

Homework 2

Jack Kai Lim

April 17, 2022

Problem 1

a)

As the absolute risk minimized gives us the median of a data set, we can conclude that the value of c closer to the median will give the smaller value. In this case, it would be $R(c+2)$. And $R(c)$ is always larger than $R(c+2)$ by a value of 1.2. It will always be a value of 1.2 as the absolute loss function measures the median and the distance from the median of c and $c+2$ regardless of value is always inbetween y_2 and y_3 and a distance of 2 apart. Therefore the values of $R(c)$ and $R(c+2)$ always differs by a value of 1.2.

b)

The value of $R(c)$ and $R(c+2)$ when $y_5 < c < c+2 < y_6$ are always equal as both values lie in the range of the median of the set of data therefore always returning the same value.

c)

As the values of $R(c)$, $R(c+2)$ and V are all equal. We can conclude that the median of the set of data lie within the range of c to $c+2$. Therefore if a new value of $c+0.5$ is added, $c+0.5$ will then become the new median of the dataset. Hence, the value of $R_{ab}(h)$ is now minimized at $h = c+0.5$.

d)

As the R_{sq} gives a lower number the better the prediction is, and its minimizer gives the mean of the dataset. We can conclude that $R_{sq}(c)$ will be smaller than $R_{sq}(c-1)$ as it is closer to the mean of the dataset. To find by how much they

differ by we can solve the equation $R_{sq}(c-1) - R_{sq}(c)$,

$$R_{sq}(c-1) - R_{sq}(c) = \frac{1}{n} \left[\sum_{i=1}^n (y_i - (c-1))^2 - \sum_{i=1}^n (y_i - c)^2 \right] \quad (1)$$

$$= \frac{1}{n} \left[\sum_{i=1}^n y_i^2 - y_i^2 - 2y_i(c-1) + 2y_ic + (c-1)^2 - c^2 \right] \quad (2)$$

$$= \frac{1}{n} \sum_{i=1}^n 2y_i - 2c + 1 \quad (3)$$

Therefore they differ by the equation, $\frac{1}{n} \sum_{i=1}^n 2y_i - 2c + 1$.

e)

As the minimum of R_{sq} is the mean of the dataset. We can determine h^* by calculating the mean of the dataset. And as the mean is a linear function, we know that if a value changes it is reflected by the change. In this example we have that,

$$mean(D) = \frac{y_1 + y_2 + \dots + y_{10}}{n}$$

So if one of the values of the dataset increased by 2 we get,

$$mean(D) = \frac{y_1 + y_2 + \dots + y_i + 2 + \dots + y_{10}}{n} = \frac{2}{n} + \frac{y_1 + y_2 + \dots + y_{10}}{n}$$

Therefore we can tell that the change in the value of h^* is equal too 0.2.

Problem 2

a)

For the mean of the dataset, as it is a linear function the mean of the dataset of t_i is given by,

$$mean_t = f(mean(d_i)) = a \cdot mean_d + b$$

For the variance of the dataset, as it measures the distribution of the data in the dataset, constant addition does nothing to the variance. Therefore the only thing that affect it is multiplication which is varied by a square.

$$var_t = a^2 \cdot var_d$$

As the standard deviation is just the square root of the variance, the difference is just multiplying a.

$$std_t = a \cdot std_d$$

b)

The same does not hold for the standard deviation, and the restrictions, it that addition does not affect the standard deviation as it is a measure of difference. Therefore only multiplication affects it. In the case of the standard deviation it affects it linearly such that if the data is multiplied by a factor of 10, the standard deviation scales by a factor of 10 as well.

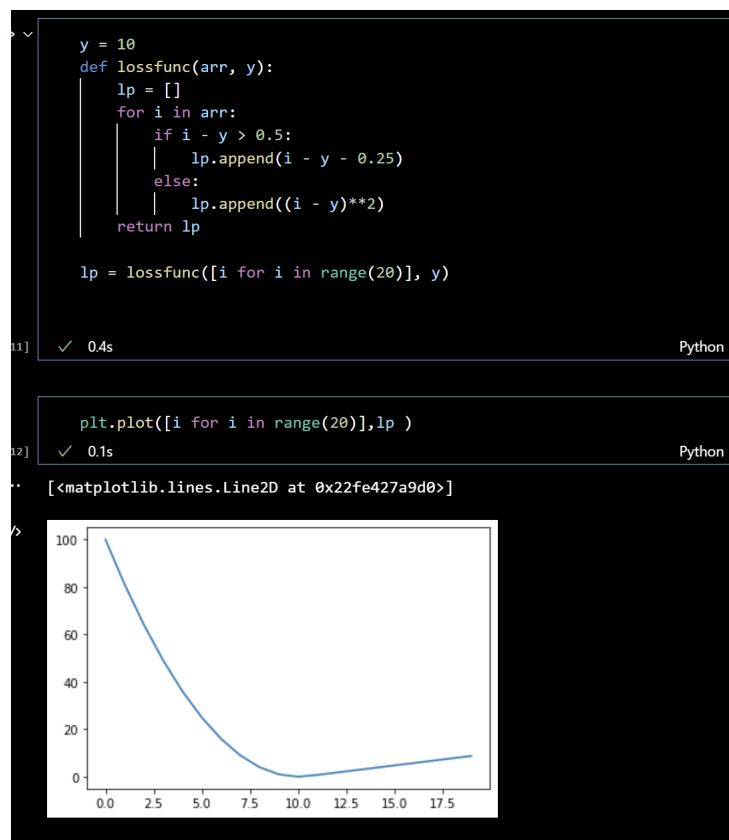
In the case of the variance, as it is the square of the standard deviation, the variance scale with the square of the scalar, eg if the data is multiplied by a factor of 10, the variance is scaled by a factor of 10^2

Problem 3

I would rank the distributions from smallest to largest as $A < C < B < D$. My reasoning for this is because A nearly every single datapoint is in the median bin on the histogram, therefore it will produce the smallest value of MAD. Next I chose C, as although it is skewed to the right, the median is still contained within the cluster of data points on the right. Therefore still having little to no deviation from the median. Then I chose B as the dataset is evenly spread with the median in the center. I think it is larger than C as although means are sensitive to outlier values, I think that due to the number of data points that are deviated from the median, it makes up for it and causes B to give a larger value than C. Finally D is the largest as again, the mean is sensitive to outliers, and in this scenario, there are a lot of outlier datapoints which deviate by a large distance from the median, therefore giving the largest value of MAD.

Problem 4

a)



b)

First we look at the piecewise function when $h - y > \frac{1}{2}$,

$$\lim_{x \rightarrow a} x - y - \frac{1}{4} = a - y - \frac{1}{4} \quad (4)$$

$$a - y - \frac{1}{4} = a - y - \frac{1}{4} \quad (5)$$

That proves continuity for the function, now to prove differentiability,

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \rightarrow 0} \frac{x+h-y-\frac{1}{4} - x+y+\frac{1}{4}}{h} \quad (6)$$

$$= \frac{h}{h} \quad (7)$$

$$= 1 \quad (8)$$

As the limit exist, the function is therefore continuous at all points. Now we look at the function when $h - y \leq \frac{1}{2}$

$$\lim_{x \rightarrow a} (x - y)^2 = (a - y)^2 \quad (9)$$

$$(a - y)^2 = (a - y)^2 \quad (10)$$

As the equations are equal, the function is therefore continuous. Now to prove differentiability,

$$\lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h} = \lim_{h \rightarrow 0} \frac{(x + h - y)^2 - (x - y)^2}{h} \quad (11)$$

$$= \lim_{h \rightarrow 0} \frac{x^2 - 2x(h - y) + (h - y)^2 - x^2 + 2xy - y^2}{h} \quad (12)$$

$$= \lim_{h \rightarrow 0} \frac{h^2 - 2hx - 2hy}{h} \quad (13)$$

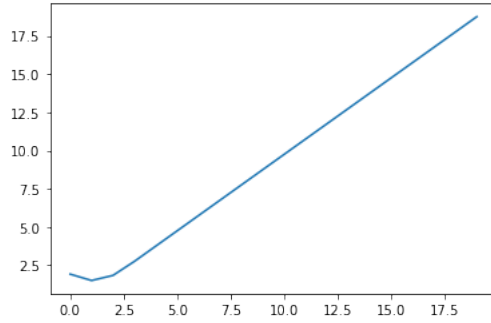
$$= -2x - 2y \quad (14)$$

As the limit exist the function is therefore differentiable. Now we can write the function derivative.

$$\frac{\delta}{\delta h} L_p(h; y) \begin{cases} 1 & h - y > \frac{1}{2} \\ 2(h - y) & h - y \leq \frac{1}{2} \end{cases}$$

c)

This is a plot for the empirical risk using the loss function with the arr for values of h ranging from 1 to 19



Problem 5

a)

When $(h - y), (h' - y) \leq \frac{1}{2}$,

$$= 4(L_p(h; y_i) - L_p(h'; y_i)) + 4 \frac{d}{dh} L_p(h'; y_i)(h' - h) \quad (15)$$

$$= 4((h - y)^2 - (h' - y)^2) + 4(2(h' - y)(h' - h)) \quad (16)$$

$$= 4(h^2 - h'^2 + 2h'^2 - 2hh') \quad (17)$$

$$= 4(h^2 + h'^2 - 2hh') \quad (18)$$

and,

$$= (\frac{d}{dh} L_p(h; y_i) - \frac{d}{dh} L_p(h'; y_i))^2 \quad (19)$$

$$= (2(h - y) - 2(h' - y))^2 \quad (20)$$

$$= 4(h^2 - 2hy + y^2 - 2(hh' - hy - h'y + y^2) + h'^2 - 2h'y + y^2) \quad (21)$$

$$= 4(h^2 + h'^2 - 2hh') \quad (22)$$

As both equations show equality, the inequality is true.

When $(h - y) \leq \frac{1}{2}$ and $(h' - y) > \frac{1}{2}$,

$$= 4(L_p(h; y_i) - L_p(h'; y_i)) + 4 \frac{d}{dh} L_p(h'; y_i)(h' - h) \quad (23)$$

$$= 4((h - y)^2 - h' + y + \frac{1}{4}) + 4(h' - h) \quad (24)$$

$$= 4h^2 - 8hy - 4h + 4y^2 + 4y + 1 \quad (25)$$

$$= 4(h^2 - 2hy + y^2) - 4(h - y) + 1 \quad (26)$$

$$= (2(h - y) - 1)^2 \quad (27)$$

$$= (\frac{d}{dh} L_p(h; y_i) - \frac{d}{dh} L_p(h'; y_i))^2 \quad (28)$$

As the equation shows equality, the inequality holds.

When $(h - y), (h' - y) > \frac{1}{2}$

$$= 4(L_p(h; y_i) - L_p(h'; y_i)) + 4 \frac{d}{dh} L_p(h'; y_i)(h' - h) \quad (29)$$

$$= 4(h - y - \frac{1}{4} - h' + y + \frac{1}{4}) + 4(h' - h) \quad (30)$$

$$= 4h - 4h' + 4h' - 4h \quad (31)$$

$$= 0 \quad (32)$$

and,

$$= \left(\frac{d}{dh} L_p(h; y_i) - \frac{d}{dh} L_p(h'; y_i) \right)^2 \quad (33)$$

$$= (1 - 1)^2 \quad (34)$$

$$= 0 \quad (35)$$

As equality holds, the inequality is true.

When $(h - y) > \frac{1}{2}$ and $(h' - y) \leq \frac{1}{2}$

$$\begin{aligned} &= \left(\frac{d}{dh} L_p(h; y_i) - \frac{d}{dh} L_p(h'; y_i) \right)^2 - [4(L_p(h; y_i) - L_p(h'; y_i)) + 4 \frac{d}{dh} L_p(h'; y_i)(h' - h)] \leq 0 \\ &= (1 - 2(h' - y))^2 - 4(h - y) + 1 + 4(h' - y)^2 - 8(h' - y)(h' - h) \leq 0 \\ &= 2(h' - y)[2(h' - y) - 1 - 2(h' - h)] - 4(h - y) + 1 \leq 0 \\ &= 2(h' - y)[2h - 2y - 1] - 4h - 4y + 1 \leq 0 \\ &= 4h[h' - y' - 1] - 4y[h' - y' - 1] - 2[h' - y' - 1] \leq 0 \\ &= 4h' - 4y - 2 \leq 0 \\ &= h' - y \leq \frac{1}{2} \end{aligned}$$

As the inequality is achieved, the inequality equation is true.

b)

$$\begin{aligned} &= \left[\frac{d}{dh} R_p(h) - \frac{d}{dh} R_p(h') \right]^2 \\ &= \left[\frac{1}{n} \sum_{i=1}^n \frac{d}{dh} L_p(h; y_i) - \frac{d}{dh} L_p(h'; y_i) \right]^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n [L_p(h; y_i) - L_p(h'; y_i)] \end{aligned}$$

Therefore using all the facts that we have gotten we can prove the inequality,

$$\begin{aligned} &= \left(\frac{d}{dh} R_p(h) - \frac{d}{dh} R_p(h^*) \right)^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \left[\frac{d}{dh} L_p(h; y_i) - \frac{d}{dh} L_p(h^*; y_i) \right]^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \left[4(L_p(h; y_i) - L_p(h^*; y_i)) - \frac{d}{dh} L_p(h^*; y_i)(h^* - h) \right] \\ &= \frac{4}{n} \sum_{i=1}^n [L_p(h; y_i) - L_p(h^*; y_i)] - \frac{1}{n} \sum_{i=1}^n L_p(h^*; y_i)(h^* - h) \\ &= 4[R_p(h; y_i) - R_p(h^*)] \end{aligned}$$