HW 2 - solution

CS-189

Zubo Gu

1. expected value $= 4 \cdot \int_0^{\frac{1}{\sqrt{3}}} \frac{2}{\pi(1+x^2)} dx + 3 \cdot \int_{\frac{1}{\sqrt{3}}}^{1} \frac{2}{\pi(1+x^2)} dx + 2 \cdot \int_1^{\sqrt{3}} \frac{2}{\pi(1+x^2)} dx$

$= \frac{8}{\pi} \arctan x \Big|_0^{\frac{1}{\sqrt{3}}} + \frac{6}{\pi} \arctan x \Big|_{\frac{1}{\sqrt{3}}}^{1} + \frac{4}{\pi} \arctan x \Big|_1^{\sqrt{3}}$

$= \frac{8}{\pi}(\frac{\pi}{6} - 0) + \frac{6}{\pi}(\frac{\pi}{4} - \frac{\pi}{6}) + \frac{4}{\pi}(\frac{\pi}{3} - \frac{\pi}{4})$

$= \frac{4}{3} + \frac{1}{2} + \frac{1}{3} = \frac{13}{6}$

2. $L(\theta) = \theta \cdot e^{-0.9\theta} \cdot \theta \cdot e^{-1.7\theta} \cdot \theta \cdot e^{-0.4\theta} \cdot \theta \cdot e^{-0.3\theta} \cdot \theta e^{-2.4\theta}$

$= \theta^5 \cdot e^{-5.7\theta}$

$\log(L(\theta)) = 5\log\theta - 5.7\theta$

let $\log(L(\theta))' = 0$

$\theta = \frac{5}{5.7}$

3. a.) $K(\vec{x}, \vec{y}) = (\vec{x}^T \vec{y} + c)^2 = (\sum_{i=1}^{n} x_i y_i + c)^2$

$= (\sum_{i=1}^{n} x_i y_i)^2 + 2 \cdot c \cdot \sum_{i=1}^{n} x_i y_i + c^2$

$= \sum_{i=1}^{n} x_i^2 y_i^2 + 2 \cdot \sum_{1 \le i < j \le n} x_i y_i x_j y_j + 2 \cdot c \sum_{i=1}^{n} x_i y_i + c^2$

Thus. $\Phi(\vec{z}) = \langle z_1^2, z_2^2, \cdots, z_n^2, \sqrt{2} z_1 z_2, \sqrt{2} z_1 z_3, \cdots, \sqrt{2} z_1 z_n, \sqrt{2} z_2 z_3, \cdots \sqrt{2} z_2 z_n$

$\cdots, \sqrt{2} z_{n-1} z_n, \sqrt{2c} z_1, \sqrt{2c} \cdot z_2, \cdots, \sqrt{2c} z_n, c \rangle$.

b) When increasing d. the model will be more fit for the training data. But it will be overfit for the test data. which means the accuracy for test data falling. Thus. we increasing d value, until the accuracy for the test data reach the highest value. ~~for~~. The corresponding d value is the optimal value for a given dataset.

4. a.) $\vec{x}^T A \vec{x} = [x_1\ x_2\ \cdots\ x_n] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$= x_1(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n) + x_2(a_{21}x_1 + a_{22}x_2 + \cdots a_{2n}x_n)$

$+ x_n(a_{n1}x_1 + \cdots + a_{nn}x_n)$

$= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ij} x_j$

b.) If A is positive definite, $\vec{x}^T A \vec{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ij} x_j > 0$ $\forall \vec{x} \in \mathbb{R}^n$

Thus, if we set $x_k = 1$ in $\vec{x}$, and all other $x_i = 0$ for $i \neq k$.

Then, $\vec{x}^T A \vec{x} = a_{kk} > 0$.

let k be 1 to n, we will get that $a_{ii} > 0$ for all $1 \le i \le n$

Hence, the entries on the diagonal of A are positive.

5. $\vec{x}^T (B + \lambda I) \vec{x} = \vec{x}^T B \vec{x} + \vec{x}^T \lambda I \vec{x}$

and $\vec{x}^T \lambda I \vec{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i \lambda a_{ij} x_j$ (I is identity)

$= \sum_{i=1}^{n} x_i \lambda \cdot x_i = \sum_{i=1}^{n} \lambda x_i^2$

since $\vec{x} \neq 0$, exist some $x_i \neq 0$, $\Rightarrow \lambda x_i^2 > 0$ for $\lambda > 0$

Thus, $\vec{x}^T \lambda I \vec{x} > 0$.

And B is a positive semidefinite matrix.

Thus, $\vec{x}^T B \vec{x} \geq 0$

Hence, we get $\vec{x}^T (B + \lambda I) \vec{x} > 0$.

which means $B + \lambda I$ is positive definite for any $\lambda > 0$.

6. a.) $\frac{\partial (\vec{x}^T \vec{a})}{\partial \vec{x}}$, let $y = \vec{x}^T \vec{a} = x_1 a_1 + \cdots + x_n a_n$

$\frac{\partial y}{\partial x_i} = a_i$

Thus $\frac{\partial (\vec{x}^T \vec{a})}{\partial \vec{x}} = \left[\frac{\partial y}{\partial x_1} \quad \frac{\partial y}{\partial x_2} \quad \cdots \quad \frac{\partial y}{\partial x_n} \right]$

$= [a_1 \quad a_2 \quad \cdots \quad a_n]$

$= \boxed{\vec{a}^T}$

b.) let $y = \vec{x}^T A \vec{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$

$\frac{\partial y}{\partial x_k} = \sum_{i=1,}^{n} a_{ki} \cdot x_i + \sum_{i=1,}^{n} a_{ik} x_i$

6. a.) Let $y = \vec{x}^T \vec{a} = x_1 a_1 + \cdots + x_n a_n$, $\frac{\partial y}{\partial x_i} = a_i$

Thus, $\frac{\partial(\vec{x}^T\vec{a})}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \boxed{\vec{a}}$

b.) Let $y = \vec{x}^T A \vec{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ij} x_j$

$\frac{\partial y}{\partial k} = \sum_{i=1}^{n} a_{ki} x_i + \sum_{i=1}^{n} a_{ik} x_i = \sum_{i=1}^{n} x_i (a_{ki} + a_{ik})$

Thus, $\frac{\partial(\vec{x}^T A \vec{x})}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i (a_{ik} + a_{ki}) \\ \vdots \\ \sum_{i=1}^{n} x_i (a_{nk} + a_{kn}) \end{bmatrix}$

$= \begin{bmatrix} a_{1k} + a_{k1} & \cdots & a_{1n} + a_{n1} \\ \vdots & \ddots & \vdots \\ a_{n1} + a_{1n} & \cdots & a_{nn} + a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \boxed{(A + A^T) \cdot \vec{x}}$

c) Let $y = \text{trace}(XA)$, $X = \begin{bmatrix} \vec{x}_1^T \\ \vec{x}_2^T \\ \vdots \\ \vec{x}_n^T \end{bmatrix}$, $A = [\vec{a}_1 \cdots \vec{a}_n]$

$y = \text{trace}(XA) = \vec{x}_1^T \vec{a}_1 + \vec{x}_2^T \vec{a}_2 + \cdots + \vec{x}_n^T \vec{a}_n$

and, $\frac{\partial y}{\partial x_{ij}} = a_{ji}$

Thus, $\frac{\partial \, \text{trace}(XA)}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{n1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{n1}} & \cdots & \frac{\partial y}{\partial x_{nn}} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{nn} \end{bmatrix} = \boxed{A^T}$

d.) Let $y = \vec{a}^T X \vec{b} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i x_{ij} b_j$

and $\frac{\partial y}{\partial x_{ij}} = a_i b_j$

Thus, $\frac{\partial(\vec{a}^T X \vec{b})}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ \vdots & \ddots & & \vdots \\ \frac{\partial y}{\partial x_{m1}} & \cdots & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix} = \begin{bmatrix} a_1 b_1 & a_1 b_2 & \cdots & a_1 b_n \\ \vdots & \ddots & & \vdots \\ a_m b_1 & \cdots & \cdots & a_m b_n \end{bmatrix} = \boxed{\vec{a} \cdot \vec{b}^T}$

e.) Proof: need to show $\|\vec{x}\|_2 \le \|\vec{x}\|_1 \le \sqrt{n}\|\vec{x}\|_2$

$\iff \sqrt{\sum_{i=1}^{n} x_i^2} \le \sum_{i=1}^{n} |x_i| \le \sqrt{n} \cdot \sqrt{\sum_{i=1}^{n} x_i^2}$

$\iff \sum_{i=1}^{n} x_i^2 \le \left(\sum_{i=1}^{n} |x_i|\right)^2 \le n \cdot \sum_{i=1}^{n} x_i^2$ (square all side, since each side is positive)

For left and middle side:

$\left(\sum_{i=1}^{n} |x_i|\right)^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n}\sum_{j=1, i\ne j}^{n} |x_i||x_j| \ge \sum_{i=1}^{n} x_i^2$ Done

For middle and right side:

$\left(\sum_{i=1}^{n} |x_i|\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} |x_i| \cdot |x_j| = 2\sum_{1\le i\le j\le n} |x_i||x_j|$

For each $|x_i| \cdot |x_j|$, since $(|x_i| - |x_j|)^2 \ge 0 \iff 2|x_i||x_j| \le x_i^2 + x_j^2$

Thus $\left(\sum_{i=1}^{n} |x_i|\right)^2 \le \sum_{1\le i\le j\le n} x_i^2 + x_j^2 = n \cdot \sum_{i=1}^{n} x_i^2$

Hence, we show $\sum_{i=1}^{n} x_i^2 \le \left(\sum_{i=1}^{n} |x_i|\right)^2 \le n \cdot \sum_{i=1}^{n} x_i^2$

which mean $\|\vec{x}\|_2 \le \|\vec{x}\|_1 \le \sqrt{n}\|\vec{x}\|_2$ ∎

7. a.) $R[W] = (X\cdot W - Y)^T \cdot \Lambda (XW - Y)$

b.) $\frac{\partial (R[W])}{\partial W} = \frac{\partial (WX^T\Lambda XW - Y^T\Lambda XW - W^TX^T\Lambda Y + Y^T\Lambda Y)}{\partial W}$

$= \frac{\partial}{\partial W}(WX^T\Lambda XW - 2W^TX^T\Lambda Y + Y^T\Lambda Y)$

$= 2X^T\Lambda XW - 2X^T\Lambda Y = 0$ ($\Lambda = A^T$ for Λ is diagonal matrix)

$\Rightarrow X^T\Lambda X \cdot W = X^T\Lambda Y$

Since $X^T\Lambda X$ is full rank

Thus $W = (X^T\Lambda X)^{-1}X^T\Lambda Y$

c.) ~~$R[W] = 2X^T\Lambda XW - 2X^T\Lambda Y + \gamma W^T I W$ (I is dxd)~~

~~let $\frac{\partial(R[W])}{\partial W}$~~ $R[W] = (X\cdot W - Y)^T \cdot \Lambda (XW - Y) + \gamma W^T W$

Thus $\frac{\partial (R[W])}{\partial W} = 2X^T\Lambda XW - 2X^T\Lambda Y + 2\gamma I \cdot W$ (I is dxd matrix)

$= 2(X^T\Lambda X + \gamma I)\cdot W - 2X^T\Lambda Y = 0$

need $X^T\Lambda X + \gamma I$ invertible

since Λ is diagonal. $X^T\Lambda X$ will be diagonal

and entries will be $\lambda^{(i)}(X_{ii})^2$

since $\lambda^{(i)} > 0$, Thus all entries for $X^T\Lambda X$ greater or equal to 0.

Hence $X^T\Lambda X$ will be positive semi-definite

And by Q5, with $\gamma > 0$. ~~and~~

$X^T\Lambda X + \gamma I$ will be positive definite

Hence invertible

Thus, $W = (X^T\Lambda X + \gamma I)^{-1} X^T\Lambda Y$

d.) $X^T\Lambda X$ is positive semidefinite, but it maybe not full rank, Thus, not invertible.

However, if adding the term in c.)

we can get $X^T\Lambda X + \gamma I$ which is positive definite as we proved in Q5.

Hence we can invertible it and solve for W.

8. a) For $i = 1, \cdots, c$, $R(\alpha_i | x) = \sum_{j=1}^{c} \lambda(f(x)=i, i \neq j)\, P(w_j | x)$

$$= \lambda_s \sum_{j=1, j\neq i}^{c} P(w_j | x)$$

$$= \lambda_s (1 - P(w_i | x))$$

For $i = c+1$, $R(\alpha_{c+1} | x) = \lambda_r$

Thus, for the policy ① if $P(w_i|x) \geq P(w_j|x)$ for all j.

Thus, $1 - P(w_i|x)$ will be the minimum.

Then, the choose of i will be the minimum of $R(\alpha_i | x)$ for $i \in \{1, 2, \cdots, c\}$

In the mean-time, $P(w_i|x) \geq 1 - \lambda_r/\lambda_s$

$$\Leftrightarrow 1 - P(w_i|x) \leq \frac{\lambda_r}{\lambda_s}$$

$$\Leftrightarrow \lambda_s(1 - P(w_i|x)) \leq \lambda_r$$

Thus, $R(\alpha_i|x) \leq R(\alpha_{c+1}|x)$ for any $i \in \{1, 2, \cdots, c\}$

In this case, the i we chose will be the minimum risk.

for the policy ②. We can always find the i such that $P(w_i|x) \geq P(w_j|x)$ for all j.

But, if $P(w_i|x) \geq 1 - \lambda_r/\lambda_s$ not satisfy.

which means $\lambda_s(1 - P(w_i|x)) > \lambda_r$

Thus, pick $i = c+1$ will be the minimum

Thus, choose doubt otherwise.

b.) if $\lambda_r = 0$. we ~~always~~ always choose doubt.

Since $\lambda_r = 0$, $R(\alpha_{c+1}|x) = 0$, But $R(\alpha_i|x) = \lambda_s(1 - P(w_i|x))$ with $\lambda_s \geq 0$. $P(w_i|x) \leq 1 \Rightarrow R(\alpha_i|x) \geq 0$ for $i \in \{1, 2, \cdots c\}$.

Thus, choose doubt will be minimum.

If $\lambda_r > \lambda_s$, we will never choose doubt

since, $R(\alpha_i|x) = \lambda_s(1 - P(w_i|x)) \leq \lambda_s$ $(0 \leq P(w_i|x) \leq 1)$

if $\lambda_r > \lambda_s$. which means $R(\alpha_i|x) < R(\alpha_{c+1}|x)$.

Thus, we will never choose doubt.