\eta)}]$$

$$= \log b(y^{(i)}) + \eta^T T(y^{(i)}) - a(\eta)$$

$$= \log\frac{1}{y^{(i)}!} + x^{T(i)}\theta \cdot y^{(i)} - \exp(\theta^T x^{(i)})$$

so $\frac{\partial\ell(\theta)}{\partial\theta} = y^{(i)}x^{(i)} - \exp(\theta^T x^{(i)}) \cdot x^{(i)}$

$$\Rightarrow \theta^{(i)} := \theta^{(i)} + \alpha\frac{\partial\ell(\theta)}{\partial\theta}$$

$$= \theta^{(i)} + \alpha[y^{(i)}x^{(i)} - \exp(\theta^T x^{(i)})x^{(i)}]$$

$$= \theta^{(i)} + \alpha[y^{(i)} - \exp(\theta^T x^{(i)})] \cdot x^{(i)}, \quad \text{where } i = 1, \cdots, n$$