1.

(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

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

$$\nabla^2 J(\theta) = \nabla \nabla J(\theta)$$

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

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

$$+, \frac{\partial}{\partial x_i} y^{(i)} \log(h_\theta(x^{(i)})) = y^{(i)} \frac{1}{h_\theta(x^{(i)})} \cdot \frac{\partial}{\partial x^{(i)}} \cdot h_\theta(x^{(i)})$$

$$-, \frac{\partial}{\partial x_j}(1-y^{(i)}) \cdot \log(1-h_\theta(x^{(i)})) = (1-y^{(i)}) \times \frac{1}{1-h_\theta(x^{(i)})} \cdot \frac{\partial}{\partial x_i} h_\theta(x^{(i)})$$

$$\Rightarrow \nabla J(\theta) = \frac{-1}{m}\sum_{i=1}^{m}\left(\frac{y^{(i)}}{h_\theta(x^{(i)})} - \frac{1-y^{(i)}}{1-h_\theta(x^{(i)})}\right) \cdot \frac{\partial}{\partial x_i} h_\theta(x^{(i)})$$

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

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

$$\Rightarrow \nabla^2 J(\theta) = \frac{\partial}{\partial \theta_k} \frac{-1}{m}\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))\, x^{(i)} = \frac{-1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial x_k} -h_\theta(x^{(i)}) \cdot x_j^{(i)}$$

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

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

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

$$= \boxed{\frac{1}{m} \cdot X^T \cdot g(X \cdot \theta) \cdot (1 - g(X\theta)) \cdot X}$$

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

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

≥ 0 $\quad \geq 0$ $\quad \geq 0$

* $x_j \cdot z_j \Rightarrow x^{(i)T} \cdot z$

$x : n \times 1$

$z : n \times 1$

C1

[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.

Have $P(y=1|x) = \dfrac{P(x|y=1) \cdot P(y)}{P(x)}$ (Bayes's rule)

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

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

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

have $(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)$

$= x^T\Sigma^{-1}x - x_k^T\Sigma^{-1}\mu_k - \mu_k^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}\mu_k$

$= x^T\Sigma^{-1}x - 2\mu_k^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}\mu_k$

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

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

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

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

$\Rightarrow$

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

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

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.)

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

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

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

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

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

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

$$= \frac{-m \cdot n}{2} \cdot \log(2\pi) - \frac{m}{2}\log(\Sigma) - \frac{1}{2}\sum_{i=1}^{m} (x - \mu_{y^{(i)}})^T \Sigma^{-1} (x - \mu^{(i)})$$

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

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

Here $\mu_{y^{(i)}} = 1\{y^{(i)}=0\}\mu_0 + 1\{y^{(i)}=1\}\mu_1$

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

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

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

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

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

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

Maximize $\Rightarrow$

$$\begin{cases} \frac{\partial l}{\partial \phi} = 0 \\ \frac{\partial l}{\partial \mu_0} = 0 \\ \frac{\partial l}{\partial \mu_1} = 0 \\ \frac{\partial l}{\partial \Sigma} = 0 \end{cases}$$

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

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

}

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

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

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

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

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

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

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

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

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

2.

a. have $P(y=1, t=1, x) = P(t=1 | y=1, x) \cdot P(y=1 | x) \cdot P(x)$

$$= P(y=1 | t, x) \cdot P(t=1 | x) \cdot P(x)$$

$$\Rightarrow P(t=1 \mid y=1, x) \cdot P(y=1 \mid x) \cdot P(a) = P(y=1 \mid t=1, x) \cdot P(t=1 \mid x) \cdot P(a)$$

$P(y=1 \mid t=1)$

$\Rightarrow P(y=1 \mid x) = P(t=1 \mid x) \cdot P(y=1 \mid t=1)$

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

b. V: held-out validation

V_+: set of labled (and hence positive) example in V

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

$$h(x^{(i)}) \approx p(y^{(i)}=1 \mid x^{(i)}) \approx P(t=1 \mid x) \cdot \alpha \approx \alpha \quad \text{for all } x^{(i)} \in V^+$$

2.

a. $$p(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = \frac{1}{y!} \cdot \exp\left(\log\left(e^{-\lambda}\lambda^y\right)\right)$$

$$= \frac{1}{y!} \cdot \exp\left(y \log \lambda - \lambda\right)$$

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

$$\eta = \log \lambda$$

$$T(y) = y$$

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

b. $x^{(i)} \in V_+ \Rightarrow P(t=1 \mid x^{(i)}) = 1$

$h_\theta(x^{(i)}) \approx P(y^{(i)}=1|x) = P(+=1|x^{(i)}) \cdot \alpha \approx 1$

3.

a. $p(y;\lambda) = \frac{1}{y!} \cdot \exp\left(y\log(\lambda) - \lambda\right)$

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

$\eta = \log(\lambda)$

$T(y) = y$

$a(\eta) = \lambda = \exp(\eta)$

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

$g(\eta) = a(\eta)' = \exp(\eta)' = \exp(\theta^T x)$

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

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

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

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

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

4.

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

$= \int \frac{\partial}{\partial \eta} b(y) \cdot \exp\left(\eta y - a(\eta)\right) dy$

$= \int b(y) \cdot \exp(\eta y - a(\eta) \cdot y - \frac{\partial}{\partial \eta} a(\eta)\, dy$

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

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

$= E[y;\eta] - \frac{\partial}{\partial \eta} a(\eta) = 0$

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

b. $E[y;\eta] = \frac{\partial}{\partial \eta} a(\eta)$

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

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

$$= \frac{\partial}{\partial \eta} \int y \cdot b(y) \cdot \exp(\eta y - a(\eta)) \, dy$$

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

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

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

$$= E[y^2;\eta] - (E[y;\eta])^2 = Var[Y|\eta]$$

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

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

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

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

By chain rule:

$$\frac{\partial(-\theta^T x^{(i)} y^{(i)})}{\partial \theta_j} = -y^{(i)} \cdot x^{(i)}_j$$

$$\frac{\partial a(\theta^T x^{(i)})}{\partial \theta_j} = a'(\theta^T x^{(i)}) \cdot x^{(i)}_j$$

$$\Rightarrow \boxed{\frac{\partial l(\theta)}{\partial \theta_j} = \sum_{i=1}^{m} \left(a'(\theta^T x^{(i)}) - y^{(i)}\right) x_j^{(i)}}$$

GLM Gradient

$$H_{jk} = \frac{\partial^2 l(\theta)}{\partial \theta_j \partial \theta_k} = \sum_{i=1}^{m} a''(\theta^T x^{(i)}) x_j^{(i)} x_k^{(i)}$$

$$H = a''(X\theta) \cdot x^{(i)T} \cdot x^{(i)}$$

have $z^T H z = \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{i=1}^{m} a'' z_j a''(\theta^T x^{(i)}) x_j^{(i)} x_k^{(i)} z_k$

$$= \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{i=1}^{m} a''(\theta^T x^{(i)}) (x^{(i)T} \cdot z)^2$$

$a''(\theta^T x^{(i)}) = \mathrm{Var}[Y|x;\theta] \geq 0$, $(x^{(i)T} \cdot z)^2 \geq 0$

$\Rightarrow$ H is a PSD $\Rightarrow$ NLL is a convex function of θ

5、

a、

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

W is a diagonal weight matrix:

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

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

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

ii, $\nabla_\theta J(\theta) = \nabla (X\theta - y)^T W (X\theta - y)$

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

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

$= 2X^T W X\theta - 2y^T W X = 0$

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