學號：B10705025　系級：資管三　姓名：彭鈞道

1. (1%) 解釋什麼樣的 data preprocessing 可以 improve 你的 training/testing accuracy，e.g.，你怎麼挑掉你覺得不適合的 data points。請提供數據(例如 kaggle public score RMSE)以佐證你的想法。

計算各個 feature 與 PM2.5 的相關係數，只選擇相關係數絕對值大於 0.5 的 feature 放入模型中，分別為 CO, NO, NOx, PM10，與放入前 8 個 feature 的模型在相同條件下比較，可發現前者表現較好。(上圖為挑選過 feature, 下圖為直接使用前 8 個 feature)

my_sol (3).csv Complete · 2d ago	3.88222	3.25476
my_sol (6).csv Complete (after deadline) · 12s ago	5.89725	6.14305

2. (1%) 請實作 2nd-order polynomial regression model (不用考慮交互項)。

a. 貼上 polynomial regression 版本的 Gradient descent code 內容

```python
import numpy as np
import math

def minibatch(x, y, config):

    # Randomize the data in minibatch
    index = np.arange(x.shape[0])
    np.random.shuffle(index)
    x = x[index]
    y = y[index]

    # Initialization
    batch_size = config.batch_size
    lr = config.lr
    lam = config.lam
    epoch = config.epoch

    beta_1 = 0.9
    beta_2 = 0.99
```

```
    # Polynomial regression: three parameters (z, w, b)
    z = np.full(x[0].shape, 0.1).reshape(-1, 1)
    w = np.full(x[0].shape, 0.1).reshape(-1, 1)
    bias = 0.1

    m_t_z = np.zeros(z.shape)
    v_t_z = np.zeros(z.shape)
    m_t_w = np.zeros(w.shape)
    v_t_w = np.zeros(w.shape)
    m_t_b = 0.0
    v_t_b = 0.0
    t = 0
    epsilon = 1e-8

    # Training loop
    for num in range(epoch):
        for b in range(int(x.shape[0] / batch_size)):
            t += 1
            x_batch = x[b * batch_size:(b + 1) * batch_size]
            y_batch = y[b * batch_size:(b + 1) *
batch_size].reshape(-1, 1)

            # Prediction of polynomial regression
            pred = np.dot(x_batch ** 2, z) + np.dot(x_batch, w) +
bias

            # Loss
            loss = y_batch - pred

            # Compute gradients
            g_t_z = -2 * np.dot(np.square(x_batch).T, loss)
            g_t_w = -2 * np.dot(x_batch.T, loss)
            g_t_b = -2 * loss.sum()

            m_t_z = beta_1 * m_t_z + (1 - beta_1) * g_t_z
            v_t_z = beta_2 * v_t_z + (1 - beta_2) * (g_t_z ** 2)
            m_cap_z = m_t_z / (1 - (beta_1 ** t))
            v_cap_z = v_t_z / (1 - (beta_2 ** t))
```

```
            m_t_w = beta_1 * m_t_w + (1 - beta_1) * g_t_w
            v_t_w = beta_2 * v_t_w + (1 - beta_2) * (g_t_w ** 2)
            m_cap_w = m_t_w / (1 - (beta_1 ** t))
            v_cap_w = v_t_w / (1 - (beta_2 ** t))

            m_t_b = 0.9 * m_t_b + (1 - 0.9) * g_t_b
            v_t_b = 0.99 * v_t_b + (1 - 0.99) * (g_t_b ** 2)
            m_cap_b = m_t_b / (1 - (0.9 ** t))
            v_cap_b = v_t_b / (1 - (0.99 ** t))

            # Update parameters
            z -= ((lr * m_cap_z) / (np.sqrt(v_cap_z) +
epsilon)).reshape(-1, 1)
            w -= ((lr * m_cap_w) / (np.sqrt(v_cap_w) +
epsilon)).reshape(-1, 1)
            bias -= (lr * m_cap_b) / (math.sqrt(v_cap_b) +
epsilon)

    return z, w, bias
```

b. 在只使用 NO 數值作為 feature 的情況下，紀錄該 model 所訓練出的 parameter 數值（w2, w1, b）以及 kaggle public score.
在如下圖的設定，random seed = 9487，最後得到的結果為所附的 kaggle public score.

```
train_config = Namespace(
        batch_size = 512,
        lr = 1e-1,
        lam = 0.001,
        epoch = 1,
)
```

```
feats = [2]
```

my_sol_test_no.csv Complete (after deadline) · now	12.99882	21.00481	

平方項係數（w2）為[[0.0324149],[-0.10955948],[-0.01614332],[0.09518239],[-0.051995],[0.0658696],[-0.07257947],[0.08801479]] (8*1 的矩陣)
一次項係數（w1）為[[0.38487713],[0.27877768],[0.26185372],[0.24145826],[0.22826823],[0.29885901],[0.26746609],[0.36222515]] (8*1 的矩陣)
常數（b）為 0.7981775406162891

Math 1.a, 1.b

Problem 1.

(a) apply first order approximation

$$f(w+\Delta w) = f(w) + (\Delta w)^T \cdot \nabla_w f(w)$$

$$f(w) = w^T A w,$$

$$f(w+\Delta w) = (w+\Delta w)^T A (w+\Delta w)$$

$$= (w^T + \Delta w^T) A \cdot (w+\Delta w)$$

$$= w^T A w + w^T A \Delta w + \Delta w^T A w + \Delta w^T A \Delta w$$

since $\Delta w \to 0$, $\Delta w^T A \Delta w \to 0$

$$f(w+\Delta w) = w^T A w + w^T A \Delta w + \Delta w^T A w$$

since $w^T A \Delta w \in R^{1\times 1}$, $w^T A \Delta w = (w^T A \Delta w)^T$
$= \Delta w^T A^T w$

$$f(w+\Delta w) = f(w) + \Delta w^T A^T w + \Delta w^T A w$$

$$= f(w) + \Delta w^T (A^T w + A w)$$

$$= f(w) + \Delta w^T \cdot \nabla_w f(w)$$

$\nabla_w f(w) = A^T w + A w$, if A is symmetric $A^T = A$

$$\nabla_w f(w) = 2Aw$$

(b) let $A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix}$, $B = [B_1 B_2 \dots B_m]$

$A_i \in R^{1\times m}$ $\quad$ $B_i \in R^{m\times 1}$

$$(AB)_{ij} = A_i \cdot B_j$$

$$tr(AB) = \sum_{i=1}^{m} \sum_{j=1}^{m} A_{ij} B_{ji}$$

$$\frac{\partial\, tr(AB)}{\partial A_{ij}} = \frac{\sum_{k=1}^{m} \sum_{w=1}^{m} A_{kw} B_{kw}}{\partial A_{ij}} = \frac{A_{ij} B_{ji} + \sum_{k=1}^{m} \sum_{w=1}^{m} A_{kw} B_{kw}}{\partial A_{ij}}$$

when $k \neq i$ and $w \neq j$

$$= B_{ji} + 0 = B_{ji}$$

Math 1.c

1. (c) we have $X^{-1} = \frac{1}{\det A}(\operatorname{adj} A)$, $\operatorname{adj} A = (C_A)^T$

adjucate matrix

cofactor matrix

by cofactor expansion

$$\det A = \sum_{k=1}^{n} a_{ik} C_{ik}$$

$$\frac{\partial \det A}{\partial a_{ij}} = \sum_{k=1}^{n} \left(\frac{\partial a_{ik}}{\partial a_{ij}} C_{ik} + a_{ik} \frac{\partial C_{ik}}{\partial a_{ij}} \right)$$

If $k=j$, $\frac{\partial a_{ik}}{\partial a_{ij}} = 1$, $k \neq j$, $\frac{\partial a_{ik}}{\partial a_{ij}} = 0$

C_{ik} does not affect by a_{ij}, $\frac{\partial C_{ik}}{\partial a_{ij}} = 0$

$$\frac{\partial \det A}{\partial a_{ij}} = C_{ij}$$

$$\frac{\partial \ln(\det A)}{\partial a_{ij}} = \frac{1}{\det A} \cdot C_{ij} = \frac{1}{\det}(\operatorname{adj} A)^T_{ij}$$

$$= (A^{-1})^T_{ij} = e_j^T A^{-1} e_i$$

Reference

https://statisticaloddsandends.wordpress.com/2018/05/24/derivative-of-log-det-x/

Math 2.a1

2. (a) 不失一般性，設 $y_1 \sim y_j$ 為 c_1, $y_j \sim y_n$ 為 c_2

(i) $L(\theta) = \prod_{i=1}^{n} P_\theta[X=x_i, Y=y_i]$

$$= \prod_{i=1}^{j} P_{(\pi_1,\mu_1,\Sigma_1)}[X=x_i, Y=y_i] \times \prod_{i=j+1}^{n} P_{(\pi_2,\mu_2,\Sigma_2)}[X=x_i, Y=y_i]$$

$$= \prod_{i=1}^{j} \pi_1 \frac{\exp\left(-\frac{1}{2}(x_i-\mu_1)^T \Sigma_1^{-1}(x_i-\mu_1)\right)}{\sqrt{(2\pi)^d |\Sigma_1|}}$$

$$\times \prod_{i=j+1}^{n} \pi_2 \frac{\exp\left(-\frac{1}{2}(x_i-\mu_2)^T \Sigma_2^{-1}(x_i-\mu_2)\right)}{\sqrt{(2\pi)^d |\Sigma_2|}}$$

$$= (\pi_1)^j (2\pi)^{-\frac{dj}{2}} |\Sigma_1|^{-\frac{j}{2}} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{j}(x_i-\mu_1)^T \Sigma_1^{-1}(x_i-\mu_1)\right)$$

$$\cdot (\pi_2)^{n-j} (2\pi)^{-\frac{d(n-j)}{2}} \cdot |\Sigma_2|^{-\frac{n-j}{2}} \cdot \exp\left(-\frac{1}{2}\sum_{i=j+1}^{n}(x_i-\mu_2)^T \Sigma_2^{-1}(x_i-\mu_2)\right)$$

(ii) maximize $L(\theta) \Rightarrow$ maximize $\ln(L(\theta))$

$$\ln(L(\theta)) = j \cdot \ln(\pi_1) - \frac{dj}{2}\ln(2\pi) - \frac{j}{2}\ln(|\Sigma_1|) - \frac{1}{2}\left(\sum_{i=1}^{j}(x_i-\mu_1)^T \Sigma_1^{-1}(x_i-\mu_1)\right) + (n-j)\ln(\pi_2) - \frac{d(n-j)}{2}\ln(2\pi) - \frac{(n-j)}{2}\ln|\Sigma_2| - \frac{1}{2}\sum_{i=j+1}^{n}(x_i-\mu_2)^T \Sigma_2^{-1}(x_i-\mu_1)$$

since d, j, π are constant

maximize $\ln(L(\theta))$ means

maximize $j \cdot \ln(\pi_1) + (n-j)\ln(\pi_2)$,

$-\frac{1}{2}\left(j\ln(|\Sigma_1|) + \sum_{i=1}^{j}(x_i-\mu_1)^T \Sigma_1^{-1}(x_i-\mu_1)\right)$,

$-\frac{1}{2}\left((n-j)\ln|\Sigma_2| + \sum_{i=j+1}^{n}(x_i-\mu_2)^T \Sigma_2^{-1}(x_i-\mu_2)\right)$

to maximize $j \cdot \ln(\pi_1) + (n-j)\ln(\pi_2)$, we knew $\pi_1 + \pi_2 = 1$

because sum of propability $= 1$,

$$\frac{d}{d\pi_1}\left(j \cdot \ln(\pi_1) + (n-j)\ln(1-\pi_1)\right) = j \cdot \frac{1}{\pi_1} + (j-n)\frac{1}{1-\pi_1} = 0$$

$$\pi_1^* = \frac{j}{n}, \quad \pi_2^* = \frac{n-j}{n} \quad \#$$

Math 2.a2

接 2. (a)
(ii) maximize $-\frac{1}{2}\left(j\cdot \ln|\Sigma_1| + \sum_{i=1}^{j}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1)\right)$

we already knew $\nabla_A \ln(\det(A)) = (A^{-1})^T$ $\quad \nabla_A tr(BA) = B^T$

$\nabla_x(x^TAx) = 2Ax$ when A is symmetric
Σ_1 is symmetric

$$\nabla_{\mu_1}\left(-\frac{1}{2}\left(j\cdot\ln|\Sigma_1| + \sum_{i=1}^{j}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1)\right)\right)$$

$$= -\frac{1}{2}\left(0 + \sum_{i=1}^{j} 2\Sigma_1^{-1}(x_i-\mu_1)\right)$$

$$= -\sum_{i=1}^{j}\Sigma_1^{-1}(x_i-\mu_1) = -\Sigma_1^{-1}\left(\sum_{i=1}^{j}(x_i) - j\mu_1\right)$$

since Σ_1 is non-singular $-\Sigma^{-1}\left(\sum_{i=1}^{j}(x_i) - j\mu_1\right) = 0$

only when $\sum_{i=1}^{j} x_i - j\mu_1 = 0 \Rightarrow \mu_1^* = \frac{1}{j}\sum_{i=1}^{j} x_i$ ※

$|\Sigma_1| = \frac{1}{|\Sigma_1^{-1}|}$

$$-\frac{1}{2}(j\ln|\Sigma_1|) - \frac{1}{2}\sum_{i=1}^{j}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1)$$

$$\hookrightarrow = \frac{1}{2}j\cdot\ln|\Sigma_1^{-1}| - \frac{1}{2}\sum_{i=1}^{j}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1)$$

$$\nabla_{\Sigma^{-1}}\frac{1}{2}\left(j\cdot\ln|\Sigma_1^{-1}| - \sum_{i=1}^{j}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1)\right)$$

if c is scalar, $tr(c) = c$

$$= \frac{1}{2}\left(j\cdot\Sigma_1^T - \nabla_{\Sigma^{-1}}\left(\sum_{i=1}^{j} tr\left((x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1)\right)\right)\right)$$

$tr(AB) = tr(BA)$

$$= \frac{j}{2}\Sigma^T - \frac{1}{2}\sum_{i=1}^{j}\nabla_{\Sigma^{-1}}\left(tr\left((x_i-\mu_1)(x_i-\mu_1)^T\Sigma_1^{-1}\right)\right)$$

$$= \frac{j}{2}\Sigma_1^T - \frac{1}{2}\left(\sum_{i=1}^{j}(x_i-\mu_1)(x_i-\mu_1)^T\right)^T = 0$$

$$j\Sigma_1 = \sum_{i=1}^{j}(x_i-\mu_1)(x_i-\mu_1)^T$$

$$\Sigma_1^* = \frac{1}{j}\sum_{i=1}^{j}(x_i-\mu_1)(x_i-\mu_1)^T \text{ ※}$$

since Σ_1 and μ_1 does not interfere Σ_2 and μ_2, by the same steps we can find $\mu_2^* = \frac{1}{n-j}\sum_{i=j+1}^{n} x_i$, $\Sigma_2^* = \frac{1}{n-j}\sum_{i=j+1}^{n}(x_i-\mu_2)(x_i-\mu_2)^T$

※

Reference

https://www.statlect.com/fundamentals-of-statistics/multivariate-normal-distribution-maximum-likelihood

Math 2.a3

2、(a)

(iii) $P_\theta[Y=C_1 \mid X=x] = \dfrac{P_\theta(X=x, Y=C_1)}{P_\theta(X=x, Y=C_1) + P_\theta(X=x, Y=C_2)}$

↳ when $X=x$, the probability of $Y=C_1$

$$= \frac{\pi_1 \cdot \frac{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))}{\sqrt{(2\pi)^d|\Sigma_1|}}}{\pi_1 \cdot \frac{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))}{\sqrt{(2\pi)^d|\Sigma_1|}} + \pi_2 \cdot \frac{\exp(-\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2))}{\sqrt{(2\pi)^d|\Sigma_1|}}}$$

$P_\theta[X=x \mid Y=C_1]$

↳ when $Y=C_1$, the probability of $X=x$

$$= \frac{P_\theta(X=x, Y=C_1)}{P(Y=C_1)} = \frac{\pi_1 \frac{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))}{\sqrt{(2\pi)^d|\Sigma_1|}}}{\frac{j}{n}}$$

recall $\pi_1^* = \frac{j}{n}$, $P_\theta[X=x|Y=C_1] = \dfrac{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))}{\sqrt{(2\pi)^d|\Sigma_1|}}$

2.(a)

(iv) $P_\theta[Y=C_1 \mid X=x] = \dfrac{P_\theta(X=x, Y=C_1)}{P_\theta(X=x, Y=C_1) + P_\theta(X=x, Y=C_2)}$

$$= \frac{1}{1 + \frac{P(X=x, Y=C_2)}{P(X=x, Y=C_1)}} = \sigma(z) = \frac{1}{1+\exp(-z)}$$

$$z = \ln \frac{P(X=x, Y=C_1)}{P(X=x, Y=C_2)} = \ln \frac{P(x|C_1)}{P(x|C_2)} + \ln \frac{\pi_1}{\pi_2}$$

↳ in a.(iii), we know $p(x|C_1)$

$$\ln \frac{P(x|C_1)}{P(x|C_2)} = \ln \frac{\sqrt{|\Sigma_2|}}{\sqrt{|\Sigma_1|}} \cdot \exp\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right\}$$

$$= \ln \frac{\sqrt{|\Sigma_2|}}{\sqrt{|\Sigma_1|}} - \frac{1}{2}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right]$$

Math 2.a4

2.(a)
(iv)

$$Z = \ln\frac{\pi_1}{\pi_2} + \ln\sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}} - \frac{1}{2}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right]$$

$$\begin{aligned}
&(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) \\
&= (x^T\Sigma_1^{-1} - \mu_1^T\Sigma_1^{-1})(x-\mu_1) \\
&= x^T\Sigma_1^{-1}x - \underbrace{x^T\Sigma_1^{-1}\mu_1 - \mu_1^T\Sigma_1^{-1}x}_{\text{same}} + \mu_1^T\Sigma_1^{-1}\mu_1 \\
&= x^T\Sigma_1^{-1}x - 2\mu_1^T\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}\mu_1
\end{aligned}$$

and so is $(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)$

$$= x^T\Sigma_2^{-1}x - 2\mu_2^T\Sigma_2^{-1}x + \mu_2^T\Sigma_2^{-1}\mu_2$$

$$Z = \ln\frac{\pi_1}{\pi_2} + \ln\frac{\sqrt{|\Sigma_2|}}{\sqrt{|\Sigma_1|}} - \frac{1}{2}x^T\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}x$$
$$- \frac{1}{2}\mu_1^T\Sigma_1^{-1}\mu_1 + \frac{1}{2}x^T\Sigma_2^{-1}x - \mu_2^T\Sigma_2^{-1}x + \frac{1}{2}\mu_2^T\Sigma_2^{-1}x$$

※

Math 2.b

2. (b) rewrite $\ln(L(\theta)) = (n-j)\ln(\pi_2) + j\ln\pi_1$

$$-\frac{1}{2}\Big(j\ln|\Sigma| + \sum_{i=1}^{j}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)$$

$$+(n-j)\ln|\Sigma| + \sum_{i=j+1}^{n}(x_i-\mu_2)^T\Sigma^{-1}(x_i-\mu_2)\Big)$$

π_1, π_2 are not affected by Σ, have the same result

$$\pi_1^* = \frac{j}{n} \quad \pi_2^* = \frac{n-j}{n}$$

$$\nabla_{\mu_1}\ln(L(\theta)) = -\frac{1}{2}\nabla_{\mu_1}\Big(\sum_{i=1}^{j}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\Big)$$

Σ^{-1} are symmetric

$$= -\frac{1}{2}\sum_{i=1}^{j} 2\Sigma^{-1}(x_i-\mu_1) = 0$$

$$\sum_{i=1}^{j}\Sigma^{-1}(x_i-\mu_1) = 0, \quad \sum_{i=1}^{j}(x_i-\mu_1) = 0$$

$$\mu_1^* = \frac{1}{j}\sum_{i=1}^{j}x_i,$$ 同理 $$\mu_2^* = \frac{1}{n-j}\sum_{i=j+1}^{n}x_i$$

$$\nabla_{\Sigma^{-1}}\ln(L(\theta)) = \frac{1}{2}\nabla_{\Sigma^{-1}} n\ln|\Sigma^{-1}| - \frac{1}{2}\nabla_{\Sigma^{-1}}\sum_{i=1}^{j}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)$$

$$-\frac{1}{2}\nabla_{\Sigma^{-1}}\sum_{i=j+1}^{n}(x_i-\mu_2)^T\Sigma^{-1}(x_i-\mu_2)$$

$$= \frac{n}{2}\Sigma^T - \frac{1}{2}\nabla_{\Sigma^{-1}}\Big(\sum_{i=1}^{j} tr\big[(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\big]\Big)$$

$$-\frac{1}{2}\nabla_{\Sigma^{-1}}\Big(\sum_{i=j+1}^{n} tr\big[(x_i-\mu_2)^T\Sigma^{-1}(x_i-\mu_2)\big]\Big)$$

by the same computation in 2-(a).(ii)

$$= \frac{n}{2}\Sigma^T - \frac{1}{2}\Big(\sum_{i=1}^{j}(x_i-\mu_1)(x_i-\mu_2)^T\Big)^T - \frac{1}{2}\Big(\sum_{i=j+1}^{n}(x_i-\mu_2)(x_i-\mu_2)^T\Big)^T$$

$$\Sigma^* = \frac{\sum_{i=1}^{j}(x_i-\mu_1)(x_i-\mu_1)^T + \sum_{i=j+1}^{n}(x_i-\mu_2)(x_i-\mu_2)^T}{n}$$

Math 3

Let c2 = class 0 in the dataset

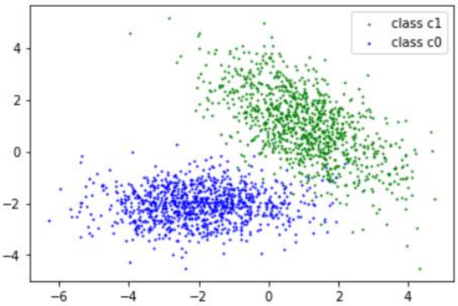
a.

ϑ* = (π1*, π2*, μ1*, μ2*, Σ*) = (0.5, 0.5, [1.011436374828497, 1.004931936779685],
[-2.025716971154811, -2.0461950110868674],
[[1.85889712, -0.51610136],
[-0.51610136, 1.14373928]])

b.

ϑ* = (π1*, π2*, μ1*, μ2*, Σ1*, Σ2*) = (0.5, 0.5, [1.011436374828497, 1.004931936779685], [-2.025716971154811, -2.0461950110868674],
[[1.70649036, -1.06606724], [-1.06606724, 1.82770502]],
[[2.0133172, 0.03389842], [0.03389842, 0.46023378]])

c. I would choose the method in (a) because we can simply draw a straight line to distinguish class 1 and class 0.



Math 4.a

4. (a) rewrite $\sum_i k_i(y_i - x_i\theta)^2 + \lambda \sum_j w_j^2, \lambda > 0$

as $(y - X\theta)^T K (y - X\theta) + \lambda \|\theta\|_2^2$

$= (y - X\theta)^T K (y - X\theta) + \lambda \cdot \theta^T \cdot \theta$

$\nabla_\theta ((y - X\theta)^T K (y - X\theta) + \lambda \cdot \theta^T \theta)$

$= \lambda I \cdot 2\theta - 2 X^T K (y - X\theta) = 0$

$-X^T K y + X^T K X\theta + \lambda I \theta = 0$

$(\lambda I + X^T K X)\theta = X^T K y$

$\theta^* = (X^T K X + \lambda I)^{-1} X^T K y$

$(\theta + \Delta\theta)^T (\theta + \Delta\theta) - \theta^T \theta = (\Delta\theta)^T \cdot \nabla(\theta^T \theta)$

$= (\theta^T + \Delta\theta)^T (\theta + \Delta\theta) - \theta^T \theta$

$= \theta^T \theta + 2(\Delta\theta)^T \theta + (\Delta\theta)^T \Delta\theta - \theta^T \theta$ $\quad (\Delta\theta)^T \theta$

$= (\Delta\theta)^T \cdot 2\theta + (\Delta\theta)^T \cdot \cancel{\Delta\theta}$ $\quad = \theta^T \Delta\theta$

$(y - X(\theta + \Delta\theta))^T K (y - X(\theta + \Delta\theta)) - (y - X\theta)^T K (y - X\theta)$

$= (y^T - \theta^T X^T - (\Delta\theta)^T X^T) K (y - X\theta - X\Delta\theta)$

$- (y - X\theta)^T K (y - X\theta)$

$= (y - X\theta)^T K (-X\Delta\theta) - (\Delta\theta)^T X^T K (y - X\theta)$

$+ (\Delta\theta)^T X^T K \cancel{X\Delta\theta} \to 0 \quad \Delta\theta \to 0$

$= -(\Delta\theta)^T X^T K^T (y - X\theta) - (\Delta\theta)^T X^T K (y - X\theta)$

$K^T = K$ since K is diagonal $\quad = -(\Delta\theta)^T \, 2X^T K (y - X\theta)$

Math 4.b

4. (b)

$$L = \sum_i k_i (y_i - x_i\theta)^2 + \lambda \sum_j w_j^2$$

$$= (y_i - x_i\theta)^T k (y_i - x_i\theta) + \lambda \|w\|_2$$

$$\theta = \begin{bmatrix} w \\ b \end{bmatrix} \quad x = [\,\tilde{x}, 1\,]$$ below the x has become $\tilde{x}$

$$= \left(y - [\tilde{x}, 1]\begin{bmatrix} w \\ b \end{bmatrix}\right)^T k \left(y - [\tilde{x}, 1]\begin{bmatrix} w \\ b \end{bmatrix}\right)^T + \lambda \|w\|_2^2 + \lambda b^2$$

$$= (y - \tilde{x}w + eb)^T k (y - \tilde{x}w + eb) + \lambda \|w\|_2^2 + \lambda b^2$$

let $B = eb = \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix}$

$$= (y^T - w^T\tilde{x}^T - B^T) k (y - \tilde{x}w + B) + \lambda \|w\|_2^2 + \lambda b^2$$

$$= (y^T k - w^T \tilde{x}^T k - B^T k)(y - \tilde{x}w - B) + \lambda w^T w + \lambda b^2$$

$$= y^T k y - y^T k \tilde{x} w - y^T k B - w^T \tilde{x}^T k y + w^T \tilde{x}^T k \tilde{x} w + w^T \tilde{x}^T k B - B^T k y + B^T k \tilde{x} w + B^T k B + \lambda w^T w + \lambda b^2$$

$$= y^T k y - 2 y^T k \tilde{x} w - 2 y^T k B + 2 B^T k \tilde{x} w + w^T \tilde{x}^T k \tilde{x} w + B^T k B + \lambda w^T w + \lambda b^2$$

$$\nabla_b L = \underbrace{2b\,tr(k)}_{\frac{\partial B^T k B}{\partial b}} - \underbrace{2(y_1 k_1 + y_2 k_2 + \cdots + y_n k_n)}_{\frac{\partial y^T k B}{\partial b}} + \underbrace{2((\tilde{x}w)_1 k_1 + (\tilde{x}w)_2 k_2 + \cdots + (\tilde{x}w)_n k_n)}_{\frac{\partial B^T k \tilde{x} w}{\partial b}}$$

$$= 2b\,tr(k) - 2 e^T k y + 2 e^T k \tilde{x} w = 0$$

$$b^* = \frac{(e^T k y - e^T k \tilde{x} w)}{tr(k)}$$

$$\nabla_w L = -2\frac{\partial y^T K \tilde{x} w}{\partial w} + 2\frac{\partial B^T K \tilde{x} w}{\partial w} + \frac{\lambda w^T w}{\partial w} + \frac{w^T \tilde{x}^T K \tilde{x} w}{\partial w}$$

$$= -2\tilde{x}^T K y + 2\tilde{x}^T K B + 2\lambda w + 2\tilde{x}^T K \tilde{x} w$$

4-(b) $= 0 \quad -\tilde{x}^T K y + x^T K B + \lambda I w + \tilde{x}^T K \tilde{x} w = 0$

$$(\tilde{x}^T K \tilde{x} + \lambda I) w = \tilde{x}^T K (y - B)$$

$$B = \begin{bmatrix} \frac{e^T K y - e^T K \tilde{x} w}{tr(K)} \\ \vdots \\ \frac{e^T K y - e^T K \tilde{x} w}{tr(K)} \end{bmatrix} \in R^n = e\frac{e^T K y - e^T K \tilde{x} w}{tr(K)}$$

$$(\tilde{x}^T K \tilde{x} + \lambda I) w = \tilde{x}^T K \left(y - e\frac{e^T K y - e^T K x w}{tr(K)}\right)$$

$$(\tilde{x}^T K x + \lambda I) w - \tilde{x}^T K \frac{e e^T K \tilde{x} w}{tr(K)}$$

$$= \tilde{x}^T K \left(y - \frac{e e^T K y}{tr(K)}\right)$$

$$\left(\tilde{x}^T K \tilde{x} + \lambda I - \frac{\tilde{x}^T K e e^T K \tilde{x}}{tr(K)}\right) w = \tilde{x}^T K \left(y - \frac{e e^T K y}{tr(K)}\right)$$

$$w^* = \left(\tilde{x}^T K \tilde{x} + \lambda I - \frac{\tilde{x}^T K e e^T K \tilde{x}}{tr(x)}\right)^{-1} \tilde{x}^T K \left(y - \frac{e e^T K y}{tr(K)}\right) \#$$

Math5

5. $f_{w,b}(x) = w^T x + b$

$$\hat{C}_{ss}(w,b) = \mathbb{E}\left[\frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i+\eta_i)-y_i\right)^2\right]$$

$$= \frac{1}{2N}\mathbb{E}\left[\sum_{i=1}^{N}\left(f_{w,b}(x_i+\eta_i)-y_i\right)^2\right]$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[\left(w^T(x_i+\eta_i)-y_i\right)^2\right]$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[\left(w^T x - y_i + w^T\eta_i\right)^2\right]$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[\left((f_{w,b}(x_i)-y_i)+w^T\eta_i\right)^2\right]$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left((f_{w,b}(x_i)-y_i)^2 + 2(f_{w,b}(x_i)-y_i)\mathbb{E}(w^T\eta_i) + \mathbb{E}(w^T\eta_i)^2\right)$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i)-y_i\right)^2 + \frac{1}{2N}\sum_{i=1}^{N}2\left(f_{w,b}(x_i-y_i)\right)\cdot w^T\mathbb{E}(\eta_i) + \frac{1}{2N}\sum_{i=1}^{N}\|w\|^2\cdot\mathbb{E}[\eta_i^2]$$

since $\mathbb{E}[\eta_i]=0$, $\mathbb{E}[x^2] = E(x)^2 + Var(x)$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i)-y_i\right)^2 + 0 + \frac{1}{2N}\sum_{i=1}^{N}\|w\|^2\cdot\left((\mathbb{E}(\eta_i))^2+\sigma^2\right)$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i)-y_i\right)^2 + \frac{1}{2N}\sum_{i=1}^{N}\|w\|^2\cdot(0+\sigma^2)$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i)-y_i\right)^2 + \frac{\sigma^2}{2\cancel{N}}\cdot\cancel{N}\cdot\|w\|^2$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i)-y_i\right)^2 + \frac{\sigma^2}{2}\|w\|^2$$