Assignment 3
COMP3670

September 27, 2020

name: Xuecheng Zhang
UID:u6284513

31/8/2020

Exercise 1

1. Find the gradient ∇_Θ $\mathcal{L}_{A,\wedge}(\Theta)$.

$$\begin{aligned}\mathcal{L}_{A,\wedge}(\Theta) &=< y - X\Theta, y - X\Theta >_A + < \Theta, \Theta >_\wedge \\ &= (y - X\Theta)^T A(y - X\Theta) + \Theta^T \wedge \Theta \\ &= (y^T - \Theta^T X^T)A(y - X\Theta) + \Theta^T \wedge \Theta \\ &= (y^T Ay - \Theta^T X^T Ay - y^T AX\Theta + \Theta^T X^T AX\Theta) + \Theta^T \wedge \Theta \end{aligned} \tag{1}$$

As $y^T Ay$ has no element of Θ, the derivative of $y^T Ay$ is 0. What is more, using the first identity of computing gradients, we can get $\frac{\partial \Theta^T X^T Ay}{\partial \Theta} = (X^T Ay)^T$. Using the fourth identity of computing gradients $\frac{\partial x^T Bx}{\partial x} = x^T(B + B^T)$, we can get $\frac{\partial \Theta^T X^T AX\Theta}{\partial \Theta} = \Theta^T(X^T AX + (X^T AX)^T)$ and $\frac{\partial \Theta^T \wedge \Theta}{\partial \Theta} = \Theta^T(\wedge + \wedge^T)$

$$\begin{aligned}\nabla_\Theta \mathcal{L}_{A,\wedge}(\Theta) &= -(X^T Ay)^T - y^T AX + \Theta^T(X^T AX + (X^T AX)^T) + \Theta^T(\wedge + \wedge^T) \\ &= -2y^T A^T X + \Theta^T(X^T AX + X^T A^T X) + \Theta^T(\wedge + \wedge^T) \\ &= -2y^T A^T X + 2\Theta^T X^T AX + 2\Theta^T \wedge \end{aligned} \tag{2}$$

2. **Let ∇_Θ $\mathcal{L}_{A,\wedge}(\Theta) = \mathbf{0}$, and solve for Θ**
As A, $\wedge$ are symmetric, positive definite so that A = A^T, $\wedge^T = \wedge$

$$\begin{aligned}\nabla_\Theta \mathcal{L}_{A,\wedge}(\Theta) &= 0 \\ 2y^T A^T X &= 2\Theta^T X^T AX + 2\Theta^T \wedge \\ \Theta^T &= y^T A^T X(X^T AX + \wedge)^{-1} \\ \Theta &= ((X^T AX + \wedge)^{-1})^T X^T Ay \\ \Theta &= ((X^T AX + \wedge)^T)^{-1} X^T Ay \\ \Theta &= (X^T AX + \wedge)^{-1} X^T Ay \end{aligned} \tag{3}$$

3. **Show that if we choose A = I, and $\wedge = \lambda$I, where $\lambda \in$ R, your answer for 2. agrees with the analytic solution for the standard least squares regression problem with regularization, given by $\Theta = (X^T X + \lambda I)^{-1} X^T y$.**

When A = I, and $\wedge = \lambda$I, the Θ is

$$\Theta = (X^T IX + \lambda I)^{-1} X^T Iy = (X^T X + \lambda I)^{-1} X^T y \tag{4}$$

4.**What advantages are there in choosing $||\Theta||^2_\wedge$ over the more standard regularization term $\lambda||\Theta||^2_2$?**

As $\lambda||\Theta||^2_2 = \lambda\Theta^T\Theta = \Theta^T\lambda I\Theta$
When $\wedge = \lambda I$, $||\Theta||^2_\wedge = \lambda||\Theta||^2_2$

Therefore, $\lambda||\Theta||_2^2$ is a special case of $||\Theta||_\wedge^2$ which means $||\Theta||_\wedge^2$ can make loss function more generally than standard regularization term $\lambda||\Theta||_2^2$.
Using regular term $\lambda||\Theta||_2^2$, sometimes it is difficult to find the optimal value λ, as the result may stuck in local minimal. However, using this $||\Theta||_\wedge^2$ will be more comprehensive because it covers $\lambda||\Theta||_2^2$ result, and its results will only have better or the same results and not worse than it.
In this case, we are more likely and easily to find $\wedge$ make the regularisation term not particularly important to the value of loss function, but it limits the weight of θ and avoids overfitting.

Exercise 2

1. **For the L_2 loss function, L_2(μ, D) $= \sum_{i=1}^N (\mu - x_i)^2$, show that the optimal choice for μ is the mean, $\mu = \frac{1}{N}\sum_{i=1}^N x_i$.**

Define F(μ) = L_2(μ,D) = $\sum_{i=1}^N (\mu - x_i)^2$
The smallest loss is obtained when we find μ which makes the derivative of the loss function to zero.

$$\begin{aligned} \frac{dF(\mu)}{d\mu} &= 2(\mu - x_1) + 2(\mu - x_2) + 2(\mu - x_3) + \cdots + 2(\mu - x_N) \\ &= 2N\mu - 2(x_1 + x_2 + \cdots + x_N) = 0 \end{aligned} \tag{5}$$

When $2N\mu = 2(x_1 + x_2 + \cdots + x_N)$, i.e. $\mu = \frac{x_1+x_2+\cdots+x_N}{N} = \frac{\sum_{i=1}^N x_i}{N}$.
When $\mu < \frac{\sum_{i=1}^N x_i}{N}$, $\frac{dF(\mu)}{d\mu}$ <0 indicates that $F(\mu)$ is decreasing to approach to the minimum. When $\mu > \frac{\sum_{i=1}^N x_i}{N}$, $\frac{dF(\mu)}{d\mu}$ >0 indicates that $F(\mu)$ is moving away from the minimum. Therefore, when $\mu = \frac{\sum_{i=1}^N x_i}{N}$ which is mean value of in the set, $F(\mu)$ has the minimum value.

2. For the L_1 loss function, L_1(μ, D) $= \sum_{i=1}^N |\mu - x_i|$, show that the optimal choice for μ is the median value in the set.

Define $F(\mu) = L_1(\mu, D) = \sum_{i=1}^N |\mu - x_i|$

As $\mu \in \mathbb{R}$, the derivative of $F(\mu)$ is:

When $\mu \leq x_1$,

$$\frac{dF(\mu)}{d\mu} = \frac{d(x_1 - \mu + x_2 - \mu + \cdots + x_N - \mu)}{d\mu} = -N \tag{6}$$

When $\mu \geq x_n$,

$$\frac{dF(\mu)}{d\mu} = \frac{d(\mu - x_1 + \mu - x_2 + \cdots + \mu - x_N)}{d\mu} = N \tag{7}$$

The minimum value exists among x_1 and x_n

When $x_1 < \mu < x_n$,
Assume that $x_1 < x_j < \mu < x_{j+1} < x_n$, n = 2k, k, j $\in \mathbb{R}$

$$\begin{aligned}\frac{dF(\mu)}{d\mu} = \frac{d(\mu - x_1 + \mu - x_2 + \cdots + \mu - x_j + x_{j+1} - \mu + \cdots + x_n - \mu)}{d\mu} \\ = j - (2k - j - 1 + 1) = 2j - 2k \qquad (8)\end{aligned}$$

$$F(\mu) = \begin{cases} j < k \text{ and } x_j < \mu < x_{j+1} & F(\mu) < 0. \text{ The function is decreasing} \\ j > k \text{ and } x_j < \mu < x_{j+1} & F(\mu) > 0. \text{ The function is increasing} \\ j = k \text{ and } x_j < \mu < x_{j+1} & F(\mu) = 0. \text{ The function is constant} \end{cases}$$

The function exist minimum value when $x_{\frac{n}{2}} < \mu < x_{\frac{n}{2}+1}$:

$$\begin{aligned}\sum_{i=1}^{N} F(\mu) = \mu - x_1 + \mu - x_2 + \cdots + \mu - x_{\frac{n}{2}} + x_{\frac{n}{2}+1} - \mu + \cdots + x_n - \mu \\ = (\frac{n}{2})\mu - (\frac{n}{2})\mu - \sum_{i=1}^{\frac{n}{2}} x_i + \sum_{i=\frac{n}{2}+1}^{n} x_i = -\sum_{i=1}^{\frac{n}{2}} x_i + \sum_{i=\frac{n}{2}+1}^{n} x_i \qquad (9)\end{aligned}$$

When $x_{\frac{n}{2}} < \mu < x_{\frac{n}{2}+1}$ have the minimum value $-\sum_{i=1}^{\frac{n}{2}} x_i + \sum_{i=\frac{n}{2}+1}^{n} x_i$
The median of the set is $x_{\frac{n}{2}} < \frac{1}{2}$ $(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) < x_{\frac{n}{2}+1}$. so when n = 2k, μ is the median value having the minimum L_1 loss value.

Assume that $x_1 < x_j \leq \mu \leq x_{j+1} < x_n$, n = 2k+1, k, j $\in \mathbb{R}$

$$\begin{aligned}\frac{dF(\mu)}{d\mu} = \frac{d(\mu - x_1 + \mu - x_2 + \cdots + \mu - x_j + x_{j+1} - \mu + \cdots + x_n - \mu)}{d\mu} \\ = j - (n - j) = 2j - (2k + 1) = 2j - 2k - 1 \qquad (10)\end{aligned}$$

$$F(\mu) = \begin{cases} j <= k \text{ and } x_j < \mu < x_{j+1} & F(\mu) < 0. \text{ The function is decreasing} \\ j > k \text{ and } x_j < \mu < x_{j+1} & F(\mu) > 0. \text{ The function is increasing} \end{cases}$$

When j = k or j = k+1, F(μ) have the minimum value since the gradient $\frac{dF(\mu)}{d\mu}$ is approaching to 0.

When j = k

$$\begin{aligned}F(x_k) = x_k - x_1 + x_k - x_2 + \cdots + x_k - x_k + x_{k+1} - x_k + \cdots + x_{2k+1} - x_k \\ = x_k - x_1 + x_k - x_2 + \cdots + x_k - x_{k-1} + x_{k+1} - x_k + \cdots + x_{2k+1} - x_k \\ = -2x_k - x_1 - x_2 - \cdots - x_{k-1} + x_{k+1} + x_{k+2} + \cdots + x_{2k+1} \qquad (11)\end{aligned}$$

When j = k + 1

$$F(x_{k+1}) = x_{k+1} - x_1 + x_{k+1} - x_2 + \cdots + x_{k+1} - x_k + x_{k+1} - x_{k+1} + \cdots + x_{2k+1} - x_{k+1}$$
$$= x_{k+1} - x_1 + x_{k+1} - x_2 + \cdots + x_{k+1} - x_k + x_{k+2} - x_{k+1} \cdots + x_{2k+1} - x_{k+1}$$
$$= -x_1 - x_2 - \cdots - x_k + x_{k+2} + x_{k+3} + \cdots x_{2k+1} \tag{12}$$

$$F(x_{k+1}) - F(x_k) = -x_1 - x_2 - \cdots - x_k + x_{k+2} + x_{k+3} + \cdots + x_{2k+1} -$$
$$(-2x_k - x_1 - x_2 - \cdots - x_{k-1} + x_{k+1} + x_{k+2} + \cdots + x_{2k+1}) \tag{13}$$
$$= x_k - x_{k+1} < 0 \tag{14}$$

Therefore, $F(x_{k+1}) < F(x_k)$. Thus, when $\mu = x_{k+1} = x_{\frac{1+(2k+1)}{2}}$, n = 2k + 1, μ is the median value having the minimum L_1 loss value.
In a conclusion, the optimal choice for μ is the median value in the set.

4. **For the L_∞ loss function, $L_\infty(\mu$, D) = $max_{1 \leq i \leq n}$ $|\mu - x_i|$, show that the optimal choice for μ is $\frac{x_1 + x_n}{2}$,the average of only the largest and smallest data point in the set.**

When $\mu \leq x_1$,
As $\mu \leq x_1$, the maximum difference between μ and x_i is when x_i has the max value, which $(x_i = x_n)$.

$$L_\infty(\mu, D) = max_{1 \leq i \leq n} |\mu - x_i| = x_n - \mu \tag{15}$$

To get the minimum of $L_\infty(\mu, D)$, we have to maximize μ. When $\mu = x_1$ we get the minimum value.

When $\mu \geq x_n$,
As $\mu \geq x_n$, the maximum difference between μ and x_i is when x_i has the min value, which $(x_i = x_1)$.

$$L_\infty(\mu, D) = max_{1 \leq i \leq n} |\mu - x_i| = \mu - x_1 \tag{16}$$

To get the minimum of $L_\infty(\mu, D)$, we have to minimize μ. When $\mu = x_n$ we get the minimum value.
Above two conditions, we have the minimum $L_\infty(\mu$, D) $= x_n - x_1$

When $x_1 < \mu < x_n$,
Assume that μ is closer to x_1 than x_n (i.e. $|\mu - x_1| < |\mu - x_n|$). Therefore, the maximum of the value is not $|\mu - x_1|$.
(Situation 1) If there exists $x_k, x_1 < x_k < \mu$ has the maximum distance between x_k and μ. However, it is impossible since the distance between x_1 and μ is farther than the distance between x_k and μ. Also, the maximum of the value is

not $|\mu - x_1|$. Therefore, the maximum x_k can only exists between μ and x_n.

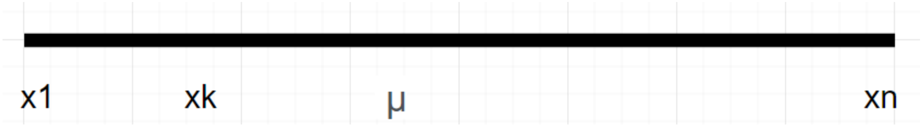

Figure 1: Situation 1

(Situation 2) If there exists x_k, $\mu < x_k < x_n$ has the maximum distance between x_k and μ. However, it is impossible since the distance between x_n and μ is farther than the distance between x_k and μ.

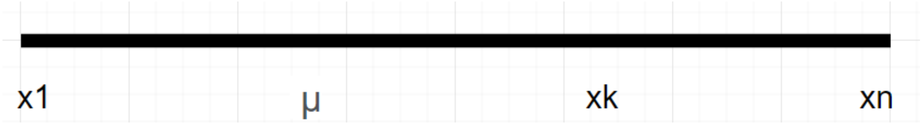

Figure 2: Situation 2

Therefore, when $x_k = x_n$ has the maximum value between a_k and μ.
As $F(\mu) = L_\infty(\mu, D) = |\mu - x_n| = x_n - \mu$ and μ is closer to x_1, when the distance between μ and x_1 is large but it does not exceed the distance between μ and x_n, $F(\mu)$ is the minimum value. That's when the distance between μ and x_1 is equal to the distance between μ and x_n, i.e. $\mu = \frac{x_1+x_n}{2}$.

Assume that μ is farther to x_1 than x_n (i.e. $|\mu - x_1| > |\mu - x_n|$).Therefore, the maximum of the value is not $|\mu - x_n|$.
(Situation 3) If there exists $x_k, \mu < x_k < x_n$ has the maximum distance between x_k and μ. However, it is impossible since the distance between x_n and μ is farther than the distance between x_k and μ. Also, the maximum of the value is not $|\mu - x_n|$. Therefore, the maximum x_k can only exists between x_1 and μ.

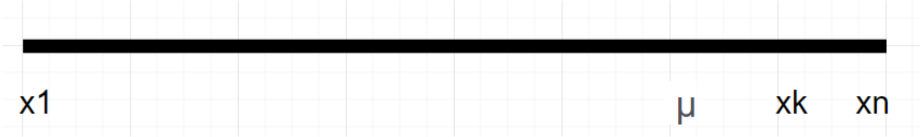

Figure 3: Situation 3

(Situation 4) If there exists x_k, $x_1 < x_k < \mu$ has the maximum distance between x_k and μ. However,it is impossible since the distance between x_1 and μ is farther than the distance between x_k and μ.

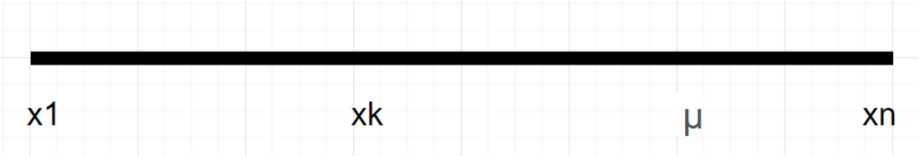

Figure 4: Situation 4

Therefore, when $x_k = x_1$ has the maximum value between x_k and μ.
As $F(\mu) = L_\infty(\mu, D) = |\mu - x_1| = \mu - x_1$ and μ is closer to x_n, when the distance between μ and x_n is large enough but it does not exceed the distance between μ and x_1, $F(\mu)$ has the minimum value. That's when the distance between μ and x_1 is equal to the distance between μ and x_n, i.e. $\mu = \frac{x_1+x_n}{2}$.

Above all, the minimum value is $L_\infty(\mu, \text{D}) = \frac{x_n-x_1}{2}$, when $\mu = \frac{x_1+x_n}{2}$

Therefore, for the L_∞ loss function, $L_\infty(\mu, \text{D}) = max_{1\leq i\leq n}|\mu - x_i|$, show that the optimal choice for μ is $\frac{x_1+x_n}{2}$

4. **For the following data sets, compute the optimal representative µ with respect to the L_2, L_1 and L_∞ loss function. Draw your result in a table as shown below. (You don't need to show the calculations for this.) Comment on how sensitive each loss function is to outliers.**

μ	L_1	L_2	L_∞
$\mathcal{D}_1$	2	2.2	2.5
$\mathcal{D}_2$	2	21.2	50
$\mathcal{D}_3$	30	35	50

L_1 is the least sensitive to the data because when $\mathcal{D}_2$ include 100 which includes a large outlier in the list, the result remains the same as the $\mathcal{D}_1$ which excludes any outlier in the list.
However,L_∞ is the most sensitive to the data because when $\mathcal{D}_2$ include 100 which includes a large outlier in the list, the result obtained is more than 10 times that of other normal data. Even if the outlier is not obvious, it will produce results that deviate from the normal data
Talking to L_2, it is not sensitive when the outlier is not very far away from normal data. However, when the outlier is extremely far away from the normal data, it is sensitive and the result is more than 7 times that of other normal data.