ML page 1

1. feature $x \in \mathbb{R}^n$

posterior $p(G|\vec{x})$

prior $\pi_1 + \pi_2 = 1$

$f_1(\vec{x}) = p(\vec{X}=\vec{x} \mid G=1)$

$f_2(\vec{x}) = p(\vec{X}=\vec{x} \mid G=2)$

If class2 ~~is~~ classified. it means

$$f_2(\vec{x}) > f_1(\vec{x}) \Rightarrow \log\frac{f_2(\vec{x})}{f_1(\vec{x})} > 0$$

since $f_k(\vec{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_k)^T \Sigma_k^{-1} (\vec{x}-\vec{\mu}_k)\right\}$

$\Sigma_1 = \Sigma_2$ by assumption.

$$f_1(\vec{x}) = (2\pi)^{-p}\Sigma^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_1)\Sigma^{-1}(\vec{x}-\vec{\mu}_1)\right\}.$$

$$\cancel{\log\frac{\log f_2(\vec{x})}{\log f_1(\vec{x})}} \quad \log\frac{f_2(\vec{x})\pi_2}{f_1(\vec{x})\pi_1} = \log[(2\pi)^{-p}\Sigma^{-\frac{1}{2}}] - \frac{1}{2}(\vec{x}-\vec{\mu}_1)\Sigma^{-1}(\vec{x}-\vec{\mu}_1) - \log[(2\pi)^{-p}\Sigma^{-\frac{1}{2}}] + \frac{1}{2}(\vec{x}-\vec{\mu}_2)\Sigma^{-1}(\vec{x}-\vec{\mu}_2) - \log\frac{\pi_2}{\pi_1}.$$

since $\pi_2 = \frac{N_2}{N}$

$\pi_1 = \frac{N_1}{N}$. $\Rightarrow \log\frac{\pi_2}{\pi_1} = \log\frac{N_2}{N_1}$

$$\log\frac{f_2(\vec{x})\pi_2}{f_1(\vec{x})\pi_1} > 0 \iff \boxed{-\frac{1}{2}\left[(\vec{x}-\vec{\mu}_1)^T\Sigma^{-1}\vec{x} - (\vec{x}-\vec{\mu}_1)^T\Sigma^{-1}\vec{\mu}_1 - (\vec{x}-\vec{\mu}_2)^T\Sigma^{-1}\vec{x} + (\vec{x}-\vec{\mu}_2)^T\Sigma^{-1}\vec{\mu}_2\right] - \log\frac{N_2}{N_1} > 0.}$$

$$\cancel{\Rightarrow = \frac{1}{2}\left[(-\vec{\mu}_1+\vec{\mu}_2)\Sigma^{-1}\vec{x}\right.}$$

① Prove $(\vec{\mu_2}-\vec{\mu_1})^T\Sigma^{-1}\vec{X} = \vec{X}\Sigma^{-1}(\vec{\mu_2}-\vec{\mu_1})$

$\mu_1 . \mu_2 \in \mathbb{R}^p$ $\quad \vec{X}\in\mathbb{R}^{p\times 1}$ $\quad \Sigma^{-1}\in\mathbb{R}^{p\times p}$

$$\begin{bmatrix} \mu_{21}-\mu_{11} \\ \mu_{22}-\mu_{12} \\ \vdots \\ \mu_{2p}-\mu_{1p} \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & & \\ \vdots & & \ddots & \\ \sigma_{2p} & & & \sigma_p^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

$$\begin{bmatrix} \mu_{21}-\mu_{11} & \mu_{22}-\mu_{12} & \cdots & \mu_{2p}-\mu_{1p} \end{bmatrix} \begin{bmatrix} a & b & c & \cdots \\ \vdots & & & \\ \vdots & & \cdots & \cdots \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

$$\uparrow \vec{a_1} \quad \uparrow \vec{a_2} \quad \cdots \quad \uparrow \vec{a_p}$$

$$= \begin{bmatrix} (\mu_{21}-\mu_{11})\vec{a_1} & (\mu_{22}-\mu_{12})\vec{a_2} & \cdots & (\mu_{2p}-\mu_{1p})a_p \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

$$\Longleftrightarrow \begin{bmatrix} x_1\vec{a_1} & x_2\vec{a_2} & \cdots & x_p\vec{a_p} \end{bmatrix} \begin{bmatrix} \mu_{21}-\mu_{11} \\ \vdots \\ \mu_{2p}-\mu_{1p} \end{bmatrix}$$

$$\Rightarrow (\vec{\mu_2}-\vec{\mu_1})^T\Sigma^{-1}\vec{X} = \vec{X}\Sigma^{-1}(\vec{\mu_2}-\vec{\mu_1})$$

② ~~Prove~~ Similarly $\quad \vec{\mu_2}^T\Sigma^{-1}\vec{\mu_1} = \vec{\mu_1}^T\Sigma^{-1}\vec{\mu_2}$

Continous prove the final result from last page:

$$\Rightarrow -\frac{1}{2}\left[(\mu_2^T-\mu_1^T)\Sigma^{-1}\vec{X} + \vec{X}\Sigma^{-1}(\vec{\mu_2}-\vec{\mu_1}) + \mu_1^T\Sigma^{-1}\vec{\mu_1} - \mu_2^T\Sigma^{-1}\mu_2\right] - \log\frac{N_2}{N_1} > 0$$

$$\Rightarrow X^T\Sigma^{-1}(\mu_2-\mu_1) \geq \frac{1}{2}\left[\vec{\mu_2}^T\Sigma^{-1}\vec{\mu_2} - \vec{\mu_1}^T\Sigma^{-1}\vec{\mu_1} - \vec{\mu_2}^T\Sigma^{-1}\vec{\mu_1} + \vec{\mu_1}\Sigma^{-1}\vec{\mu_2}\right] - \log\frac{N_2}{N_1}$$

$$\Rightarrow X^T\Sigma^{-1}(\vec{\mu_2}-\vec{\mu_1}) \geq \frac{1}{2}(\hat{\mu_2}+\hat{\mu_1})^T\Sigma^{-1}(\hat{\mu_2}-\hat{\mu_1}) - \log\frac{N_2}{N_1}$$

b). Consider the normal equation for minimizing $\sum_{i=1}^{N}(y_i-\beta_0-\beta^T x_i)^2$

i.e. $X^TX\vec{\beta}=X^TY$ $\quad x_i\in\mathbb{R}^p.\quad \mu_1, \mu_2\in\mathbb{R}^p$

where
$$X=\begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_{N_1}^T \\ 1 & x_{N_1+1}^T \\ 1 & x_{N_1+N_2}^T \end{bmatrix}\begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\1\\1\end{matrix}}\right\} k=1 \\ \left.\vphantom{\begin{matrix}1\\1\end{matrix}}\right\} k=2 \end{matrix} \qquad Y=\begin{bmatrix} -\frac{N}{N_1} \\ \vdots \\ -\frac{N}{N_1} \\ \frac{N}{N_2} \\ \vdots \\ \frac{N}{N_2} \end{bmatrix}$$

x_1^T to $x_{N_1}^T$ are corresponding to the data in class 1.

$x_{N_1+1}^T$ to $x_{N_1+N_2}^T$ are corresponding to the data in class 2.

$$X^TX=\begin{bmatrix} 1 & \cdots & \cdots & & 1 \\ x_1 & x_2 & x_3 & \cdots & x_{N_1+N_2}\end{bmatrix}\begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_{N_1+N_2}^T\end{bmatrix}=\begin{bmatrix} N_1+N_2 & \sum_{i=1}^{N}x_i^T \\ \sum_{i=1}^{N}x_i & \sum_{i=1}^{N}x_ix_i^T\end{bmatrix}$$

Let $\hat{\Sigma}$ denotes the estimation for covariance ~~between X and X~~ of N sample points.

By definition
$$\hat{\Sigma}=\frac{1}{N-2}\left[\sum_{i=1}^{N_1}(x_i-\mu_1)(x_i-\mu_1)^T+\sum_{i=N_1+1}^{N_1+N_2}(x_i-\mu_2)(x_i-\mu_2)^T\right]$$
$$=\frac{1}{N-2}\left[\sum_{i=1}^{N_1}\left(x_ix_i^T-x_i\mu_1^T-\mu_1x_i^T+\mu_1\mu_1^T\right)+\sum_{i=N_1+1}^{N_1+N_2}\cancel{x_i}\left(x_ix_i^T-x_i\mu_2^T-\mu_2x_i^T+\mu_2\mu_2^T\right)\right]$$

Note ~~that $x_i\mu_1^T=\mu_1x_i^T$~~

since
$$x_i\mu_1^T=\begin{bmatrix}x_{i1}\\x_{i2}\\\vdots\\x_{ip}\end{bmatrix}\begin{bmatrix}\mu_{11} & \mu_{12} & \cdots & \mu_{1p}\end{bmatrix}=\begin{bmatrix}x_{i1}\mu_{11} & x_{i1}\mu_{12} & \cdots & x_{i1}\mu_{1p}\\ x_{i2}\mu_{11} & & & \\ x_{ip}\mu_{11} & & & x_{ip}\mu_{1p}\end{bmatrix}.$$

$$=\begin{bmatrix}\mu_{11}x_{i1} & \mu_{12}x_{i1} & \cdots & \mu_{1p}x_{i1}\\ \mu_{11}x_{i2} & & & \\ \mu_{11}x_{ip} & & & \mu_{1p}x_{ip}\end{bmatrix}=$$

$$\begin{bmatrix}\mu_{11}x_{i1}\\ \mu_{12}x\\ \\ \end{bmatrix}$$

$$\begin{bmatrix}\mu_{11}\\ \mu_{12}\\ \vdots\\ \mu_{1p}\end{bmatrix}\begin{bmatrix}x_{i1} & x_{i2} & \cdots & x_{ip}\end{bmatrix}=\mu_1x_i^T$$

$$= \frac{1}{N-2}\left[\sum_{i=1}^{N_1} x_i x_i^T - \sum_{i=1}^{N_1} x_i \mu_1^T - \sum_{i=1}^{N_1} \mu_1 x_i^T + N_1 \mu_1 \mu_2^T + \right.$$
$$\left. \sum_{i=N_1+1}^{N_1+N_2} x_i x_i^T - \sum_{i=N_1+1}^{N_1+N_2} x_i \mu_2^T - \sum_{i=N_1+1}^{N_1+N_2} \mu_2 x_i^T + N_2 \mu_2 \mu_2^T \right]$$

$$= \frac{1}{N-2}\left[\sum_{i=1}^{N_1} x_i x_i^T - \cancel{N_1\mu_1^T}\ N_1\mu_1\mu_1^T - N_1\mu_1\mu_1^T + N_1\mu_1\mu_2^T + \right.$$
$$\left. \sum_{i=N_1+1}^{N_1+N_2} x_i x_i^T - N_2\mu_2\mu_2^T - N_2\mu_2\mu_2^T - N_2\mu_2\mu_1^T \right]$$

$$= \frac{1}{N-2}\left[\sum_{i=1}^{N_1} x_i x_i^T + \sum_{i=N_1+1}^{N_1+N_2} x_i x_i^T - N_1\mu_1\mu_1^T - N_2\mu_2\mu_2^T \right]$$

$$\Rightarrow \hat{\Sigma} = \frac{1}{N-2}\left[\sum_{i=1}^{N} x_i x_i^T - N_1\mu_1\mu_1^T - N_2\mu_2\mu_2^T \right]$$

$$\Rightarrow \sum_{i=1}^{N} x_i x_i^T = (N-2)\hat{\Sigma} + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T$$

We have proved $X^TX = \begin{bmatrix} N & \sum_{i=1}^{N} x_i^T \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i x_i^T \end{bmatrix}$

Since $\sum_{i=1}^{N} x_i = \sum_{i=1}^{N_1} x_i + \sum_{i=N_1+1}^{N_1+N_2} x_i = N_1\mu_1 + N_2\mu_2$

Similary $\sum_{i=1}^{N} x_i^T = N_1\mu_1^T + N_2\mu_2^T$

$$\Rightarrow X^TX = \begin{bmatrix} N & N_1\mu_1^T + N_2\mu_2^T \\ N_1\mu_1 + N_2\mu_2 & (N-2)\hat{\Sigma} + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T \end{bmatrix}$$

$\cancel{X^TX}$

For right side of normal equation:

$$X^TY = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{N_1+N_2} \end{bmatrix} \begin{bmatrix} -\frac{N}{N_1} \\ -\frac{N}{N_1} \\ -\frac{N}{N_1} \\ +\frac{N}{N_2} \\ \frac{N}{N_2} \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{N}{N_1}\cdot N_1\mu_1 + \frac{N}{N_2} N_2\mu_2 \end{bmatrix}$$
$$= \begin{bmatrix} 0 \\ -N\mu_1 + N\mu_2 \end{bmatrix}$$

By $X^TX\begin{bmatrix}\beta_0\\ \beta\end{bmatrix} = X^TY$. i.e.

① $$N\beta_0 + (N_1\mu_1^T + N_2\mu_2^T)\beta = 0$$

② $$(N_1\mu_1 + N_2\mu_2)\beta_0 + \left((N-2)\hat{\Sigma} + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T\right)\beta = -\cancel{\tfrac{N}{N_1}}N - N\mu_1 + N\mu_2$$

By ① $\Rightarrow$ $$\beta_0 = \frac{-(N_1\mu_1^T + N_2\mu_2^T)\beta}{N}$$

Plug into ② $\Rightarrow$ $$(N_1\mu_1 + N_2\mu_2)\left(-\frac{N_1}{N}\mu_1^T - \frac{N_2}{N}\mu_2^T\right)\beta + \left[(N-2)\hat{\Sigma} + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T\right]\beta = -N\mu_1 + N\mu_2$$
$$= \underline{N(\mu_2 - \mu_1)}$$

$$\Rightarrow \Big[\underline{-\frac{N_1^2}{N}\mu_1\mu_1^T} - \frac{N_1N_2}{N}\mu_1\mu_2^T - \frac{N_1N_2}{N}\mu_2\mu_1^T \underline{- \frac{N_2^2}{N}\mu_2\mu_2^T}$$
$$\underline{+N_1\mu_1\mu_1^T} + \underline{N_2\mu_2\mu_2^T} + (N-2)\hat{\Sigma}\Big]\beta = N(\mu_2 - \mu_1)$$

Note $\underline{\cancel{A}\ \mu_1\mu_2^T}$ $$= \begin{bmatrix}\mu_{11}\\ \mu_{12}\\ \vdots\\ \mu_{1p}\end{bmatrix}\begin{bmatrix}\mu_{21} & \mu_{22} & \cdots & \mu_{2p}\end{bmatrix} = \begin{bmatrix}\mu_{11}\mu_{21} & \mu_{11}\mu_{22} & \cdots & \mu_{11}\mu_{2p}\\ \mu_{12}\mu_{21} & \mu_{12}\mu_{22} & & \vdots\\ \mu_{13}\mu_{21} & \vdots & & \\ \mu_{1p}\mu_{21} & \mu_{1p}\mu_{22} & & \mu_{1p}\mu_{2p}\end{bmatrix}$$

$$\mu_2\mu_1^T = \begin{bmatrix}\mu_{21}\mu_{11} & \mu_{21}\cancel{\mu_{22}}\mu_{12} & \cdots\\ \mu_{22}\cancel{\mu_{11}}\mu_{11} & \mu_{22} & \\ \vdots & \vdots & \\ \mu_{2p}\cancel{\mu_{11}}\mu_{11} & \mu_{2p}\ \mu_{12} & \end{bmatrix}$$

We have:
$$\left[\left(-\frac{N_1^2}{N} + N_1\right)\mu_1\mu_1^T + \left(-\frac{N_2^2}{N} + N_2\right)\mu_2\mu_1^T - \frac{N_1N_2}{N}\mu_1\mu_2^T - \frac{N_1N_2}{N}\mu_2\mu_1^T + (N-2)\hat{\Sigma}\right]$$
$$\times\beta = N(\mu_2 - \mu_1)$$

$$\Rightarrow \left[\frac{N_1}{N}(-N_1 + N)\mu_1\mu_1^T + \frac{N_2}{N}(-N_2 + N)\mu_2\mu_2^T - \frac{N_1N_2}{N}\mu_1\mu_2^T - \frac{N_1N_2}{N}\mu_2\mu_1^T + (N-2)\hat{\Sigma}\right]$$
$$\times\beta = N(\mu_2 - \mu_1)$$

$$\left(\frac{N_1N_2}{N}\mu_1\mu_1^T + \frac{N_2N_1}{N}\mu_2\mu_2^T - \frac{N_1N_2}{N}\mu_1\mu_2^T - \frac{N_1N_2}{N}\mu_2\mu_1^T + (N-2)\hat{\Sigma}\right)\times\beta = N(\mu_2 - \mu_1)$$

Since $(\mu_2-\mu_1)(\mu_2-\mu_1)^T$

$= \mu_2\mu_2^T - \mu_2\mu_1^T - \mu_1\mu_2^T + \mu_1\mu_1^T$

continue ~~proofing~~

$\Rightarrow \left[\frac{N_1N_2}{N}\left[\mu_1\mu_1^T + \mu_2\mu_2^T - \mu_2\mu_1^T - \mu_1\mu_2^T\right] + (N-2)\hat{\Sigma}\right] * \beta = N(\mu_1-\mu_2)$

$\Rightarrow \left[\frac{N_1N_2}{N}\left[(\mu_2-\mu_1)(\mu_2-\mu_1)^T\right] + (N-2)\hat{\Sigma}\right] \times \beta = N(\mu_1-\mu_2)$

Let $\hat{\Sigma}_B = (\mu_2-\mu_1)(\mu_2-\mu_1)^T$

$\left(\frac{N_1N_2}{N}\hat{\Sigma}_B + (N-2)\hat{\Sigma}\right]\beta = N(\mu_1-\mu_2)$

$\Rightarrow \left((N-2)\hat{\Sigma} + \frac{N_1N_2}{N}\hat{\Sigma}_B\right)\beta = N(\hat{\mu}_1-\hat{\mu}_2)$ where $\hat{\Sigma}_B = (\hat{\mu}_2-\hat{\mu}_1)(\hat{\mu}_2-\hat{\mu}_1)^T$ all estimated value.

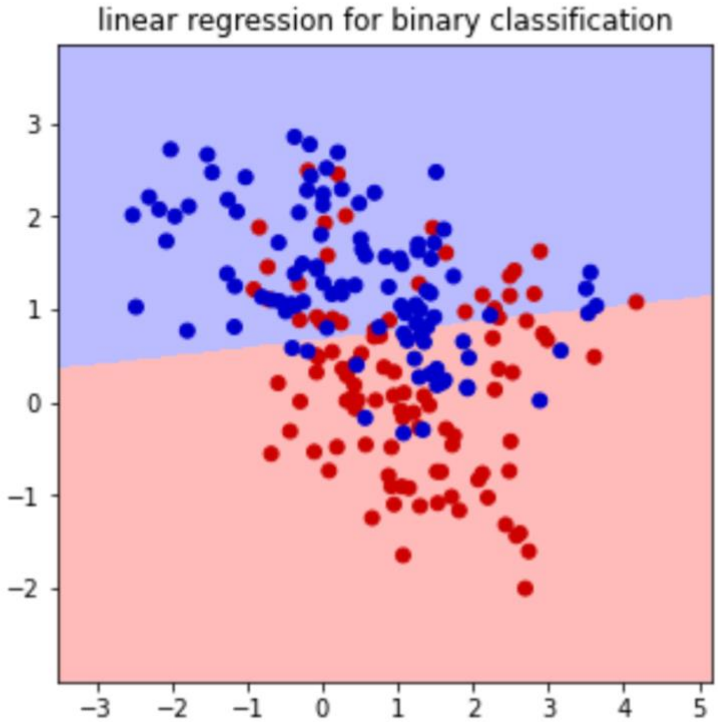
c) : Since $\hat{\Sigma}_B\beta = (\mu_2-\mu_1)\underbrace{(\mu_2-\mu_1)^T\beta}_{\text{scaler}}$

$\Rightarrow \hat{\Sigma}_B\beta$ is in the direction of $(\hat{\mu}_2-\hat{\mu}_1)$

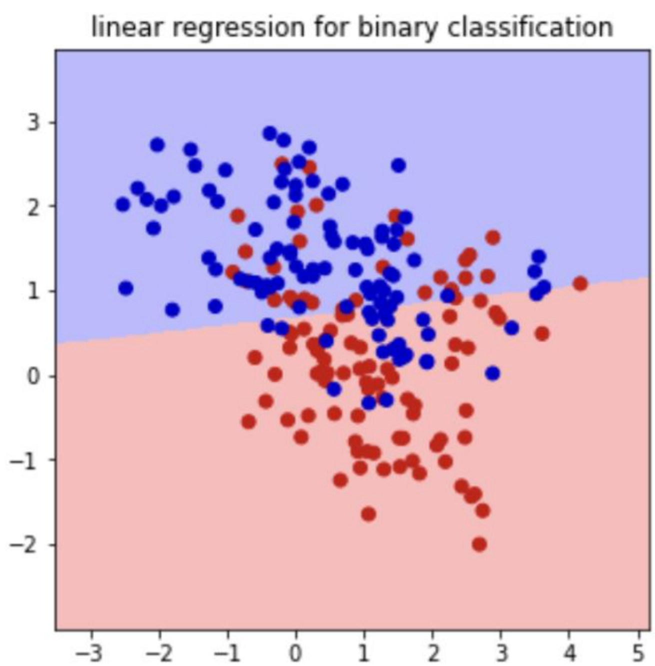
Since the right hand side $N(\hat{\mu}_2-\hat{\mu}_1)$ is in direction of $(\hat{\mu}_2-\hat{\mu}_1)$

$\Rightarrow$ ~~$\hat{\Sigma}$~~ β must be in direction $\Sigma^{-1}(\mu_2-\mu_1)$.

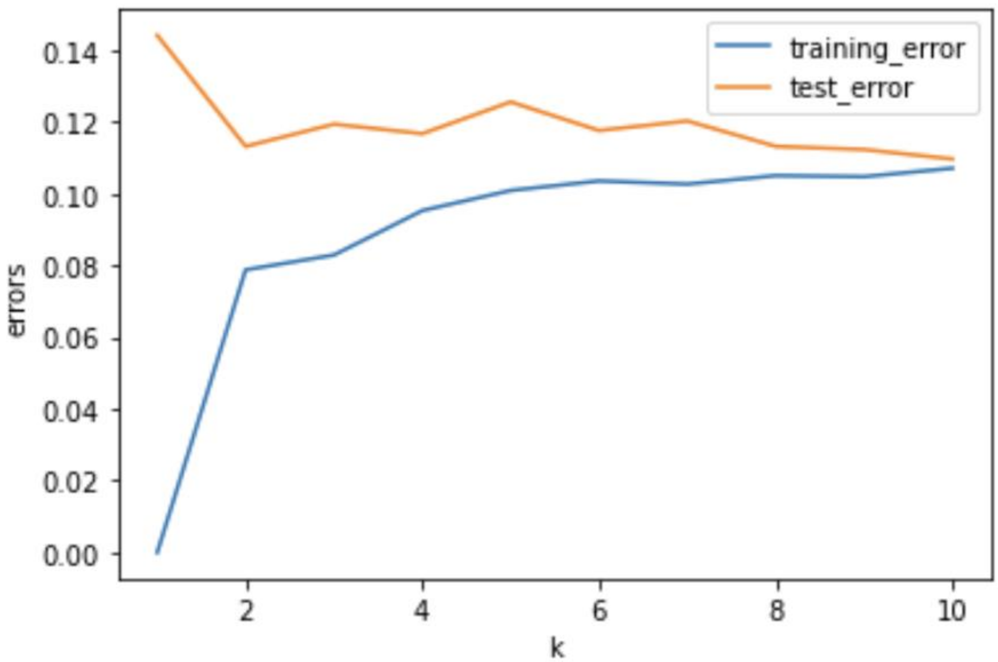
2 a


linear regression for binary classification
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5

2 b


linear regression for binary classification
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5

3b)



When k is small, the model is more complex, the bias is small, but variance is large.
When k gets bigger, the model is simpler. The bias is large by underfit, but variance is small.
That is the bias and variance tradeoff between different ks.
Typically, we would like to choose our complexity to trade bias off with variance in such a way as to minimize the test error. K=2 would be a good choice here.

4.
Error for LDA is 5.48268398
Error for QDA is 3.41774892

Error for LOGISTIC lbfgs is 5.59090909
Error for LOGISTIC sag is 5.79220779
Error for LOGISTIC liblinear is 7.51298701
Error for LOGISTIC saga is 6.31385281

'lbfgs' gives the smallest error