

1.

(a)

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_j} &= -\frac{1}{m} \sum_{i=1}^m y^{(i)} \frac{g(\theta^T x^{(i)})[1 - g(\theta^T x^{(i)})]}{g(\theta^T x^{(i)})} x_j^{(i)} - (1 - y^{(i)}) \frac{g(\theta^T x^{(i)})[1 - g(\theta^T x^{(i)})]}{1 - g(\theta^T x^{(i)})} x_j^{(i)} \\
&= -\frac{1}{m} \sum_{i=1}^m y^{(i)} [1 - g(\theta^T x^{(i)})] x_j^{(i)} - (1 - y^{(i)}) g(\theta^T x^{(i)}) x_j^{(i)} \\
&= \frac{1}{m} \sum_{i=1}^m [g(\theta^T x^{(i)}) - y^{(i)}] x_j^{(i)}
\end{aligned}$$

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

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

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

$$\begin{aligned}
z^T H z &= \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^n g(\theta^T x^{(i)}) [1 - g(\theta^T x^{(i)})] x_j^{(i)} x_k^{(i)} z_j z_k \\
&= \frac{1}{m} \sum_{i=1}^m g(\theta^T x^{(i)}) [1 - g(\theta^T x^{(i)})] [(x^{(i)})^T z]^2 \geq 0
\end{aligned}$$

(c)

$$\begin{aligned}
p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1) + p(x|y=0)p(y=0)} \\
&= \frac{\exp\{-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\} \phi}{\exp\{-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\} \phi + \exp\{-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\} (1 - \phi)} \\
&= \frac{1}{1 + \exp\{\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1) - \frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\} \frac{1-\phi}{\phi}} \\
&= \frac{1}{1 + \exp\{-(\Sigma^{-1}(\mu_1 - \mu_0))^T x + \frac{1}{2}(\mu_0 + \mu_1)^T \Sigma^{-1}(\mu_0 - \mu_1) - \ln(\frac{1-\phi}{\phi})\}}
\end{aligned}$$

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

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

(d)

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

$$p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}})\right\}$$

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

$$\begin{aligned}
\ell &= \sum_{i=1}^m \log p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^m \log p(y^{(i)}; \phi) \\
&= \sum_{i=1}^m \log \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \right\} + \sum_{i=1}^m \log \phi^{1\{y^{(i)}=1\}} (1-\phi)^{1-1\{y^{(i)}=1\}} \\
&= -\frac{mn}{2} \log(2\pi) - \frac{m}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \\
&\quad + \sum_{i=1}^m 1\{y^{(i)} = 1\} \log \phi + \left(m - \sum_{i=1}^m 1\{y^{(i)} = 1\} \right) \log(1 - \phi)
\end{aligned}$$

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

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

$$\frac{\partial \mu_{y^{(i)}}}{\partial \mu_0} = 1\{y^{(i)} = 0\}, \quad \frac{\partial \mu_{y^{(i)}}}{\partial \mu_1} = 1\{y^{(i)} = 1\}$$

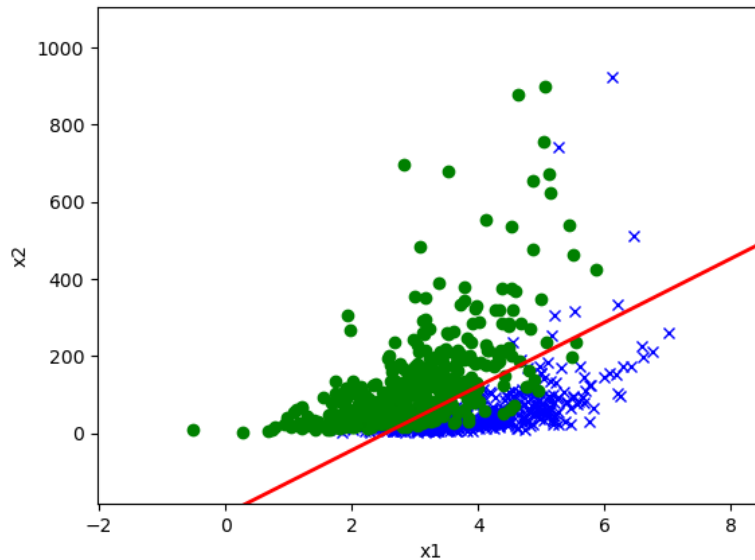
$$\frac{\partial \ell}{\partial \mu_0} = \frac{\partial \ell}{\partial \mu_{y^{(i)}}} \frac{\partial \mu_{y^{(i)}}}{\partial \mu_0} = \Sigma^{-1} \sum_{i=1}^m (x^{(i)} 1\{y^{(i)} = 0\} - \mu_0 1\{y^{(i)} = 0\})$$

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

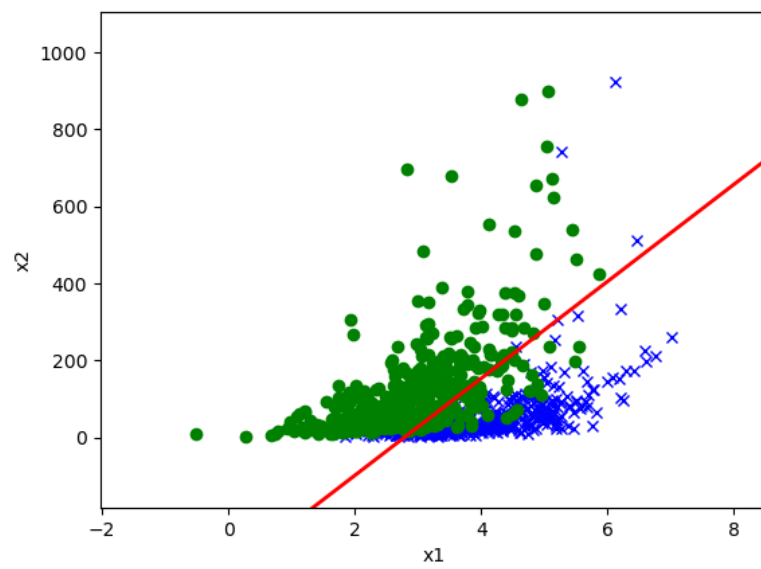
$$\frac{\partial \ell}{\partial \Sigma} = -\frac{m}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} \left(\sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \right) \Sigma^{-1}$$

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

(f)

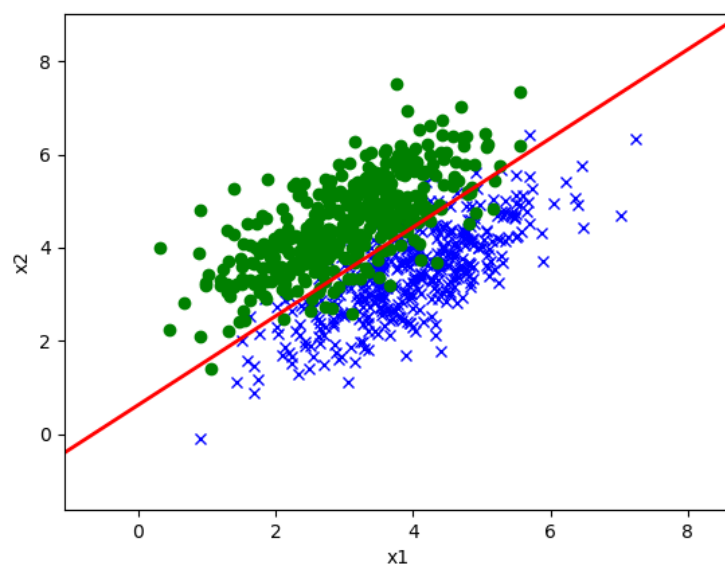


logistic regression

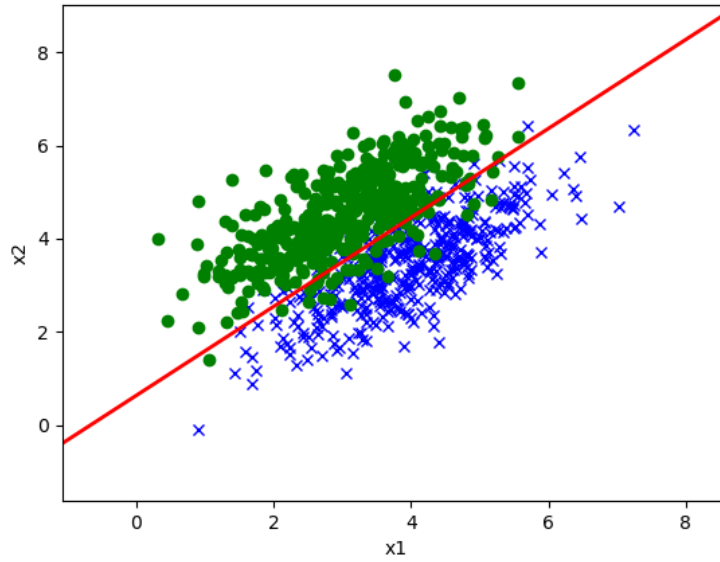


GDA

(g)



logistic regression



GDA

On Dataset 1 GDA perform worse than logistic regression.

Because $p(x|y)$ may be not Gaussian distribution.

(h)

Box-Cox transformation.

2.

(a)

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

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

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

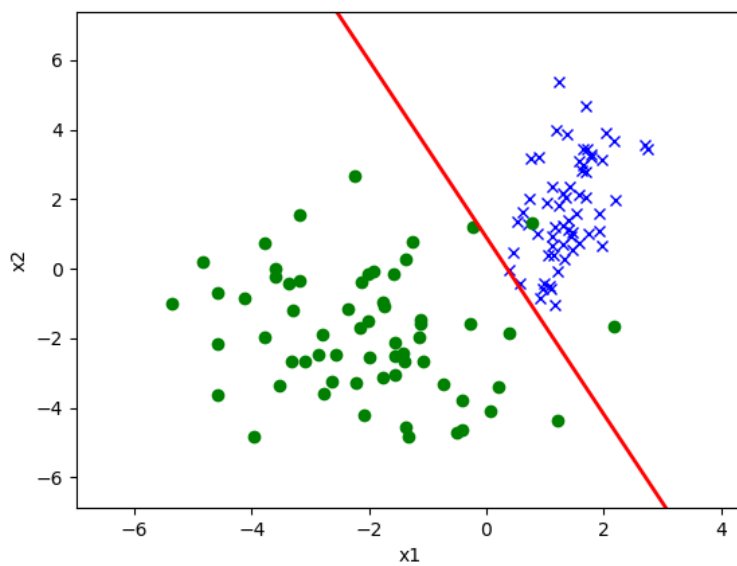
$$P(t = 1|x) = \frac{P(y = 1|x)}{P(y = 1|t = 1)}$$

$$P(y = 1|t = 1) = \alpha$$

(b)

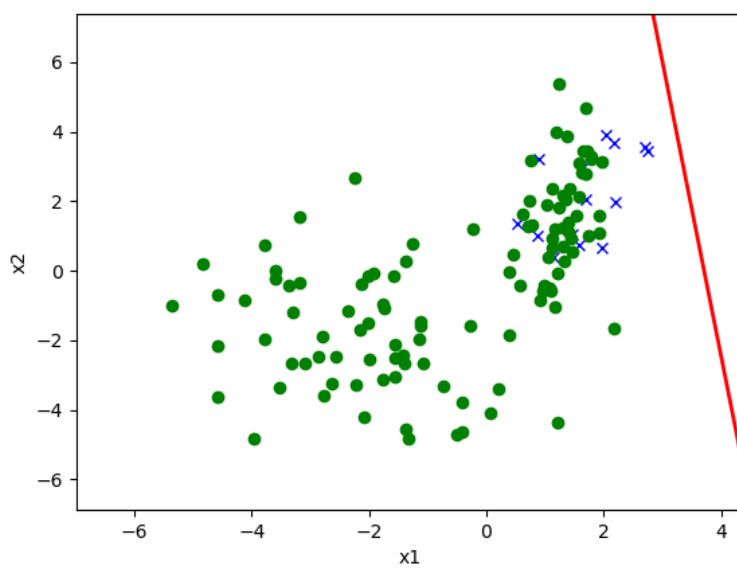
$$h(x) \approx p(y = 1|x) = p(t = 1|x)\alpha \approx \alpha \quad \text{for all } x \in V_+$$

(c)



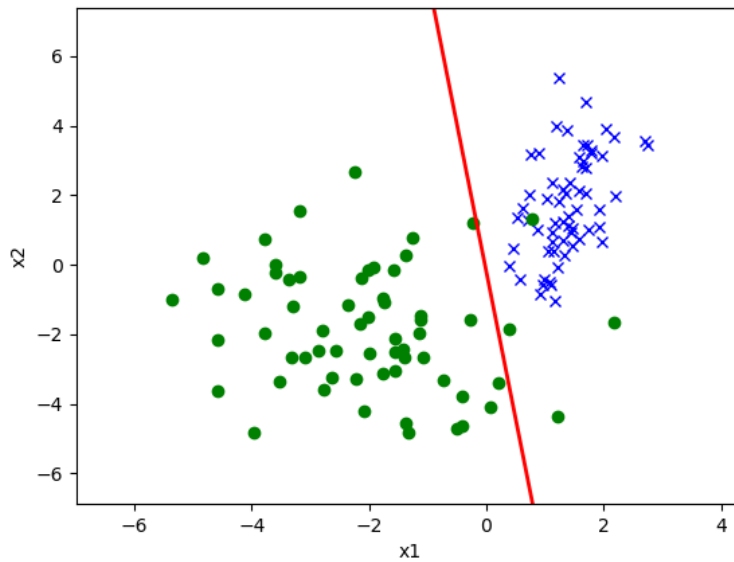
train use t-label

(d)



train use y-label

(e)



train use y-label, rescale by α

3.

(a)

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

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

(b)

$$h_\theta(x) = E(y|x; \theta) = \lambda = e^\eta = e^{\theta^T x}$$

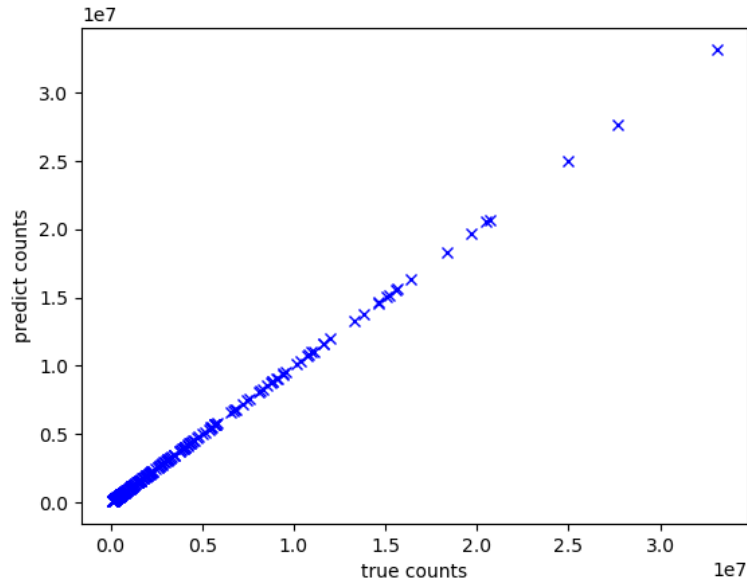
(c)

$$\begin{aligned} \log p(y^{(i)} | x^{(i)}; \theta) &= \log \frac{1}{y^{(i)}!} \exp\{\theta^T x^{(i)} y^{(i)} - e^{\theta^T x^{(i)}}\} \\ &= -\log y^{(i)}! + \theta^T x^{(i)} y^{(i)} - e^{\theta^T x^{(i)}} \end{aligned}$$

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

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

(d)



4.

(a)

$$\begin{aligned}
 \frac{\partial}{\partial \eta} \int p(y; \eta) dy &= 0 \\
 \frac{\partial}{\partial \eta} \int p(y; \eta) dy &= \int \frac{\partial}{\partial \eta} p(y; \eta) dy \\
 &= \int b(y) \exp\{\eta y - a(\eta)\} \left(y - \frac{\partial a(\eta)}{\partial \eta}\right) dy \\
 &= \int p(y; \eta) \left(y - \frac{\partial a(\eta)}{\partial \eta}\right) dy \\
 &= \int y p(y; \eta) dy - \frac{\partial a(\eta)}{\partial \eta} \int p(y; \eta) dy \\
 &= E[Y; \eta] - \frac{\partial a(\eta)}{\partial \eta} \\
 E[Y; \eta] &= E[Y|X; \theta] = \frac{\partial a(\eta)}{\partial \eta}
 \end{aligned}$$

(b)

$$\begin{aligned}
 \frac{\partial}{\partial \eta} \int y p(y; \eta) dy &= \frac{\partial^2 a(\eta)}{\partial \eta^2} \\
 \frac{\partial}{\partial \eta} \int y p(y; \eta) dy &= \int y \frac{\partial}{\partial \eta} p(y; \eta) dy \\
 &= \int y p(y; \eta) \left(y - \frac{\partial a(\eta)}{\partial \eta}\right) dy \\
 &= \int y^2 p(y; \eta) dy - \frac{\partial a(\eta)}{\partial \eta} \int y p(y; \eta) dy \\
 &= E[Y^2; \eta] - E^2[Y; \eta] \\
 &= \text{Var}[Y; \eta] \\
 \text{Var}[Y; \eta] &= \text{Var}[Y|X; \theta] = \frac{\partial^2 a(\eta)}{\partial \eta^2}
 \end{aligned}$$

(c)

$$\begin{aligned}\ell(\theta) &= -\sum_{i=1}^m \log p(y^{(i)} | x^{(i)}; \theta) \\ &= \sum_{i=1}^m -\log b(y^{(i)}) - \theta^T x^{(i)} y^{(i)} + a(\theta^T x^{(i)}) \\ \frac{\partial \ell(\theta)}{\partial \theta_j} &= \sum_{i=1}^m [a'(\theta^T x^{(i)}) - y^{(i)}] x_j^{(i)} \\ H_{jk} &= \frac{\partial^2 \ell(\theta)}{\partial \theta_j \partial \theta_k} = \sum_{i=1}^m a''(\theta^T x^{(i)}) x_j^{(i)} x_k^{(i)} \\ z^T H z &= \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^n a''(\theta^T x^{(i)}) x_j^{(i)} x_k^{(i)} z_j z_k \\ &= \sum_{i=1}^m a''(\theta^T x^{(i)}) [(x^{(i)})^T z]^2 \\ a''(\theta^T x) &= \text{Var}[Y|X; \theta] \geq 0 \Rightarrow z^T H z \geq 0\end{aligned}$$

5.

(a)

i.

$$\begin{aligned}W &\in \mathbb{R}^{m \times m} \\ W_{ij} &= \begin{cases} \frac{1}{2} w^{(i)} & i = j \\ 0 & i \neq j \end{cases}\end{aligned}$$

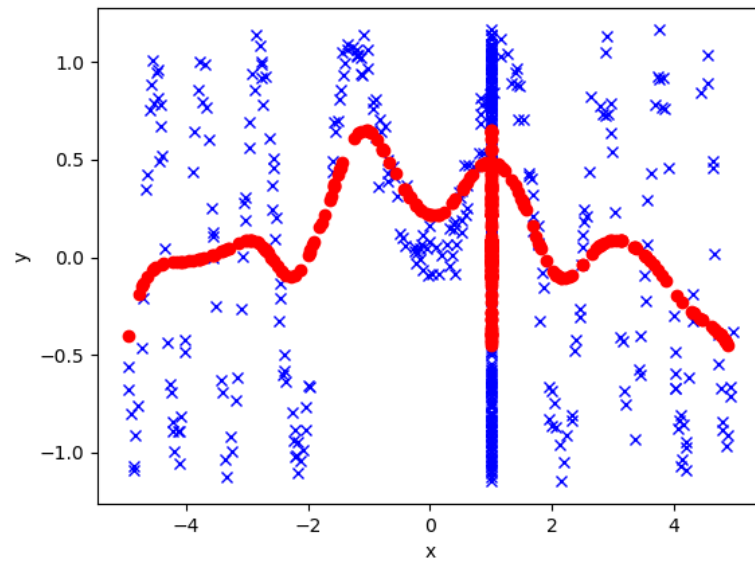
ii.

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \nabla_{\theta} (X\theta - y)^T W (X\theta - y) \\ &= \nabla_{\theta} (\theta^T X^T - y^T) W (X\theta - y) \\ &= \nabla_{\theta} (\theta^T X^T W X \theta - y^T W X \theta - \theta^T X^T W y + y^T W y) \\ &= \nabla_{\theta} (\theta^T X^T W X \theta - 2y^T W X \theta) \\ &= 2X^T W X \theta - 2X^T W y \\ \nabla_{\theta} J(\theta) &= 0 \Rightarrow \theta = (X^T W X)^{-1} X^T W y\end{aligned}$$

iii.

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^m \log p(y^{(i)} | x^{(i)}; \theta) \\ &= \sum_{i=1}^m -\log(\sqrt{2\pi}\sigma^{(i)}) - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2} \\ w^{(i)} &= -\frac{1}{(\sigma^{(i)})^2} \\ \frac{\partial \ell(\theta)}{\partial \theta_j} &= \sum_{i=1}^m \frac{y^{(i)} - \theta^T x^{(i)}}{(\sigma^{(i)})^2} x_j^{(i)}\end{aligned}$$

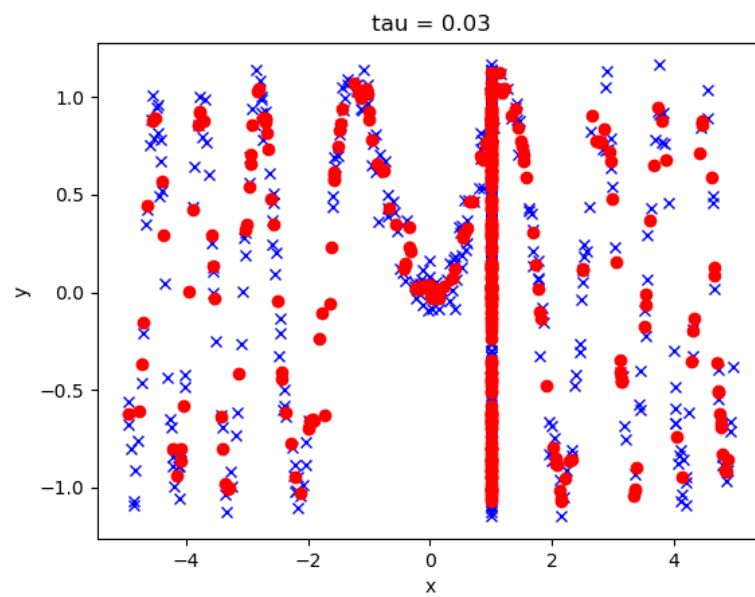
(b)

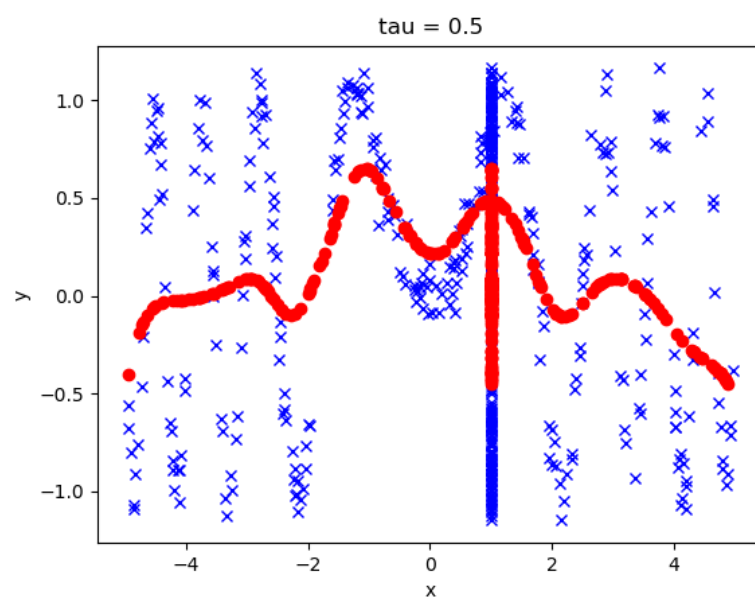
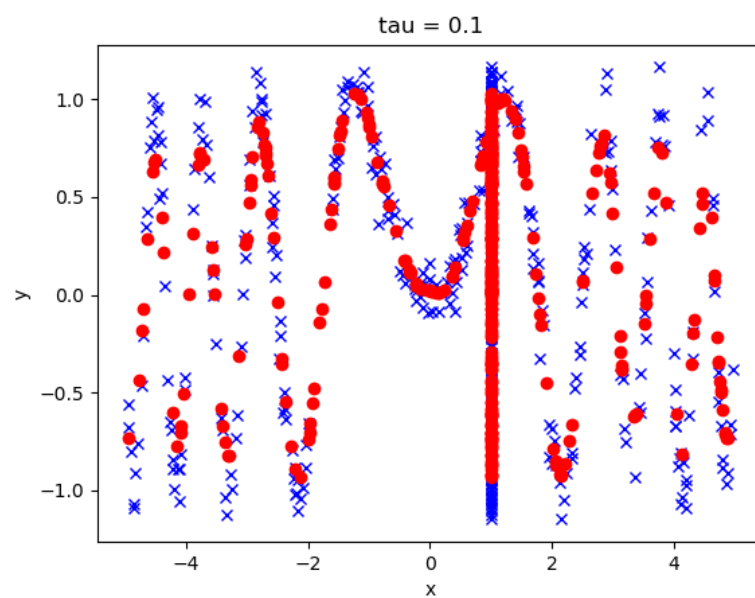
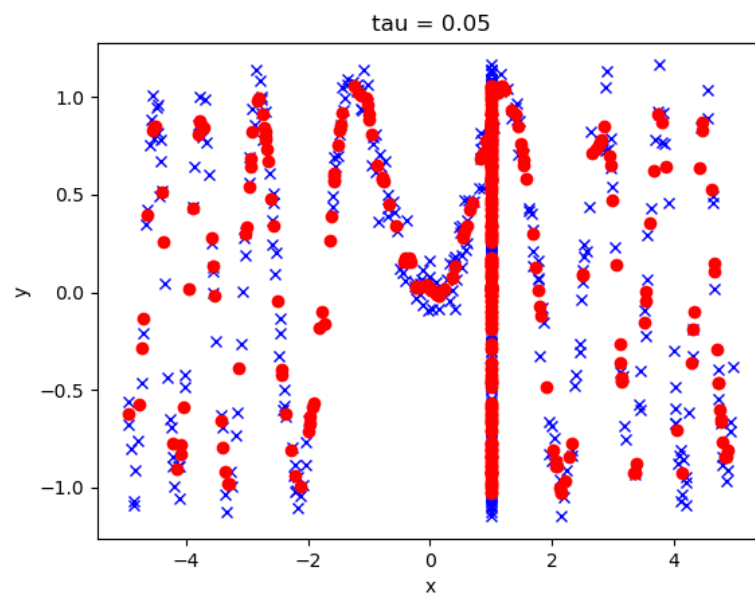


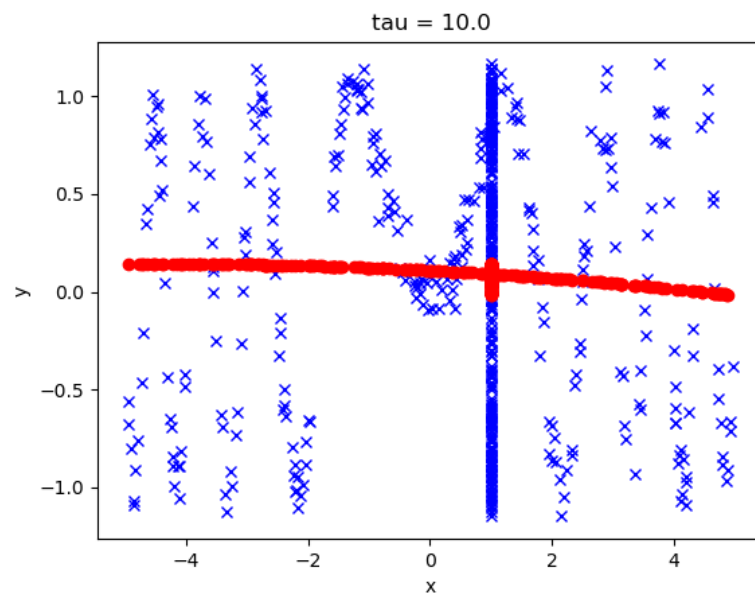
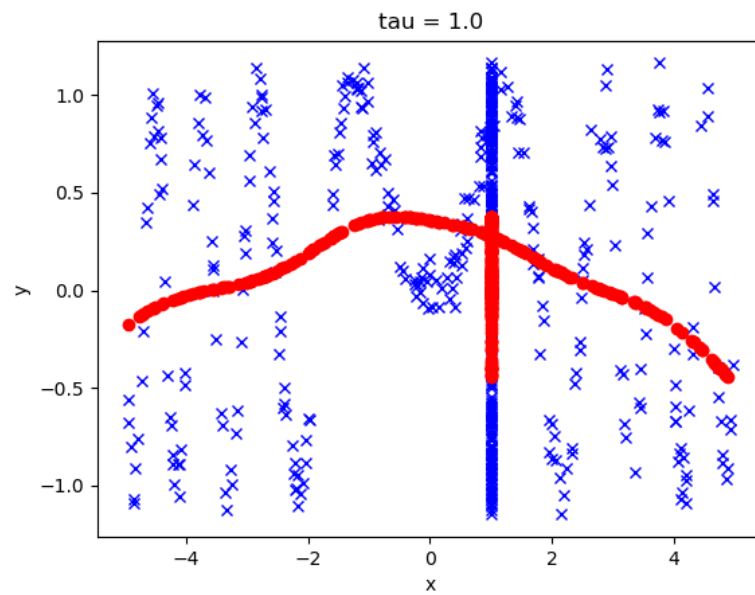
MSE=0.331.

The model seems to be underfitting.

(c)







$\tau = 0.05$ achieves the lowest MSE on the valid set.

MSE=0.012 on the valid set, MSE=0.017 on the test set.