PROBLEM - 14.4 =

Multi class classifier via matrix least Squares.

Show that the coefficient vectors $\beta_1, \beta_2 -- \beta_K$ can be found by solving the matrix least squares problem of minimizing $\|X^T\beta - Y\|^2$ where β is the $n \times K$ Matrix with columns $\beta_1 ---- \beta_K$ and Y is an $N \times K$ Matrix

While training the model =

1 = N (n-vectors) are given for training.

2 = correct outcomes of labels for these Inputs are also given.

3 = for these values of X and Y we can say least square Problem is minimized.

4 = Hence differentiating the problem and putting $\hat{\beta}$ in the equation will be equals to 0.

(3) = Hence using the given X and Y we can calculate β.

Part. a =

Matrix Y tells us the Prediction result of multiclass classifier labels.

The ith row of matrix Y represent the Prediction value of the ith datapoint among N- datapoints for all the K-labels which the classifier classifier.

Similarly Y_{ij} represent that whether its ith datapoint is near to Jth label or not.

Part. b =

Its given that the rows of X are linearly independent.

It means that columns of transposed matrix X^T will be linearly independent.

we know that at minima of any function the value of differentiation is equal to 0.

It means to find minima, we have solve (differentiation of $f(x) = 0$)

here to minimize $f(x) = \|X^T B - Y\|^2$

for known X and Y values. we can calculate $\hat{\beta}$ which will minimize the problem.

$\hat{\beta}$ is the minima, Hence $\nabla f(\hat{x}) = 0$

$$\nabla f(\hat{x}) = 2(X^T)^T (X^T\hat{\beta} - Y) = 0$$

$$\Rightarrow 2X(X^T\hat{\beta} - Y) = 0$$

$$\Rightarrow 2XX^T\hat{\beta} - 2Y = 0$$

$$= XX^T\hat{\beta} = Y$$

$$\Rightarrow \hat{\beta} = (XX^T)^{-1} Y$$

$$\Rightarrow \hat{\beta} = (X^T)^{\dagger} Y$$

Hence $\boxed{\hat{\beta} = (X^T)^{\dagger} Y}$

PROBLEM 15.4 =

Robust approximate solution of linear equation.

given that $Ax = b$

if A is invertible then $x = A^{-1}b$

for given k versions of A =

$$A^{(1)}, A^{(2)} \; --- \; A^{(k)}$$

Choose x to minimize =

$$\|A^{(1)}x - b\|^2 + \; ---- + \|A^{(k)}x - b\|^2$$

we can represent this equation in form of Norm of another matrix L.

$$L = \left\| \begin{bmatrix} A^{(1)}x - b \\ A^{(2)}x - b \\ \vdots \\ A^{(k)}x - b \end{bmatrix} \right\|$$

we can define -

$$A' = \begin{bmatrix} A^{(1)} \\ A^{(2)} \\ A^{(3)} \\ \vdots \\ A^{(k)} \end{bmatrix}_{k \times 1}, \quad b' = \begin{bmatrix} b \\ b \\ b \\ \vdots \\ b \end{bmatrix}_{k \times 1}$$

for a general equation $Cy - d = 0$

we have $y = (C^T C)^{-1} C^T d$

Hence in our can we can write

$$x^{rob} = \left(A'^T A'\right)^{-1} A'^T b'$$

which reduces to =

$$x^{rob} = \left(\sum_{k=1}^{K} (A^k)^T A^k\right)^{-1} \left(\sum_{k=1}^{K} (A^k)^T b\right)$$

when $\boxed{k=1}$ then

$$x^{rob} = \left(\left(\left(A^{(1)}\right)^T A^{(1)}\right)^{-1} \left(\left(A^{(1)}\right)^T b\right)\right)$$

$$x^{rob} = \left[\left(A^{(1)}\right)^{-1} \left(A^{(1)T}\right)^{-1}\right] \left[A^{(1)T} b\right]$$

$$x^{rob} = \left(A^{(1)}\right)^{-1} \left[\left(A^{(1)T}\right)^{-1} \left(A^{(1)T}\right)\right] b$$

$$x^{rob} = \left(A^{(1)}\right)^{-1} b$$

<u>proved</u>

PROBLEM 15.7) — General Pseudo Inverse

Part. a = what is the Pseudo Inverse of $m \times n$ zero matrix.

given that A is zero matrix of $m \times n$

$$A = [0]_{m \times n}$$

also pseudo Inverse By kernel trick identity is given as =

$$A^{\dagger} = \lim_{\lambda \to 0} (A^T A + \lambda I)^{-1} A^T$$

$$A^{\dagger} = \lim_{\lambda \to 0} \left[[0]^T_{m \times n} [0]_{m \times n} + \lambda I \right]^{-1} [0]^T_{m \times n}$$

$$A^{\dagger} \Rightarrow \lim_{\lambda \to 0} [0_{n \times n} + \lambda I]^{-1} [0]_{n \times m}$$

$$A^{\dagger} \Rightarrow 0_{n \times m}$$

Since $A^{\dagger} = O_{n\times m} = [0]^{T}_{m\times n}$

we can say that for a zero matrix the pseudo Inverse could be defined as its transpose.

Part. b = given that A has linearly independent columns.

$$A^{\dagger} = \lim_{\lambda \to 0} (A^T A + \lambda I)^{-1} A^T$$

as λ reaches 0, objective function

$$\|Ax - b\|^2 + \lambda \|x - x^{des}\|^2$$

reaches to $\|Ax - b\|^2$

So if $[A]_{m\times n}$ matrix has linearly independent columns.

Then it means that,

$$\hat{x} = A^{+}b$$

and $A^{+} = (A^{T}A)^{-1}A^{T}$

Matrix inverse is continuous function

Hence $\lim\limits_{\lambda \to 0} (A^{T}A + \lambda I)^{-1}$

$$= \left(\lim_{\lambda \to 0} A^{T}A + \lim_{\lambda \to 0} \lambda I\right)^{-1}$$

$$= \left(\lim_{\lambda \to 0} A^{T}A + 0\right)^{-1}$$

$$= (A^{T}A)^{-1} \longrightarrow \text{eqtn } 1$$

NOTE - given that -

$$\lim (f(x))^{-1} = (\lim f(x))^{-1}$$

if $f(x)$ = continuous function

as $\lambda \to 0$, also given that A has linearly independent columns.

$$A^{+} = \lim_{\lambda \to 0} (A^T A + \lambda I)^{-1} A^T$$

$$A^{+} = (A^T A)^{-1} A^T$$ [putting values from equation ①]

Part. C – A has linearly Independent rows.

As we solved in part b for linearly independent columns. using the same method we can solve this part.

Here AA^T is invertible, $\lambda \to 0$.

also $\|Ax-b\|^2 + \lambda \|x - x^{des}\|^2 \approx \|Ax-b\|^2$

will be having =

$$\hat{x} = A^{+} b = A^T (AA^T)^{-1} b$$

we can use the same idea in part b that since matrix inverse is continuous Hence limit of the inverse of matrix will be equal to inverse of limit.

$$\text{So } \lim_{\lambda \to 0} (AA^T + \lambda I)^{-1}$$

$$= \left(\lim_{\lambda \to 0} AA^T + \lim_{\lambda \to 0} \lambda I \right)^{-1}$$

$$= (AA^T)^{-1} \longrightarrow \text{eqt}^n \text{ } 1.$$

Since, $A^{+} = \lim_{\lambda \to 0} A^T (AA^T + \lambda I)^{-1}$

using eqt^n 1 =

$$A^{+} = A^T (AA^T)^{-1}$$

PROBLEM 16.11 = least distance Problem.

minimizing $\|x-a\|^2$
is subject to $Cx = d$.

we have to determin n-vector x.

Given = n-vector a.
2) = $p \times n$ matrix C
3) = p vector d.

we have to prove that =

Solution of this Problem is =

$$\hat{x} = a - C^{\dagger}(Ca-d)$$

lets assume $\boxed{y = x-a}$

if, $(y = x - a)$ Then $(x = y + a)$

and we have to minimize $= \|y\|^2$

Then according to the Problem it will be subjected to,

$$\Rightarrow Cx = d$$

$$\Rightarrow C(y + a) = d$$

$$\Rightarrow Cy + Ca = d$$

$$\Rightarrow Cy = d - Ca$$

After using least norm solution we get =

$$\hat{y} = C^{\dagger}(d - Ca)$$

$$\Rightarrow \hat{x} - a = C^{\dagger}(d - Ca)$$

$$\Rightarrow \hat{x} = a + C^{\dagger}(d - Ca)$$

$$\Rightarrow \hat{x} = a - C^{\dagger}(Ca - d)$$

Proved

PROBLEM. 5 =

Solving least square Problem ⇒ minimize $\|Ax-b\|^2$

given that columns of A are linearly independent.

Solution for the Problem is given by =

$$\hat{x} = A^{\dagger}b = (A^TA)^{-1}A^Tb$$

$A^{\dagger}$ = Pseudo Inverse of A.

(1) = QR factorization = The least square problem could be solved using QR factorization. In order to get $\hat{x}$ we need to figure out Pseudo inverse of matrix A first which is $(A^{\dagger})$.

Using QR factorization we can determine the $A^{\dagger}$. and then $\hat{x}$ values could be calculated.

we know that A could be written as

$$A = QR.$$

So $A^{\dagger}$ could be found as follow =

$$A^{\dagger} = (A^T A)^{-1} A^T$$

$$A^{\dagger} = \left((QR)^T (QR)\right)^{-1} (QR)^T$$

$$A^{\dagger} = (R^T Q^T Q R)^{-1} (QR)^T$$

$$A^{\dagger} = (R^T R)^{-1} (QR)^T$$

$$A^{\dagger} = (R^{-1} R^{-T} R^T Q^T)$$

$$A^{\dagger} = R^{-1} Q^T$$

Hence $\hat{x} = A^{-1} b$

$$\hat{x} = (R^{-1} Q^T) b$$

$$\hat{x} = R^{-1} Q^T b$$

So QR-factorization algorithm will have following steps =

① = find the QR factorization of matrix A. $(A = QR)$

(requires $2n^3$ flops)

② = Algorithm will calculate $Q^T b$ first. (requires $2n^2$ flops)

③ = Using Back substitution method solve the equation $Rx = Q^T b$ (requires n^2 flops)

This will give us $\hat{x}$ value.

Complexity of QR-factorization algo =

$= (2n^3 + 2n^2 + n^2)$ flops

$= (2n^3 + 3n^2)$ flops

Complexity $= O(n^3)$ flops.

(2) = CHOLESKY factorization Algo =

we can see that A is positive definite. Hence cholesky factorization can be used to solve the least squares problem.

It breaks given matrix A as follows -

$$A = LL^T$$

where L is lower triangular matrix.

In order to calculate $\hat{x} = A^T b$ the algorithm follows below steps =

1) = Do cholesky factorization.

$$A = LL^T$$

(requires $\left(\frac{1}{3}\right) n^3$ flops)

2) = Do forward substitution and solve

$$Lv = b$$

(requires n^2 flops)

3) = Do backward substitution and solve

$$L^T x = v$$

(requires n^2 flops)

complexity of Cholesky Algorithm =

$$= \left(\frac{1}{3}n^3 + n^2 + n^2\right) \text{ flops.}$$

$$= \left(\frac{1}{3}n^3 + 2n^2\right) \text{ flops}$$

$$= O(n^3) \text{ flops.}$$

PROBLEM.6 = Solving Regularized Least squares.

The least square Problem

$$\underset{x}{\text{minimize}} \ \|Ax-b\|^2 + \lambda\|x\|^2$$

could be solved in $O(mn^2 + n^3)$.

if A is wide $(m<n)$ we have to reduce it to $O(m^2n + m^3)$.

we know that,

$$\hat{x} = (A^TA + \lambda I)^{-1}A^Tb$$

also matrix inversion lemma tells us that =

$$(A^TA + \lambda I)^{-1}A^T = A^T(AA^T + \lambda I)^{-1}$$

Hence we can rewrite $\hat{x}$ as =

$$\hat{x} = A^T(AA^T + \lambda I)^{-1} b$$

This method uses following =

① = It takes $2(m+n)m^2$ flops for QR factorization.

② = A^T is a matrix
$(AA^T + \lambda I)^{-1} b \rightarrow$ a vector.

Hence to calculate $\hat{x}$ it takes mn flops for matrix vector multiplication.

Total complexity =

$$2(m+n)m^2 + mn$$
$$= 2m^3 + 2nm^2 + mn$$

Hence complexity $= O(m^2 n + m^3)$

PROBLEM 7 = Companion Matrix.

Part a = what is characteristic polynomial of C?

given $C = \begin{bmatrix} 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & 0 & 0 & -c_2 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & -c_{n-1} \end{bmatrix}$

as given in the slides the characteristic polynomial for any matrix A is given as $P(x) = \det(xI - C)$

Hence for companion matrix we can write =

$$P(y) = \det(yI - C)$$

where I = Identity matrix.

Matrix $(yI - C) =$

$$\begin{bmatrix} y & 0 & 0 & \cdots & 0 \\ 0 & y & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & 0 & \cdots & & y \end{bmatrix} - \begin{bmatrix} 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & 0 & 0 & -c_2 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & -c_{n-1} \end{bmatrix}$$

$$= \begin{bmatrix} y & 0 & \cdots & 0 & c_0 \\ -1 & y & \cdots & 0 & c_1 \\ 0 & -1 & \cdots & 0 & c_2 \\ \vdots & \vdots & & & \\ 0 & 0 & \cdots & -1 & y+c_{n-2} \end{bmatrix}$$

Then

$$\det(yI - C) = \begin{vmatrix} y & 0 & \cdots & 0 & c_0 \\ -1 & y & \cdots & 0 & c_1 \\ 0 & -1 & \cdots & 0 & c_2 \\ \vdots & \vdots & & & \\ 0 & 0 & \cdots & -1 & y+c_{n-2} \end{vmatrix}$$

using cofactor expansion corresponding to first row we obtain that =

$\det(yI - C) =$

$$\Rightarrow y \begin{vmatrix} y & 0 & \cdots & 0 & c_1 \\ -1 & y & \cdots & 0 & c_2 \\ \vdots & \vdots & \ddots & & \\ 0 & 0 & \cdots & y & y + c_{n-1} \end{vmatrix}$$

$$+ (-1)^{n+1} c_0 \begin{vmatrix} -1 & y & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 \end{vmatrix}$$

By Induction we know the first determinant value is =

$$y^{n-1} + c_{n-1} y^{n-2} + \cdots + c_2 y + c_1$$

the second determinant is $(-1)^{n-1}$ Since it is an $(n-1) \times (n-1)$ triangular matrix determinant will be product of diagonal entries.

2nd determinant $= (-1)^{n+1} c_0 (-1)^{n-1}$

Therefore

$\det(\gamma I - C) =$

$= \gamma(\gamma^{n-1} + c_{n-1}\gamma^{n-2} + \cdots c_2\gamma + c_1)$
$+ (-1)^{n+1} c_0 (-1)^{n-1}$

$\Rightarrow \gamma^n + c_{n-1}\gamma^{n-1} + \cdots + c_1\gamma + c_0$

$$\boxed{P(x) = \gamma^n + c_{n-1}\gamma^{n-1} + \cdots + c_1\gamma + c_0}$$

Part b =

given that C has n-distinct eigen-values $\lambda_1, \lambda_2, \lambda_3 --- \lambda_n$.

Show that V is the matrix of left eigen vectors of C.

i.e. $VC = \Lambda V$

lets calculate VC =

$$VC = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_n & \lambda_n^2 & \cdots & \lambda_n^{n-1} \end{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & 0 & 0 & -c_2 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & -c_{n-1} \end{bmatrix}$$

$$VC = \begin{bmatrix} \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{n-1} & -c_0 - c_1\lambda_1 - c_2\lambda_1^2 \cdots - c_{n-1}\lambda_1^{n-1} \\ \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{n-1} & -c_0 - c_1\lambda_2 \cdots - c_{n-1}\lambda_2^{n-1} \\ \vdots & \vdots & & & \\ \lambda_n & \lambda_n^2 & \cdots & \lambda_n^{n-1} & -c_0 - c_1\lambda_n - c_2\lambda_n^2 \cdots - c_{n-1}\lambda_n^{n-1} \end{bmatrix}$$

Here λ_i are roots of characteristic equation hence characteristic equation will be 0 at λ_i.

So VC reduces to =

$$VC = \begin{bmatrix} \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^n \\ \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^n \\ \vdots & \vdots & & \\ \lambda_n & \lambda_n^2 & \cdots & \lambda_n^n \end{bmatrix}$$

also

$$\Lambda V = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \cdots & \lambda_2^{n-1} \\ \vdots & \vdots & & \\ 1 & \lambda_n & \cdots & \lambda_n^{n-1} \end{bmatrix}$$

$$= \begin{bmatrix} \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^n \\ \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^n \\ \vdots & \vdots & & \\ \lambda_n & \lambda_n^2 & \cdots & \lambda_n^n \end{bmatrix} = VC.$$

Hence $VC = \Lambda V$

Problem. 6 = Circulant Matrix:

given are circulant matrix and DFT matrix. C and F.

Part a =

from Previous problem. 7 we proved that.

$$YC = \Lambda Y$$

also $\Lambda = \mathrm{Diag}(\lambda_1 - - - - \lambda_n)$

Using the same analogy we can say that =

$$\Rightarrow \quad FC = \Lambda F$$

where Λ should be $\mathrm{Diag}(FC)$.

$$\Rightarrow \quad C = F^{-1} \Lambda F$$

we are given that $F^{-1} = F^*$

Hence $\boxed{C = F^* \Lambda F}$ proved

Part-b = Solve linear equation — $Cx = b$ where C is circulant Matrix.

$Cx = b$

[from Part. a $C = F^{-1}\Lambda F$]

$= (F^{-1}\Lambda F)x = b$

$\Rightarrow x = (F^{-1}\Lambda F)^{-1} b$

$\Rightarrow x = (F)^{-1}(F^{-1}\Lambda)^{-1} b$

$x = F^{-1}\Lambda^{-1}(F^{-1})^{-1} b$

$x = F^{-1}\Lambda^{-1} F b$

b/c $\boxed{(AB)^{-1} = B^{-1}A^{-1}}$

To calculate $\Rightarrow$ $\overbrace{F^{-1}\underbrace{\Lambda^{-1}\overbrace{Fb}^{P}}_{Q}}^{R}$

(1) $P = Fb$ (multiplication)

$n(\log n)$ flops are required b/c FFT is used in calculation.

(2) = Then $Q = \Lambda^{-1} P$

This multiplication requires n flops.

(3) = Then $R = F^{-1} B$

b/c inverse FFT is used so that's why we will need $n\log n$ flops.

Total complexity = for $\boxed{x = F^{-1}\Lambda^{-1}Fb}$

$n\log(n) + n + n\log n$

$= 2n\log n + n$

$= O(n\log n)$

PROBLEM-9 = General least Squares.

Part-a = for thin SVD of $A = U\Sigma V^T$

we can give the solution as =

$$\hat{x} = A^{\dagger} b = V\Sigma^{-1}U^T b \rightarrow ①$$

Transforming equation ① by multiplying By A^TA. we get =

$$A^TA\hat{x} = A^TA(A^{\dagger}b) = A^TA(V\Sigma^{-1}U^Tb)$$

now we know that =

$$A^TA\hat{x} = A^TA(V\Sigma^{-1}U^Tb)$$

Substituting the value of $A = U\Sigma V^T$ in the right hand side we get,

$$A^TA\hat{x} = A^T.(U\Sigma V^T).(V\Sigma^{-1}U^Tb)$$

$$A^TA\hat{x} = A^T \left[U\Sigma V^T V \Sigma^{-1} U^T b \right]$$

$$A^TA\hat{x} = A^T \left[U\Sigma \Sigma^{-1} U^T b \right]$$

$$A^TA\hat{x} = A^T \left[U U^T b \right]$$

$$A^TA\hat{x} = A^T b$$ (Proved)

Part.b - verify that it minimizes the objective function.

lets solve $\|Ax-b\|^2$

$$\boxed{\begin{aligned} \|Ax-b\|^2 &= \|Ax - A\hat{x} + A\hat{x} - b\|^2 \\ &= \|Ax - A\hat{x}\|^2 + \|A\hat{x} - b\|^2 \\ &\quad + 2\cdot(Ax - A\hat{x})^T (A\hat{x} - b) \end{aligned}}$$

↳ equation (1).

Simplify, $2(Ax - A\hat{x})^T (A\hat{x} - b)$

$$= 2(Ax - A\hat{x})^T (A\hat{x} - b)$$

$$= 2(x - \hat{x})^T A^T \cdot (A\hat{x} - b)$$

$$= 2(x - \hat{x}) (A^T A\hat{x} - A^T b)$$

$$= 2(x - \hat{x}) (A^T b - A^T b)$$

$$\left[b/c \quad A^T A\hat{x} = A^T b \right]$$

$$= 2(x - \hat{x}) (0)$$

$$= 0$$

Hence putting this value in equation ① =

$$\|Ax - b\|^2 = \|Ax - A\hat{x}\|^2 + \|A\hat{x} - b\|^2$$

This implies that =

$$\|A\hat{x} - b\|^2 \leq \|Ax - b\|^2$$

It means $\hat{x}$ has minimized the Objective function. Hence $\hat{x}$ minimizes the objective function.

Problem.10 = QRSVD.

given $m \times n$ matrix A. $(m \geq n)$

The QRSVD algorithm is given.

Part-b =

Show that $Q_1(R_1R_0)$ is the QR decomposition of $R_0^T R_0$ where Q_1 is the orthonormal factor and R_1R_0 is upper triangular factor.

from the given algorithm =

QR factorization is given as =

$$R_0^T = Q_1 R_1 \longrightarrow (1)$$

$$R_1^T = Q_2 R_2$$

$$R_2^T = Q_3 R_3 \quad \text{and so on.}$$

So lets calculate $R_0^T R_0$.

$$R_0^T R_0 = (Q_1 R_1) R_0 \quad (\text{from eqtn } ⑦)$$

$$R_0^T R_0 = Q_1 R_1 R_0$$ proved