Question 1:

1.1 $A \in \mathbb{R}^{m \times n}$ & $B \in \mathbb{R}^{n \times m}$. Prove:

a) if $m = n$, A & A^T have same eigenvalues.

→ eigenvalue & eigenvector eqn: $Av = \lambda v$; $\lambda \rightarrow$ eigenvalues, $v \rightarrow$ eigenvector

$$(Av)^T = (\lambda v)^T$$

$$v^T A^T = \lambda \cdot v^T \Rightarrow v^T A^T v = \lambda v^T v \quad (v \neq 0)$$

$$\frac{v^T A^T v}{v^T v} = \lambda$$

Char eqn $\Rightarrow \det(A - \lambda I) = 0$

for $A \Rightarrow \det(A - \lambda I) = 0$

for $A^T \Rightarrow \det(A^T - \lambda I) = 0$

and $m = n \Rightarrow$ dimensions of A & A^T are the same.

$$A^T - \lambda I = (A - \lambda I)^T$$

therefore; $m = n$; and

$$\det(A - \lambda I) = \det(A^T - \lambda I)$$

$\Rightarrow$ so characteristic equations are the same for A & A^T; therefore, this implies that they have the same roots and the roots are the eigenvalues.

$\Rightarrow$ so if $m = n$; then A and A^T have the same set of eigenvalues.

let $m = n = 2$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad A^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

$$\det(A - \lambda I) = 0$$

$$\det \begin{bmatrix} a - \lambda I & b \\ c & d - \lambda I \end{bmatrix} = 0$$

$$\det(A^T - \lambda I) = 0$$

$$\det \begin{bmatrix} a - \lambda I & c \\ b & d - \lambda I \end{bmatrix} = 0$$

so we see that

$$\det(A - \lambda I) = \det(A^T - \lambda I)$$

$Ax = \lambda x$

1.1

b) AB and BA have the same set of eigenvalues.

Let λ_1 eigenvalue of AB, v_1 eigenvector of AB:

Then

$(AB)v_1 = \lambda_1 v_1$

$BAB v_1 = B\lambda_1 v_1$

$(BA)(Bv_1) = \lambda_1(Bv_1)$ $\Rightarrow$ so λ_1 is eigenvalue of BA.

(Bv_1 = x, Bv_1 = x)

can look at it from BA's perspective as well.

Let λ_2 eigenval of BA, v_2 eigenvec. of BA;

Then

$(BA)v_2 = \lambda_2 v_2$

$ABA v_2 = A\lambda_2 v_2$

$(AB)(Av_2) = \lambda_2(Av_2)$ $\Rightarrow$ so λ_2 is eigenvalue of AB.

(Av_2 = y, Av_2 = y)

$\Rightarrow$ Then, every eigenvalue of AB is also an eigenvalue of BA; and every eigenvalue of BA is also an eigenvalue of AB. So, we proved that AB & BA have the same set eigenvalues.

1.2 Rank;

Prove if $A \in \mathbb{R}^n$ is a full matrix, matrices B and $A^{-1}BA$ have same eigenvalues.

$\Rightarrow A \in \mathbb{R}^n$, A is full rank $\Rightarrow$ so, its determinant is nonzero, and it is invertable.

$\Rightarrow \det(ABA^{-1}) = \det(B)$ because of the multiplication property of determinants.

$\Rightarrow \det(A^{-1}) = \frac{1}{\det(A)}$.

use the characteristic eqn $\det(X - \lambda I) = 0$ below.

$\det(A^{-1}BA - \lambda I) = 0$

$\Rightarrow \det(A^{-1}BA - \lambda I) = \det(A(A^{-1}BA - \lambda I)A^{-1}) = \det((BA - \lambda AI)A^{-1})$

$= \det(B - \lambda AIA^{-1}) = \det(B - \lambda I)$.

$\Rightarrow$ so that means

$\det(A^{-1}BA - \lambda I) = \det(B - \lambda I)$. Therefore, we proved that both $A^{-1}BA$ and B have the same set eigenvalues to satisfy the equation above.

1.3 symm. matrix $A \in \mathbb{R}^n$ is +ve definite if for all non-zero vectors $x \in \mathbb{R}^n$, $x^TAx > 0$

symm. matrix $A \in \mathbb{R}^n$ is +ve semidefinite if for all vectors $x \in \mathbb{R}^n$, $x^TAx \geq 0$

Prove

a) if every eigenvalue of A is +ve, A is +ve definite matrix.

for positive definite matrix, we need to prove the following:

Any nonzero vector $x \in \mathbb{R}^n$ such that $x^TAx > 0$

Since A is symmetric, we have n linearly independent eigenvectors and eigenvalues

$v_1, v_2, v_3 \ldots v_n$ and $\lambda_1, \lambda_2 \ldots \lambda_n$.

x can be written as a linear combination of these eigenvectors of A.

$x = c_1v_1 + c_2v_2 + \ldots + c_nv_n$ ($c_1, c_2 \ldots c_n$ are coefficients)

$x^TAx > 0$ for positive definite:

$(c_1v_1 + c_2v_2 + \ldots + c_nv_n)^TA(c_1v_1 + c_2v_2 + \ldots + c_nv_n) = x^TAx$

$= c_1^2\lambda_1 + c_2^2\lambda_2 + \ldots + c_n^2\lambda_n > 0$.

Each of these terms (like $c_1^2\lambda_1, c_2^2\lambda_2 \ldots$) are positive since all eigenvalues of A are positive. Therefore, the sum is also positive;

so $x^TAx > 0$ for any nonzero vector $x \in \mathbb{R}^n$; so A is positive definite.

b) $\det(A) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$ and $\lambda_i > 0$ for positive definite A.

if matrix A is positive definite; then $x^T A x > 0$ for any non zero $x \in \mathbb{R}^n$.

To prove A is full rank and invertible, let's assume $A \in \mathbb{R}^n$ is not full rank for now. Then, assume jth column can be expressed as linear combination of other $n-1$ columns.

$a_j = \sum_{i \neq j} x_i a_i$ for some $x_1, x_2 \ldots x_{j-1}, x_{j+1} \ldots x_n \in \mathbb{R}$. Setting $x_j = -1$, we have $Ax = \sum_{i=1}^{n} x_i a_i = 0$. But this implies that $x^T A x = 0$ for some non-zero vector x; so A must be neither positive nor negative definite. Therefore, A has to be either positive or negative definite, it must be full rank. We know A is positive definite; therefore A has to be full rank and hence, invertible.

1.4 $A \in \mathbb{R}^{m \times n}$, $G = A^T A$ is a Gram matrix. Prove

a) G is always +ve semidefinite

Assume x is non-zero column vector in $\mathbb{R}^n$:

$x^T G x = x^T (A^T A) x = (xA)^T (Ax) = (Ax)^T (Ax) = \|Ax\|^2$ because x is a column vector.

So $\|Ax\|^2$ is a scalar and is always non negative; either 0 or positive

but we also know Ax is non zero. $\|Ax\|^2 \geq 0$

if $\|Ax\|^2 > 0$, then $x^T G x \geq 0$ for any x.

and if $x > 0$; then $\|Ax\| > 0$ and therefore $\|Ax\|^2 = x^T G x > 0$

But in our case:

We also know G (Gram matrix) is positive semidefinite, since $x^T G x \geq 0$ for any x and $x^T G x \geq 0$ for any non-zero x, Therefore, we proved that G is always positive semidefinite matrix

b) if $m \geq n$ and A is full rank, G is +ve definite matrix.

let x be non-zero vector in $\mathbb{R}^n$.

$x^T G x = x^T (A^T A) x = (Ax)^T (Ax) = \|Ax\|^2$; again $\|Ax\|$ is the norm of Ax.

A is full rank $\Rightarrow$ only vector $x \in \mathbb{R}^n$ to satisfy this is $Ax = 0$.

$\Leftrightarrow$ null space (A) is trivial.

$Ax \neq 0$ when A is full rank

So, $\|Ax\| > 0$ for any non-zero x and $x^T G x > 0$.

So, we proved that $G = A^T A$ is positive definite when A is full rank and $m \geq n$.

1.5 $f(x,y) = \begin{bmatrix} f_1(x,y) \\ f_2(x,y) \\ f_3(x,y) \end{bmatrix} = \begin{bmatrix} x^2y^3 \\ 4x^2 + \cos y \\ 4y^2 - 2x^2 \end{bmatrix}$

a) dim of Jacobian Matrix?

$$J = \begin{bmatrix} \partial f_1/\partial x & \partial f_1/\partial y \\ \partial f_2/\partial x & \partial f_2/\partial y \\ \partial f_3/\partial x & \partial f_3/\partial y \end{bmatrix} \Rightarrow \text{So dimensions are } \underline{3 \times 2}$$

b) $\dfrac{\partial f_2}{\partial y} = \dfrac{\partial(4x^2 + \cos y)}{dy} = \underline{-\sin(y)}$

c) $\dfrac{\partial^2 f_1}{\partial x \partial y} = \dfrac{\partial}{\partial x}\left(\dfrac{\partial f_1}{\partial y}\right) = \dfrac{\partial}{\partial x}\left(3y^2x^2\right) = 3y^2 \cdot 2x = \boxed{6xy^2}$

1.6 A $n \times n$, B $n \times m$, $(\lambda I + AB)$ is $n \times m$ and invertable.

•) Prove $n \times n$ $(\lambda I + BA)$ is invertable. Hint: Show $(\lambda I + BA)x = 0$ implies $(\lambda I + AB)y = 0$, where $y = Ax$

1.6 a)

Using the hint, we can say that $(\lambda I + BA)x = 0$ then

$\Rightarrow \lambda Ix + BAx = 0 \Rightarrow BAx = -\lambda Ix = -\lambda x$

$ABAx = -\lambda Ax$

$ABAx + \lambda Ax = 0 \Rightarrow \underline{A(\lambda I + BA)x = 0}$

This means that $Ax = 0$; because $\lambda I + AB$ is invertible. So;

$BAx = -\lambda Ix = -\lambda x = 0$

$= (\lambda I + BA)x = 0$

Only possible solution is $x = 0$; therefore, $\lambda I + BA$ must be invertable because the equation has only the trivial solution (0)

1.6 b

b) Prove $B(\lambda I + AB)^{-1} = (\lambda I + BA)^{-1} B$

Use this: $B(\lambda I + AB) = (\lambda I + BA) B$ ⇒ prove this first.

$B(\lambda I + AB) = \lambda BI + BAB$

$(\lambda I + BA) B = \lambda BI + BAB$

} so they are equal, so we prove that $B(\lambda I + AB) = (\lambda I + BA) B$

use this idea here

$(\lambda I + BA) B (\lambda I + AB)^{-1} = (\lambda I + BA)^{-1} B (\lambda I + AB)$

let y (left) let y (right)

$y(\lambda I + AB)^{-1} = (\lambda I + BA)^{-1} y$ ⇒ since $(\lambda I + AB)$ and $(\lambda I + BA)$ are invertible.

we can the following:

Using the proof from above, and knowing $(\lambda I + AB)$ and $(\lambda I + BA)$ are both invertible; we prove that $y(\lambda I + AB)^{-1} = (\lambda I + BA)^{-1} y$

and

that $(\lambda I + BA) B (\lambda I + AB)^{-1} = (\lambda I + BA)^{-1} B (\lambda I + AB)$

and $\boxed{B(\lambda I + AB)^{-1} = (\lambda I + BA)^{-1} B}$

2.1 Regression Coefficients

c) It is not possible to comment on the importance of this feature without additional information, because the coefficient value is not enough to determine the importance of this feature. There are other factors that also affect our understanding of this feature such as the quality of the data, normalization/standardization steps in the pre-processing of the data, fit of the model and maybe more importantly whether there is multi-collinearity which could lead to unreliable coefficients when there is highly correlated features, or depends on what you are modelling with your regression model. Example: assume there is a line approximation with the following: $0x_1 + 0x_2 + 0x_3 + \dots 0x_n - 1000 = 0$. It is just a straight line. Does it mean it has strong effect or it should be ignored? Not really.

2.2

n training w m features, target $y = [y^{(0)}, \dots y^{(n)}] \in \mathbb{R}^n$ and $X = [x^{(0)}, \dots x^{(n)}]$ $\in \mathbb{R}^{n \times m}$. x_j denotes col of this matrix.

a) Show training regressor on one feature, then we have $w_j = \frac{x_j^T y}{x_j^T x_j}$

We know the optimal weights for regressor: $w = (X^T X)^{-1} X^T y$

if we have only 1 feature, then $X = [x^{(0)}, \dots x^{(n)}]$ with shape $(n, 1)$.

Then x^T has the shape $(1, n)$;

$$x^T x \rightarrow \underset{(1,n)}{X^T} \underset{(n,1)}{X} = \text{scalar with shape } (1,1) \text{ or size } 1 \times 1.$$

Because X was given to us as $X \in \mathbb{R}^{n \times m}$; with one feature $x_j \in \mathbb{R}^{n \times 1}$

Then we can express the equation by taking the inverse of a scalar. (reciprocal)

$$\Rightarrow w_j = (x_j^T x_j)^{-1} x_j^T y = \boxed{\frac{x_j^T y}{x_j^T x_j} = w_j}$$

b) Suppose columns of X orthogonal. Prove optimal params from training regressor on all features are the same as optimal params resulting from training on each feature independently.

columns of X are orthogonal, so their internal multiplication is zero.

$$X^TX = \text{diagonal}\left(x_1^Tx_1, x_2^Tx_2, \dots x_m^Tx_m\right) \Rightarrow (X^TX)^{-1} = \text{diagonal}\left((x_1^Tx_1)^{-1}, \dots (x_m^Tx_m)^{-1}\right)$$

$\Rightarrow w = (X^TX)^{-1}X^Ty \rightarrow$ optimal parameters eqn.

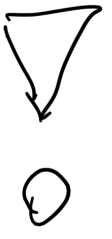
$$w = \text{diagonal}\left((x_1^Tx_1)^{-1}, \dots (x_m^Tx_m)^{-1}\right)X^Ty$$

$$w_j = \left(\text{diagonal}\left((x_1^Tx_1)^{-1}, \dots (x_m^Tx_m)^{-1}\right)X^Ty\right)_j = \left(\text{diagonal}\left((x_1^Tx_1)^{-1}, \dots (x_m^Tx_m)^{-1}\right)\right)_j (X^Ty)_j$$

$$= (x_j^Tx_j)^{-1}(X^Ty)_j = \frac{x_j^Ty}{x_j^Tx_j} = \frac{(X^Ty)_j}{x_j^Tx_j} = \frac{x_j^Ty}{x_j^Tx_j}$$

The eqn from part a) combined with this shows that optimal parameters from training the regressor on all features is the same as optimal parameters resulting from training the regressor on each feature independently.

Solution for Question 2.3 in the next page



2.3 $X \in \mathbb{R}^{n \times d}$ as training data w n samples, d features, $y \in \mathbb{R}^n$ as labels.

Using L-2 norm w L-2 regularization, minimization problem becomes.

$$\arg\min_{w \in \mathbb{R}^d} \frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2 \quad ; \lambda \rightarrow \text{tuning hyperparam. Setting}$$

gradient to zero yields following:

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

a) Is $(X^T X + \lambda I_d)$ is always invertible. Prove it, otherwise, what conditions for X?

We can use the Gram matrix identity and positive definite case. If $(X^T X + \lambda I_d)$ is positive definite, then it is full rank and invertible.

$$A^T(X^T X + \lambda I_d)A > 0 \quad , \quad G = A^T A \text{ for Gram Matrix}$$

$X^T X$ is a symmetric matrix and it can be written with their eigen decomposition

$X^T X = \Theta \Lambda \Theta^{-1} = V \Lambda V^T$ where V is the orthogonal matrix of eigenvectors and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\Lambda_{ii} = d_i$. Since $X^T X$ is semi positive definite, its eigenvalues are non negative.

The eigenvalues of $(X^T X + \lambda I_d)$ are sum of eigenvalues of $X^T X$ and λI_d. I is positive definite. Since $\lambda > 0$, all eigenvalues of $(X^T X + \lambda I_d)$ are positive; and we proved earlier that in this case, $(X^T X + \lambda I_d)$ is positive definite, and therefore invertable.

b) If $n < d$ solution w can be computed faster using eqv. expression for w by calculating the inverse of $n \times n$ matrix instead of $d \times d$ matrix. Find efficient expression for w.

Push Through Identity: $(\lambda I + AB)$ is invertable, for $A_{m \times n}$, $B_{n \times m}$

$$B(\lambda I + AB)^{-1} = (\lambda I + BA)^{-1} B$$

let $X \in \mathbb{R}^{n \times d}$ and $X^T \in \mathbb{R}^{d \times n}$ given above. $y \in \mathbb{R}\ \mathbb{R}^n$

if $n < d$; w can be calculated as the inverse of $n \times n$ matrix.

$w = (X^T X + \lambda I_d)^{-1} (X^T y)$ but from the push through identity

w becomes the following $\Rightarrow$ $w = X^T (X X^T + \lambda I_d) y$

$$\boxed{w = X^T (X X^T + \lambda I_d) y}$$

Solution for Question 2.4 is in the next page

2.4 Binary classifier h, a function from feature domain to $\{0,1\}$: $h: X \to \{0,1\}$ error rate is $R(h) = P[Y \neq h(X)]$ prove that optimal h that minimizes $R(h)$ as follows: $h^*(x) = 1$ if $m(x) > 1/2$, otherwise 0; $m(x) = E(Y|X=x) = P(Y=1|X=x)$.

$m(x)$ denotes regression function.

h^* is called Bayes rule. Risk $R^* = R(h^*)$ → Bayes risk; $\{x \in X : m(x) = 1/2\}$ ⇒ Bayes decision boundary

Show that $R(h) - R(h^*) \geq 0$. Note $R(h) = P(\{Y \neq h(X)\}) = \int P(Y \neq h(X) | X=x) dP_X(x)$

it suffices to show that

$P(Y \neq h(X) | X=x) - P(Y \neq h^*(X) | X=x) \geq 0$ for all $x \in X$

$P(Y \neq h(X) | X=x) = 1 - P(Y = h(X) | X=x)$ → comes from Basic Probability

$= 1 - \left(P(Y=1, h(X)=1 | X=x) + P(Y=0, h(X)=0 | X=x) \right)$

$= 1 - \left(h(x) P(Y=1 | X=x) + (1-h(x)) P(Y=0 | X=x) \right)$

$= 1 - \left(h(x) m(x) + (1-h(x))(1-m(x)) \right)$

Therefore: $P(Y \neq h(X) | X=x) - P(Y \neq h^*(X) | X=x)$

$= \left(h^*(x) m(x) + (1-h^*(x))(1-m(x)) \right) - \left(h(x) m(x) + (1-h(x))(1-m(x)) \right)$

$= (2m(x) - 1)(h^*(x) - h(x)) = 2\left(m(x) - \frac{1}{2}\right)(h^*(x) - h(x))$ ← this equation

when $m(x) \geq 1/2$ and $h^*(x) = 1$, equation above is non-negative. When $m(x) < 1/2$ and $h^*(x) = 0$, leads to the equation above being also non-negative. This proves that $P(Y \neq h(X) | X=x) - P(Y \neq h^*(X) | X=x) \geq 0$ for all $x \in X$.