王若怡 2021533135

Hw4

1.(a) $\{y_i \underbrace{(x_i \cdot w + \psi)}_{f(x_i)} \geq 0, \forall i = 1 \cdots n\}$ as a set of n constraints to ensure that all n points are correctly classified.

(b) We have known $H = \{x \in R^d : x \cdot w + \psi = 0\}$ and $x_i \in R^d$

$\therefore$ we assume another two points: v_1 is the point which is the projection of x_i on H; v_2 is any point on the H ($v_2 \neq v_1$)

And then the angle between $\overrightarrow{x_i v_1}$ and $\overrightarrow{x_i v_2}$ is θ

$\Rightarrow d = \|\overrightarrow{x_i v_1}\| = \|\overrightarrow{x_i v_2}\| \cdot \cos\theta$ (1) and $\cos\theta = \dfrac{\overrightarrow{x_i v_1} \cdot \overrightarrow{x_i v_2}}{\|\overrightarrow{x_i v_1}\| \|\overrightarrow{x_i v_2}\|}$ (2)

$\therefore \overset{(1)(2)}{\Longrightarrow} d = \dfrac{\overrightarrow{x_i v_1} \cdot \overrightarrow{x_i v_2}}{\|\overrightarrow{x_i v_1}\|}$ (5)

And then because v_1 and v_2 both are on the H

$\therefore \Rightarrow \begin{cases} v_1 w + \psi = 0 \\ v_2 w + \psi = 0 \end{cases} \Rightarrow (v_1 - v_2) w = 0 \quad (3)$

$\overrightarrow{v_1 v_2} \cdot w = 0$

And then because $\overrightarrow{x_i v_1} \perp \overrightarrow{v_1 v_2}$ $\therefore \overrightarrow{x_i v_1} \cdot \overrightarrow{v_1 v_2} = 0$ (4)

$\therefore \overset{(3)(4)}{\Longrightarrow} \overrightarrow{x_i v_1} = k \cdot w$ (6) because $\overrightarrow{v_1 v_2} \neq \vec{0}$

$\therefore \overset{(5)(6)}{\Longrightarrow} d = \dfrac{|k| \cdot |w \cdot \overrightarrow{x_i v_2}|}{|k| \|w\|} = \dfrac{|w \cdot (x_i - v_2)|}{\|w\|} \overset{\because w v_2 + \psi = 0}{=} \dfrac{|w x_i + \psi|}{\|w\|}$

$\therefore$ Finally, we get it $d = \dfrac{|w x_i + \psi|}{\|w\|}$

(c) Because all the points have been classified correctly, then we can calculate all the distances between points and decision boundary (H).

$\Rightarrow \dfrac{|w x_i + \psi|}{\|w\|} \geq r_w, \forall x_i \in \{x_1 \cdots x_n\}$. where r_w is the min distance between points and H given w.

(d) Because support vector is the training points closest to the decision boundary then we can use these points to calculate the distance.

That is $\dfrac{|w x_i + \psi|}{\|w\|}$, in which x_i is $\in$ closet set (support vector)

(e) $\max\limits_{w \in R^d} r_w = \max\limits_{w \in R^d} \min\limits_{x_i \in \{x_1 \cdots x_n\}} \dfrac{|w x_i + \psi|}{\|w\|}$, s.t. $\forall x_i \in \{x_1 \cdots x_n\}$ $y_i (x_i w + \psi) \geq 0$.

2 ① Assume all the observations in X have been normalized, which is let $X_{new\,i} = X_i - \mu$. for every X_i in X. And $\mu = \frac{1}{n}\sum_{1}^{n} X_i$.

Then we can have $\max_v \frac{1}{n}\sum_{i=1}^{n}(x_i^T v - \mu^T v)^2$ s.t. $\|v\| = 1$

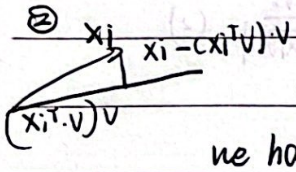
$$= \max_v \frac{1}{n} v^T \Big(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\Big)\cdot v$$

$$= \max_v \frac{1}{n} v^T X X^T v$$

$\because n$ is scalar (constant number

$\therefore \Rightarrow \underline{\max_v v^T X X^T v \quad s.t.\ v^T v = 1}$ max projected variance

②



we have this projection relation. so from the Pythagoras theorem, we have $\|x_i\|^2 = \|(x_i^T v)v\|^2 + \|x_i - (x_i^T v)v\|^2$.

Therefore, we can sum it up and divide by n and have

$$\underbrace{\boxed{\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2}}_{\text{obviously this is a constant}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\|(x_i^T v)v\|^2}_{\text{max projected variance}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\|x_i - (x_i^T v)v\|^2}_{\text{our goal: min projected error}}$$

$\therefore$ Finally we get it because $\underline{\text{max projected var}} + \underbrace{\text{our goal}}_{\frac{1}{n}\sum_{i=1}^{n}\|x_i-(x_i^Tv)v\|^2} = \text{constant}$

$$|||$$

$$\therefore \quad \min \frac{1}{n}\sum_{i=1}^{n}\|x_i-(x_i^T v)v\|^2 \ //$$

No.

Date. / /

3. (a) compute $\mu = \frac{1}{n}\sum_{1}^{n} x_i = \frac{1}{6}\begin{bmatrix} 0+0+1+1+2+2 \\ 0+1+0+2+1+2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\dot{X} = X - \mu = \begin{bmatrix} -1 & -1 \\ -1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$ $R^{6\times 2}$

(b) $\dot{X}^T\dot{X} = \begin{bmatrix} -1 & -1 & 0 & 0 & 1 & 1 \\ -1 & 0 & -1 & 1 & 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & -1 \\ -1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 4, & 2 \\ 2, & 4 \end{bmatrix}$

$|\dot{X}^T\dot{X} - \lambda I| = \begin{vmatrix} 4-\lambda, & 2 \\ 2, & 4-\lambda \end{vmatrix} = (4-\lambda)^2 - 4 = 16 + \lambda^2 - 8\lambda - 4 = \lambda^2 - 8\lambda + 12$
$= (\lambda - 2)(\lambda - 6)$

① when $\lambda_1 = 2$ $\begin{bmatrix} 2. & 2 \\ 2. & 2 \end{bmatrix} = \dot{X}^T\dot{X} - \lambda I$

$\Rightarrow \begin{bmatrix} 1. & 1 \\ 0. & 0 \end{bmatrix} v_1 = 0 \Rightarrow v_1 = \begin{bmatrix} +\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$ or $\begin{bmatrix} -\frac{\sqrt{2}}{2} \\ +\frac{\sqrt{2}}{2} \end{bmatrix}$ $\because \|v_1\| = \|v_2\| = 1$

② when $\lambda_2 = 6$ $\begin{bmatrix} -2, & 2 \\ 2, & -2 \end{bmatrix} = \dot{X}^T\dot{X} - \lambda I$

$\Rightarrow \begin{bmatrix} 1, & -1 \\ 0, & 0 \end{bmatrix} v_2 = 0 \Rightarrow v_2 = \begin{bmatrix} +\frac{\sqrt{2}}{2} \\ +\frac{\sqrt{2}}{2} \end{bmatrix}$ or $\begin{bmatrix} -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$

(c) substitude v_1 and v_2 $\Longrightarrow$

$$\begin{cases} \frac{1}{n} v_{1_1}^T \cdot \dot{X}^T\dot{X} \cdot v_{1_1} = \frac{2}{3} & \frac{1}{n} v_{1_2}^T \cdot \dot{X}^T\dot{X} \cdot v_{1_2} = \frac{2}{3} \\ \frac{1}{n} v_{2_1}^T \cdot \dot{X}^T \cdot \dot{X} \cdot v_{2_1} = 2 & \frac{1}{n} v_{22}^T \dot{X}^T\dot{X} \cdot v_{22} = 2 \end{cases}$$

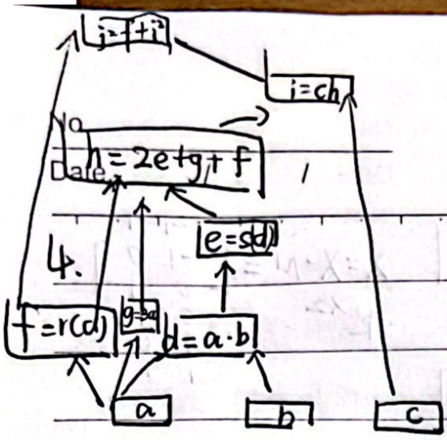
$\therefore$ ① we may choose $\lambda_2 = 6$ as the main principal component

(or you can calculate one by one) → ② we can use the value of eignvalue λ to choose the best. The ~~more~~ bigger λ, the better principal component.

③ Because it will max the projected variance and then it can make the information loss minimize during the dimension reduce process.

Wengu

4.



① $\frac{\partial j}{\partial e} = \frac{\partial j}{\partial h}\frac{\partial h}{\partial e} = 2ic \cdot 2 = 4ic$

② $\frac{\partial j}{\partial f} = \frac{\partial j}{\partial i}\frac{\partial i}{\partial h}\frac{\partial h}{\partial f} + \frac{\partial j}{\partial f} = 2ic \cdot 1 + 1 = 2ic+1$

③ $\frac{\partial j}{\partial g} = \frac{\partial j}{\partial i}\frac{\partial i}{\partial h}\frac{\partial h}{\partial g} = 2ic.$

④ $\frac{\partial j}{\partial d} = \frac{\partial j}{\partial i}\frac{\partial i}{\partial h}\frac{\partial h}{\partial e}\frac{\partial e}{\partial d} = 2ic \cdot \left(\frac{1}{1+e^{-d}}\right)' = 4ic \cdot \frac{+e^{-d}}{(1+e^{-d})^2} = +4ic \cdot s(d)(1-s(d))$

⑤ $\frac{\partial j}{\partial c} = \frac{\partial j}{\partial i}\frac{\partial i}{\partial c} = 2ih$

⑥ $\frac{\partial j}{\partial b} = \frac{\partial j}{\partial i}\frac{\partial i}{\partial h}\frac{\partial h}{\partial e}\frac{\partial e}{\partial d}\frac{\partial d}{\partial b} = +4ic \cdot s(d)(1-s(d)) \cdot a$

$= +4ica\, s(d)(1-s(d))$

⑦ $\frac{\partial j}{\partial a} = \frac{\partial j}{\partial f}\frac{\partial f}{\partial a} + \frac{\partial j}{\partial g}\cdot\frac{\partial g}{\partial a} + \frac{\partial j}{\partial d}\frac{\partial d}{\partial a}$

$= (2ic+1)\cdot \mathbb{I}(a \geq 0) + 2ic \cdot 3 + (+4ic \cdot s(d)(1-s(d))) \cdot b$

$= (2ic+1)\mathbb{I}(a \geq 0) + 6ic + 4icb\, s(d)(1-s(d))$