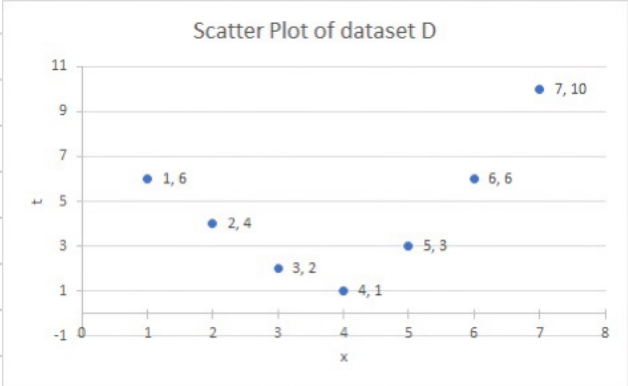
ECE421: Assignment #2

Won-Jin Kim
(1003611424)

P1

Problem 1 Assume we collected a dataset $D = \{(x^{(i)}, t^{(i)})\}_{i\in 1..7}$ of $N = 7$ points (i.e., observations) with inputs $\{x^{(i)}\}_{i\in 1..7} = (1, 2, 3, 4, 5, 6, 7)$ and outputs $\{t^{(i)}\}_{i\in 1..7} = (6, 4, 2, 1, 3, 6, 10)$ for a regression problem with both scalar inputs and outputs.

1. (1 point) Draw a scatter plot of the dataset on a spreadsheet software (e.g., Excel).



2. (6 points) Let us use a linear regression model $g_{w,b}(x) = wx + b$ to model this data. Write down the analytical expression of the least squares loss (covered in Video 6) of this model on dataset D. Your loss should take the form of

$$\frac{1}{2N}\sum_{i\in 1..N} A_i w^2 + B_i b^2 + C_i wb + D_i w + E_i b + F_i$$

where A_i, B_i, C_i, D_i, E_i, and F_i are expressed only as a function of $x^{(i)}$ and $t^{(i)}$ or constants. Do not fill-in any numerical values yet.

2. $g_{w,b} = wx + b \rightarrow$ regression model

Mean square loss : $\mathcal{L}(g,t) = \frac{1}{2}(g-t)^2$

loss over model:

$$\mathcal{E}(w,b) = \frac{1}{2N}\sum_{i=1}^{N}(g^{(i)} - t^{(i)})^2$$

$$= \frac{1}{2N}\sum_{i=1}^{N}(wx^{(i)} + b - t^{(i)})^2 \quad (wx^{(i)} + b - t^{(i)})$$

expand. ↓

$$= \frac{1}{2N}\sum_{i=1}^{N}\begin{pmatrix} w^2x^{(i)2} + bwx^{(i)} - wx^{(i)}t^{(i)} \\ +bwx^{(i)} + b^2 - bt^{(i)} \\ -wx^{(i)}t^{(i)} - bt^{(i)} + t^{(i)2} \end{pmatrix}$$

By rearranging

$$= \frac{1}{2N}\sum_{i=1}^{N}\begin{pmatrix} x^{(i)2}w^2 + 1b^2 + 2x^{(i)}wb \\ -2x^{(i)}t^{(i)}w - 2t^{(i)}b + t^{(i)2} \end{pmatrix}$$

$\therefore A_i = x^{(i)2}$, $B_i = 1$, $C_i = 2x^{(i)}$,

$D_i = -2x^{(i)}t^{(i)}$, $E_i = -2t^{(i)}$, $F_i = t^{(i)2}$

3. (4 points) Derive the analytical expressions of w and b by minimizing the mean squared loss from the previous question. Your expressions for parameters w and b should only depend on $A = \sum_i A_i$, $B = \sum_i B_i$, $C = \sum_i C_i$, $D = \sum_i D_i$ and $E = \sum_i E_i$. Do not fill-in any numerical values yet.

3

from previous question we obtained MSL (mean square loss)

minimizing MSL refers to having their partial derivative with respect to both w, b.

$$\rightarrow \frac{\partial MSL}{\partial w} = \frac{\partial MSL}{\partial b} = 0$$

$$\frac{\partial MSL}{\partial w} = \frac{\partial}{\partial w}\left(\frac{1}{2N}\sum_{i=1}^{N} A_i w^2 + B_i b^2 + C_i wb + D_i w + E_i b + F_i\right)$$

$$= \frac{1}{2N}\sum_{i=1}^{N}(2A_i w + C_i b + D_i) = 0 \quad \text{——— (1)}$$

$$\frac{\partial MSL}{\partial b} = \frac{\partial}{\partial b}\left(\frac{1}{2N}\sum_{i=1}^{N} A_i w^2 + B_i b^2 + C_i wb + D_i w + E_i b + F_i\right)$$

$$= \frac{1}{2N}\sum_{i=1}^{N}(2B_i b + C_i w + E_i) = 0 \quad \text{——— (2)}$$

for equation (1)

$$\frac{1}{2N}\left(2\underbrace{\sum_{i=1}^{N} A_i}_{A} w + \underbrace{\sum_{i=1}^{N} C_i}_{C} b + \underbrace{\sum_{i=1}^{N} D_i}_{D}\right) = 0$$

$$2Aw + Cb + D = 0$$

$$w = \frac{-Cb - D}{2A} \quad \text{——— (3)}$$

for equation (2)

$$\frac{1}{2N}\left(2\sum_{i=1}^{N} B_i b + \sum_{i=1}^{N} C_i w + \sum_{i=1}^{N} E_i\right) = 0$$

$$2Bb + Cw + E = 0$$

insert equation (3) ↓

$$2Bb + C\left(\frac{-Cb - D}{2A}\right) + E = 0$$

$$2Bb + \frac{-bC^2 - CD}{2A} + E = 0$$

→

$$2Bb + \frac{-bC^2 - CD}{2A} + E = 0$$

$$2Bb + \left(-\frac{C^2}{2A}\right)b - \frac{CD}{2A} + E = 0$$

$$\left(2B - \frac{C^2}{2A}\right)b = \frac{CD}{2A} - E$$

$$\frac{4AB - C^2}{2A}\, b = \frac{CD - 2AE}{2A}$$

$$b = \frac{\frac{CD-2AE}{\cancel{2A}}}{\frac{4AB-C^2}{\cancel{2A}}} = \frac{CD-2AE}{4AB-C^2}$$

similarly we input the expression of b to equation (3)

$$w = \frac{-Cb - D}{2A} = \frac{-C\left(\frac{CD-2AE}{4AB-C^2}\right) - D}{2A}$$

$$= \frac{\frac{-C^2D + 2ACE}{4AB-C^2} - \frac{D(4AB-C^2)}{4AB-C^2}}{2A}$$

$$= \frac{\cancel{-C^2D}^{\,0} + \cancel{2A}CE - \cancel{4}^{\,2}ABD + \cancel{C^2D}^{\,0}}{(4AB-C^2)\,\cancel{2A}}$$

$$= \frac{CE - 2BD}{(4AB-C^2)}$$

$$\boxed{\therefore\ b = \frac{CD-2AE}{4AB-C^2} \qquad w = \frac{CE-2BD}{(4AB-C^2)}}$$

4. (2 points) Give approximate numerical values for w and b by plugging in numerical values from the dataset D.

$\boxed{4}$ from Question #2, we found that

$A_i = x^{(i)^2}$, $B_i = 1$, $C_i = 2x^{(i)}$,

$D_i = -2x^{(i)}t^{(i)}$, $E_i = -2t^{(i)}$, $F_i = t^{(i)^2}$

$$A = \sum_i A_i = (1)^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2$$
$$= 1 + 4 + 9 + 16 + 25 + 36 + 49$$
$$= 140$$

$$B = \sum_i B_i = 1 \cdot i = 7$$

$$C = \sum_i C_i = 2(1+2+3+4+5+6+7)$$
$$= 56$$

$$D = \sum_i D_i = -2\big((1\cdot 6) + (2\cdot 4) + (3\cdot 2) + (4\cdot 1) + (5\cdot 3) + (6\cdot 6) + (7\cdot 10)\big)$$
$$= -2(6 + 8 + 6 + 4 + 15 + 36 + 70)$$
$$= -290$$

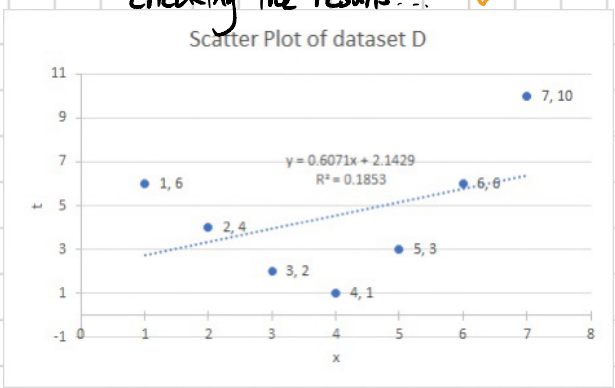
$$E = \sum_i E_i = -2(6+4+2+1+3+6+10) = -64$$

$$F_i = \sum_i F_i = 6^2 + 4^2 + 2^2 + 1^2 + 3^2 + 6^2 + 10^2$$
$$= 202$$

$$b = \frac{(56)(-290) - 2(140)(-64)}{4(140)(7) - (56)^2}$$
$$= \frac{-16240 + 17920}{3920 - 3136} = \frac{1680}{784} \approx \boxed{2.149}$$

$$w = \frac{(56)(-64) - 2(7)(-290)}{784} = \frac{-3584 + 4060}{784}$$
$$\approx \boxed{0.607}$$

checking the results... ✓



P2

Problem 2 The goal of this problem is to revisit Problem 1, but solving it with a different technique known as the method of least squares. This will serve as a "warm-up" to Problem 3. In the rest of this problem, any reference to a dataset refers to the dataset described in Problem 1.

1. (1 point) Verify that one can rewrite the linear regression model $g_{w,b}(x) = wx + b$ in the simpler form of

$$g_w(\vec{x}) = \vec{x}\vec{w}$$

if one assumes each input $\vec{x}$ is a two-dimensional row vector such that a point in our dataset is now $\vec{x^{(i)}} = (x^{(i)}, 1)$ where $x^{(i)}$ is the scalar input described in Problem 1. Write the components of the new column vector $\vec{w}$ as a function of w and b from Problem 1.

1 Previously $g_{w,b}(x) = wx + b$

we could rewrite this assuming

row vector $\vec{x}^{(i)}$ as $\vec{x}^{(i)} = (x^{(i)}, 1)$

column vector $\vec{w}$ as (w, b)

$$\rightarrow \vec{w}^T = \begin{bmatrix} w \\ b \end{bmatrix}$$

$$\vec{g}_w(\vec{x}^{(i)}) = \begin{bmatrix} | & | \\ x^{(i)}, & | \\ | & | \end{bmatrix} \begin{bmatrix} w \\ b \end{bmatrix}$$

$$= x^{(i)} w^T + b = g_{w,b}(x)$$

2. (4 points) Derive analytically $\nabla_{\vec{w}} \|X\vec{w} - \vec{t}\|^2$ where X is a $N \times 2$ matrix such that each row of X is a vector $\vec{x^{(i)}}$ described in the previous question, and $\vec{t} = \{t^{(i)}\}_{i \in 1..7}$.

2

$$\nabla_{\vec{w}} \|X\vec{w} - \vec{t}\|^2$$

$$X\vec{w} = \begin{bmatrix} x^{(1)} & 1 \\ x^{(2)} & 1 \\ x^{(3)} & 1 \\ \vdots & \vdots \\ x^{(i)} & 1 \end{bmatrix} \begin{bmatrix} w \\ b \end{bmatrix} = \vec{g}_w$$

$$\|X\vec{w} - \vec{t}\|^2 = (X\vec{w} - t)^T (X\vec{w} - t)$$

$$= \vec{w}^T X^T X \vec{w} - \vec{w}^T X^T t - t^T X \vec{w} + t^T t$$

we know $\vec{w}^T X^T t = t^T X \vec{w}$ since they are just transpose of scalar.

$$\therefore \|X\vec{w} - \vec{t}\|^2 = \vec{w}^T X^T X \vec{w} - 2t^T X \vec{w} + t^T t$$

$$\nabla_{\vec{w}} \|X\vec{w} - \vec{t}\|^2 = \frac{\partial}{\partial w}(\vec{w}^T X^T X \vec{w} - 2t^T X \vec{w} + t^T t)$$

$$= \frac{\partial}{\partial w}(\vec{w}^T X^T X \vec{w}) + \frac{\partial}{\partial w}(-2t^T X \vec{w})$$

$$= 2\vec{w}^T X^T X - 2t^T X$$

$$= 2X^T X \vec{w} - 2Xt$$

3. (1 point) Conclude that the model's weight value $\vec{w}^*$ which minimizes the least squares loss (covered in Video 6) must satisfy

$$2X^\top X\vec{w}^* - 2X^\top\vec{t} = 0$$

$\boxed{3}$

from previous question we concluded that

$$\nabla_{\vec{w}}\|X\vec{w} - \vec{t}\|^2 = 2X^\top X\vec{w} - 2X^\top\vec{t}$$

setting this to 0 will lead to getting minimized least square loss with respect to $\vec{w}$.

4. (1 point) Assuming that $X^\top X$ is invertible, derive analytically the value of $\vec{w}^*$.

$\boxed{4}$

Using Previous equation,

$$2X^\top X\vec{w}^* - 2X^\top\vec{t} = 0$$

$$\Leftrightarrow \quad \cancel{2}X^\top X\vec{w}^* = \cancel{2}X^\top\vec{t}$$

$$\Leftrightarrow \quad X^\top X\vec{w}^* = X^\top\vec{t}$$

assuming $(X^\top X)$ is invertible

$$(X^\top X)^{-1}X^\top X\vec{w}^* = (X^\top X)^{-1}X^\top\vec{t}$$

$$\therefore \vec{w}^* = (X^\top X)^{-1}X^\top\vec{t}$$

P3

notation in the previous questions.

Problem 3 Let us now assume that D is a dataset with d features per input and $N > 0$ inputs. We have $D = \{((\vec{x_j^{(i)}})_{j\in 1..d}, t_i)\}_{i\in 1..N}$. In other words, each $\vec{x^{(i)}}$ is a column vector with d components indexed by j such that $x_j^{(i)}$ is the jth component of $\vec{x^{(i)}}$. The output $\vec{t^{(i)}}$ remains a scalar (real value).

Let us assume for simplicity that we have a simplified linear regression model, as presented in the Question 1 of Problem 2. We would like to train a regularized linear regression model, where the mean squared loss is augmented with an ℓ_2 regularization penalty $\frac{1}{2}\|\vec{w}\|_2^2$ on the weight parameter $\vec{w}$:

$$\varepsilon(\vec{w}, D) = \frac{1}{2N}\sum_{i\in 1..N}(g_{\vec{w}}(\vec{x^{(i)}}) - t^{(i)})^2 + \frac{\lambda}{2}\|\vec{w}\|_2^2$$

where $\lambda > 0$ is a hyperparameter that controls how much importance is given to the penalty.

1. (3 points) Let $A = \sum_{i\in 1..N}\vec{x^{(i)}}\vec{x^{(i)}}^\top$. Give a simple analytical expression for the components of A.

1

$$A = \sum_{i=1}^{N}\vec{x^{(i)}}\vec{x^{(i)}}^T = \vec{x^{(1)}}\vec{x^{(1)}}^T + \vec{x^{(2)}}\vec{x^{(2)}}^T + \dots + \vec{x^{(N)}}\vec{x^{(N)}}^T$$

$$= \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_d^{(1)} \end{pmatrix}\begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \end{pmatrix} + \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \\ \vdots \\ x_d^{(2)} \end{pmatrix}\begin{pmatrix} x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \end{pmatrix}$$

$$\cdots \begin{pmatrix} x_1^{(N)} \\ x_2^{(N)} \\ \vdots \\ x_d^{(N)} \end{pmatrix}\begin{pmatrix} x_1^{(N)} & x_2^{(N)} & \cdots & x_d^{(N)} \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(N)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ x_d^{(1)} & x_d^{(2)} & \cdots & x_d^{(N)} \end{pmatrix}}_{d\times N}\begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \cdots & x_d^{(N)} \end{pmatrix}$$

$$= XX^T$$

2. (6 points) Let us write $\vec{b} = \sum_{i\in 1..N} t^{(i)}\vec{x^{(i)}}$, prove that the following holds:

$$\nabla\varepsilon(\vec{w}, D) = \frac{1}{N}\left(A\vec{w} - \vec{b}\right) + \lambda\vec{w}$$

2

$$\nabla\varepsilon(\vec{w}, D) = \frac{\partial}{\partial w}\left(\frac{1}{2N}\sum_{i=1}^{N}(g_{\vec{w}}(\vec{x^{(i)}}) - t^{(i)})^2 + \frac{\lambda}{2}\|\vec{w}\|_2^2\right)$$

$$= \frac{1}{2N}\left(\frac{\partial}{\partial w}\left(\sum_{i=1}^{N}(g_{\vec{w}}(\vec{x^{(i)}}) - t^{(i)})^2\right)\right) + \frac{\lambda}{2}\frac{\partial}{\partial w}\|\vec{w}\|_2^2$$

$$\downarrow \sum_{i=1}^{N} g_{\vec{w}}(\vec{x^{(i)}}) = x^{(i)}w$$

$$= \frac{1}{2N}\frac{\partial}{\partial w}\underbrace{\sum_{i=1}^{N}(x^{(i)}w - t^{(i)})^2}_{①} + \frac{\lambda}{2}\underbrace{\frac{\partial}{\partial w}\|\vec{w}\|_2^2}_{②}$$

① $\rightarrow \sum_{i=1}^{N}(x^{(i)}w - t^{(i)})^T(x^{(i)}w - t^{(i)})$

$$= \sum_{i=1}^{N}\left(w^Tx^{(i)T}x^{(i)}w - w^Tx^{(i)T}t^{(i)} - t^{(i)T}x^{(i)}w + t^{(i)T}t^{(i)}\right)$$

$$= \sum_{i=1}^{N}\left(w^Tx^{(i)T}x^{(i)}w - 2t^{(i)T}x^{(i)}w + t^{(i)T}t^{(i)}\right)$$

$$= w^TAw - 2\vec{b}w + \sum_{i=1}^{N}t^{(i)T}t^{(i)}$$

$$\frac{\partial}{\partial w}① = 2Aw - 2\vec{b}$$

② $\rightarrow \frac{\partial}{\partial w}\vec{w}^T\vec{w} = 2\vec{w}$

$$\therefore \nabla_\varepsilon(\vec{w}, D) = \frac{1}{2N}(2A\vec{w} - 2\vec{b}) + \frac{\lambda}{2}(2\vec{w})$$

$$= \frac{1}{N}(A\vec{w} - \vec{b}) + \lambda\vec{w}$$

3. (2 points) Write down the matrix equation that $\vec{w}^*$ should satisfy, where:

$$\vec{w}^* = \arg\min_{\vec{w}} \varepsilon(\vec{w}, D)$$

Your equation should only involve $A, \vec{b}, \lambda, N$, and $\vec{w}^*$.

$\boxed{3}$

$$\vec{w}^* = \arg\min_{\vec{w}} \varepsilon(\vec{w}, D)$$

$$\Rightarrow \nabla \varepsilon(\vec{w}, D) = 0$$

from previous equation

$$\vec{w}^* = \left\{ w \,\middle|\, \nabla \varepsilon(\vec{w}, D) = \frac{1}{N}(A\vec{w} - \vec{b}) + \lambda \vec{w} = 0 \right\}$$

$$\frac{1}{N}(A\vec{w} - \vec{b}) + \lambda \vec{w} = 0$$

$$A\vec{w} - \vec{b} + N\lambda \vec{w} = 0$$

$$(A + N\lambda I_d)\vec{w} - b = 0$$

$$\therefore (A + N\lambda I_d)\vec{w}^* - b = 0$$

4. (3 points) Prove that all eigenvalues of A are non-negative.

$\boxed{4}$

$A = XX^T$ is positive semidefinite matrix

$$v^T A v = v^T X X^T v = (X^T v)^T (X^T v)$$

$$= \|X^T v\|^2 \geq 0$$

Consider λ to be real eigenvalue & x to be corresponding eigenvector for A

$$Ax = \lambda x$$

$$x^T A x = \lambda x^T x = \lambda \|x\|^2$$

since left side is positive, right side is positive only if λ is positive

5. (3 points) Demonstrate that matrix $A + \lambda N I_d$ is invertible by proving that none of its eigenvalues are zero. Here, I_d is the identity matrix of dimension d.

$\boxed{5}$ from previous question we know that matrix A has non-negative eigen values.

from question we know that $N > 0$, $\lambda > 0$

thus its eigenvalues should be always positive

Note that a real symmetric matrix is diagonalizable by an orthogonal matrix.

So there exists orthogonal matrix Q s.t $Q^T(A + \lambda N I_d) Q = D$ where

$$D = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \\ & & & \lambda_d \end{bmatrix}$$

where λ_i $i \in 1 \ldots d$ are all positive.

$$(A + \lambda N I_d) = Q D Q^T$$

$$(A + \lambda N I_d) = (Q D Q^T)^{-1} = Q^{-1} D^{-1} (Q^T)^{-1} = Q D^{-1} Q^T$$

Note: transpose of orthogonal matrix is its inverse.

6. (2 points) Using the invertibility of matrix $A + \lambda N I_d$, solve the equation stated in question 3 and deduce an analytical solution for $\vec{w}^*$. You've obtained a linear regression model regularized with an ℓ_2 penalty.

$\boxed{6}$

$$\vec{w}^* = \left\{ w \,\middle|\, \nabla \varepsilon(\vec{w}, D) = \frac{1}{N}(A\vec{w} - \vec{b}) + \lambda \vec{w} = 0 \right\}$$

$$\frac{1}{N}(A\vec{w} - \vec{b}) + \lambda \vec{w} = 0$$

$$(A\vec{w} - \vec{b}) + N\lambda \vec{w} = 0$$

$$(A + N\lambda I)\vec{w} - b = 0$$

$$(A + N\lambda I)\vec{w} = b$$

$$\vec{w} = (A + N\lambda I)^{-1} b$$

$\therefore \vec{w}^* = (A + N\lambda I)^{-1} b$ if $(A + N\lambda I)$ is invertible.