Problem Set 1

Q1 (a) $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\left(\frac{-1}{m}\sum_{i=1}^{m} y^{(i)} \log\left(h_\theta(x^{(i)})\right) + \left(1-y^{(i)}\right)\log\left(1-h_\theta(x^{(i)})\right)\right)$

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

$\frac{\partial^2 J(\theta)}{\partial\theta_i \partial\theta_j} = \frac{-1}{m}\sum_{k=1}^{m}\frac{\partial}{\partial\theta_i}\left(y^{(k)} - g\,h_\theta(x^{(k)})\right)x_j^{(k)}$

$= \frac{-1}{m}\sum_{k=1}^{m}\frac{\partial}{\partial\theta_i}\left(y^{(k)} - g(\theta^T x^{(k)})\right)x_j^{(k)}$

$= \frac{-1}{m}\sum_{k=1}^{m} -g'(\theta^T x^{(k)})\frac{\partial}{\partial\theta_i}(\theta^T x^{(k)})\,x_j^{(k)}$

$= \frac{-1}{m}\sum_{k=1}^{m} -g(\theta^T x^{(k)})\left(1-g(\theta^T x^{(k)})\right)x_i^{(k)}\,x_j^{(k)}$

$= \frac{1}{m}\sum_{k=1}^{m} g(\theta^T x^{(k)})\left(1-g(\theta^T x^{(k)})\right)x_i^{(k)}\,x_j^{(k)}$

$(z^T H z) = \sum_i \sum_j z_i \frac{1}{m}\sum_{k=1}^{m} g(\theta^T x^{(k)})\left(1-g(\theta^T x^{(k)})\right)x_i^{(k)}\,x_j^{(k)}\,z_j$

$= \frac{1}{m}\sum_{k=1}^{m} g(\theta^T x^{(k)})\left(1-g(\theta^T x^{(k)})\right)\sum_i\sum_j z_i\,x_i^{(k)}\,x_j^{(k)}\,z_j$

$= \frac{1}{m}\sum_{k=1}^{m} \underbrace{g(\theta^T x^{(k)})}_{\in(0,1)}\underbrace{\left(1-g(\theta^T x^{(k)})\right)}_{\in(0,1)}(x^T z)^2 \geq 0$

$z^T H z = 0$ iff $x^T z = 0$

(b) Coding problem.

Requirements → write Hessian $H = \left[\frac{\partial^2 J(\theta)}{\partial\theta_i\partial\theta_j}\right]_{\substack{i=1,2,3\\ j=1,2,3}}$ in form of matrix multiplication

Define a diagonal matrix.

$$diag = [d_{ij}]_{i=1,2,\cdots,k}$$

$$= \begin{bmatrix} g(\theta^T x^{(1)})(1-g(\theta^T x^{(1)})) & & \\ & \ddots & \\ & & g(\theta^T x^{(k)})(1-g(\theta^T x^{(k)})) \end{bmatrix}_{k \times k}$$

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} \end{bmatrix}^T$$

Thus $H = X^T \, diag \, X$ (can use matmul)

Similarly,

$$J = +\frac{1}{m} \left(X^T (diag - y) \right)$$

$$\frac{1}{1+e^{-z}} \qquad g(\theta^T x) \qquad |\theta_k - \theta| < \epsilon$$

$$\text{loss at } n^{th} \text{ iteration} = -\frac{1}{m} \left(\sum_{i=1}^{m} y^{(i)} \log \left[g(\theta^T x^{(i)}) \right] + (1-y^{(i)}) \log \left[1 - g(\theta^T x^{(i)}) \right] \right)$$

Step 1 θ^0 is updated to $\theta^{(1)}$

Step k $\theta^{(k-1)}$ is updated to θ^k at $\rightarrow$ loss $= \frac{1}{m} \sum_{i=1}^{m} y^{(i)}$

(c). Show that $p(y=1 \mid x; \phi, \mu_0, \mu_1, \Sigma) = \frac{1}{1+\exp(-(\theta^T x + \theta_0))}$ (since done separately don't add intercept to x)

where $\theta \in \mathbb{R}^n$ & $\theta_0 \in \mathbb{R}$ are appropriate functions of ϕ, Σ, μ_0 and μ_1

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

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

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

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

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

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

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

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

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

Thus $\theta_0 = -\left[(\mu_0 - \mu_1)^T\Sigma^{-1}\right]^T = -\Sigma^{-1}(\mu_0 - \mu_1) = \Sigma^{-1}(\mu_1 - \mu_0)$

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

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

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

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

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

$$\ell(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^{m} \phi^{1\{y^{(i)}=1\}} \exp -\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu$$

$$p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma)\, p(y^{(i)}; \phi) = \begin{cases} \frac{\phi}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp -\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) & \text{when } y^{(i)} = 1 \\ \frac{(1-\phi)}{(2\pi)^{n/2}} \exp -\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) & \text{when } y^{(i)} = 0 \end{cases}$$

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

$$= \sum_{i=1}^{m} \frac{1}{p(y^{(i)};\phi)} \underbrace{\frac{\partial}{\partial \phi} p(y^{(i)}; \phi)}$$

$$\begin{cases} 1 & \text{when } y^{(i)} = 1 \\ -1 & \text{when } y^{(i)} = 0 \end{cases}$$

$$= \frac{m_1}{\phi} - \frac{m_2}{1-\phi}$$

where $m_1 + m_2 = m$ and m_1 are # of 1's & m_2 are no. of 0's

$$\frac{m_1}{\phi} - \frac{m_2}{1-\phi} = 0 \Rightarrow \frac{m_1}{\phi} = \frac{m_2}{1-\phi} \Rightarrow \frac{1-\phi}{\phi} = \frac{m_2}{m_1}$$

$$\Rightarrow \frac{1}{\phi} - 1 = \frac{m_2}{m_1} \Rightarrow \frac{1}{\phi} = 1 + \frac{m_2}{m_1} = \frac{m_1+m_2}{m_1}$$

$$\Rightarrow \hat{\phi} = \frac{m_1}{m_1+m_2} = \frac{m_1}{m} = \frac{1}{m}\sum_{i=1}^{n} 1\{y^{(i)}=1\}.$$

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

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

$$= \sum_{i=1}^{m} \begin{cases} \sqrt{2\pi}\sigma \exp \frac{1}{2\sigma^2}(x^{(i)}-\mu_1)^2 & y^{(i)}=1 \\ \sqrt{2\pi}\sigma \exp \frac{1}{2\sigma^2}(x^{(i)}-\mu_0)^2 & y^{(i)}=0 \end{cases} \begin{cases} 0 & \text{when } y^{(i)}=1 \\ \frac{+2(x^{(i)}-\mu_0)}{2\sigma^2} \exp \frac{1}{2\sigma^2}(x^{(i)}-\mu_0)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} & \text{when } y^{(i)}=0. \end{cases}$$

$$= \sum_{i=1}^{m} 1\{y^{(i)}=0\} \frac{\sqrt{2\pi}\sigma}{\exp \frac{1}{2\sigma^2}(x^{(i)}-\mu_0)^2} \times \frac{(x^{(i)}-\mu_0)}{\sqrt{2\pi}\sigma^3} \exp \frac{1}{2\sigma^2}(x^{(i)}-\mu_0)^2$$

$$= \sum_{i=1}^{m} 1\{y^{(i)}=0\} \frac{(x^{(i)}-\mu_0)}{\sigma^2}$$

$$\sum_{i=1}^{m} 1\{y^{(i)}=0\} \frac{(x^{(i)}-\mu_0)}{\sigma^2} = 0$$

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

$$\Rightarrow \hat{\mu}_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=0\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=0\}}$$

Similarly by symmetry, $\hat{\mu}_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=0\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=0\}}$

$$\frac{\partial \ell(\phi, \mu_0, \mu_1, \Sigma)}{\partial \sigma} \quad (\sigma^2 = \Sigma)$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial \sigma} \left[\log p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \sigma) + \log p(y^{(i)}; \phi) \right]$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial \sigma} \left[\log p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \sigma) + 0 \right]$$

$$= \sum_{i=1}^{m} \frac{1}{p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \sigma)} \frac{\partial}{\partial \sigma} p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \sigma)$$

$$= \sum_{i=1}^{m} \frac{(2\pi)^{1/2} \sigma}{\exp -\frac{1}{2\sigma^2}(x - \mu_0)^2} \left[\frac{1}{\sqrt{2\pi}\sigma} - \frac{(x^{(i)} - \mu_0)^2 (-2)}{2\sigma^3} \exp -\frac{1}{2\sigma^2}(x^{(i)} - \mu_0)^2 \right] 1\{y^{(i)} = 0\}$$

$$+ 1\{y^{(i)} = 1\} \frac{1}{\sqrt{2\pi}\sigma^2} \exp -\frac{1}{2\sigma^2}(x^{(i)} - \mu_0)^2 \Big] \cdots$$

$$= \sum_{i=1}^{m} \left(1\{y^{(i)} = 0\} \left[\frac{(x^{(i)} - \mu_0)^2}{\sigma^3} - \frac{1}{\sigma} \right] + 1\{y^{(i)} = 1\} \left[\frac{(x^{(i)} - \mu_1)^2}{\sigma^3} - \frac{1}{\sigma} \right] \right)$$

$$= \left(\frac{(x^{(i)} - \mu_0)^2}{\sigma^3} - \frac{1}{\sigma} \right) m_1 + \left(\frac{(x^{(i)} - \mu_1)^2}{\sigma^3} - \frac{1}{\sigma} \right) m_2$$

$$= \frac{1}{\sigma^3} \left(m_1 (x - \mu_0)^2 + m_2 (x - \mu_1)^2 \right) + \frac{m}{\sigma}$$

$$= \sum_{i=1}^{m} -\frac{1}{\sigma} + \sum_{i=1}^{m} \left(1\{y^{(i)} = 0\} \frac{(x^{(i)} - \mu_0)^2}{\sigma^3} + 1\{y^{(i)} = 1\} \frac{(x^{(i)} - \mu_1)^2}{\sigma^3} \right)$$

$\frac{\partial}{\partial \sigma}(ab)$

Equating to 0.

$$\frac{1}{\sigma}\left(\frac{m_1(x^{(i)}-\mu_0)^2 + m_2(x^{(i)}-\mu_1)^2}{\sigma^2} - m\right) = 0.$$

$$\Rightarrow \frac{m_1(x^{(i)}-\mu_0)^2 + m_2(x^{(i)}-\mu_1)^2}{m} = \hat{\sigma}^2$$

$$\Rightarrow \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \mu_{y^{(i)}}\right)\left(x^{(i)} - \mu_{y^{(i)}}\right)^T = \hat{\sigma}^2$$

Second derivative tests to show that indeed the critical pts are pts of maximas.

$$\frac{\partial^2 \ell(\phi, \mu_0, \mu_1, \Sigma)}{\partial \phi^2} = \frac{-m_1}{\phi_1^2} - \frac{m_2}{(1-\phi_1)^2} < 0.$$

$$\frac{\partial^2 \ell(\phi, \mu_0, \mu_1, \Sigma)}{\partial \mu_0^2} = \frac{\partial}{\partial \mu_0}\sum_{i=1}^{m}\left[1\{y^{(i)}=0\}\frac{(x^{(i)}-\mu_0)}{\sigma^2}\right]$$

$$= \sum_{i=1}^{m} 1\{y^{(i)}=0\}\left(\frac{-1}{\sigma^2}\right) < 0.$$

Similary by symmetry

$$\frac{\partial^2 \ell}{\partial \mu_1^2} < 0$$

$$\frac{\partial^2 \ell}{\partial \sigma^2} = \frac{\partial}{\partial \sigma}\left(\left(\frac{1}{\sigma^3}\right)\left(m_1(x-\mu_0)^2 + m_2(x-\mu_1)^2\right) - \frac{m}{\sigma}\right).$$

$$= \frac{-3}{\sigma^4}\left(m_1(x-\mu_0)^2 + m_2(x-\mu_1)^2\right) + \frac{m}{\sigma^2}$$

$$= \frac{1}{\sigma^2}\left(m - \frac{3m_1(x-\mu_0)^2 + 3m_2(x-\mu_1)^2}{\sigma^2}\right)$$

$$\begin{bmatrix} \mu_1^{(1)} & \mu_2^{(1)} & \mu_3^{(1)} \\ \mu_1^{(2)} & \mu_2^{(2)} & \mu_3^{(2)} \\ \vdots & \vdots & \vdots \\ \mu_1^{(k)} & \mu_2^{(k)} & \mu_3^{(k)} \end{bmatrix}$$

Note → In python & numpy
let m = np.array([1,2,3]) then m.T is same as m. Both are (3,) ~~vectors~~. lists.
Thus to use transpose. first convert to (m,n) form. If 3,1 needed then use:
m = np.array([1,2,3], ndims=2).T

(f) run script .../linearclass/gf.py →
logreg theta estimate → (1) [-2.4078467 1.0343113 0.24463071]
(2) [-2.0967324 0.8663494 0.828187]
gda theta estimate → (1) [-2.10437509 1.59905448 0.01081916]
→ (2) [-2.11390101 0.85879252 0.84927843]

(g) In ~~case~~ dataset 1, logistic regression produces better results than GDA because the distribution is not gaussian while in 2 it is much nearer to gaussian distribution.

(h) Transform is box-cox transformation. Python scipy.stat has a boxcox function. ~~But it can't have~~ but it requires positive values as input. I tried subtracting minimum value & then add a small value to ~~non~~ zero entries but it did not work. Giving some error.

(b) Box-cox transformation.

Q2 (a) $p(t^{(i)}=1 \mid x^{(i)}) = \frac{p(t^{(i)}=1, x^{(i)})}{p(x^{(i)})}$

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

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

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

$\Rightarrow p(t^{(i)}=1 \mid x^{(i)}) \left[1 - p(y^{(i)}=0 \mid t^{(i)}=1)\right] = p(y^{(i)}=1 \mid x^{(i)})$

$\Rightarrow p(t^{(i)}=1 \mid x^{(i)}) = \frac{p(y^{(i)}=1 \mid x^{(i)})}{1 - p(y^{(i)}=0 \mid t^{(i)}=1)}$

$\Rightarrow p(t^{(i)}=1 \mid x^{(i)}) = \frac{p(y^{(i)}=1 \mid x^{(i)})}{p(y^{(i)}=1 \mid t^{(i)}=1)} \times \alpha$

(b) $V+ = \{x^{(i)} \in V \mid y^{(i)}=1\}$.

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

$\Rightarrow p(y^{(i)}=1 \mid x^{(i)}) \approx h(x^{(i)})$

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

for all $x^{(i)} \in V+ \Rightarrow y^{(i)}=1 \Rightarrow t^{(i)}=1 \Rightarrow p(t^{(i)}=1 \mid x^{(i)})=1$

$\Rightarrow \alpha \approx h(x^{(i)}) \quad \forall x^{(i)} \in V+$

(c) ✓

(d) ✓

(e) ✓

Q3 (a) $p(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$

Exponential family

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

$$\frac{e^{-\lambda}\lambda^y}{y!} = \frac{e^{-\lambda} e^{\log_e \lambda^y}}{y!}$$

$$= \frac{e^{-\lambda} e^{y\log_e\lambda}}{y!}$$

$$= \frac{1}{y!}\exp(-\lambda + y\log_e\lambda)$$

$$= \frac{1}{y!}\exp((\log_e\lambda)\,y - \lambda)$$

$\eta = \log_e\lambda$

$\Rightarrow e^{\eta} = \lambda$

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

$\eta = \log_e\lambda$

$T(y) = y$

$a(\eta) = \lambda = e^{\eta}$

(b) 1) $y|x;\theta \sim \text{Poisson}(\lambda)$

2) $h_\theta(x) = E[y|x;\theta]$

$= \lambda$

$= e^{\eta}$

$= e^{\theta^T x}$

$g(\eta) = E[y;\eta] = e^{\eta}$

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

$= \log \frac{1}{y^{(i)}!} + \left[\eta^T y^{(i)} - e^{\eta}\right]$

$= \log \frac{1}{y^{(i)}!} + \left[(\theta^T x^{(i)})^T y^{(i)} - e^{\theta^T x^{(i)}}\right]$

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

$= \frac{\partial}{\partial \theta_j}\left[(\theta_0 + \theta_1 x_1^{(i)} + \cdots + \theta_d x_d^{(i)}) y^{(i)} - e^{\theta_0 + \theta_1 x_1^{(i)} + \cdots + \theta_d x_d^{(i)}}\right]$

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

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

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

Thus stochastic gradient ascent rule becomes

$$\theta_j := \theta_j + \alpha \frac{\partial \ell(\theta)}{\partial \theta_j}$$

ϕ reduces to

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

(d) ✓

Q4 Convexity of GLM

$$p(y;\eta) = b(y)\exp(\eta y - \alpha(\eta))$$

(a) $\frac{\partial}{\partial \eta} p(y;\eta) = \frac{\partial}{\partial \eta}\left[b(y)\exp(\eta y - \alpha(\eta))\right]$

$= b(y)\left[y - \frac{\partial}{\partial \eta}\alpha(\eta)\right]\exp(\eta y - \alpha(\eta))$

$= p(y;\eta)\left[y - \frac{\partial}{\partial \eta}\alpha(\eta)\right]$

$$\int \frac{\partial}{\partial \eta} p(y;\eta)\,dy = \int p(y;\eta)\left[y - \frac{\partial}{\partial \eta}\alpha(\eta)\right]dy$$

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

$$\Rightarrow \frac{\partial}{\partial \eta}\int p(y;\eta)\,dy = E[y;\eta] - \frac{\partial}{\partial \eta}\alpha(\eta)\int p(y;\eta)\,dy$$

$$\Rightarrow E[y;\eta] = \frac{\partial}{\partial \eta}\alpha(\eta)$$

(b) Show that $\mathrm{Var}(Y|x;\theta) = \frac{\partial^2}{\partial \eta^2}\alpha(\eta)$

$$\frac{\partial}{\partial \eta} p(y;\eta) = p(y;\eta)\left[y - \frac{\partial}{\partial \eta}\alpha(\eta)\right]$$

$$\frac{\partial^2}{\partial \eta^2} p(y;\eta) = \left[\frac{\partial}{\partial \eta} p(y;\eta)\right]\left[y - \frac{\partial}{\partial \eta}\alpha(\eta)\right] + p(y;\eta)\left[-\frac{\partial^2}{\partial \eta^2}\alpha(\eta)\right]$$

$$\frac{\partial^2}{\partial \eta^2} p(y;\eta) = p(y;\eta)\left[y - \frac{\partial}{\partial \eta}\alpha(\eta)\right]^2 + p(y;\eta)\left[-\frac{\partial^2}{\partial \eta^2}\alpha(\eta)\right]$$

$$\frac{\partial^2}{\partial \eta^2} p(y;\eta) = y^2 p(y;\eta) + \left[\frac{\partial}{\partial \eta}\alpha(\eta)\right]^2 p(y;\eta) - 2y \frac{\partial}{\partial \eta}\alpha(\eta)\, p(y;\eta) + p(y;\eta)\left[\frac{-\partial^2}{\partial \eta^2}\alpha(\eta)\right]$$

$$\int \frac{\partial^2}{\partial \eta^2} p(y;\eta)\, dy = \int y^2 p(y;\eta)\, dy + \left[\frac{\partial}{\partial \eta}\alpha(\eta)\right]^2 \int p(y;\eta)\, dy - \frac{\partial}{\partial \eta}\alpha(\eta) \int 2y\, p(y;\eta)\, dy - \frac{\partial^2}{\partial \eta^2}\alpha(\eta) \int p(y;\eta)\, dy$$

$$\Rightarrow \int \frac{\partial^2}{\partial \eta^2} p(y;\eta)\, dy = \int y^2 p(y;\eta)\, dy + \left[\frac{\partial}{\partial \eta}\alpha(\eta)\right]^2 - \frac{\partial}{\partial \eta}\alpha(\eta)\, 2\, \frac{\partial}{\partial \eta}\alpha(\eta) - \frac{\partial^2}{\partial \eta^2}\alpha(\eta)$$

$$\Rightarrow 0 = \int y^2 p(y;\eta)\, dy - \left[\frac{\partial}{\partial \eta}\alpha(\eta)\right]^2 - \frac{\partial^2}{\partial \eta^2}\alpha(\eta)$$

$$\Rightarrow \left[\frac{\partial}{\partial \eta}\alpha(\eta)\right]^2 =$$

$$\Rightarrow \frac{\partial^2}{\partial \eta^2}\alpha(\eta) = \int y^2 p(y;\eta)\, dy - [E(y;\eta)]^2 = \text{var}(y|x;\theta)$$

(c) $l(\theta)$ = ~~exp~~ - log-likelihood.

$$= -\log\left[b(y^{(1)}) \exp(\eta y^{(1)} - \alpha(\eta)) \ldots b(y^{(n)}) \exp(\eta y^{(n)} - \alpha(\eta))\right]$$

$$= -\log\left[b(y^{(1)}) \ldots b(y^{(n)}) \exp(\eta y^{(1)} - \alpha(\eta)) \ldots \exp(\eta y^{(n)} - \alpha(\eta))\right]$$

=

$$\nabla_\theta(\theta^T x) = \left[\frac{\partial}{\partial \theta_j} \sum_{i=1}^{d+1} \theta_i x_i\right]_{j=1,2,\ldots,d+1}$$

$$= \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{d+1} \end{bmatrix} = x.$$

$$\nabla_\theta(x^T\theta) = \left[\frac{\partial}{\partial \theta_j}\right] \nabla_\theta(\theta^T x) = x.$$

$$l(\theta) = -\log p(y;\eta)$$
$$= -\log[b(y)\exp(\eta y - a(\eta))]$$
$$= a(\eta) - \eta(y) - \log b(y).$$

$$\nabla_\theta l(\theta) = \frac{\partial}{\partial \eta} a(\eta)\, \nabla_\theta \eta - y \nabla_\theta \eta$$
$$= \frac{\partial}{\partial \eta} a(\eta)\, x - yx.$$

$$\nabla^2_\theta l(\theta) = \frac{\partial}{\partial \eta}\left[\frac{\partial}{\partial \eta} a(\eta)\, x - yx\right]$$
$$= \frac{\partial^2}{\partial \eta^2} a(\eta) \nabla_\theta \eta + \frac{\partial}{\partial \eta} a(\eta) \frac{\partial x}{\partial \eta} = x \frac{\partial^2}{\partial \eta^2} a(\eta) \nabla_\theta \eta$$
$$= x \frac{\partial^2}{\partial \eta^2} a(\eta) \nabla_\theta \eta \cdot \qquad = \frac{\partial^2}{\partial \eta^2} a(\eta)\, x x^T$$
$$= x \frac{\partial}{\partial \eta} \text{var}(Y;\eta) \qquad = \frac{\partial}{\partial \eta} \text{var}(Y;\eta)\, x x^T$$

Another way

$$\ell(\theta) = -\left[\log \prod_{i=1}^{n} b(y^{(i)}) + \sum_{i=1}^{n} (\theta^T x^{(i)}) y^{(i)} - \alpha(\theta^T x^{(i)})\right]$$

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = -\sum_{i=1}^{n}\left[x_j^{(i)} y^{(i)} - \frac{\partial \alpha}{\partial \eta}(x_j^{(i)})\right]$$

$$\frac{\partial^2 \ell(\theta)}{\partial \theta_k \partial \theta_j} = \sum_{i=1}^{n} x_j^{(i)} \frac{\partial^2 \alpha}{\partial \eta^2} x_k^{(i)} = \text{Var}(Y|x;\eta) \sum_{i=1}^{n} x_j^{(i)} x_k^{(i)}$$

$k = 1, 2, \dots, d+1$, $j = 1, 2, \dots, d+1$

~~$= \text{Var}(Y|x)$~~

here $x = [x_i^{(j)}] = [a_{ij}]$

$$\Rightarrow H(\theta) = \text{Var}(Y; \eta)\, x x^T$$

$n \times d+1$ $\quad d+1 \times n$

Since for $z \neq 0$ +ve.

$z^T x x^T z = (x^T z)^T x^T z = (x^T z)^2 \geq 0$

Thus $H(\theta)$ is Positive semi definite. Equals zero when $x^T z = 0$ or $\text{Var}(Y;\eta) = 0$. When $\text{Var}(Y;\eta) \neq 0$ then $x^T z = 0$. Thus it might happen that the ~~matrix is positive~~ value of $x^T z = 0$ even when $z \neq 0$. Thus the function is convex (not strictly ie can have a flat region)

d+1

$\frac{1}{2}$ (...) .

Q5 in ~~the~~ new version of the tutorial is on linear regression..

Q5 (a) $J(\theta) = \frac{1}{2} \sum_{i=1}^{N} (\theta^T \hat{x}^{(i)} - y^{(i)})^2$

$\nabla_\theta J(\theta) = \frac{1}{2} \sum_{i=1}^{N} 2(\theta^T \hat{x}^{(i)} - y^{(i)}) \hat{x}_j^{(i)}$

Update Rule :

$\theta_j := \theta_j - \lambda \sum_{i=1}^{N} (\theta^T \hat{x}^{(i)} - y^{(i)}) (\hat{x}_j^{(i)})$

#

(b) Normal eq^n.

$\nabla_\theta J(\theta) = \hat{X}^T \hat{X} \theta - \hat{X}^T y = 0$

$\Rightarrow \hat{X}^T \hat{X} \theta = \hat{X}^T y$

(c) ~~As k increases fit~~ higher degree polynomials tend to fit data better, ~~high~~ very high degree polynomials can ~~cause overfitting~~. be numerically unstable

(d) On the presence of sin(x) feature, the models seem to fit the data better/more robustly. However the numerical instability with high degree polynomials remain.

(e) ~~High~~ When dataset is small, higher degree polynomials tend to pass through all the points, but qualitatively seem like a poor fit. Numerical instability remains.