247. HW1.

Yiwen Zhang. UID: 805899489.

1. Linear Algebra.

(a). Q is a real orthogonal matrix.

(i). Since Q is orthogonal, then $Q^TQ = QQ^T = I$.

For Q^T: $\begin{cases} (Q^T)^T \cdot Q^T = Q \cdot Q^T = I \\ Q^T \cdot (Q^T)^T = Q^TQ = I \end{cases}$, then Q^T is orthogonal.

For Q^{-1}: $\begin{cases} (Q^{-1})^T \cdot Q^{-1} = (Q^T)^{-1} \cdot Q^{-1} = (Q^TQ)^{-1} = I^{-1} = I \\ Q^{-1}(Q^{-1})^T = Q^{-1}(Q^T)^{-1} = (Q \cdot Q^T)^{-1} = I^{-1} = I \end{cases}$, then Q^{-1} is orthogonal.

(ii). For eigenvalue λ, eigenvector x for Q: (Any random λ and x).

$Q \cdot x = \lambda x$.

$\Rightarrow$ Premultiply Q^T: $Q^TQx = \lambda \cdot Q^Tx$

$\Rightarrow I \cdot x = \lambda \cdot Q^Tx$.

L2 norm of both sides: $\|x\|_2 = \|\lambda \cdot Q^Tx\| = \|\lambda\| \, \|Q^Tx\|_2$ (Since λ is a scalar).

$\Rightarrow \|x\|_2^2 = |\lambda|^2 \cdot \|Q^Tx\|_2^2$ and $\|Q^Tx\|_2^2 = (Q^Tx)^TQ^Tx = x^T \cdot QQ^T \cdot x = x^Tx = \|x\|_2^2$.

$\Rightarrow \|\lambda\|^2 = 1 \Rightarrow \|\lambda\| = 1 \Rightarrow$ all eigenvalues of A have norm of 1.

(iii). Since $\begin{cases} \det(Q^TQ) = \det(I) = 1 \\ \det(Q^T) \cdot \det(Q) = \det(Q)^2 = \det(Q^TQ) \end{cases}$

$\Rightarrow \det(Q)^2 = 1 \Rightarrow \det(Q) = \pm 1$

(iv). For $\forall b \in \mathbb{R}^n$, $\|A \cdot b\|_2^2 = (A \cdot b)^T \cdot Ab = b^T \cdot A^TA \cdot b = b^Tb = \|b\|_2^2$.

$\Rightarrow$ After transformation of A, the transformed vector still has the same legth as the origional vector.

(b). (i). For $A \in \mathbb{R}^{m\times n}$, we can have SVD decomposition as:

$A = U\Sigma V^T$, where $U \in \mathbb{R}^{m\times m}$, $\Sigma \in \mathbb{R}^{m\times n}$, $V^T \in \mathbb{R}^{n\times n}$

- Left singular vectors of A = Columns of U = Eigenvectors (Orthonormal) of AA^T
- Right singular vectors of A = Columns of V = Eigenvectors (Orthonormal) of A^TA.

(ii).

Let $A = U\Sigma V^T$, $\Sigma = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}$, $S = \text{diag}(\sigma_1, \cdots \sigma_r)$.

$A^T = V\Sigma^T U^T$.

$\Rightarrow AA^T = U\Sigma V^T V\Sigma^T U^T = U\Sigma\Sigma^T U^T = \underbrace{U \Delta_1 U^T}_{\text{Eigenvalue Decomposition}}$, $\Delta_1 = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}^T$

(Since V is orthogonal) $= \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} S^T & 0 \\ 0 & 0 \end{bmatrix}$

$= \begin{bmatrix} SS^T & 0 \\ 0 & 0 \end{bmatrix} = \text{diag}(\sigma_1^2, \cdots \sigma_r^2, 0 \cdots 0)$ (Orthogonal)

$A^TA = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T\Sigma V^T = \underbrace{V \cdot \Delta_2 V^T}_{\text{Eigenvalue Decomposition}}$, $\Delta_2 = \Delta_1 = \text{diag}(\sigma_1^2 \cdots \sigma_r^2, 0 \cdots 0)$.

(Since U is orthogonal) (Orthogonal)

$\Rightarrow$ Singular value of $A = \sqrt{\text{Eigenvalue of } AA^T \text{ or } A^TA}$

(c). (i). False.

Because eigenvalues can be equal to each other.

i.e. For $I \in \mathbb{R}^{n\times n}$, $I = U \cdot \Delta \cdot U^T$, where $\Delta = I$. U is orthogonal,

here I has n same eigenvalues: 1.

(ii). False.

i.e. $A \cdot v_1 = \lambda_1 \cdot v_1$. $A \cdot v_2 = \lambda_2 \cdot v_2$.

let $v_3 = av_1 + bv_2$.

$A \cdot v_3 = A \cdot av_1 + A \cdot bv_2$

$= a\lambda_1 \cdot v_1 + b\lambda_1 \cdot v_2$

If $Av_3 = \lambda_3 v_3$, then $a(\lambda_1 - \lambda_3)v_1 = b(\lambda_2 - \lambda_3)v_2$, In general, this doesn't always hold.

(iii). True.

For $Av_1 = \lambda_1 \cdot v_1 \Rightarrow v_1^T A v_1 = \lambda_1 \cdot v_1^T v_1$.

Since $b^T A b \geq 0 \Rightarrow \lambda_1 \cdot v_1^T v_1 = \lambda_1 \cdot \|v_1\|_2^2 \geq 0$, where $\|v_1\|_2^2 > 0$

$\Rightarrow \lambda_1 \geq 0$.

(iv). False.

"r" represents the number of linear independent eigenvectors,

so it couldn't be larger than the number of distinct eigenvalues.

(v). True.

For $\begin{cases} Av_1 = \lambda v_1 \\ Av_2 = \lambda v_2 \end{cases}$, let $v_3 = av_1 + bv_2$.

$$Av_3 = Aa \cdot v_1 + A \cdot b \cdot v_2 = a \cdot \lambda v_1 + b \cdot \lambda v_2$$
$$= \lambda(av_1 + bv_2)$$
$$= \lambda \cdot v_3. \text{ Still an eigenvector.}$$

2. Probability.

(a). Suppose we define following events:

H5: Get coins of H50

H6: Get coins of H60

D: Heads up

T: Tails up.

(i). $P(H5|T) = \frac{P(T \cap H5)}{P(T)} = \frac{P(H5) \cdot P(T|H5)}{\sum P(T|Hi) \cdot P(Hi)} = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{2}{5}} = \frac{0.25}{0.25 + 0.2} = \frac{5}{9}$

(ii) $P(H5|THHH) = \frac{P(H5 \cap THHH)}{P(THHH)}$ $\quad P(H5 \cap THHH) = P(H5) \cdot P(THHH|H5) = \frac{1}{32} \approx 0.3125$

$P(THHH) = P(THHH|H5) \cdot P(H5) + P(THHH|H6) \cdot P(H6)$

$= \frac{1}{2} \times (\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}) + \frac{1}{2} \times (\frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{3}{5})$

$= \frac{1}{32} + \frac{27}{625} = 0.03125 + 0.0432 = 0.07445$

$\Rightarrow P(H5|THHH) = \frac{0.03125}{0.07445} \approx 0.4197$

(iii). $P(H50|H9T1) = \frac{P(H50 \cap H9T1)}{P(H9T1)} = \frac{P(H50)\cdot P(H9T1|H50)}{P(H9T1)}$.

$P(H55|H9T1) = \frac{P(H55)\cdot P(H9T1|H55)}{P(H9T1)}$

$P(H60|H9T1) = \frac{P(60)\cdot P(H9T1|H60)}{P(H9T1)}$

$P(H9T1) = \sum P(Hi)\cdot P(H9T1|Hi)$

$= P(H50)\cdot P(H9T1|H50) + P(H55)\cdot P(H9T1|H55) + P(60)\cdot P(H9T1|H60)$

$= \frac{1}{3}\times(\frac{1}{2})^9\times(\frac{1}{2}) + \frac{1}{3}\times(0.55)^9\times 0.45 + \frac{1}{3}\times(0.6)^9\times 0.4$.

$= a + b + c$, where $a = \frac{1}{3}\times(\frac{1}{2})^{10}$, $b = \frac{1}{3}\times 0.55^9\times 0.45$, $c = \frac{1}{3}\times 0.6^9\times 04$

$\Rightarrow P(H50|H9T1) = \frac{a}{a+b+c}$

$P(H55|H9T1) = \frac{b}{a+b+c}$

$P(H60|H9T1) = \frac{c}{a+b+c}$

(b). Suppose we define the following events:

S: student from Science. $P(S) = 15\%$.

H: student from Health care. $P(H) = 21\%$

L: student from liberal arts. $P(L) = 24\%$.

E: student from Engeering. $P(E) = 40\%$

A: student likes the lecture.

$P(S|A) = \frac{P(S\cap A)}{P(A)}$.

$P(A) = \sum_{X\in\{S,H,L,E\}} P(A|X)\cdot P(X) = 15\%\times 90\% + 21\%\times 18\% + 24\%\times 0 + 40\%\times 10\%$

$= 0.135 + 0.0378 + 0.04$

$= 0.2128$

$P(S\cap A) = P(S)\cdot P(A|S) = 15\%\times 90\% = 0.135$.

$\Rightarrow P(S|A) = 0.135/0.2128 \approx 0.634$

(c). Create a table like this:

P	Preg	Nonpreg
Postive	99%	10%
Negtive	1%	90%

$P(preg) = 1\%$.

$$P(Preg \mid pos) = \frac{P(preg \cap pos)}{P(pos)} = \frac{P(preg) \times P(pos \mid preg)}{P(pos \mid preg) \times P(preg) + P(pos \mid unpreg) \times P(unpreg)}$$

$$= \frac{1\% \times 99\%}{1\% \times 99\% + 99\% \times 10\%} = \frac{1}{11}$$

It means the testing result couldn't define that people really get pregnant or not. because in all women, probability of pregnant is just 1%.
To get more precise answer, people need to do a second test.

(d). $E(AX+b) = E(AX) + E(b)$

$= A \cdot E(X) + E(b) = A \cdot E(X) + b$

If vector x has mean m, $m \in \mathbb{R}^n$, then $E(AX+b) = A \cdot m + b \in \mathbb{R}^n$.

(e). $cov(X) = E\left((X - E(X))(X - E(X))^T\right)$, Let $y = Ax + b$

$cov(y) = E\left((AX + b - E(AX+b))(y^T - E(y)^T)\right)$

$= E\left((AX - A \cdot E(X))(x^T A^T + b^T - (A \cdot E(X) + b)^T)\right)$

$= E\left((AX - A \cdot E(X))(X^T A^T + b^T - E(X)^T \cdot A^T - b^T)\right)$

$= E\left(A(X - E(X)) \cdot (X^T - E(X)^T) \cdot A^T\right)$

$= A \cdot E\left((X - E(X))(X - E(X))^T\right) \cdot A^T$

$= A \cdot cov(X) \cdot A^T$.

3. Multivariate derivatives.

(a) $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times m}$, Find $\nabla_x x^T A y$.

$f(x) = x^T A y \in \mathbb{R}^{1 \times 1}$ is a scalar, $\nabla_x f(x) \in \mathbb{R}^n$.

Since $\frac{\partial x^T \cdot a}{\partial x} = \frac{\partial a^T \cdot x}{\partial x} = a$, (referring to "matrixcook" and lecture notes)

then $\nabla_x f(x) = \frac{\partial x^T \cdot (Ay)}{\partial x} = \frac{\partial (Ay)^T \cdot x}{\partial x} = A \cdot y \in \mathbb{R}^n$.

(b). Find $\nabla_y x^T A y$.

Since $\frac{\partial \theta^T x}{\partial x} = \theta$, then $\nabla_y x^T A y = \frac{\partial (A^T x)^T y}{\partial y} = A^T x \in \mathbb{R}^{m \times 1}$.

(c). Find $\nabla_A x^T A y$.

Since $\frac{\partial a^T x b}{\partial x} = a b^T$ (referring to matrixcook book).

Then $\nabla_A x^T A y = x y^T \in \mathbb{R}^{n \times m}$.

(d). $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$, $f = x^T A x + b^T x$, find $\nabla_x f(x)$.

f is a scalar, $x \in \mathbb{R}^n$. $\nabla_x f(x) \in \mathbb{R}^n$,

$$\begin{aligned} \nabla_x f(x) &= \nabla_x (x^T A x + b^T x) \\ &= \nabla_x (x^T A x) + \nabla_x (b^T x) \\ &= (A + A^T) x + b \quad \in \mathbb{R}^n. \end{aligned}$$

(e). $A, B \in \mathbb{R}^{n \times n}$, $f = tr(AB)$, find $\nabla_A f$.

Since $\frac{\partial Tr(XA)}{\partial x} = A^T$ (referring to matrixcook),

f is scalar, $A \in \mathbb{R}^{n \times n}$, so $\nabla_A f \in \mathbb{R}^{n \times n}$,

$$\nabla_A f = \nabla_A Tr(AB) = B^T.$$

(f). $f = tr(BA + A^TB + A^2B)$, find $\nabla_A f$.

Referring to $\frac{\partial tr(AXB)}{\partial X} = A^TB^T$, $\frac{\partial tr(X^TA)}{\partial X} = A$, $\frac{\partial tr(X^2A)}{\partial X} = (XA + AX)^T$.

$$\nabla_A f = \nabla_A (tr(BAI + A^TB + A^2BI)$$

$$= \nabla_A (tr(B \cdot A \cdot I)) + \nabla_A (tr(A^TBI)) + \nabla_A (tr(A^2BI))$$

$$= B^T \cdot I^T + B + (AB + BA)^T$$
$$= B^T + B + (AB + BA)^T$$

(g). $f = \|A + \lambda B\|_F^2$, find $\nabla_A f$.

Referring to $\frac{\partial}{\partial X}\|X\|_F^2 = \frac{\partial}{\partial X} Tr(XX^H) = 2X$.

then $\frac{\partial f}{\partial A} = \frac{\partial}{\partial A} Tr(A + \lambda B)(A + \lambda B)^T$ (since A、B are both real matrix).

$$= \frac{\partial}{\partial A} Tr(AA^T + \lambda AB^T + \lambda BA^T + \lambda^2 BB^T)$$

$$= \frac{\partial}{\partial A} Tr(AA^T) + \frac{\partial}{\partial A} Tr(\lambda AB^T) + \frac{\partial}{\partial A} Tr(\lambda BA^T) + \frac{\partial}{\partial A} Tr(\lambda^2 BB^T)$$

$$= 2A + \lambda B + \lambda B + 0$$
$$= 2(A + \lambda B) \in \mathbb{R}^{n \times n}$$

4. Derive least-square with matrix derivatives.

$\min \frac{1}{2}\sum_{i=1}^{n} \|y^{(i)} - W\cdot x^{(i)}\|^2$ $(*)$.

Let $y^{(i)}, x^{(i)} \in \mathbb{R}^m$ for $i \in [n]$, then $W \in \mathbb{R}^{m\times m}$

$\|y^{(i)} - W\cdot x^{(i)}\|^2 = (y^{(i)} - W\cdot x^{(i)})^T \cdot (y^{(i)} - W\cdot x^{(i)})$

Make $\tilde{Y} = [y_1, y_2 \cdots y_n]$, $\tilde{X} = [Wx_1, Wx_2 \cdots Wx_n]$.

Then $\tilde{Y} - \tilde{X} = [y_1 - Wx_1,\ y_2 - Wx_2,\ \cdots y_n - Wx_n] \in \mathbb{R}^{m\times n}$

$(\tilde{Y} - \tilde{X})^T = \begin{bmatrix} (y_1 - Wx_1)^T \\ \vdots \\ (y_n - Wx_n)^T \end{bmatrix} \in \mathbb{R}^{n\times m}$.

$$\Rightarrow A = (\tilde{Y} - \tilde{X})^T \cdot (\tilde{Y} - \tilde{X}) = \begin{bmatrix} (y_1 - Wx_1)^T(y_1 - Wx_1) & \cdots & (y_1 - Wx_1)^T(y_n - Wx_n) \\ (y_2 - Wx_2)^T(y_1 - Wx_1) & \ddots & \vdots \\ \vdots & \cdots & (y_n - Wx_n)^T(y_n - Wx_n) \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Then $\min \frac{1}{2}\sum_{1}^{n} \|y^{(i)} - Wx^{(i)}\|^2 = \min \frac{1}{2} tr(A)$.

where $A = (Y - WX)^T(Y - WX) = Y^TY - Y^TWX - X^TW^TY + X^TW^TWX$.

$\frac{\partial tr(A)}{\partial W} = \frac{\partial}{\partial W}(tr Y^TY - tr(Y^TWX) - tr(X^TW^TY) + tr(X^TW^TWX))$

$= 0 - YX^T - YX^T + \frac{\partial}{\partial W} tr(WXX^TW^T)$ $(tr(AB) = tr(BA))$

$= W(XX^T + XX^T) - 2YX^T$

$\Rightarrow W = YX^T \cdot (XX^T)^{-1}$

5. Regularized least squares.

$L(\theta) = \underbrace{\frac{1}{2}\Sigma (y^{(i)} - \theta^T x^{(i)})^2}_{L_1(\theta)} + \underbrace{\frac{\lambda}{2}\|\theta\|_2^2}_{L_2(\theta)}$, need to minimize $L(\theta)$.

$\partial L(\theta)/\partial\theta = \partial L_1(\theta)/\partial\theta + \partial L_2(\theta)/\partial\theta$, $\frac{\lambda}{2}\|\theta\|_2^2 = \frac{\lambda}{2}\theta^T\cdot\theta$. $\Rightarrow \frac{\partial}{\partial\theta}(\frac{\lambda}{2}\theta^T\theta) = \frac{\lambda}{2}\frac{\partial}{\partial\theta}(\theta^T\theta)$.

let $S = \theta^T\theta = \sum_{i=1}^{n}\theta_i^2$, $\frac{\partial S}{\partial\theta_1} = 2\theta_1$, $\cdots \frac{\partial S}{\partial\theta_n} = 2\theta_n \Rightarrow \frac{\partial S}{\partial\theta} = \begin{bmatrix} \partial S/\partial\theta_1 \\ \vdots \\ \partial S/\partial\theta_n \end{bmatrix} = 2\theta$

$\Rightarrow \frac{\partial L(\theta)}{\partial\theta} = X^TX\theta - X^TY + \frac{\lambda}{2}\cdot 2\theta = (X^TX + \lambda\cdot I)\theta - X^TY$

Let $\partial L(\theta)/\partial\theta = 0$, $\Rightarrow \theta = (X^TX + \lambda I)^{-1}\cdot X^TY$.