1. $x \in \mathbb{R}^n$ $X \in \mathbb{R}^{m \times n}$ $\theta \in \mathbb{R}^n$ $Y \in \mathbb{R}^m$

(a) [10 points] In lecture we saw the average empirical loss for logistic regression:

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

where $y^{(i)} \in \{0, 1\}$, $h_\theta(x) = g(\theta^T x)$ and $g(z) = 1/(1 + e^{-z})$.
Find the Hessian H of this function, and show that for any vector z, it holds true that

$$z^T H z \geq 0.$$

$$\frac{\partial}{\partial \theta_j} J(\theta) = \left(y \frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right) \frac{\partial}{\partial \theta_j} g(\theta^T x)$$

$$= \left(y \frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right) g(\theta^T x)(1-g(\theta^T x)) \frac{\partial}{\partial \theta_j}(\theta^T x)$$

$$= [y(1-g(\theta^T x)) - (1-y)g(\theta^T x)] x_j$$

$$= [y - \cancel{y g(\theta^T x)} - g(\theta^T x) + \cancel{y g(\theta^T x)}] x_j$$

$$= (y - h_\theta(x)) x_j$$

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

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

$$\nabla J(\theta) = -\frac{1}{m} X^T (Y - h_\theta(X)) \longrightarrow -\frac{1}{m}\begin{bmatrix} \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x_1^{(i)} \\ \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x_2^{(i)} \\ \vdots \\ \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x_n^{(i)} \end{bmatrix}$$

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

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

$h_\theta(x) \in \mathbb{R}^m$

$H \in \mathbb{R}^{n \times n}$

$$\frac{1}{m}\begin{bmatrix} \sum_{i=1}^{m} h_\theta(x^{(i)})(1-h_\theta(x^{(i)})) x_1^{(i)} x_1^{(i)} & \cdots \\ \vdots & \ddots \\ \cdots & \sum_{i=1}^{m} h_\theta(x^{(i)})(1-h_\theta(x^{(i)})) x_n^{(i)} x_n^{(i)} \end{bmatrix}$$

$$z^T H z = \frac{1}{m}\sum_{j=1}^{n}\sum_{k=1}^{n} H_{jk} z_j z_k$$

$$= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} h_\theta(x^{(i)})\left(1-h_\theta(x^{(i)})\right) x_j^{(i)} x_k^{(i)} z_j z_k$$

$$= \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)})\left(1-h_\theta(x^{(i)})\right)\sum_{j=1}^{n}\sum_{k=1}^{n} x_j^{(i)} x_k^{(i)} z_j z_k$$

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

$\because h_\theta(x^{(i)}) \& 1-h_\theta(x^{(i)}) > 0$

$\therefore z^T H z \geq 0$

(c) [5 points] Recall that in GDA we model the joint distribution of (x, y) by the following equations:

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

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

Suppose we have already fit ϕ, μ_0, μ_1, and Σ, and now want to predict y given a new point x. To show that GDA results in a classifier that has a linear decision boundary, show the posterior distribution can be written as

$$p(y = 1 \mid x; \phi, \mu_0, \mu_1, \Sigma) = \frac{1}{1 + \exp(-(\theta^T x + \theta_0))},$$

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

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

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

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

divided by numerator

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

$$\left[x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 - \right. \quad \because (\Sigma^{-1})^T = \Sigma^{-1}$$

$$\left.(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_0 - \mu_0^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0)\right]$$

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

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

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

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

Let

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

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

(d) [7 points] For this part of the problem only, you may assume n (the dimension of x) is 1, so that $\Sigma = [\sigma^2]$ is just a real number, and likewise the determinant of Σ is given by $|\Sigma| = \sigma^2$. Given the dataset, we claim that the maximum likelihood estimates of the parameters are given by

$$\begin{aligned}
\phi &= \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} = 1\} \\
\mu_0 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} \\
\mu_1 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} \\
\Sigma &= \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T
\end{aligned}$$

The log-likelihood of the data is

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

By maximizing ℓ with respect to the four parameters, prove that the maximum likelihood estimates of ϕ, μ_0, μ_1, and Σ are indeed as given in the formulas above. (You may assume that there is at least one positive and one negative example, so that the denominators in the definitions of μ_0 and μ_1 above are non-zero.)

$$p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma)$$

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

$$p(y^{(i)}; \phi) = \phi^{1\{y^{(i)}=1\}} (1-\phi)^{1-1\{y^{(i)}=1\}}$$

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

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

$$\frac{\partial}{\partial \mu_0}\ell = \left[\sum_{i=1}^{m} \left(-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}})\right)\right]'$$

$$= \left[\sum_{i=1}^{m} \left(-\frac{1}{2}(x^{(i)2} + \mu_{y^{(i)}}^2 - 2x^{(i)}\mu_{y^{(i)}})\right)\right]'$$

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

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

$$\frac{\partial}{\partial \Sigma} l = \left(-\frac{1}{2}\sum_{i=1}^{m} \log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(x-\mu_{y^{(i)}})^T\Sigma^{-1}(x-\mu_{y^{(i)}})\right)'$$

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

$$\left.\begin{aligned}
&\frac{\partial}{\partial \phi} l = 0\\
&\frac{\partial}{\partial \mu_0} l = 0\\
&\frac{\partial}{\partial \mu_1} l = 0\\
&\frac{\partial}{\partial \Sigma} l = 0
\end{aligned}\right\} \Rightarrow$$

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

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

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

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

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

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

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

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

(h)

The dataset where GDA performed wrose is far away from Gaussian distribution. GDA has stronger assumptions than Logistic Regression's. It will perform better if dataset if Gaussian distributed.

2.

(a) [5 points] Suppose that each $y^{(i)}$ and $x^{(i)}$ are conditionally independent given $t^{(i)}$:

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

Note this is equivalent to saying that labeled examples were selected uniformly at random from the set of positive examples. Prove that the probability of an example being labeled differs by a constant factor from the probability of an example being positive. That is, show that $p(t^{(i)} = 1 \mid x^{(i)}) = p(y^{(i)} = 1 \mid x^{(i)})/\alpha$ for some $\alpha \in \mathbb{R}$.

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

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

$$P(t^{(i)}=1 \mid x^{(i)}) = P(y^{(i)}=1 \mid x^{(i)})/2 \quad (1)$$

where $2 = P(y^{(i)}=1 \mid t^{(i)}=1)$

(b) [5 points] Suppose we want to estimate α using a trained classifier h and a held-out validation set V. Let V_+ be the set of labeled (and hence positive) examples in V, given by $V_+ = \{x^{(i)} \in V \mid y^{(i)} = 1\}$. Assuming that $h(x^{(i)}) \approx p(y^{(i)} = 1 \mid x^{(i)})$ for all examples $x^{(i)}$, show that

$$h(x^{(i)}) \approx \alpha \quad \text{for all } x^{(i)} \in V_+.$$

You may assume that $p(t^{(i)} = 1 \mid x^{(i)}) \approx 1$ when $x^{(i)} \in V_+$.

from (1), $2 \approx P(y^{(i)}=1 \mid x^{(i)}) = h(x^{(i)})$ for all $x^{(i)} \in V_+$

(e) $\frac{1}{1+e^{-\theta^T x}} = 0.52$

$1 + e^{-\theta^T x} = \frac{2}{0.52}$

$\theta^T x = \ln\left(\frac{2}{2-2}\right)$

$\theta_1 x_1 + \theta_2 x_2 + \theta_0 = \ln\left(\frac{2}{2-2}\right)$

$\theta_2 x_2 = -(\theta_1 x_1 + \ln(\frac{2}{2}-1) + \theta_0)$

$x_2 = -\left(\frac{\theta_1}{\theta_2} x_1 + \frac{\ln(\frac{2}{2}-1)+\theta_0}{\theta_2}\right)$

$\frac{\theta_0}{\theta_2} \cdot C = \frac{\ln(\frac{2}{2}-1)+\theta_0}{\theta_2}$

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

3.

(a) [5 points] Consider the Poisson distribution parameterized by λ:

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

Show that the Poisson distribution is in the exponential family, and clearly state the values for $b(y)$, η, $T(y)$, and $a(\eta)$.

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

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

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

(b) [3 points] Consider performing regression using a GLM model with a Poisson response variable. What is the canonical response function for the family? (You may use the fact that a Poisson random variable with parameter λ has mean λ.)

$$g(\eta) = E(y^{(i)}|\eta) = \frac{\partial}{\partial\eta} a(\eta) = e^{\eta}$$

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

(c) [7 points] For a training set $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$, let the log-likelihood of an example be $\log p(y^{(i)}|x^{(i)}; \theta)$. By taking the derivative of the log-likelihood with respect to θ_j, derive the stochastic gradient ascent update rule for learning using a GLM model with Poisson responses y and the canonical response function.

let $\eta = \theta^T x$

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

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

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

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

4.

(a) [5 points] Derive an expression for the mean of the distribution. Show that $\mathbb{E}[Y \mid X; \theta]$ can be represented as the gradient of the log-partition function a with respect to the natural parameter η.

Hint: Start with observing that $\frac{\partial}{\partial \eta} \int p(y;\eta)dy = \int \frac{\partial}{\partial \eta} p(y;\eta)dy$.

To prove:

$$E(Y|X;\theta) = E(y;\theta^T x) = \int y\, p(y;\eta)\, dy$$

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

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

$$\frac{\partial}{\partial \eta} p(y;\eta) = y\, p(y;\eta) - \frac{\partial}{\partial \eta} a(\eta)\, p(y;\eta) \quad (1)$$

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

($\int y\, p(y;\eta)\, dy$: $\uparrow E(y;\theta^T x)$; $\int p(y;\eta)\, dy = 1$)

$$0 = E(y;\theta^T x) - \frac{\partial}{\partial \eta} a(\eta)$$

$$E(y;\theta^T x) = \frac{\partial}{\partial \eta} a(\eta)$$

(b) [5 points] Next, derive an expression for the variance of the distribution. In particular, show that $\text{Var}(Y \mid X; \theta)$ can be expressed as the derivative of the mean w.r.t η (*i.e.*, the second derivative of the log-partition function $a(\eta)$ w.r.t the natural parameter η.)

$$\text{Var}(Y \mid X; \theta) = \text{Var}(Y; \theta^T x) = \text{Var}(Y; \eta)$$

$$= E(y^2; \eta) - \left(E(y; \eta)\right)^2$$

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

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

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

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

$E(y)$ $E(y)$

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

(c) [5 points] Finally, write out the loss function $\ell(\theta)$, the NLL of the distribution, as a function of θ. Then, calculate the Hessian of the loss w.r.t θ, and show that it is always PSD. This concludes the proof that NLL loss of GLM is convex.

Hint: Use the chain rule of calculus along with the results of the previous parts to simplify your derivations.

For a single training example:

$$\ell(\eta) = -\log p(y^{(i)};\eta) = \log p(y^{(i)} \mid x^{(i)};\theta)$$

$$\ell(\eta) = -\log b(y^{(i)}) - (\eta \cdot y^{(i)} - a(\eta))$$

$$\frac{\partial}{\partial \eta}\ell(\eta) = \frac{\partial}{\partial \eta} a(\eta) - y^{(i)} = E(Y \mid \eta) - y^{(i)}$$

$$\frac{\partial^2}{\partial \eta}\ell(\eta) = \frac{\partial^2}{\partial \eta} a(\eta) = \mathrm{Var}(Y \mid \eta)$$

$$\frac{\partial^2}{\partial \theta_j \partial \theta_K}\ell(\theta) = \frac{\partial}{\partial \eta}\ell(\eta) \cdot 0 - \frac{\partial^2}{\partial \eta}\ell(\eta)\,\frac{\partial}{\partial \theta_j}(\eta)\,\frac{\partial}{\partial \theta_K}(\eta)$$

$$= \mathrm{Var}(Y \mid \eta)\, X_j X_K$$

$$z^T H z = \mathrm{Var}(Y \mid \eta) \cdot \sum_j \sum_K z_j X_j X_K z_K$$

$$= \mathrm{Var}(Y \mid \eta)\,(X^T z)^2 \geq 0$$

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

i. [2 points] Show that $J(\theta)$ can also be written

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

for an appropriate matrix W, and where X and y are as defined in class. Clearly specify the value of each element of the matrix W.

$W \in \mathbb{R}^{m \times m}$

$W = \text{diag}(\frac{1}{2} w^{(i)})$ for $i \in \{0, 1, 2 \dots m\}$

ii. [4 points] If all the $w^{(i)}$'s equal 1, then we saw in class that the normal equation is

$$X^T X\theta = X^T y,$$

and that the value of θ that minimizes $J(\theta)$ is given by $(X^T X)^{-1} X^T y$. By finding the derivative $\nabla_\theta J(\theta)$ and setting that to zero, generalize the normal equation to this weighted setting, and give the new value of θ that minimizes $J(\theta)$ in closed form as a function of X, W and y.

$J_0(\theta) = (X\theta - y)^T W (X\theta - y)$ $\quad W^T = W$

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

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

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

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

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

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

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

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

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

iii. [4 points] Suppose we have a dataset $\{(x^{(i)}, y^{(i)}); i = 1 \ldots, m\}$ of m independent examples, but we model the $y^{(i)}$'s as drawn from conditional distributions with different levels of variance $(\sigma^{(i)})^2$. Specifically, assume the model

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

That is, each $y^{(i)}$ is drawn from a Gaussian distribution with mean $\theta^T x^{(i)}$ and variance $(\sigma^{(i)})^2$ (where the $\sigma^{(i)}$'s are fixed, known, constants). Show that finding the maximum likelihood estimate of θ reduces to solving a weighted linear regression problem. State clearly what the $w^{(i)}$'s are in terms of the $\sigma^{(i)}$'s.

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

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

$w^{(i)}$