1. (a)

$$\nabla_\theta J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left(\frac{y^{(i)}\, g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))x^{(i)}}{h_\theta(x^{(i)})} - \frac{(1-y^{(i)})\, g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))x^{(i)}}{1-h_\theta(x^{(i)})}\right)$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}(1-h_\theta(x^{(i)}))x - (1-y^{(i)})h_\theta(x^{(i)})x\right)$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}x^{(i)} - h_\theta(x^{(i)})x^{(i)}\right) = -\frac{1}{m}X^T(y-h_\theta(X))$$

where $X = \begin{pmatrix} x^{(1)T} \\ \vdots \\ x^{(m)T} \end{pmatrix}$, $h_\theta(x) = g(X\theta)$

$$\nabla_\theta^2 J(\theta) = \left[\frac{\partial \nabla_\theta J(\theta)}{\partial \theta_1}, \cdots \frac{\partial \nabla_\theta J(\theta)}{\partial \theta_n}\right]$$

$$\frac{\partial \nabla_\theta J(\theta)}{\partial \theta_j} = \frac{\frac{1}{m}\sum_{i=1}^{m} g(\theta^T x^{(i)})x^{(i)}}{\partial \theta_j}$$

$$= \frac{1}{m}\sum_{i=1}^{m} g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))x_j^{(i)}x^{(i)}$$

So $H = \nabla_\theta^2 J(\theta) = \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))x^{(i)}x^{(i)T}$

$$= \frac{1}{m}\left[X^T h_\theta(X)\right]\left[\left(\vec{1} - h_\theta(X)\right)X\right]$$

$$z^T H z = \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))z^T x^{(i)}x^{(i)T}z$$

$$= \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))(x^{(i)T}z)^2 \geq 0$$

(c)

$$P(y=1|x) = \frac{P(x|y=1)P(y=1)}{P(x)}$$

$$\frac{P(x|y=1)P(y=1)}{P(x|y=1)P(y=1) + P(x|y=0)P(y=0)}$$

$$= \frac{\phi \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}{\phi \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) + (1-\phi)\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)}$$

$$= \frac{1}{1+\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) + \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \left(\ln\frac{\phi}{1-\phi}\right)\right)}$$

$$= \frac{1}{1+\exp\left(-\left(\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \frac{1}{2}x^T\Sigma^{-1}(\mu_0-\mu_1) - \frac{1}{2}(\mu_0-\mu_1)^T\Sigma^{-1}x + \ln\frac{\phi}{1-\phi}\right)\right)}$$

$$= \frac{1}{1+\exp\left(-\left(-(\mu_0-\mu_1)^T\Sigma^{-1}x + \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_0 - \mu_0^T\Sigma^{-1}\mu_1) + \ln\frac{\phi}{1-\phi}\right)\right)}$$

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

(d)

$$\ell = \sum_{i=1}^{m} \log \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_{y^{(i)}})^T\Sigma^{-1}\right.$$

$$\left.(x-\mu_{y^{(i)}})\right) + \sum_{i=1}^{m} \log \phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}}$$

$$\mu_{y^{(i)}} = y^{(i)}\mu_1 + (1-y^{(i)})\mu_0$$

$$\ell = -\frac{mn}{2}\log 2\pi - \frac{m}{2}\log|\Sigma| + \sum_{i=1}^{m} -\frac{1}{2}(x-(y^{(i)}\mu_1+(1-y^{(i)})\mu_0))^T\Sigma^{-1}$$

$$(x-(y^{(i)}\mu_1+(1-y^{(i)})\mu_0)) + y^{(i)}\log\phi + (1-y^{(i)})\log(1-\phi)$$

$$\frac{\partial \ell}{\partial \phi} = \sum_{i=1}^{m}\left(\frac{y^{(i)}}{\phi} - \frac{1-y^{(i)}}{1-\phi}\right) = 0$$

$$\sum_{i=1}^{m}(y^{(i)} - y^{(i)}\phi) = \sum_{i=1}^{m}(\phi - y^{(i)}\phi)$$

$$\phi = \frac{1}{m}\sum_{i=1}^{m} y^{(i)} = \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)}=1\}$$

$$\frac{\partial \ell}{\partial \mu_0} = \sum_{i=1}^{m} \frac{1}{2}(1-y^{(i)})\Sigma^{-1}(x^{(i)}-(y^{(i)}\mu_1+(1-y^{(i)})\mu_0))$$

$$+ \frac{1}{2}(x^{(i)}-(y^{(i)}\mu_1+(1-y^{(i)})\mu_0))\Sigma^{-1}(1-y^{(i)}) = 0$$

$$\sum_{i=1}^{m}(1-y^{(i)})(x^{(i)}-(y^{(i)}\mu_1+(1-y^{(i)})\mu_0)) = 0$$

the same reasoning leads to

$$\sum_{i=1}^{m} y^{(i)}(x^{(i)}-(y^{(i)}\mu_1+(1-y^{(i)})\mu_0) = 0$$

so $\sum_{i=1}^{m}(1-y^{(i)})(x^{(i)}-\mu_0) = 0$ (when $y^{(i)}=1$, it comes to 0)

$$\sum_{i=1}^{m} y^{(i)}(x^{(i)}-\mu_1) = 0$$

$$\mu_0 = \frac{\sum_{i=1}^{m}(1-y^{(i)})x^{(i)}}{\sum_{i=1}^{m}(1-y^{(i)})}$$

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

$$\nabla_\Sigma \ell = -\frac{m}{2}\Sigma^{-1} - \frac{1}{2}\sum_{i=1}^{m}\nabla(u_i^T\Sigma^{-1}u_i) = 0$$

$$\sum_{i=1}^{m}\nabla(u_i^T\Sigma^{-1}u_i) = -m\Sigma^{-1}$$

$$(\Sigma+h)^{-1} = (\Sigma(I+\Sigma^{-1}h))^{-1}$$

$$= (I+\Sigma^{-1}h)^{-1}\Sigma^{-1}$$

$$= (I-\Sigma^{-1}h)\Sigma^{-1}$$

$$= \Sigma^{-1} - \Sigma^{-1}h\Sigma^{-1}$$

$$u^T(\Sigma+h)^{-1}u$$

$$= u^T\Sigma^{-1}u - u^T\Sigma^{-1}h\Sigma^{-1}u$$

$$= u^T\Sigma^{-1}u + \langle\nabla(u^T\Sigma^{-1}u), h^T\rangle + o(\|h\|)$$

$$\mathrm{tr}[\nabla(u^T\Sigma^{-1}u)h] = -u^T\Sigma^{-1}h\Sigma^{-1}u$$

$$= -\mathrm{tr}(u^T\Sigma^{-1}h\Sigma^{-1}u) = -\mathrm{tr}(\Sigma^{-1}uu^T\Sigma^{-1}h)$$

so $\nabla u^T\Sigma^{-1}u = -\Sigma^{-1}uu^T\Sigma^{-1}$

$$\sum_{i=1}^{m}\Sigma^{-1}u_iu_i^T\Sigma^{-1} = m\Sigma^{-1}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}u_iu_i^T$$

where $u_i = x^{(i)} - \mu_{y^{(i)}}$

g) dataset 1. because the distribution of the $P(x|y)$ is not multi-gaussians distribution

h) the box-cox transformation

We should add 1 to every data in x to avoid the negative or 0 that makes the boxcox trasformation disabled

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{if } \lambda = 0 \end{cases}$$

The function stats.boxcox() can help us to find to optimised λ.

2.

(a) $P(t|y)=1$

$$\frac{P(y,t|x)}{P(t|x)}=\frac{P(y|t)}{P(t)}$$

$$P(t)P(y,t|x)=P(y|t)P(t|x)$$

$$\frac{P(t|x)}{P(y|x)}=\frac{P(t)P(y,t|x)}{P(y|t)P(t|x)}$$

$$P(y|t)=P(t|y)P(y)=P(y)$$

$$P(y,t|x)=P(y,t|x)P(x)$$

$$=P(y|x)P(x)=P(x,y)$$

so $\frac{P(t|x)}{P(y|x)}=\frac{P(t)P(x,y)}{P(y)P(y|x)}=\frac{P(t,y)}{P(y)}=\frac{1}{\alpha}$

(b) $h(x^{(i)})\simeq p(y^{(i)}=1|x^{(i)})=\alpha p(t^{(i)}=1|x^{(i)})$

$=\alpha$

(c) for the boundary

$\theta^T x=0$

now we need correction

$\exp(\theta'^T x)+1=\frac{2}{\alpha}$

$\theta'^T x=-\ln(\frac{2}{\alpha}-1)$ contribute to θ_0

$\theta_0'=\theta_0+\ln(\frac{2}{\alpha}-1)=c_0\cdot\theta_0$

$c_0=1+\frac{\ln(\frac{2}{\alpha}-1)}{\theta_0}$

3. (a) $p(y;\lambda)=\frac{1}{y!}\exp(-\lambda+y\ln\lambda)$

$b(y)=\frac{1}{y!}$, $\eta=\ln\lambda$, $T(y)=y$, $a(\eta)=e^{\eta}$

(b) $E[T(y);\eta]=\lambda=e^{\eta}=e^{\theta^T x}$

(c) $\frac{\partial}{\partial\theta_j}\log p(y^{(i)}|x^{(i)};\theta)$

$=\frac{\partial}{\partial\theta_j}(\log\frac{1}{y^{(i)}!}+(-e^{\theta^T x^{(i)}}+y^{(i)}\theta^T x^{(i)}))$

$=-e^{\theta^T x^{(i)}}\cdot x_j^{(i)}+y^{(i)}x_j^{(i)}$

$=(y^{(i)}-e^{\theta^T x^{(i)}})x_j^{(i)}$

$\theta_j:=\theta_j+\alpha(y^{(i)}-e^{\theta^T x^{(i)}})x_j^{(i)}$

$\theta:=\theta+\alpha X^T(Y-\exp(X\theta))$

4. (a)

$$\frac{\partial}{\partial \eta}\int p(y;\eta)\,dy = 0$$

$$\frac{\partial}{\partial \eta}\int p(y;\eta)\,dy$$

$$= \int \frac{\partial}{\partial \eta} p(y;\eta)\,dy$$

$$= \int b(y)\exp(\eta y - a(\eta))(y - a'(\eta))\,dy$$

$$= \int p(y;\eta) y\,dy - a'(\eta)\int p(y;\eta)\,dy$$

$$= E[Y|X;\theta] - a'(\eta) = 0$$

so $E[Y|X;\theta] = \frac{da(\eta)}{d\eta}$

(b) $\frac{\partial^2}{\partial \eta^2}\int p(y;\eta)\,dy = 0$

$$\int \frac{\partial}{\partial \eta} p(y;\eta)\, y\,dy - a''(\eta) = 0$$

$$\int p(y;\eta)(y - a'(\eta))\, y\,dy - a''(\eta) = 0$$

$$\int p(y;\eta)(y - a(\eta))^2 + a'(\eta)\underbrace{\int p(y;\eta)(y - a(\eta))\,dy}_{=0} - a''(\eta) = 0$$

$\text{Var}(Y|X;\theta) = a''(\eta)$

(c)

$$-\log p(y;\eta) = -\log b(y) - \eta y + a(\eta)$$

$$NLL = -\log b(y) - \theta^T x y + a(\theta^T x)$$

$f(\theta + h)^T x = \theta^T x + \langle \nabla f(\theta^T x), h \rangle + o(\|h\|)$

$x^T h = \nabla(\theta^T x)^T h$

$\nabla(\theta^T x) = x$

$\nabla_\theta NLL = -xy + a'(\theta^T x)\, x$

$= x(a'(\theta^T x) - y)$

Hessian NLL =

$$\left[\frac{\partial \nabla_\theta NLL}{\partial \theta_1}, \frac{\partial \nabla_\theta NLL}{\partial \theta_2}, \cdots \frac{\partial \nabla_\theta NLL}{\partial \theta_m}\right]$$

$= x\left[a''(\theta^T x)x_1, a''(\theta^T x)x_2, \cdots\right]$

$= a''(\theta^T x)\, x x^T$

h^T Hessian NLL h

$= a''(\theta^T x)\, h^T x\, x^T h$

$= a''(\theta^T x)(h^T x)^2$

$= \text{Var}(Y|X;\theta)(h^T x)^2 \geq 0$

so Hessian NLL is PSD

so it's convex

5.(a)

(i) suppose $M = X\theta - y = \begin{bmatrix} \theta^T x^{(1)} - y^{(1)} \\ \theta^T x^{(2)} - y^{(2)} \\ \vdots \\ \theta^T x^{(m)} - y^{(m)} \end{bmatrix}$

$$M^T W M = \sum_{i=1}^{m} \sum_{j=1}^{m} (\theta^T x^{(i)} - y^{(i)}) w_{ij} (\theta^T x^{(i)} - y^{(i)})$$

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

so $w_{ij} = \begin{cases} \frac{1}{2} w^{(i)} & (i = j) \\ 0 & (i \neq j) \end{cases}$

(ii) $J(\theta) = \theta^T X^T W X \theta - \theta^T X^T W y - y^T W X \theta + y^T W y$

$$\nabla_\theta J(\theta) = \nabla_\theta \operatorname{tr} J(\theta) = \nabla_\theta \operatorname{tr} (\theta^T X^T W X \theta)$$

$$-2\operatorname{tr}(\theta^T X^T W y) = [\nabla_{\theta^T} \operatorname{tr} (\theta^T X^T W X \theta)]^T$$

$$-2 X^T W y = (\theta^T X^T W X + \theta^T X^T W X)^T - 2 X^T W y$$

$$= 2 (X^T W X \theta - X^T W y)$$

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

(iii) $-l(w) = -\log \prod_{i=1}^{m} p(y^{(i)} | x^{(i)}; \theta) = \sum_{i=1}^{m} -\log p(y^{(i)} | x^{(i)}; \theta)$

$$= \sum_{i=1}^{m} \left(\log \frac{1}{\sqrt{2\pi}\sigma^{(i)}} - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2} \right)$$

$$= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma^{(i)}} - \sum_{i=1}^{m} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}$$

to maximum $-l(\theta)$, that is to minimum $\sum_{i=1}^{m} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}$

which is a weighted linear regression problem

$w^{(i)}$ is $\frac{1}{(\sigma^{(i)})^2}$ in this case