1.

(a) $E[(Y - \hat{f}(x;D))^2] = E[(Y - f(x) + f(x) - \hat{f}(x;D))^2]$

$= E[(Y - f(x))^2] + E[(f(x) - \hat{f}(x;D))^2]$

$+ E[(Y - f(x))(f(x) - \hat{f}(x;D))]$

① : $E[(Y - f(x))^2] = \mathrm{Var}(\varepsilon)$

② : $E[(f(x) - \hat{f}(x;D))] = E[(f(x) - E_D[\hat{f}(x;D)|x] + E_D[\hat{f}(x;D)|x] - \hat{f}(x;D))^2]$

$= E[(f(x) - E_D[f(x;D)|x])^2]$ ⟶ bias

$+ E[(E_D[\hat{f}(x;D)|x] - \hat{f}(x;D))^2]$ ⟶ var.

③ $= E[(Y - f(x))(f(x) - \hat{f}(x;D))]$

$= E[E[(Y - f(x))(f(x) - \hat{f}(x;D)) \mid X;D]]$

$= E[(f(x) - f(x))(f(x) - \hat{f}(x;D))]$

$= 0$

⇒ ① + ② + ③ $= \mathrm{Var}(\varepsilon) + E[f(x) - E[f(x) - E[\hat{f}(x;D)|x$

$+ E[(\hat{f}(x;D) - E[\hat{f}(x;D)|x])^2]$ #

(b) As the flexibility increase, the model has smaller bias and higher variance. #

2.

(a) $L = (Y - X\beta)^T (Y - X\beta)$

$\frac{\partial L}{\partial \beta} = 0 \longrightarrow -2X^TY + 2X^TX\beta + 2\lambda\beta = 0$

$\longrightarrow \hat{\beta} = (X^TX)^{-1}XY$ #

(b)

unbased: $E[\hat{\beta} - \beta] = E[\hat{\beta}] - \beta = E[(X^TX)^{-1}X(X\beta + \varepsilon)] - \beta$

zero mean Gaussian.

$= \beta - \beta = 0$ #

variance:

$Var(\hat{\beta}) = SE(\hat{\beta})^2$

$= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T]$

$= E[((X^TX)^{-1}X^TY - \beta)((X^TX)^{-1}X^TY - \beta)^T]$

$= E[((X^TX)^{-1}X^T(X\beta + \varepsilon) - \beta)((X^TX)^{-1}X^T(X\beta - \varepsilon) - \beta)^T]$

$= E[((X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\varepsilon - \beta)((X^TX)^{-1}X^TX\beta - (X^TX)^{-1}X^T\varepsilon - \beta)^T]$

$= E[(\beta + (X^TX)^{-1}X^T\varepsilon - \beta)(\beta + (X^TX)^{-1}X^T\varepsilon - \beta)^T]$

$= E[((X^TX)^{-1}X^T\varepsilon)((X^TX)^{-1}X^T\varepsilon)^T]$

$= \sigma^2(X^TX)^{-1}$ #

3.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

(a) $D \xrightarrow{3\text{-fold}} D_1, D_2, D_3$ with size $\frac{n}{k} = \frac{3}{3} = 1$

Let $D_1 = \{(1,2)\}$ be validation set

Training set : $D \setminus D_1 = \{(3,5), (-2,-5)\}$

$$\hat{\beta}_1^{(D\setminus D_1)} = \frac{25 - 2\times\frac{1}{2}\times 0}{13 - 2\times\frac{1}{2}\times\frac{1}{2}} = 2, \quad \hat{\beta}_0^{(D\setminus D_1)} = 0 - 2\times\frac{1}{2} = -1, \quad MSE_1 = \left(y - \hat{\beta}_0^{(D\setminus D_1)} - \hat{\beta}_1^{(D\setminus D_1)} x\right)^2$$

$$= \left(2 - (-1) - 2\times 1\right)^2 = 1$$

Let $D_2 = \{(3,5)\}$ be validation set

Training set : $D \setminus D_2 = \{(1,2), (-2,-5)\}$

$$\hat{\beta}_1^{(D\setminus D_2)} = \frac{12 - 2\times -\frac{1}{2}\times -\frac{3}{2}}{5 - 2\times -\frac{1}{2}\times -\frac{1}{2}} = \frac{10.5}{4.5} = \frac{7}{3}, \quad \hat{\beta}_0 = -\frac{3}{2} - \frac{7}{3}\times\left(-\frac{1}{2}\right) = -\frac{1}{3}, \quad MSE_2 = \left(y - \hat{\beta}_0^{(D\setminus D_2)} - \hat{\beta}_1^{(D\setminus D_2)} x\right)^2$$

$$= \left(5 - \left(-\frac{1}{3}\right) - \frac{7}{3}\times 3\right)^2 = \frac{25}{9}$$

Let $D_3 = \{(-2,-5)\}$ be validation set

Training set : $D \setminus D_3 = \{(1,2), (3,5)\}$

$$\hat{\beta}_1^{(D\setminus D_3)} = \frac{17 - 2\times 2\times\frac{7}{2}}{10 - 2\times 2\times 2} = \frac{3}{2}, \quad \hat{\beta}_0^{(D\setminus D_3)} = \frac{7}{2} - \frac{3}{2}\times 2 = \frac{1}{2}, \quad MSE_3 = \left(y - \hat{\beta}_0^{(D\setminus D_3)} - \hat{\beta}_1^{(D\setminus D_3)} x\right)^2$$

$$= \left(-5 - \frac{1}{2} - \frac{3}{2}\times(-2)\right)^2 = \frac{25}{4}$$

$$CV_{(3)} = \frac{1}{3}\sum_{j=1}^{3} MSE_j = \frac{1}{3}\left(1 + \frac{25}{9} + \frac{25}{4}\right) = \frac{361}{108}$$

(b)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\beta}_1^{*1} = \frac{19 - 3\times\frac{5}{3}\times 3}{11 - 3\times\frac{5}{3}\times\frac{5}{3}} = \frac{4}{\frac{8}{3}} = \frac{3}{2}, \quad \hat{\beta}_1^{*2} = \frac{40 - 3\times\frac{4}{3}\times\frac{5}{3}}{22 - 3\times\frac{4}{3}\times\frac{4}{3}} = 2, \quad \hat{\beta}_1^{*3} = \frac{22 - 3\times 1\times\frac{8}{3}}{9 - 3\times 1\times 1} = \frac{14}{6} = \frac{7}{3}$$

$$Var(\hat{\beta}_1) = SE_B(\hat{\beta}_1)^2 = \frac{1}{3-1}\sum_{r=1}^{3}\left(\hat{\beta}_1^{*r} - \underbrace{\frac{1}{3}\sum_{r=1}^{3}\hat{\beta}_1^{*r}}_{\frac{35}{18}}\right)^2 = \frac{1}{2}\cdot\left(\frac{64}{18^2} + \frac{1}{18^2} + \frac{49}{18^2}\right) = \frac{57}{324}$$

4.

$$l(B_0, B_1) = \prod_{i=1}^{n} P(Y = y_i \mid X = x_i; B_0, B_1)$$

$$= \prod_{i:y_i=1}^{n} P(x_i; B_0 B_1) \prod_{i:y_i=0}^{n} (1 - P(x_i; B_0 B_1))$$

$$\longrightarrow \log l(B_0, B_1) = \sum_{i=1}^{n}\left[y_i \log P(x_i; B_0, B_1) + (1-y_i)\log(1 - P(x_i; B_0, B_1))\right]$$

$$= \sum_{i=1}^{n}\left[y_i \log \frac{1}{1 + e^{-(2y_i-1)(B_0+B_1 x_i)}} + (1-y_i)\log \frac{1}{1 + e^{-(2y_i-1)(B_0+B_1 x_i)}}\right]$$

$$= \sum_{i=1}^{n} \log \frac{1}{1 + e^{-(2y_i-1)(B_0+B_1 x_i)}}$$

$$\longrightarrow \hat{B} = \arg\max_{\beta} \sum_{i=1}^{n} \log \frac{1}{1 + e^{-(2y_i-1)(B_0+B_1 x_i)}}$$

$$= \arg\min_{\beta} \sum_{i=1}^{n} \log 1 + e^{-(2y_i-1)(B_0+B_1 x_i)}$$

#

5.

Start at $\underline{\beta}(k)$. Take a small step of size $\eta$ in the direction $\underline{v}$, where $\|\underline{v}\| = 1$

$\Rightarrow$ Arrive at $\underline{\beta}(k) + \eta \underline{v}$

Find $\underline{v}$ such that the reduction from $J(\beta(k))$ to $J(\beta(k) + \eta \underline{v})$ is largest

$\Rightarrow J(\underline{\beta}(k)) - J(\beta(k) + \eta \underline{v})$

$= J(\beta(k)) - [J(\beta(k)) + \eta \underline{v}^T \nabla J(\beta(k)) + O(\eta^2)]$

$= -\eta \underline{v}^T \nabla J(\beta(k)) - O(\eta^2) \leq \eta \frac{\nabla J(\beta(k))^T}{\|\nabla J(\beta(k))\|} \nabla J(\beta(k)) - O(\eta^2) \Rightarrow \underline{v} = \frac{-\nabla J(\beta(k))}{\|\nabla J(\beta(k))\|}$

$J(\beta_0, \beta_1) = (y - \beta_0 - \beta_1 x)^2$

$\frac{\partial J}{\partial \beta_0} = -2(y - \beta_0 - \beta_1 x)$

$\frac{\partial J}{\partial \beta_1} = -2x(y - \beta_0 - \beta_1 x)$

$\frac{\partial J}{\partial \beta} = \left[\frac{\partial J}{\partial \beta_0}, \frac{\partial J}{\partial \beta_1}\right]^T$, $\left\|\frac{\partial J}{\partial \beta}\right\| = \sqrt{\left(\frac{\partial J}{\partial \beta_0}\right)^2 + \left(\frac{\partial J}{\partial \beta_1}\right)^2} = \sqrt{4(y - \beta_0 - \beta_1 x)^2 + 4x^2(y - \beta_0 - \beta_1 x)^2}$

$= \sqrt{4(y - \beta_0 - \beta_1 x)^2 (1 + x^2)}$

$= 2|y - \beta_0 - \beta_1 x|\sqrt{(1 + x^2)}$

$\underline{v} = \frac{-\frac{\partial J}{\partial \beta}}{\left\|\frac{\partial J}{\partial \beta}\right\|} = \left[\frac{2(y - \beta_0 - \beta_1 x)}{2|y - \beta_0 - \beta_1 x|\sqrt{1 + x^2}}, \frac{2x(y - \beta_0 - \beta_1 x)}{2|y - \beta_0 - \beta_1 x|\sqrt{1 + x^2}}\right]^T$

$\Rightarrow \underline{v} = \begin{cases} \left[\frac{1}{\sqrt{1 + x^2}}, \frac{x}{\sqrt{1 + x^2}}\right]^T, & \text{if } y - \beta_0 - \beta_1 x > 0 \\ \left[\frac{-1}{\sqrt{1 + x^2}}, \frac{-x}{\sqrt{1 + x^2}}\right]^T, & \text{if } y - \beta_0 - \beta_1 x < 0 \end{cases}$

6.

(a) Probability of classification error:

$$P(\hat{Y}(X) \neq Y) = \sum_{k=1}^{K} P(Y=k)\left(1 - \underbrace{P(\hat{Y}(X)=k \mid Y=k)}_{=P(X \in R_k \mid Y=k)}\right)$$

$$= 1 - \sum_{k=1}^{K} P(Y=k) \int_{R_k} f_k(x)\, dx$$

$$= 1 - \sum_{k=1}^{K} \int_{R_k} P(Y=k \mid X=x) f(x)\, dx$$

To minimize the probability of classification error, we should let $x \in R_k$ whenever $P(\hat{Y}(x)=k \mid X=x) > P(\hat{Y}(x)=l \mid X=x), \forall l \neq k$

(b)

$$\hat{Y}(x) = \arg\max_{k \in \{0,1\}} P_k(x) = \arg\max_{k \in \{0,1\}} \frac{\pi_k \frac{1}{(2\pi)^{p/2}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right)}{\sum_{k=0}^{1} \pi_k \frac{1}{(2\pi)^{p/2}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right)}$$

$$= \arg\max_{k \in \{0,1\}} \log \pi_k - \underbrace{\log (2\pi)^{p/2}|\Sigma|^{\frac{1}{2}}}_{\text{定值}} - \frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)$$

$$= \arg\max_{k \in \{0,1\}} -\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) + \log \pi_k$$

$$= \arg\max_{k \in \{0,1\}} -\frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_k + \log \pi_k$$

$$= \arg\max_{k \in \{0,1\}} x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k$$

class 1: $(\frac{3}{2}, \frac{3}{2})$, $(1,1)$, $(\frac{1}{2}, \frac{1}{2})$

class 0: $(0,1)$, $(0,0)$, $(0,-1)$

$\mu_1 = [1,1]^T$, $\mu_0 = [0,0]^T$, $\pi_1 = \pi_0 = \frac{1}{2}$

$$\hat{\Sigma} = \frac{1}{6-2}\left( \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}\begin{bmatrix} \frac{1}{2}, \frac{1}{2} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}\begin{bmatrix} 0, 0 \end{bmatrix} + \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}\begin{bmatrix} -\frac{1}{2}, -\frac{1}{2} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}\begin{bmatrix} 0, 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}\begin{bmatrix} 0, 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}\begin{bmatrix} 0, -1 \end{bmatrix} \right)$$

$$= \frac{1}{4}\left( \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix} + \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

$$= \begin{bmatrix} \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{5}{8} \end{bmatrix} \qquad \hat{\Sigma}^{-1} = 16\begin{bmatrix} \frac{5}{8} & -\frac{1}{8} \\ -\frac{1}{8} & \frac{1}{8} \end{bmatrix} = \begin{bmatrix} 10 & -2 \\ -2 & 2 \end{bmatrix}$$

$$\hat{\delta}_0(x) = \left[\frac{2}{3}, -1\right]\begin{bmatrix} 10 & -2 \\ -2 & 2 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 0, 0 \end{bmatrix}\begin{bmatrix} 10 & -2 \\ -2 & 2 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \log\frac{1}{2}$$

$$= \log\frac{1}{2}$$

$$\hat{\delta}_1(x) = \left[\frac{2}{3}, -1\right]\begin{bmatrix} 10 & -2 \\ -2 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 1, 1 \end{bmatrix}\begin{bmatrix} 10 & -2 \\ -2 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \log\frac{1}{2}$$

$$= \frac{28}{3} - 4 + \log\frac{1}{2} = \frac{16}{3} + \log\frac{1}{2}$$

$$\hat{\delta}_1 > \hat{\delta}_0 \quad \Rightarrow \quad \hat{\gamma}(x) = 1$$

(c)

$$\sqrt{\left(\frac{2}{3} - \frac{3}{2}\right)^2 + \left(-1 - \frac{3}{2}\right)^2} = \frac{5\sqrt{10}}{6} \qquad \sqrt{\left(\frac{2}{3} - 0\right)^2 + (-2)^2} = \frac{\sqrt{40}}{3}$$

$$\sqrt{\left(\frac{2}{3} - 1\right)^2 + (-2)^2} = \frac{\sqrt{37}}{3} \qquad \sqrt{\left(\frac{2}{3} - 0\right)^2 + (-1 - 0)^2} = \frac{\sqrt{13}}{3} \qquad \hat{\gamma}(x) = 0$$

$$\sqrt{\left(\frac{2}{3} - \frac{1}{2}\right)^2 + \left(-1 - \frac{1}{2}\right)^2} = \frac{\sqrt{82}}{6} \qquad \sqrt{\left(\frac{2}{3} - 0\right)^2 + 0^2} = \frac{2}{3}$$

7.

(a) As k increases, the bias of the test error estimate decreases, but the variance increases.

(b) QDA will perform better

Since the solution of logistic regression is $\hat{\beta}^T \underline{x}$, which is different from the true solution $\beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_3 x_1 x_2$

QDA will be more flexible on this dataset.

(c) If j-th predictor is essential in predicting $Y$, $t_j$ should be larger, since the key idea of t-statistic is to observe how far $\hat{\beta}_1$ is to 0 relative to $\hat{SE}(\hat{\beta}_j)$