Mohammed Alshurbaji

Adv. ML.

Assignment 1

Q1: Lasso (Least Absolute Shrinkage and Selection operator)

$$\frac{1}{2}\|Ax-y\|_2^2 + \lambda|x|_1$$

Algorithm:

1) Initial guess for the coeff. X. (start with zeros)

2) Iteration for each coordinate: (X_j)
   - → Fix all other coordinates X_{-j}.
   - → Update X_j by minimizing the obj. w.r.t. X_j.

$$\frac{1}{2}\|Ax-y\|_2^2 + \lambda|X_j|$$

3) Solve this as observed in class :- (Soft Thresholding)

$$= \begin{cases} -\frac{b+\lambda}{2A} & \text{if } b > \lambda \\ 0 & \text{if } |b| \le \lambda \\ -\frac{b+\lambda}{2A} & \text{if } b \le -\lambda \end{cases}$$

$$\boxed{ax^2 + bx + \lambda|x|}$$

Summary: It's supposed that Min. coordinate to be much faster, but it's not the case here. maybe because sklearn is using more optimization while I am using the basic one.

Q2: Least Squares Extension

Assume:

$A, X, C, Y, U \in \mathbb{R}^{n \times n}$

Given: $vec(AUC) = (C^T \otimes A)\, vec(U)$ ... ①

Use Least Squares: to solve:

$$\min_X \|AX + XC - Y\|_F^2$$

Now Review Basics:

$vec(AUC) \equiv$ Vectorization; $AUC \equiv [n \times n][n \times n][n \times n] \equiv [n \times n]$

Kronecker Product: multiply each element of first Matrix with the second Matrix.

Ex: $A \otimes B$; $A = [n \times m]$, $B = [q \times p]$

$\therefore A \otimes B \equiv [nq \times mp]$

To prove ①:

$vec(AUC) \equiv [n^2 \times 1]$

$C^T \otimes A \equiv [n^2 \times n^2]$; $vec(U) = [n^2 \times 1]$;

$\therefore (C^T \otimes A)\, vec(U) = [n^2 \times 1]$ ✱

Frobenius: $\|A\|_F$ (It's the same as Euclidean but for matrices)

$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2}$$

Note That: $\|A\|_F = Tr(A^T A)$

(#) $vec(A + B) = vec(A) + vec(B)$

Starting with the obj. Fnaction:-

$$\min_X \|AX + XC - Y\|_F^2 = (AX+XC-Y)^T (AX+XC-Y) \quad \cdots (1)$$

Using Vectorization:-

$$\begin{aligned} Vec(AX+XC-Y) &= Vec(AX) + Vec(XC) - Vec(Y) \\ &= (I \otimes A)\, Vec(X) + (C^T \otimes I)\, Vec(X) - Vec(Y) \\ &= (I \otimes A + C^T \otimes I)\, Vec(X) - Vec(Y) \quad \cdots (2) \end{aligned}$$

Subs. (2) in the obj. Fnact. :-

$$\min \|(I \otimes A + C^T \otimes I)\, Vec(X) - Vec(Y)\|_F^2$$

Take gradient & set to zero :-

$$\underbrace{(I \otimes A + C^T \otimes I)^T (I \otimes A + C^T \otimes I)}_{II}\, Vec(X) = \underbrace{(I \otimes A + C^T \otimes I)^T}_{III} Vec(Y)$$

Solving II :-

First:

$$\begin{aligned} II &= (I \otimes A^T + (C \otimes I))(I \otimes A + C^T \otimes I) \\ &= (I \otimes A^T)(I \otimes A) + (I \otimes A^T)(C^T \otimes I) + (C \otimes I)(I \otimes A) \\ &\quad + (C \otimes I)(C^T \otimes I) \\ &\overset{I\otimes}{=} (A^T A) + I \otimes (A^T C^T) + I \otimes (CA) + I \otimes (C C^T) \end{aligned}$$

Secondly

Solving III :-

$$III = (I \otimes A + C^T \otimes I)^T = I \otimes A^T + C \otimes I$$

∴ Solution of Vec(X) :

$$Vec(X) = II^{-1}\, III\, Vec(Y) \quad \#$$

For Simplicity:-

$$W = (I \otimes A) + (C^T \otimes I)$$

$$Vec(X) = (W^T W)^{-1} W^T Vec(Y)$$

Q3: Ridge Regression :-

$$J(x) = \underbrace{\min_x \|Ax - y\|_2^2}_{I} + \underbrace{\lambda \|x\|_2^2}_{II}$$

I : Using Composite :-

$$g(x) = Ax - y$$
$$J = \|g\|_2^2$$
$$\frac{\partial J}{\partial x} = \frac{\partial g}{\partial x} \cdot \frac{\partial J}{\partial g} \quad \text{(Chain Rule)}$$
$$= A^T \cdot 2g$$
$$= 2A^T(Ax - y)$$

II : $2\lambda x$

Then : Set $\frac{\partial J(x)}{\partial x} = 0$

$$2A^T(Ax - y) + 2\lambda x = 0$$
$$\therefore A^T(Ax - y) + \lambda x = 0$$
$$A^TAx + \lambda x = A^Ty$$

$$\therefore x = (A^TA + \lambda I)^{-1} A^Ty \quad \#$$

Now: To prove that it's monotically decreasing (As shown in class).

Eigenval. Decomp. :- $A^TA = Q \Lambda Q^T$

$$A^TA + \lambda I = Q(\Lambda + \lambda I) Q^T$$

$$\therefore x = Q(\Lambda + \lambda I)^{-1} Q^T A^T y$$
$$\|x\|_2 = \|Q(\Lambda + \lambda I)^{-1} Q^T A^T y\|_2 \; ; Q \text{ is ortho.}$$

\# Thus, as λ increases, the inverse decreases which reduces $\|x\|_2$. \#

Golden Question.

$\sqrt{6}\left(\frac{1+\sqrt{5}}{2}\right)$

Q4: Linear Regression and it's Extensions

Boston houses Datasets 506 records, 13 features (X)

14th feature ≡ median house price ≡ y

- All features are cont. except feature 4, binary. (Consider it cont.)
- Standarization: The process of transforming data to have a mean of zero and STD of 1.

section 2: All of these why Questions are related to Bias-Variance Trade off.

Basically: for the standard Lin. Reg. (High Variance) The training error at low λ is simple and as λ increases becames more generalize, the error increases. While the test error moving in an opposite way, since the model ~~#f~~ is learning with M increases.

While the meet point means that the training and testing are representitive for the same generalized Model.

section 3: Also, Bias-Variance Tradeoff. while at the beginning too simple then increasing data is helpful. Then the model start fitting the Noise (overfitting happened).

section 4:-

As the regularization operator λ increased at the begining the errors decreases then at certain point we reach the underfitting as the λ keeps the coe. to be too small. (Underfitting).