1.

(a)

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))$$

$$h_\theta(x^{(i)}) = g(\theta^T x^{(i)}) \quad . \quad g'(z) = g(z)(1-g(z))$$

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

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

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

$$\nabla_\theta J(\theta) = \frac{1}{m} X^T (g(\theta X) - y)$$

$$X = \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix} \quad m \times n$$

$$H_{jk} = \frac{\partial^2 J(\theta)}{\partial \theta_j \partial \theta_k} = \frac{1}{m}\sum_{i=1}^{m} g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)})) x_j^{(i)} x_k^{(i)}$$

$$H = \frac{1}{m} \underset{(n,m)}{X^T} \underbrace{g(X\theta)(1-g(X\theta))}_{(m,m)} \underset{(m,n)}{X}$$

element-wise.?

$$z^T H z = \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{m} g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)})) x_j^{(i)} x_k^{(i)} z_j z_k$$

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

(c) Show GDA can be written as linear classifier

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

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

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

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

$$\theta^T = (\mu_1^T - \mu_0^T)\Sigma^{-1}, \quad \theta = \Sigma^{-1}(\mu_1-\mu_0)$$

$$\theta_0 = \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \log\frac{1-\phi}{\phi}$$

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

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

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

Maximizing l: maximum likelihood estimate.

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

$= \sum_{i=1}^{m} \log \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}})\right) + \sum_{i=1}^{m} \log \phi^{1\{y^{(i)}=1\}} (1-\phi)^{1\{y^{(i)}=0\}}$

$= -\frac{mn}{2} \log 2\pi - \frac{m}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) + \sum_{i=1}^{m} 1\{y^{(i)}=1\} \log\phi + \sum_{i=1}^{m} 1\{y^{(i)}=0\} \log(1-\phi)$

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

set $\frac{\partial l}{\partial \phi} = 0$, $\phi = \frac{1}{m} \sum_{i=1}^{m} 1\{y^{(i)}=1\}$

② $\frac{\partial l}{\partial \mu_0} = \frac{\partial l}{\partial \mu_{y^{(i)}}} \cdot \frac{\partial \mu_{y^{(i)}}}{\partial \mu_0} = m\Sigma^{-1} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}}) \cdot 1\{y^{(i)}=0\}$

$\frac{\partial l}{\partial \mu_0} = \frac{\partial l}{\partial \mu_{y^{(i)}}} \cdot \frac{\partial \mu_{y^{(i)}}}{\partial \mu_0} = m\Sigma^{-1} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}}) \cdot 1\{y^{(i)}=1\}$.

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

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

$\frac{\partial a^T X b}{\partial X} = ab^T$

③ $\frac{\partial l}{\partial \Sigma} = -\frac{m}{2}\Sigma^{-1} + \frac{1}{2} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-2}$

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

(h) take logarithm of x_2. in Dataset 1

or squared root.

Box-Cox transformation.

2.

(a) Given $t^{(i)}$, $y^{(i)}$ and $x^{(i)}$ are conditionally independent.

$p(y^{(i)}=1 \mid t^{(i)}=1, x^{(i)}) = p(y^{(i)}=1 \mid t^{(i)}=1)$

labeled positive data; truth value being positive

which says labeled data are randomly selected the set of positive examples. (All $x^{(i)}$ are positive samples, so $x^{(i)}$ is irrelevent)

结论数据真实值是1，数据有标记的概率与 $x^{(i)}$ 无关.

prove: $p(t^{(i)}=1 \mid x^{(i)}) = p(y^{(i)}=1 \mid x^{(i)})/\alpha$.

$$p(y^{(i)}=1 \mid x^{(i)}) = \sum_{t^{(i)}} p(y^{(i)}=1, t^{(i)} \mid x^{(i)})$$
$$= p(y^{(i)}=1, t^{(i)}=1 \mid x^{(i)}) + p(y^{(i)}=1, t^{(i)}=0 \mid x^{(i)})$$

$p(A \mid B, C)\, p(B \mid C) = p(A, B \mid C)$

$$= p(y^{(i)}=1, t^{(i)}=1 \mid x^{(i)})$$
$$= p(y^{(i)}=1 \mid t^{(i)}=1, x^{(i)})\, p(t^{(i)}=1 \mid x^{(i)})$$
$$= p(y^{(i)}=1 \mid t^{(i)}=1)\, p(t^{(i)}=1 \mid x^{(i)})$$

$\therefore \alpha = p(y^{(i)}=1 \mid t^{(i)}=1)$

(b) $h(x^{(i)}) \approx p(y^{(i)}=1 \mid x^{(i)}) = p(y^{(i)}=1 \mid t^{(i)}=1)\, p(t^{(i)}=1 \mid x^{(i)}) \approx p(y^{(i)}=1 \mid t^{(i)}=1) = \alpha$

so $h(x^{(i)}) \approx \alpha$ for all $x^{(i)} \in V+$

$\frac{1}{\alpha} \frac{1}{1+\exp(-\theta^T x)} = \frac{1}{2}$ what is corrected theta?

$\frac{2}{\alpha} - 1 = \exp(-\theta^T x)$

$\theta^T x + \log(\frac{2}{\alpha} - 1) = 0$

add to θ_0

$\therefore \theta' = \theta + \log(\frac{2}{\alpha} - 1)[1, 0, 0]^T$

3. Poisson Regression.

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

$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$

$\therefore b(y) = \frac{1}{y!}$

$\eta = \ln\lambda$,

$a(\eta) = e^\eta$.

$T(y) = y$

(b) canonical response function: $g(\eta) = E[T(y);\eta]$

$g(\eta) = E[y;\eta] = \lambda = e^\eta$.

(c) In order to construct a GLM.

(i) $y|x;\theta \sim$ ExponentialFamily(η)

(ii) $h(x) = E(y|x)$

(iii) $\eta = \theta^T x$

$$\log p(y^{(i)}|x^{(i)};\theta) = \log b(y^{(i)})\exp(\eta^T T(y^{(i)}) - a(y^{(i)}))$$
$$= \log\left(\frac{1}{y^{(i)}!}\right)\exp((\theta^T x)^T y - e^{\theta^T x})$$
$$= -\log y^{(i)}! + (x^T\theta)y^{(i)} - e^{\theta^T x^{(i)}}$$

$h(x) = e^{\theta^T x}$

$$\frac{\partial \ell(\theta)}{\partial \theta} = x^{(i)}y^{(i)} - x^{(i)}e^{\theta^T x^{(i)}}$$
$$= x^{(i)}(y^{(i)} - h(x^{(i)}))$$

$\therefore \theta := \theta + \alpha(y^{(i)} - h_\theta(x^{(i)}))x^{(i)}$ Stocastic Gradient Ascent Rule

4. Convexity of Generalized Linear Models.

(a) $\frac{\partial}{\partial \eta} p(y;\eta) = \frac{\partial}{\partial \eta} b(y)\exp(\eta y - a(\eta))$

$= b(y)\exp(\eta y - a(\eta))(y - \frac{\partial}{\partial \eta} a(\eta))$

$= p(y;\eta)(y - \frac{\partial}{\partial \eta} a(\eta))$

$= yp(y;\eta) - p(y;\eta)\frac{\partial}{\partial \eta} a(\eta)$

$E[Y|X;\theta] = \int y p(y;\eta)\,dy$

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

$= \frac{\partial}{\partial \eta}\underbrace{\int p(y;\eta)dy}_{1} + \frac{\partial}{\partial \eta} a(\eta)\cdot \underbrace{\int p(y;\eta)dy}_{1}$

$= \frac{\partial}{\partial \eta} a(\eta)$

(b) $Var[Y|X;\theta] = Var[Y;\eta]$

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

$= \int (y^2 - 2y\frac{\partial}{\partial \eta} a(\eta) + \frac{\partial^2}{\partial \eta^2} a(\eta)) p(y;\eta)\,dy$

$= \int y^2 p(y;\eta)dy - \int 2y\frac{\partial}{\partial \eta} a(\eta)\, p(y;\eta)\,dy + \int \frac{\partial^2}{\partial \eta^2} a(\eta) p(y;\eta)dy$

$= \int y^2 p(y;\eta)dy - 2\frac{\partial}{\partial \eta} a(\eta)\underbrace{\int y\, p(y;\eta)dy}_{\frac{\partial}{\partial \eta} a(\eta)} + \frac{\partial^2}{\partial \eta^2} a(\eta)\underbrace{\int p(y;\eta)dy}_{1}$

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

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

$\therefore -\int y^2 p(y;\eta) = \frac{\partial^2}{\partial \eta^2} a(\eta)$

$\therefore Var[Y|X;\theta] = \frac{\partial^2}{\partial \eta^2} a(\eta)$.

(c) loss function of one example

NLL: $l(\theta) = -\log p(y;\eta)$

$= -\log b(y)\exp(\eta y - a(\eta))$

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

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

$\frac{\partial l(\theta)}{\partial \theta_j} = -x_j y + \frac{\partial a(\theta^T x)}{\partial \theta_j}\cdot x_j$

$\nabla_\theta l(\theta) = x\frac{\partial a(\theta^T x)}{\partial \theta} - xy$

$H_{jk} = \frac{\partial l(\theta)}{\partial \theta_j \partial \theta_k} = \frac{\partial^2}{\partial \theta_j \partial \theta_k} a(\theta^T x)\cdot x_j x_k$

$H = x^T x\cdot\frac{\partial^2}{\partial \theta^2} a(\theta^T x)$

Then, prove PSD

$\forall z \in \mathbb{R}^n$, $z^T H z = z^T x^T x\frac{\partial^2}{\partial \theta^2} a(\theta^T x) z$

$= z^T x^T x z\frac{\partial^2}{\partial \theta^2} a(\theta^T x)$

$= (x^T z)^2\frac{\partial^2}{\partial \theta^2}(a\theta^T x)$

$= (x^T z)^2 Var[Y|X;\theta]$

≥ 0

So H of GLM loss function is PSD.

The NLL loss of GLM is convex.

5. Locally Weighted Linear Regression.

(a) (i)

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

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

W : $m \times m$. $\quad w_{ij} = \begin{cases} 0, & i \neq j \\ \frac{1}{2}w^{(i)}, & i = j \end{cases}$.

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

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

$(n,1) \quad (n,m)(m,m)(m,1)$

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

(iii) $l(\theta) = \log p(y|X;\theta)$

$$= \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta)$$

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

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

Maximizing $l(\theta)$ equals to minimizing $J(\theta)$

$w^{(i)} = \frac{1}{(\sigma^{(i)})^2}$