CS229 Fall 2018 Problem Set #1

Supervised learning

1) * $x \in \mathbb{R}^n$ and $\theta \in \mathbb{R}^n$

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

* $h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$

Problem: Find the Hessian H of $J(\theta)$ and show that $\forall z \in \mathbb{R}^n$, $z^T H z \geq 0$

* $\frac{\partial g(\theta^T x)}{\partial \theta_k} = \frac{\partial h_\theta(x)}{\partial \theta_k}$ where $k \in \{0, \dots, n\}$

$$\Rightarrow \frac{\partial}{\partial \theta_k}\left(\frac{1}{1+e^{-\theta^T x}}\right) = \frac{-(-x_k)\, e^{-\theta^T x}}{(1+e^{-\theta^T x})^2} =$$

$$= \frac{x_k\, e^{-\theta^T x}}{(1+e^{-\theta^T x})^2} = x_k \left(\frac{1}{1+e^{-\theta^T x}}\right)\left(\frac{e^{-\theta^T x}}{1+e^{-\theta^T x}}\right)$$

$$= x_k\, h_\theta(x)\,(1-h_\theta(x))$$

$$\Rightarrow \boxed{\frac{\partial h_\theta(x)}{\partial \theta_k} = x_k h_\theta(x)(1-h_\theta(x))}$$

* $\nabla J(\theta) = \begin{pmatrix} \frac{\partial J(\theta)}{\partial \theta_0} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta_n} \end{pmatrix}$

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

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

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

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

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

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

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

$$\Rightarrow \nabla_\theta^2 J(\theta) = \frac{1}{m} \nabla_\theta \left(X^T g(X\theta) \right)$$

$$\frac{\partial^2 J(\theta)}{\partial \theta_k \partial \theta_l} = \frac{1}{m} \sum_{i=1}^{m} x_k^{(i)} \cdot \frac{\partial h_\theta(x^{(i)})}{\partial \theta_l}$$

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

$$\Rightarrow (HZ)_k = \sum_{l=1}^{n} \frac{1}{m} \sum_{i=1}^{m} x_k^{(i)} x_l^{(i)} z_l \left(h_\theta(x^{(i)}) \right) \left[1 - h_\theta(x^{(i)}) \right]$$

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

$$\Rightarrow z^T H z = \frac{\left(x^{(i)T} z \right)^2}{m} \sum_{i=1}^{m} h_\theta(x^{(i)}) \left(1 - h_\theta(x^{(i)}) \right)$$

≥ 0 Cause $(x^T z)^2 \geq 0$ and

$h_\theta(x^{(i)}) \geq 0$ and

$(1 - h_\theta x^{(i)}) \geq 0$

C) Gaussian Discriminant Analysis:

$$* p(y) = \begin{cases} \phi & \text{if } y = 1 \\ 1-\phi & \text{if } y = 0 \end{cases}$$

$$* p(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} * \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right.$$

$$* p(x|y=1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} * \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu\right.$$

where ϕ, μ_0, μ_1, and Σ are parameters of our model

Problem: To show that GDA results in a Classifier that has a linear boundary, show that posterior distribution can be written as:

$$p(y=1|x;\phi,\mu_0,\mu_1,\Sigma) = \frac{1}{1+\exp(-(\theta^T x+\theta$$

According to the Bayes Theorem:

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

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

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

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

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

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

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

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

$$\Rightarrow \begin{cases} \theta_0 = \ln\left(\frac{\phi}{1-\phi}\right) - \frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0) \\ \theta = -(\mu_0-\mu_1)^T\Sigma^{-1} \end{cases}$$

d) * $p(y) = \begin{cases} \phi & \text{if } y = 1 \\ 1-\phi & \text{if } y = 0 \end{cases}$

* $p(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)\right)$

* $p(x|y=1) \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right)$

Problem:

* Let's assume $n = 1 \Rightarrow \Sigma = \sigma^2 = |\Sigma|$

By maximizing

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

Prove that the maximum likelihood estimates of the parameters are

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

$$\mu_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}}, \quad \mu_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$

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

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

$$= \sum_{i=1}^{m} \log\left(p(x^{(i)} | y^{(i)}=1; \mu_0, \mu_1, \Sigma)\right) + \log\left(p(x^{(i)} | y^{(i)}=0; \mu_0, \mu_1, \Sigma)\right) + \log\left(p(y^{(i)}; \phi)\right)$$

$$= \sum_{i=1}^{m} \log(c) \left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1} (x-\mu_0)\right) 1\{y^{(i)}=0\} +$$

$$\left(-\frac{1}{2}(x^{(i)}-\mu_1)^T \Sigma^{-1} (x^{(i)}-\mu_1)\right) 1\{y^{(i)}=1\} \log(c) + \log\left(\phi^{y^{(i)}} (1-\phi)^{(1-y^{(i)})}\right)$$

where $c = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}}$

$(n=1) \Rightarrow \frac{\partial l(\phi, \mu_0, \mu_1, \Sigma)}{\partial \mu_0} = \frac{\partial}{\partial \mu_0} \left(\sum_{i=1}^{m} \log(c) \left(-\frac{1}{2}(x-\mu_0)^2\right) \right.$

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

$\Rightarrow \frac{\partial l(\phi, \mu_0, \mu_1, \Sigma)}{\partial \mu_0} = 0$ $\left(\forall\, x^{(i)} \in \mathbb{R} \right.$

$-\frac{1}{2}(x-\mu_0)^2$ is a conc

function with a

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

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

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

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

* Let's now prove for Σ

$$l(\phi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{m} \log \left(\frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1} (x-\mu_i)\right) \right) + \log\left(\phi^{y^{(i)}} (1-\phi)^{1-y^{(i)}}\right)$$

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

$$\Rightarrow \frac{\partial l}{\partial \Sigma} = \sum_{i=1}^{m} \frac{-1}{2|\Sigma|} \cdot |\Sigma| \cdot \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} (x^{(i)} - \mu_i)(x^{(i)} - \mu_i)^T \Sigma^{-1}$$

$$\Rightarrow \frac{\partial l}{\partial \Sigma} = 0 \iff \sum_{i=1}^{m} \Sigma^{-1} + \Sigma^{-1} (x^{(i)} - \mu_i)(x^{(i)} - \mu_i)^T \Sigma^{-1} = 0$$

$$\Rightarrow \sum_{i=1}^{m} \Sigma^{-1} (x^{(i)} - \mu_i)(x^{(i)} - \mu_i)^T \Sigma^{-1} = m \Sigma^{-1}$$

$$\Rightarrow \sum_{i=1}^{m} (x^{(i)} - \mu_i)(x^{(i)} - \mu_i)^T \Sigma^{-1} = m I_{n \times n}$$

$$\Rightarrow \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu_i)(x^{(i)} - \mu_i)^T = \Sigma$$

Let's now solve for ϕ

$$\frac{\partial l(\phi, \mu_0, \mu_1, \Sigma)}{\partial \phi} = \frac{\partial}{\partial \phi} \left(\sum_{i=1}^{m} y^{(i)} \log(\phi) + (1 - y^{(i)}) \log(1 - \phi \right.$$

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

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

$$\Rightarrow 0 = \sum_{i=1}^{m} \frac{y^{(i)} - y^{(i)}\phi + y^{(i)}\phi - \phi}{\phi(1-\phi)}$$

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

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

g) The gaussian discriminant analysis was worse than the logistic regression on the dataset1.

This could be due to the fact that dataset 1 is not gaussian. ($p(x|y)$ is not gaussian).

If a dataset ($p(x|y)$) is gaussian then GDA would outperform logistic regression. The smaller the dataset, greater will be the difference in performance.

h) We should find a transformation that transform $p(x|y)$ into gaussian.

2a) Incomplete, Positive only Labels

* $P(t^{(i)}=1 \mid y^{(i)}=1) = 1$

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

Problem: Prove that $P(t^{(i)}=1 \mid x^{(i)}) = \frac{P(y^{(i)}=1 \mid x^{(i)})}{\alpha}$

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

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

$\Rightarrow P(t^{(i)}=1 \mid x^{(i)}) = \frac{P(y^{(i)}=1 \mid x^{(i)})}{P(y^{(i)}=1 \mid t^{(i)}=1, x^{(i)})}$

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

$\Rightarrow \alpha = P(y^{(i)}=1 \mid t^{(i)}=1)$

b) * $\alpha = P(y^{(i)}=1 \mid t^{(i)}=1)$

* $P(y^{(i)}=1 \mid x^{(i)}) \cdot \frac{1}{\alpha} = P(t^{(i)}=1 \mid x^{(i)})$

* $P(t^{(i)}=1 \mid x^{(i)}) \approx 1, \forall x^{(i)} \in V_+$

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

* $h(x^{(i)}) \approx \alpha, \forall x^{(i)} \in V_+$

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

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

$\Rightarrow \frac{P(y^{(i)}=1 \mid x^{(i)})}{\alpha} \approx 1 \quad x^{(i)} \in V_+$

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

3 Poisson Regression

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

Problem: Prove $p(y;\lambda) \in$ exponential family

Form of Exponential families: $p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$

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

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

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

b) * $E[y] = \lambda$, $(y \sim Poi(\lambda))$

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

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

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

* log-likelihood of an example $= \log(p(y^{(i)} | x^{(i)}; \theta))$

Problem: Find the stochastic gradient ascent update rule.

$$\Rightarrow \log(p(y^{(i)} | x^{(i)}; \theta)) = \log\left(\frac{e^{-e^{\theta^T x^{(i)}}} \cdot e^{(\theta^T x^{(i)}) y^{(i)}}}{y^{(i)}!}\right)$$

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

$$\Rightarrow \frac{\partial \log(p(y^{(i)} | x^{(i)}; \theta))}{\partial x_j^{(i)}} = -x_j^{(i)} e^{\theta^T x^{(i)}} + x_j^{(i)} y^{(i)}$$

$\Rightarrow$ Stochastic gradient ascent update rule : $\theta_j := \theta_j + \alpha (y^{(i)} - e^{\theta^T x^{(i)}}) x_j^{(i)}$

(α = learning rate)

4 Convexity of GLMs

* let's assume $\eta \in \mathbb{R}$ and $T(y) = y$

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

Show that: $E[Y|X;\theta] = \frac{\partial a(\eta)}{\partial \eta}$

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

$$\Rightarrow \frac{\partial}{\partial \eta}(1) = \int \left(y - \frac{\partial a(\eta)}{\partial \eta}\right) p(y;\eta)\,dy$$

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

$$\Rightarrow \frac{\partial a(\eta)}{\partial \eta} = E(Y;\eta)$$

$$= E(Y|X;\theta)$$

b) * $E(y|x;\theta) = \frac{\partial a(\eta)}{\partial \eta}$

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

Prove that: $Var(y|x;\theta) = \frac{\partial^2 a(\eta)}{\partial \eta^2} = \partial E(y|x;\theta)$

* $Var(y;\eta) = \int (y - E(y;\eta))^2 p(y;\eta)\, dy$

$$\frac{\partial^2 a(\eta)}{\partial \eta^2} = \frac{\partial}{\partial \eta}\left(E(y;\eta)\right)$$

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

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

$$= \int (y - E(y;\eta))^2 p(y;\eta)\, dy + \int y\, E(x;\eta) p(y;\eta) dy$$

$$- \int E^2(y;\eta) p(y;\eta) dy$$

$$= \int p(y;\eta)(y - E(y;\eta))\,dy + E^2(y;\eta) - E^2(y;\eta)$$

$$= \mathrm{Var}(y;\eta)$$

c) $$l(\theta) = -\log\left(\prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta)\right)$$

$$= -\sum_{i=1}^{m} \log(b(y^{(i)})) + \sum_{i=1}^{m}(\eta^T T(y^{(i)}) - a(\eta))$$

$$= -\sum_{i=1}^{m} \log(b(y^{(i)})) - \sum_{i=1}^{m}(\theta^T x^{(i)})^T T(y^{(i)}) - a(\eta))$$

$$\Rightarrow \frac{\partial l(\theta)}{\partial \theta_j} = -\sum_{i=1}^{m} x_j^{(i)} T(y^{(i)})_j - x_j^{(i)} \frac{\partial a(\eta)}{\partial \eta}$$

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

$z \in \mathbb{R}^n$

$$\Rightarrow Hz = \begin{bmatrix} \sum_{i=1}^{m} \sum_{j=1}^{n} x_1^{(i)} x_j^{(i)} \frac{\partial^2 a(\eta)}{\partial \eta^2} z_j \\ \vdots \\ \sum_{i=1}^{m} \sum_{j=1}^{n} x_n^{(i)} x_j^{(i)} \frac{\partial^2 a(\eta)}{\partial \eta^2} z_j \end{bmatrix}$$

$$\Rightarrow z^T Hz = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{n} x_k^{(i)} x_j^{(i)} \frac{\partial^2 a(\eta)}{\partial \eta^2} z_j z_k$$

$$= \sum_{i=1}^{m} \frac{\partial^2 a(\eta)}{\partial \eta^2} \left(x^{(i)T} z\right)^2 \geq 0$$

$\Rightarrow H$ is PSD

5 a) Locally Weighted linear regression

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

Problem: Show that $J(\theta)$ can also be written $J(\theta) = (X\theta - y)^T W (X\theta - y)$

$$\Rightarrow (X\theta - y)^T W (X\theta - y) = (X\theta - y)^T \begin{bmatrix} \sum_{j=1}^{m} w_{1j} (\theta^T x^{(j)} - y^{(j)}) \\ \vdots \\ \sum_{j=1}^{m} w_{mj} (\theta^T x^{(j)} - y^{(j)}) \end{bmatrix}$$

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

$\Rightarrow J(\theta)$ can be written as $(X\theta - y)^T W (X\theta - y)$ if and only if W is a $m \times m$ diagonal matrix with coefficients $\begin{cases} w_{ij} = \frac{1}{2} w^{(i)} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$

(ii) Problem: By finding the derivative $\nabla_\theta J(\theta)$ and setting that to zero, generalize the normal equation to this weighted setting.

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

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

$$= \theta^T X^T W X\theta - y^T W X\theta - \theta^T X^T W y - y^T W y$$

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

$$\Rightarrow \nabla_\theta J(\theta) = 2X^T W X\theta - 2 X^T W^T y = 0$$

$$\Rightarrow 2X^T W^T y = 2X^T W X\theta$$

$$\Rightarrow (X^T W X)^{-1} (X^T W^T y) = \theta$$

iii $\propto p(y^{(i)} | x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma^{(i)}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}\right)$

Problem: Show that finding the maximum likelihood estimate of θ reduces to solving a weighted linear regression problem.

$$l(\theta) = \log\left(\prod_{i=1}^{m} p(y^{(i)} | x^{(i)}; \theta)\right)$$

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

$$\Rightarrow \frac{\partial l(\theta)}{\partial \theta_j} = \sum_{i=1}^{m} -\frac{(-x_j^{(i)})(y^{(i)} - \theta^T x^{(i)})}{(\sigma^{(i)})^2}$$

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