CS182 HW0

1. (a). My initial learning goal is to further exploring the world of Machine Learning. And lay a solid foundation for the courses that I want to take in the next semester, such as NLP.

(b). The LLM told me DL is a subset of ML which can solve more complex problems. So my learning goal doesn't change.

(c). enhance the understanding of ML. lay a foundation for further AI learning (such as NLP). Have the ability to do some funny and cool projects.

(d). The study group formed by my peers, the discussion, and the homework will benefit my understanding of DL.
Discuss with my peers and ask the TA if ~~we~~ I encounter
It will be helpful to some intractable problems in DL learning.

2. (a). Let $X = [x_1, x_2, \cdots x_n]^T$, $C = [c_1, c_2, \cdots c_n]^T$

$$\therefore X^TC = \sum_{i=1}^{n} x_i c_i$$

$$\therefore \frac{\partial}{\partial X}(X^TC) = [\frac{\partial X^TC}{\partial x_1}, \frac{\partial X^TC}{\partial x_2}, \cdots \frac{\partial X^TC}{\partial x_n}] = [c_1, c_2, \cdots c_n] = C^T$$

(b). $$\|X\|_2^2 = X^TX = \sum_{i=1}^{n} x_i^2$$

$$\therefore \frac{\partial}{\partial X}\|X\|_2^2 = [2x_1, 2x_2, \cdots 2x_n] = 2X^T$$

(c). $$Ax = \begin{bmatrix} \sum_{j=1}^{n} A_{1j}x_j \\ \sum_{j=1}^{n} A_{2j}x_j \\ \vdots \\ \sum_{j=1}^{n} A_{nj}x_j \end{bmatrix}$$ Let $A = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{nn} \end{bmatrix}$

$$\therefore \frac{\partial}{\partial X}(\sum_{j=1}^{n} A_{1j}x_j) = [A_{11}, A_{12}, \cdots A_{1n}]$$

$$\therefore \frac{\partial}{\partial X}AX = \begin{bmatrix} A_{11}, & \cdots & A_{1n} \\ A_{21}, & \cdots & A_{2n} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{nn} \end{bmatrix} = A.$$

(d). $$X^TAx = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i A_{ij} x_j$$

$$\therefore \frac{\partial}{\partial x_k}(X^TAx) = \frac{\partial}{\partial x_k}(\sum_{i=1}^{n}\sum_{j=1}^{n} x_i A_{ij} x_j) = \sum_{j=1}^{n} x_j A_{kj} + \sum_{i=1}^{n} x_i A_{ik} = (Ax)_k + (A^Tx)_k$$

$$\therefore \frac{\partial}{\partial X}(X^TAx) = [(Ax)_1 + (A^Tx)_1, (Ax)_2 + (A^Tx)_2, \cdots (Ax)_n + (A^Tx)_n]$$
$$= (Ax)^T + (A^Tx)^T$$
$$= A^Tx^T + x^TA$$
$$= x^T(A + A^T)$$

(e). It means: $x^T(A + A^T) = 2x^TA$

$\therefore$ $x^T(A - A^T) = 0$

$\because x^T \neq 0$

$\therefore A = A^T$

$\therefore A$ is symmetric, then $\frac{\partial}{\partial x}(x^TAx) = 2x^TA$.

3. (a). $\|Xw-y\|_2^2 = (Xw-y)^T(Xw-y) = w^TX^TXw - w^TX^Ty - y^TXw + y^Ty$

$\because w^TX^Ty$ is scalar

$\therefore (w^TX^Ty)^T = w^TX^Ty \Rightarrow y^TXw = w^TX^Ty$

$\therefore \|Xw-y\|_2^2 = w^TX^TXw - 2y^TXw + y^Ty$

$\frac{\partial}{\partial w}\|Xw-y\|_2^2 = w^T(X^TX + X^TX) - 2y^TX = 0$

$\therefore \quad w^TX^TX = y^TX.$

$\therefore \quad X^TXw = X^Ty$

$\therefore w = (X^TX)^{-1}X^Ty$, that's the w that minimizes $\|X-y\|^2$

(b). $X = U\Sigma V^T$

$\therefore w = (V\Sigma^TU^TU\Sigma V^T)^{-1}V\Sigma^TU^Ty$

$= (V\Sigma^T\Sigma V^T)^{-1}V\Sigma^TU^Ty$ $\qquad \Sigma^T\Sigma$: σ_i^2 diagonal.

$= V(\Sigma^T\Sigma)^{-1}V^TV\Sigma^TU^Ty$ $\qquad (\Sigma^T\Sigma)^{-1}$: $\frac{1}{\sigma_i^2}$

$= V(\Sigma^T\Sigma)^{-1}\Sigma^TU^Ty$ $\qquad (\Sigma^T\Sigma)^{-1}\Sigma^T$: $\frac{1}{\sigma_i^2}\cdot\sigma_i = \frac{1}{\sigma} = \Sigma^+$

$= V\Sigma^+U^Ty$

(c). $w^* = Ay$, where $A = V\Sigma^+U^T$

$AX = V\Sigma^+U^TU\Sigma V^T = V\Sigma^+\Sigma V^T$

Σ^+ is a $\frac{1}{\sigma_i}$ diagonal matrix and Σ is σ_i.

$\therefore \Sigma^+\Sigma = I_{n\times n}.$

$\therefore AX = VV^T = I_{n\times n}$

(d). $w = X^T\lambda$, $XX^T\lambda = y \Rightarrow \lambda = (XX^T)^{-1}y$

$\therefore w = X^T(XX^T)^{-1}y$

(e). $w = V\Sigma^TU^T(U\Sigma\Sigma^TU^T)^{-1}y$

$= V\Sigma^TU^TU(\Sigma\Sigma^T)^{-1}U^Ty$

$= V\Sigma^+U^Ty$

(f). $w^* = By$, $B = V\Sigma^+U^T$

$XB = U\Sigma V^TV\Sigma^+U^T = U\Sigma\Sigma^+U^T = UU^T = I_{m\times m}$

4. (a). $f = \|y - Xw\|_2^2 + \lambda \|w\|_2^2 = (y-Xw)^T(y-Xw) + \lambda w^T w$

$= y^T y - y^T X w - (Xw)^T y + w^T X^T X w + \lambda w^T w$

$= w^T X^T X w - 2 w^T X^T y + y^T y + \lambda w^T w$

$\frac{\partial f}{\partial w} = 2X^T X w - 2X^T y + 2\lambda w = 0$

$\therefore \ w = (X^T X + \lambda I)^{-1} X^T y$

(b). $X = U \Sigma V^T$

$X^T X = V \Sigma^T \Sigma V^T$

$\therefore \ X^T X + \lambda I = V(\Sigma^T \Sigma + \lambda I_d) V^T$

$w = V(\Sigma^T \Sigma + \lambda I_d)^{-1} V^T \cdot V \Sigma^T U^T y$

$= V(\Sigma^T \Sigma + \lambda I_d)^{-1} \Sigma^T U^T y$

$\frac{\sigma_i}{\sigma_i^2 + \lambda}$, when $\sigma_i << \lambda$, it becomes $\frac{\sigma_i}{\lambda} \to 0$.

when $\sigma_i >> \lambda$, it becomes $\frac{1}{\sigma_i}$

(c). We need to find $\underset{w}{\operatorname{argmax}} P(W=w \mid Y=y)$ in MAP

$\therefore \ \underset{w}{\operatorname{argmax}} P(W=w \mid Y=y) = \underset{w}{\operatorname{argmax}} P(Y=y \mid W=w) P(W=w)$

$\because \ Y = Xw + \sqrt{\lambda} N$, $\sqrt{\lambda} N \sim N(0, \lambda I)$

$\therefore \ P(Y=y \mid W=w) \propto \exp(-\frac{1}{2}(y-Xw)^T(\lambda I)^{-1}(y-Xw)) = \exp(-\frac{1}{2\lambda}\|y-Xw\|_2^2)$

$P(W=w) \propto \exp(0 - \frac{1}{2} w^T w) = \exp(-\frac{1}{2}\|w\|_2^2)$

$\therefore \ P(W=w \mid Y=y) \propto \exp(-\frac{1}{2\lambda}\|y-Xw\|_2^2 - \frac{1}{2}\|w\|_2^2)$

$\therefore \ -\log(W \mid Y) \propto \frac{1}{2\lambda}(\|y-Xw\|_2^2 + \lambda\|w\|_2^2)$

$\therefore$ The MAP estimate for w is $\|y-Xw\|_2^2 + \lambda\|w\|_2^2$, given $Y=y$.

(d). $\|\hat{y} - \hat{X}w\|_2^2 = \begin{bmatrix} y-Xw \\ -\sqrt{\lambda}w \end{bmatrix}^T \begin{bmatrix} y-Xw \\ -\sqrt{\lambda}w \end{bmatrix}$

$= (y-Xw)^T(y-Xw) + (-\sqrt{\lambda}w)^T(-\sqrt{\lambda}w)$

$= \|y-Xw\|_2^2 + \lambda\|w\|_2^2$

It's just the ridge regression objective.

OLS: $\hat{X}^T \hat{X} w = \hat{X}^T \hat{y}$

$\hat{X}^T \hat{X} = [X^T \ \sqrt{\lambda} I_d] \begin{bmatrix} X^T \\ \sqrt{\lambda} I_d \end{bmatrix} = X^T X + \lambda I_d$

$\hat{X}^T \hat{y} = [X^T \ \sqrt{\lambda} I_d] \begin{bmatrix} y \\ 0_d \end{bmatrix} = X^T y$

$\therefore\ w = (X^TX + \lambda I_d)^{-1} X^Ty$

(e). No idea.

(f). $y = \check{X}^{-1}y = \check{X}^T(\check{X}\check{X}^T)^{-1}y$

$\check{X}\check{X}^T = [X\ \sqrt{\lambda}I_n]\ [X\ \sqrt{\lambda}I_n]^T = XX^T + \lambda I_n$

$\therefore\ y = \check{X}^T \cdot (XX^T+\lambda I_n)^{-1}y = \begin{bmatrix} X^T(XX^T+\lambda I_n)^{-1}y \\ \sqrt{\lambda}(XX^T+\lambda I_n)^{-1}y \end{bmatrix}$

$\therefore$ The first d coordinates are $\hat{w} = X^T(XX^T+\lambda I_n)^{-1}y$

$\hat{w} = X^T(XX^T+\lambda I)^{-1}y$

Then, need to prove $X^T(XX^T+\lambda I)^{-1} = (X^TX+\lambda I)^{-1}X^T$

$\therefore\ (X^TX+\lambda I)X^T(XX^T+\lambda I)^{-1} = X^T$

$X^T(XX^T+\lambda I)(XX^T+\lambda I)^{-1} = X^T$

$\therefore\ X^T = X^T$

$\therefore\ \hat{w} = (XX^T+\lambda I)^{-1}X^Ty$

(g). when $\lambda \to \infty$, $X^TX+\lambda I \to \lambda I$, $(X^TX+\lambda I)^{-1} \approx \frac{1}{\lambda}I$

$\therefore\ w = \frac{1}{\lambda}X^Ty \to 0$

(h). when $\lambda \to 0$, $X^TX + \lambda I \Rightarrow X^TX$,

case 1: tall matrix: $(n>d)$. $\Rightarrow$ invertible

$\therefore\ w \to (X^TX)^{-1}X^Ty$

case 2: wide matrix: $(n<d)$. $\Rightarrow$ singular / not invertible.

$w \to X^T(XX^T)^{-1}y$

5/(a). (i). $\phi(x) = \max(0, wx+b)$

$\therefore e = -\frac{b}{w}$

(ii). $\frac{d\ell}{d\phi} = \phi(x) - y$

(iii). $\frac{d\ell}{dw} = \frac{d\ell}{d\phi} \cdot \frac{d\phi}{dw} = \begin{cases} (\phi(x)-y)x & \text{if } wx+b>0 \\ 0 & \text{if } wx+b \le 0 \end{cases}$

(iv). $\frac{d\ell}{db} = \frac{d\ell}{d\phi} \cdot \frac{d\phi}{db} = \begin{cases} \phi(x)-y & , \text{if } wx+b>0 \\ 0 & , \text{if } wx+b \le 0 \end{cases}$

(b). (i). $\phi(x) = 0$

$\therefore w' = w - \lambda \frac{\partial \ell}{\partial w} = w$, $e' = -\frac{b'}{w'} = e$

$b' = b - \lambda \frac{\partial \ell}{\partial b} = b$

$\therefore$ slop and elbow are unchanged.

(ii). $\frac{\partial \ell}{\partial w} = (wx+b-y)x = x$

$\frac{\partial \ell}{\partial b} = (wx+b-y) = 1$

$\therefore w' = w - \lambda x$, $b' = b - \lambda$

$\because x > 0$, $\lambda x > 0$, $w \downarrow$.

$\therefore$ slope decreases.

$e' = -\frac{b'}{w'} = \frac{\lambda - b}{w - \lambda x}$

numerical check: Let $w=b=1$, $\lambda = 0.1$, $x=1$

$\therefore e = -1$, $e' = \frac{-0.9}{0.9} = -1$

$\therefore$ no change in this case. but when $x=2$, $e'<e$

(iii). similarly: $w' = w - \lambda x$, $b' = b - \lambda$

$\because x < 0$

$\therefore w \uparrow$

$\therefore$ slope increases.

$e' = -\frac{b'}{w'} = \frac{\lambda - b}{w - \lambda x} = -\frac{b-\lambda}{w + \lambda |x|}$ $\quad w + \lambda|x| > w$, $b - \lambda < b$.

$\therefore e' > e$

$\therefore$ elbow increases. (shifts right) (iv). $w' = w - \lambda x, \quad b' = b - \lambda$

$\because x > 0$

$\therefore w \downarrow \Rightarrow$ slope decreases.

$e' = -\frac{b-\lambda}{w-\lambda x}$

numerical check: $w = -1, \ b = 2, \ x = 1, \ \lambda = 0.1$

$e = 2, \quad e' = -\frac{1.9}{-1.1} = 1.72 < 2 = e$

$\therefore$ elbow decreases. (shifts left).

(c). $\hat{f}(x) = W^{(2)} \Phi (W^{(1)} x + b)$ 1 hidden layer.

$W_i^{(1)} x + b_i = 0$

$\therefore e_i = -\frac{b_i}{W_i^{(1)}}$

(d). $\cancel{e_i' = e_i - \lambda \frac{\partial f}{\partial e}}$

$W_i^{(1)'} = W_i^{(1)} - \lambda \frac{\partial L}{\partial W_i^{(1)}} = W_i^{(1)} - \lambda (\hat{f}(x) - y) \cdot W^{(2)} \Phi' (W_i^{(1)} x + b)$

Let $z_i := W_i^{(1)} x + b_i$, $I(z_i > 0) = \begin{cases} 1, & z_i > 0 \\ 0, & z_i \le 0 \end{cases}$

$\therefore W_i^{(1)'} = W_i^{(1)} - \lambda (\hat{f}(x) - y) W_i^{(2)} x \, I(z_i > 0)$

Similarly: $b_i' = b_i - \lambda (\hat{f}(x) - y) W_i^{(2)} I(z_i > 0)$

$\therefore e_i' = -\frac{b_i'}{W_i^{(1)'}} = -\frac{b_i - \lambda (\hat{f}(x) - y) W_i^{(2)} I(z_i > 0)}{W_i^{(1)} - \lambda (\hat{f}(x) - y) W_i^{(2)} x \, I(z_i > 0)}, \quad z_i = W_i^{(1)} x + b_i$

7. (a). I used Grok to ask some questions about the homework.

(b). Nobody.

(c). About 6-7 hours.