problem 1

a) $w_i \sim N(0, \tau^2)$

$$p(w \mid 0, \tau^2) = \prod_{i=1}^{d} \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{w_i}{\tau})^2}$$

$$w_{map} = \underset{w}{\arg\max} \left(\sum_{i=1}^{N} \ln p(y^{(i)} \mid x^{(i)}; w) \right) + \ln p(w)$$

$$f = \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi}\delta} \exp\left(\frac{-(y^{(i)} - w^T x^{(i)})^2}{2\delta^2} \right) + \sum_{i=1}^{d} \ln\left(\frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\frac{w_i}{\tau})^2\right) \right)$$

$$= \ln \frac{n}{\sqrt{2\pi}\delta} - \frac{1}{2\delta^2} \sum_{i=1}^{N} (y^{(i)} - w^T x^{(i)})^2 + \ln \frac{d}{\tau\sqrt{2\pi}} - \frac{1}{2\tau^2} \sum_{i=1}^{d} (w_i^2)$$

$$= \ln \frac{n}{\sqrt{2\pi}\delta} - \frac{1}{2\delta^2} (y - xw)^T (y - xw) + \ln \frac{d}{\tau\sqrt{2\pi}} - \frac{1}{2\tau^2} w^T w$$

Take derivative

$$\frac{df}{dw} = \frac{1}{\delta^2}(x^T y - x^T x w) - \frac{1}{\tau^2} w$$

$$= \frac{1}{\delta^2} x^T (y - xw) - \frac{1}{\tau^2} w \overset{set}{=} 0$$

$$\frac{1}{\delta^2} x^T (y - xw) = \frac{1}{\tau^2} w$$

$$x^T y - x^T x w = \frac{\delta^2}{\tau^2} w$$

$$w_{map} = x^T y \cdot (x^T x + \frac{\delta^2}{\tau^2} I)^{-1}$$

From ridge regession. we have similar form

$$w_{RR} = (\lambda I + x^T x)^{-1} \cdot x^T y$$

b) $w^{(i)} \sim \text{Laplace}(0, b)$

$$P(w_i \mid b) = \prod_{i=1}^{d} \frac{1}{2b} e^{-\frac{|w_i|}{b}}$$

$$w_{map} = \arg\max \left(\sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi}\delta} \exp\left(\frac{-(y^{(i)} - w^T x^{(i)})^2}{2\delta^2} \right) + \sum_{i=1}^{d} \ln \frac{1}{2b} \exp\left(-\frac{|w_i|}{b}\right) \right)$$

$$= \arg\max \left(-\frac{1}{2\delta^2} \sum_{i=1}^{N} (y^{(i)} - w^T x^{(i)})^2 - \frac{1}{b} \sum_{i=1}^{d} |w_i| \right)$$

$$= \arg\min \left(\sum_{i=1}^{N} (y^{(i)} - w^T x^{(i)})^2 + \frac{2\delta^2}{b} \|w\| \right)$$

From Lasso regession, we see similar formula

$$\hat{w}_{Lasso} = \arg\min \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - \hat{w}^T x^{(i)})^2 + \lambda \|w\|$$

problem 2.

(1) From bayes rule:

$$f_{bayes}(x) = \arg\max p(x|y) \cdot pr(y)$$ where $p(x|y)$ is class conditional density and $pr(y)$ is class prior

Assume we have data
$\{(x^{(i)}, y^{(i)})\}_{i=1}^{N} = \{(x^{(1)}, y^{(1)}), \cdots (x^{(N)}, y^{(N)})\}$
$x^{(i)} \in X \subseteq R^{d}$ $y^{(i)} \in Y (:=\{0,1\})$

Then from Gaussian Assumption

$$p(x,y) = p(y)p(x|y) = \begin{cases} p_0 \frac{1}{\sqrt{2\pi}\delta_0} e^{\frac{(x-\mu_0)^2}{2\delta_0^2}} & \text{if } y=0 \\ p_1 \frac{1}{\sqrt{2\pi}\delta_1} e^{-\frac{(x-\mu_0)^2}{2\delta_1^2}} & \text{if } y=1 \end{cases}$$

The Bayes optimal one under the assumed joint distribution depends on

$$\mathbb{1}(Pr(y=1|x) \geq Pr(y=0|x)) \overset{bayes}{\Rightarrow} \mathbb{1}(p(x|y=1)\,pr(y=1) \geq p(x|y=0)\,Pr(y=0))$$

$$\Rightarrow \mathbb{1}\left(-\frac{(x-\mu_1)^2}{2\delta_1^2} - \log\sqrt{2\pi}\,\delta_1 + \log p_1 \geq -\frac{(x-\mu_0)^2}{2\delta_0^2} - \log\sqrt{2\pi}\,\delta_0 + \log p_0\right)$$

$$\Rightarrow \mathbb{1}(ax^2+bx+c \geq 0)$$ So the decision bondary is not linear

In matrix form

Gaussian Assumption

$$p(x|y;\mu_y,\Sigma_y) = \frac{1}{(2\pi)^{d/2}|\Sigma_y|^{1/2}} \exp\left(\frac{(x-\mu_y)^T(\Sigma_y)^{-1}(x-\mu_y)}{2}\right)$$

let $d_\Sigma(x,\mu) = (x-\mu)^T(\Sigma)^{-1}(x-\mu)$

$$\hat{f}(x) = \mathbb{1}(Pr(y=1|x) > Pr(y=0|x)) \Rightarrow \mathbb{1}\left[\ln\frac{Pr(y=1|x)}{pr(y=0|x)} > 0\right]$$

$$\Rightarrow \mathbb{1}\left(\ln\frac{P(x|y,\mu,\Sigma)\,Pr(y=1)}{P(x|y,\mu,\Sigma)\,Pr(y=0)}\right)$$ $*\ (Pr(y=1)=\pi_1)$

$$\Rightarrow \mathbb{1}\left[\ln\frac{\pi_1}{1-\pi_1} - \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_0|} - \frac{1}{2}(d_{\Sigma_1}(x,\mu_1) - d_{\Sigma_0}(x,\mu_0)) > 0\right]$$

Since $d_\Sigma(x,\mu)$ is in quadratic form, GDA has a quadratic decision boundary.

Further, if we assume $\Sigma_1 = \Sigma_0 = \Sigma$

$$\begin{aligned} d_{\Sigma_1}(x,\mu_1) - d_{\Sigma_0}(x,\mu_0) &= (x-\mu_1)^T(\Sigma)^{-1}(x-\mu_1) - (x-\mu_0)^T(\Sigma)^{-1}(x-\mu_0) \\ &= \cancel{x^T\Sigma^{-1}x} - x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x + \mu_1^T\mu_1\Sigma^{-1} \\ &- \cancel{x^T\Sigma^{-1}x} + x^T\Sigma^{-1}\mu_0 + \mu_0^T\Sigma^{-1}x - \mu_0^T\Sigma^{-1}\mu_0 \\ &= x^T\Sigma^{-1}(\mu_0-\mu_1) - \mu_1^T\Sigma^{-1}\mu_1 + \mu_2^T\Sigma^{-1}\mu_2 \end{aligned}$$

So it is linear in x.

(2) From Gaussian Discriminant Analysis

$$f(x) = \underset{y \in \{0,1\}}{\arg\max}\ Pr(y|x) = \underset{y \in \{0,1\}}{\arg\max}\ \underbrace{Pr(y)}_{\pi_y}\ P(x \mid y ; \mu_y, \Sigma_y)$$

$$= \arg\max_y \frac{1}{(2\pi)^{d/2} |\Sigma_y|^{1/2}} \exp\{ -\frac{1}{2} (x-\mu_y)^T \Sigma_y^{-1} (x-\mu_y) \} \cdot \pi_y$$

$$= \mathbb{I}[Pr(y=1|x) > Pr(y=0|x)]$$

$$= \mathbb{I}[\ln\frac{\pi_1}{1-\pi_1} - \frac{1}{2} \ln \frac{\Sigma_1}{\Sigma_0} - \frac{1}{2} (d_{\Sigma_1}(x, \mu_1) - d_{\Sigma_0}(x, \mu_0)) > 0]$$

Since $\Sigma_1\ \Sigma_0$ are diagonal. $|\Sigma_1| = \prod_{k=1}^{d} \sigma_{k,y=1}$ $|\Sigma_0| = \prod_{k=1}^{d} \sigma_{k,y=0}$

$$d_{\Sigma_y}(x, \mu) = \underset{1\times n}{(x-\mu)^T}\ \underset{n\times n}{(\Sigma)^{-1}}\ \underset{n\times 1}{(x-\mu)} = \sum_{k=1}^{d} (\frac{x-\mu}{\sigma})^2$$

$$= \mathbb{I}[\ln\frac{\pi_1}{1-\pi_1} - \frac{1}{2}\sum_{k=1}^{d} \ln \frac{\sigma_{k,y=1}}{\sigma_{k,y=0}} - \frac{1}{2} (\sum_{k=1}^{d} (\frac{x_k-\mu_{k,y=1}}{\sigma_{k,y=1}})^2 - \sum_{k=1}^{d} (\frac{x_k-\mu_{k,y=0}}{\sigma_{k,y=0}})^2) > 0]$$

From Gaussian Naive Bayes classifier

$$f(x) = \underset{y \in y}{\arg\max} \prod_{k=1}^{d} P(x_k \mid y ; \mu_k, \sigma_k^2) \cdot \underbrace{Pr(y)}_{\pi_y} \qquad x_k \in \{0,1\}$$

$$= \mathbb{I}[\prod_{k=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2} (\frac{x_k-\mu_{k,y=1}}{\sigma_{k,y=1}})^2\} \cdot \pi_1 > \prod_{k=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2} (\frac{x_k-\mu_{k,y=0}}{\sigma_{k,y=0}})^2\} \cdot \pi_0]$$

$$= \mathbb{I}[\ln \frac{\prod_{k=1}^{d} \frac{1}{\sigma_{y=1}\sqrt{2\pi}} \exp\{\frac{1}{2} (\frac{x_k-\mu_{y=1}}{\sigma_{y=1}})^2\} \pi_1}{\prod_{k=1}^{d} \frac{1}{\sigma_{y=0}\sqrt{2\pi}} \exp\{\frac{1}{2} (\frac{x_k-\mu_{y=0}}{\sigma_{y=0}})^2\} \pi_2} > 0]$$

$$= \mathbb{I}[\ln\frac{\pi_1}{1-\pi_1} - \frac{1}{2}\sum_{k=1}^{d} \ln \frac{\sigma_{k,y=0}}{\sigma_{k,y=1}} - \frac{1}{2} (\sum_{k=1}^{d} (\frac{x_k-\mu_{k,y=0}}{\sigma_{k,y=0}})^2 - \sum_{k=1}^{d} (\frac{x_k-\mu_{k,y=0}}{\sigma_{k,y=0}})^2) > 0]$$

Therefore, when the covariance matrix Σ_y is diagonal
Gaussian NBC is a special case of Gaussian Discriminant Analysis.

problem 3

a) optimization function :

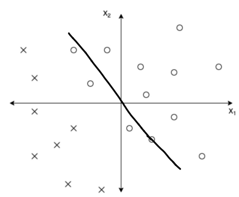
$$\sum_{i=1}^{N} \log(P(y^{(i)} \mid x^{(i)}; w_0, w_1, w_2)) - \lambda w_j^2$$

when λ is very large, the penalty term has large impact. To minimize the cost function, w_j has to be 0.

when $w_0 = 0$,

train error increase,

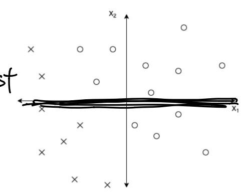
arround 3-4 points misclassified.



when $w_1 = 0$

train error increase most

arrond 6 points misclassified.



when $w_2 = 0$

train error increase

arrond 3 points misclassified.