Homework 1

Jack Kai Lim

April 9, 2022

Problem 1

a)

Since we know that the mean of the 7 salaries, we can get the summation of all the salaries by multiplying the mean by the number of salaries, $210K \times 7 = 1470K$. Substracting the incorrect value and adding the correct value we get $1470K - 200K + 270K = 1540K$ then to get the mean of the change we simply divide the value by 7. $1540K \div 7 = 220K$

b)

In general, one would be able to correct the mean if a data value changes. However, it is generally not possible for the median or the mode. As for the median, if a value changes without the values, you would not be able to locate the position of the value or the new position therefore you are unable to obtain a new mean. While for the mode, if the value changes without knowledge of the other values you will not be able to tell the number of the old value or the number of the new values or the number of the current mode therefore generally you will not be ablee to correct it.

Problem 2

a)

This statement is not always true, as all the numbers in the data set could be identical giving an average that is not larger nor smaller than the data values.

b)

This statement is not always true, as the average can be heavily skewed by extreme values.

c)

This statement is not always true, as the average of a data set can be heavily skewed by extreme values which usually ould caused the mean to not be the same value of the median.

d)

This statement is True, as if all the numbers are smaller than the average there is no possible way for the average of the numbers to be larger than the largest possible value.

Problem 3

a)

Skip is correct in his assumption as if we calculate the median for San Diego in Fahrenheit then covert it to Celsius we get,

$$median(66, 66, 66, 67, 69, 69, 70, 72, 74, 76, 77, 77) = 69.5$$

Then converting it to Celsius I get,

$$C = \frac{5}{9}(F - 32) = \frac{5}{9}(69.5 - 32) = 20.83 \text{ rounded to 2 decimal places}$$

And if I converted all the values to Celsius then calculate the median from there I would get,

$$median(18.89, 18.89, 18.89, 19.44, 20.56, 20.56, 21.11, 22.22, 23.33, 24.44, 25.0, 25.0) = 20.83$$

Therefore Skips assumption is correct.

b)

Proof.

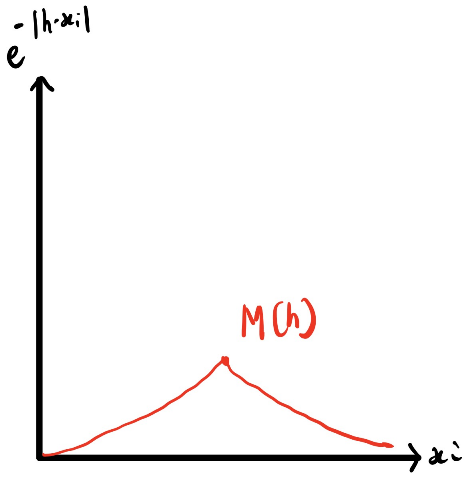
$$\begin{aligned} Median(g(t_1), ..., g(t_n)) &= g(Median(t_1, ..., t_n)) \\ Median([\frac{5}{9}(t_1 - 32), ..., \frac{5}{9}(t_n - 32)]) &= g(t_m) \\ \frac{5}{9}(t_m - 32) &= \frac{5}{9}(t_m - 32) \end{aligned}$$

□

c)

I think that the median is a better representation of the temperature in Osaka as Osaka experiences much more extreme values in temperature which the median is not affect hugely by outliers. While San Diego weather does not experience as large of changes therefore I would think that the mean would be a betetr representation of the average temperature in San Diego.

Problem 4



The graph of the function $M(h) = \sigma_{i=1}^{n} e^{-|h-x_i|}$ will produce a shape similar to the figure above. As the graph is based of the absolute loss of the prediction, the larger values of M(h) corresponds to a better prediction as the smaller the value of the power of the exponent, the larger the value of M(h) therefore the larger value of M(h) corresponds to a better prediction.

b)

As g is the function of the negative exponent, given the shape of the graph of it. We know that the max of the exponent is given by the smallest value in the domain. Therefore,

$$g(x_{max}) = g(e^{-min(cf(x))})$$

This is True as $min(cf(x)) = cf(x_{min}) \forall c \in \mathbb{R}$. Therefore for the function $g(e^{-cf(x)})$ the minimizer for f is also the maximizer for g.

c)

The value of M(h) is maximized when $h^* = x_i$ as the absolute loss is 0 corresponding to the largest value possible for M(h). As the function uses the exponential function, it is very sensitive to any changes in the data, therefore it is bad as it will be incredibly difficult to get the initial best prediction espicailly with large databases. On the other hand with smaller databases it can possibly give a very accurate prediction as it has less values to deviated by.

Problem 5

a)

Proof.

$$\begin{aligned} R_{sq}(h, D) &= \frac{1}{n}\sum_{i=1}^{n} h^2 - 2hx_i + x_i^2 \\ &= h^2 - \frac{1}{n}\sum_{i=1}^{n} 2hx_i + \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\ &= h^2 - \frac{2h(x_1, ..., x_n)}{n} + \frac{(x_1^2, ..., x_n^2)}{n} \end{aligned}$$

Let $y = \frac{x_1, ..., x_n}{n}$ and $z = \frac{(x_1^2, ..., x_n^2)}{n}$ therefore we get that,

$$R_{sq}(h, D) = \frac{1}{n}\sum_{i=1}^{n} h^2 - 2hx_i + x_i^2 = h^2 - 2hy + z$$

Then we complete the square of the equation which yields,

$$(h - y)^2 + z$$

□

b)

As the function will give us a quadratic curve, the minimizer of the functionis given by the min point of the curve which is (y, z) where y is the min for h and z is the min for the set of D.

c)

The quantity of y can be of any sign as it is just the mean of the set of D. While z is the square mean of the set of D, therefore it is geenrally a positive value as squaring any number always gives a positive value.