assignment 3

Student ID : 7205329

Exercise 1.

1. Find the gradient $\nabla_\theta \mathcal{L}_{A,\Lambda}(\theta)$

Firstly, we rewrite $\mathcal{L}_{A,\Lambda}$

$$\begin{aligned}\mathcal{L}_{A,\Lambda}(\underset{\sim}{\theta}) &= \| \underset{\sim}{y} - X\underset{\sim}{\theta} \|_A^2 + \| \underset{\sim}{\theta} \|_\Lambda^2 \\ &= \langle \underset{\sim}{y} - X\underset{\sim}{\theta}, \underset{\sim}{y} - X\underset{\sim}{\theta} \rangle_A + \langle \underset{\sim}{\theta}, \underset{\sim}{\theta} \rangle_\Lambda \\ &= (\underset{\sim}{y} - X\underset{\sim}{\theta})^T A (\underset{\sim}{y} - X\underset{\sim}{\theta}) + \underset{\sim}{\theta}^T \Lambda \underset{\sim}{\theta} \\ &= (\underset{\sim}{y} - X\underset{\sim}{\theta})^T (A\underset{\sim}{y} - AX\underset{\sim}{\theta}) + \underset{\sim}{\theta}^T \Lambda \underset{\sim}{\theta} \\ &= \underset{\sim}{y}^T(A\underset{\sim}{y}) - \underset{\sim}{y}^T(AX\underset{\sim}{\theta}) - (X\underset{\sim}{\theta})^T(A\underset{\sim}{y}) + (X\underset{\sim}{\theta})^T(AX\underset{\sim}{\theta}) + \underset{\sim}{\theta}^T\Lambda\underset{\sim}{\theta}\end{aligned}$$

A B C D E

We will calculate the derivatives individual and sum them up later using the sum rule. The following calculations will use results from section 5.5, lecture "Vector Calculus".

Part A. $\dfrac{\partial}{\partial \underset{\sim}{\theta}}\left[\underset{\sim}{y}^T(A\underset{\sim}{y})\right] = \underset{\sim}{0}^T$.

Part B. $\dfrac{\partial}{\partial \underset{\sim}{\theta}}\left[\underset{\sim}{y}^T(AX\underset{\sim}{\theta})\right] = \dfrac{\partial}{\partial \underset{\sim}{\theta}}\left[\left((\underset{\sim}{y}^TAX)^T\right)^T\underset{\sim}{\theta}\right] = \underset{\sim}{y}^TAX$

using the result $\dfrac{\partial(\underset{\sim}{a}^T\underset{\sim}{x})}{\partial \underset{\sim}{x}} = \underset{\sim}{a}^T$

Part C.

$$\frac{\partial}{\partial \underset{\sim}{\theta}}\left[(X\underset{\sim}{\theta})^T(A\underset{\sim}{y})\right] = \frac{\partial}{\partial \underset{\sim}{\theta}}\left[(A\underset{\sim}{y})^T(X\underset{\sim}{\theta})\right]$$

$$= \frac{\partial}{\partial \underset{\sim}{\theta}}\left[\underset{\sim}{y}^T A^T X\underset{\sim}{\theta}\right]$$

$$= \frac{\partial}{\partial \underset{\sim}{\theta}}\left[\underset{\sim}{y}^T AX\underset{\sim}{\theta}\right] \quad \text{since } A \text{ is symmetric } A^T = A$$

$$= \underset{\sim}{y}^T AX \quad (\text{from part b above})$$

Part D.

$$\frac{\partial}{\partial \underset{\sim}{\theta}}\left[(X\underset{\sim}{\theta})^T AX\underset{\sim}{\theta}\right] = \frac{\partial}{\partial \underset{\sim}{\theta}}\left[\underset{\sim}{\theta}^T(X^TAX)\underset{\sim}{\theta}\right] = \underset{\sim}{\theta}^T\left[X^TAX + (X^TAX)^T\right]$$

using result $\frac{\partial(\underset{\sim}{x}^T B\underset{\sim}{x})}{\partial \underset{\sim}{x}} = \underset{\sim}{x}^T(B + B^T)$

$$= \theta^T\left[X^TAX + X^TA^TX\right] = 2\theta^TX^TAX.$$

Part E.

$$\frac{\partial}{\partial \theta}\left[\underset{\sim}{\theta}^T\Lambda\underset{\sim}{\theta}\right] = \underset{\sim}{\theta}^T(\Lambda + \Lambda^T) = \theta^T(\Lambda + \Lambda) = 2\theta^T\Lambda.$$

using result $\frac{\partial(\underset{\sim}{x}^T B\underset{\sim}{x})}{\partial \underset{\sim}{x}} = \underset{\sim}{x}^T(B + B^T)$, Λ is symmetric $\Lambda = \Lambda^T$

Finally,

$$\nabla_{\underset{\sim}{\theta}}\mathcal{L}_{\Lambda,\Lambda}(\underset{\sim}{\theta}) = \underset{\sim}{0}^T - \underset{\sim}{y}^TAX - \underset{\sim}{y}^TAX + 2\underset{\sim}{\theta}^TX^TAX + 2\underset{\sim}{\theta}^T\Lambda$$

$$= -2\underset{\sim}{y}^TAX + 2\underset{\sim}{\theta}^TX^TAX + 2\underset{\sim}{\theta}^T\Lambda$$

$$= -2(\underset{\sim}{y}^T - \underset{\sim}{\theta}^TX^T)AX + 2\underset{\sim}{\theta}^T\Lambda \qquad \underset{\sim}{y}^T - \underset{\sim}{\theta}^TX^T = \underset{\sim}{y}^T - (X\underset{\sim}{\theta})^T = (\underset{\sim}{y} - X\underset{\sim}{\theta})^T$$

$$= -2(\underset{\sim}{y} - X\underset{\sim}{\theta})^TAX + 2\underset{\sim}{\theta}^T\Lambda$$

2. Let $\nabla_\theta \mathcal{L}_{A,\Lambda}(\underset{\sim}{\theta}) = 0$, we have

$$-2(\underset{\sim}{y} - X\underset{\sim}{\theta})^T AX + 2\underset{\sim}{\theta}^T \Lambda = 0$$

$$\Leftrightarrow -2\underset{\sim}{y}^T AX + 2\underset{\sim}{\theta}^T X^T AX + 2\underset{\sim}{\theta}^T \Lambda = 0$$

$$\Leftrightarrow 2\underset{\sim}{\theta}^T (X^T AX + \Lambda) = 2\underset{\sim}{y}^T AX$$

$$\Leftrightarrow \underset{\sim}{\theta}^T = \underset{\sim}{y}^T AX (X^T AX + \Lambda)^{-1}$$

$$\Leftrightarrow \underset{\sim}{\theta} = \left[(X^T AX + \Lambda)^T\right]^{-1} (\underset{\sim}{y}^T AX)^T$$

$$\Leftrightarrow \underset{\sim}{\theta} = (X^T AX + \Lambda)^{-1} X^T A \underset{\sim}{y}$$

We have $X^T AX = \langle X, X \rangle_A$ is asymetric and positive definite (property of inner product.

Since the sum of a symmetric (Λ) and asymetric ($X^T AX$) matrix is asymetric, and the sum of two positive definite matrix is positive definite.

$\Rightarrow X^T AX + \Lambda$ is an asymetric positive definite $D \times D$ matrix and thus it is invertible, making our $\underset{\sim}{\theta}$ defined.

3. Substituting $A = I$ and $\Lambda = \lambda I$, we have

$$\underset{\sim}{\theta} = (X^T IX + \lambda I)^{-1} X^T I \underset{\sim}{y} = (X^T X + \lambda I)^{-1} X^T \underset{\sim}{y}.$$

4. For standard regularisation term $\lambda \|\underset{\sim}{\theta}\|_2^2$, we are applying the same weight of λ to all θ_i's.

$$\begin{aligned}\lambda \|\underset{\sim}{\theta}\|_2^2 &= \lambda\left(\theta_1^2 + \theta_2^2 + \dots + \theta_D^2\right) \\ &= \lambda\theta_1^2 + \lambda\theta_2^2 + \dots + \lambda\theta_D^2\end{aligned}$$

On the other hand, using $\|\underset{\sim}{\theta}\|_\Lambda^2$ let us apply different weights to different parameters θ_i. This way, we can specify the sensitivity of $\hat{y}$ to each x_i.

For example, let $\Lambda = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.3 \end{bmatrix}$ and $\underset{\sim}{\theta} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \end{bmatrix}^T$

$$\Rightarrow \underset{\sim}{\theta}^T \Lambda \underset{\sim}{\theta} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \end{bmatrix} \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.3 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5\theta_1 & 0.1\theta_2 & 0.3\theta_3 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

$$= 0.5\theta_1^2 + 0.1\theta_2^2 + 0.3\theta_3^2 .$$

Exercise 2.

1. We have

$$l_2(\mu, D) = \sum_{i=1}^{n} (\mu - x_i)^2$$

$$= (\mu - x_1)^2 + (\mu - x_2)^2 + \dots + (\mu - x_n)^2$$

$$= \mu^2 - 2\mu x_1 + x_1^2 + \mu^2 - 2\mu x_2 + \dots + \mu^2 - 2\mu x_n + x_n^2$$

$$= n\mu^2 - 2\mu \Sigma x_i + \sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \frac{dL_2}{d\mu} = \frac{d(n\mu^2 - 2\mu \Sigma x_i + \Sigma x_i^2)}{d\mu}$$

$$= 2n\mu - 2\Sigma x_i$$

Setting $\frac{dL_2}{d\mu} = 0$ we have

$$2n\mu - 2\Sigma x_i = 0$$

$$2n\mu = 2\Sigma x_i$$

$$\mu = \frac{1}{n} \Sigma x_i$$

Therefore, the optimal choice for μ is the mean.

2. We have $L_1 = \sum_{i=1}^{N} |\mu - x_i|$, we consider 2 cases of data set D.

1) D has an odd number of elements.

Suppose we have μ that minimises $L_1(\mu, D)$ (★), with n_l = number of points $x_i < \mu$ and n_r = number of points $x_i > \mu$.

. If $n_r < n_l$,

For example, let $D = \{0, 1, 2, 3, 5\}$ and $\mu = 3 \Rightarrow n_l = 4$, $n_r = 1$.

$$\Rightarrow L_1 = |3-0| + |3-1| + |3-2| + |3-3| + |3-5|$$
$$= \underbrace{3 + 2 + 1}_{S_l} + 0 + \underbrace{2}_{S_r}$$

now if we move μ to be $\mu' = 2 \Rightarrow \Delta = |\mu' - \mu| = |2-3| = 1$.

$$\Rightarrow L_1' = |2-0| + |2-1| + |2-2| + |2-3| + |2-5|$$
$$= \underbrace{2 + 1 + 0}_{\text{each decreases by } \Delta} + 1 + \underbrace{3}_{\text{each increases by } \Delta}$$

$S_l' = S_l - n_l \cdot \Delta$ $\qquad$ $S_r' = S_r + n_r \cdot \Delta$

$$\Rightarrow L_1' = S_l' + S_r' = S_l - n_l\Delta + S_r + n_r \cdot \Delta$$
$$= (S_l + S_r) + \Delta(n_r - n_l)$$
$$= L_1 + \Delta(n_r - n_l)$$

Since $n_r < n_l$, the net change is negative and we can further minimise L_1 which contradicts our assumption (★).

- If $n_\ell < n_r$: Let $D = \{0, 1, 2, 3, 5\}$ and $\mu = 1$ to be the optimal choice $\Rightarrow n_\ell = 1$ and $n_r = 3$.

With $\mu = 1$, $L_1 = |1-0| + |1-1| + |1-2| + |1-3| + |1-5|$

$= \underbrace{1}_{S_\ell} \quad 0 \quad \underbrace{1 \quad 2 \quad 4}_{S_r}$

if we move μ to be $\mu' = 2$, $\Delta = |\mu' - \mu| = |2-1| = 1$

$\Rightarrow L_1' = |2-0| + |2-1| + |2-2| + |2-3| + |2-5|$

$= \underbrace{2}_{\text{increases by } \Delta} + 1 + \underbrace{0 + 1 + 3}_{\text{each decreases by } \Delta}$

$S_\ell' = S_\ell + n_\ell \cdot \Delta$ $\qquad$ $S_r' = S_r - n_r \Delta$

$L_1' = S_\ell' + S_r' = S_\ell + S_r + \Delta(n_\ell - n_r)$

$= L_1 + \Delta(n_\ell - n_r)$

since $n_\ell < n_r$, the net change is also negative and that we can further minimise L_1.

Thus, as $n_\ell \neq n_r$, if we move μ to the left (decrease μ by a value of Δ) or to the right (increase by Δ), then we will get a different μ that can further minimises L_1.

It follows the minimum L_1 can be reached with a μ value such that $n_\ell = n_r$. For D with an odd number of elements, it is easy to point out that the optimal choice for μ is the median of D (by definition).

2) For D with an even number of elements, we start off with any number between the two middle numbers.

For example, $D = \{1, 2, 4, 6\}$
If we choose μ to be any number between 2 and 4 (e.g. $\mu = 3$) then $n_r = n_l$, and thus will minimise L_1.

However, we also find that the endpoints of this interval also minimise L_1.
Indeed, $x_{n/2}$ and $x_{n/2+1}$ divide the set D into two smaller groups, each with $n/2$ elements.
$D = \{ x_1, \dots, x_{n/2}, x_{n/2+1}, \dots, x_n \}$

Let $\mu = x_{n/2}$

$$\Rightarrow L_1 = |x_{n/2} - x_1| + \dots + |x_{n/2} - x_{n/2}| + |x_{n/2} - x_{n/2+1}| + \dots + |x_{n/2} - x_n|$$

if we move μ to be $\mu' = x_{n/2+1}$, $\Delta = |x_{n/2+1} - x_{n/2}|$

$$\Rightarrow L_1' = \underbrace{|x_{n/2+1} - x_1| + \dots + |x_{n/2+1} - x_{n/2}|}_{\text{each increases by } \Delta} + \underbrace{|x_{n/2+1} - x_{n/2+1}| + \dots + |x_{n/2+1} - x_n|}_{\text{each decreases by } \Delta}$$

$\Rightarrow s_l' = s_l + \Delta \cdot (n/2)$ $\quad$ $\Rightarrow s_r' = s_r - \Delta \cdot (n/2)$

$$\Rightarrow L_1' = s_l' + s_r' = s_l + \Delta \cdot (n/2) + s_r - \Delta \cdot (n/2)$$
$$= s_l + s_r = L_1 .$$

Finally, we found that L_1 is minimised when

- μ is the median (when D has an odd number of elements), and
- μ is any number between $x_{n/2}$ and $x_{n/2+1}$ (when D has an even number of elements).

$\Rightarrow$ The best representative μ for any data set is the median of that dataset.

3. For the L_∞ loss function, we consider 3 cases of μ:

- $\mu \leq x_1$, then $L_\infty(\mu, D) = \max_{1 \leq i \leq n} |\mu - x_i| = |\mu - x_n|$ since x_n has the furthest distance from μ.

$$\Rightarrow \frac{dL_\infty}{d\mu} = \frac{d(|\mu - x_n|)}{d\mu} = \frac{(\mu - x_n)\mu}{|\mu - x_n|}$$

Setting $\frac{dL_\infty}{d\mu} = 0$, we have

$$\Leftrightarrow \quad \frac{(\mu - x_n)\mu}{|\mu - x_n|} = 0$$

$$\Leftrightarrow \quad (\mu - x_n)\mu = 0$$

$$\Leftrightarrow \quad \mu^2 - \mu x_n = 0 \quad \Leftrightarrow \quad \mu = x_n.$$

$\Rightarrow$ In this case, $\mu = x_n$ is the best representative for D.

- $\mu \geq x_n$, then $L_\infty(\mu, D) = \max_{1 \leq i \leq n} |\mu - x_i| = |\mu - x_1|$ since x_1 has the furthest distance from μ.

$$\Rightarrow \frac{dL_\infty}{d\mu} = \frac{d(|\mu - x_1|)}{d\mu} = \frac{(\mu - x_1)\mu}{|\mu - x_1|}$$

Setting $\frac{dL_\infty}{d\mu} = 0$, we have

$$\Leftrightarrow \quad \frac{(\mu - x_1)\mu}{|\mu - x_1|} = 0$$

$$\Leftrightarrow \quad (\mu - x_1)\mu = 0$$

$$\Leftrightarrow \quad \mu^2 - \mu x_1 = 0 \quad \Leftrightarrow \quad \mu = x_1.$$

Using the same argument as above, we found in the case $\mu > x_n$, $\mu = x_1$ is the optimal choice for minimising L_∞.

- $x_1 < \mu < x_n$, then $l_\infty(\mu, D) = |\mu - x_1|$ or $|\mu - x_n|$ depends on μ is closer to x_n or to x_1, respectively.

⇒ Using the same argument, the best representative for this case can be either $\mu = x_1$ or $\mu = x_n$.

Finally, for all 3 cases, we found that μ can only be either $\mu = x_1$ or $\mu = x_n$. Thus, the best representative μ for any data set is the average of x_1 and x_n

$$\Rightarrow \mu = \frac{x_1 + x_n}{2}.$$

4.

μ	L_1	L_2	L_∞
D_1	2	2.2	2.5
D_2	2	21.2	50
D_3	30	35	50

- For $L_1(\mu, D) = \sum_{i=1}^{N} |\mu - x_i|$, we see that this loss function is robust and not too sensitive to outliers (comparing to L_2).

Indeed,

D_2 : $L_1 = |2-0| + |2-1| + |2-2| + |2-3| + |2-100|$

$= 2 + 1 + 0 + 1 + 98 = 102$

- For $L_2(\mu, D) = \sum (\mu - x_i)^2$, we see that this loss function is very sensitive to outliers as it can easily be skewed towards larger values (large squared distances).

D_2 : $L_2 = (21.2-0)^2 + (21.2-1)^2 + (21.2-2)^2 + (21.2-3)^2 + (21.2-100)^2$

$= 449.44 + 408.04 + 368.64 + 331.24 + 6209.44$

$= 7766.8$

- For $L_\infty = \max_{1 \le i \le n} |\mu - x_i|$, we see that this loss function is also not too sensitive to outliers.

D_2. $L_\infty = \max(|2.5-0|, |2.5-1|, |2.5-2|, |2.5-3|, |2.5-100|)$

$= \max(2.5, 1.5, 0.5, 0.5, 97.5)$

$= 97.5.$