Assignment 1: CS 215
Solutions

Neeraj Dhake 150050022

Rohit Kumar Jena 150050061

6th August 2016

Honor Code:

- We pledge by our honor that we will complete the assignments in a legitimate way and will not provide or recieve any unauthorized help.

Instructions to run code:

- For Question 4, the corresponding files are *hw1_q4.m* (to run the file and generate plots as well as give the relative errors for both the median and the mean) and *neighbours.m* (to generate the neighbours of the array at a given index .i.e. if index is 14 then it will return **z(6:22)**).
  To run the program, simply type *hw1_q4* from the MATLAB command line.

- For Question 5, the corresponding files are *hw1_q5.m* (to read the original array from the file named *input_array.txt* and compute the mean, median, and the standard deviation). Then it prompts the user to enter the new data value. It uses the *UpdateMean.m*, *UpdateMedian.m*, and *UpdateStd.m* files to compute the new mean, new median and new standard deviation from the old values.
  To run the file, simply type *hw1_q5* from the MATLAB command line. Make sure you have a file named *input_array.txt* containing the array elements in **one line**. Or you can just use the UpdateMean, UpdateMedian and UpdateStd files as you wish. Make sure the array A is horizontal.

Solutions:

1. Let μ = mean and v = median of the dataset $\{x_i\}_{i=1}^{N}$ containing N data points.

$$|\mu - v| = \left| \frac{\sum_{i=1}^{N}(x_i - v)}{N} \right| \leq \frac{\sum_{i=1}^{N}|x_i - v|}{N}$$

By Triangular Inequality,

$$\left| \frac{\sum_{i=1}^{N} x_i - v}{N} \right| \leq \frac{\sum_{i=1}^{N}|x_i - v|}{N}$$

$$\implies \frac{\sum_{i=1}^{N}|x_i - v|}{N} \leq \frac{\sum_{i=1}^{N}|x_i - \mu|}{N}$$

(the value for which $\frac{\sum_{i=1}^{N}|x_i - x|}{N}$ is minimum is when x = median)

Now,

$$\left(\frac{\sum_{i=1}^{N}|x_i - \mu|}{N}\right)^2 \leq \frac{\sum_{i=1}^{N}|x_i - \mu|^2}{N} \leq \frac{\sum_{i=1}^{N}|x_i - \mu|^2}{N-1} = \sigma^2$$

(by RMS-AM inequality)

$$\therefore |\mu - v| \leq \sigma$$

2. Given 4 datasets $\{x_i\}_{i=1}^{N}$, $\{y_i\}_{i=1}^{N}$, $\{z_i\}_{i=1}^{N}$, and $\{w_i\}_{i=1}^{N}$ such that,

$$z_i = ax_i + b$$

$$w_i = cy_i + d$$

Using the definition of correlation coefficient, we have,

$$r(x, y) = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{(N-1)\sigma_x\sigma_y}$$

$$r(z,w) = \frac{\sum_{i=1}^{N}(z_i - \mu_z)(w_i - \mu_w)}{(N-1)\sigma_z \cdot \sigma_w}$$

Now if mean of $\{x_i\}_{i=1}^{N} = \mu_x$ then mean of $\{z_i\}_{i=1}^{N} = a\mu_x + b$, and similarly mean of $\{w_i\}_{i=1}^{N} = a\mu_y + b$.

Substituting the values of z_i, μ_z, w_i, μ_w in the second equation, we get,

$$r(z,w) = \frac{\sum_{i=1}^{N}[(ax_i + b) - (a\mu_x + b)] \cdot [(cy_i + d) - (c\mu_y + d)]}{(N-1)\,|a|\,\sigma_x \cdot |c|\,\sigma_y}$$

$$= \frac{\sum_{i=1}^{N} a(x_i - \mu_x) \cdot c(y_i - \mu_y)}{(N-1)\,|a|\,\sigma_x \cdot |c|\,\sigma_y}$$

$$= sgn(ac) \cdot \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{(N-1)\sigma_x\sigma_y}$$

Where $sgn(ac)$ denotes the signum function.
Now, substituting the definition of $r(x,y)$, we have,

$$r(z,w) = sgn(ac) \cdot r(x,y)$$

When the sign of ac is positive, i.e. when both a and c are positive or negative, we have $sgn(ac) = 1$, and hence $r(z,w) = r(x,y)$.
When the sign of ac is negative, i.e. when the sign of a and c are opposite, we have $sgn(ac) = -1$, and hence $r(z,w) = -r(x,y)$.

3. Let x_i be the datapoints of a dataset $\{x_i\}_{i=1}^{N}$ with mean μ.
So, we have the following inequality,

$$|x_i - \mu| \leq \sqrt{\sum_{i=1}^{N}|x_i - \mu|^2}$$

This can easily be shown as follows:

$$|x_i - \mu|^2 \leq \sum_{i=1}^{N}|x_i - \mu|^2$$

Since both R.H.S and L.H.S are positive, taking square root on both sides gives the required result.

$$|x_i - \mu| \leq \sigma\sqrt{n-1}$$

Q.E.D.

4.

5. Let $\bar{a}$ denote the mean of a dataset $\{a_i\}_{i=1}^{N}$ having N datapoints.
Now, we have,

$$\bar{a} = \frac{\sum_{i=1}^{N} a_i}{N}$$

After we add a new element, say z, the new mean becomes,

$$\bar{a}_{new} = \frac{(\sum_{i=1}^{N} a_i) + z}{N+1} = \frac{N\bar{a} + z}{N+1}$$

$\therefore$ We have the value of $\bar{a}_{new}$ in terms of $\bar{a}, N, z$.

For the given data, let the standard deviation be denoted by σ. So we have,

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} = \frac{(\sum_{i=1}^{N} x_i^2) - N\bar{x}^2}{N-1}$$

After adding a new value z to the data, we have a new value σ_{new} denoted by,

$$\sigma_{new}^2 = \frac{\sum_{i=1}^{N+1}(x_i - \bar{x})^2}{N} = \frac{(\sum_{i=1}^{N+1} x_i^2) - (N+1)\bar{x}_{new}^2}{N}$$

Using the value of σ^2, we apply the following substitution,

$$\sum_{i=1}^{N} x_i^2 = (N-1)\sigma^2 + N\bar{x}^2$$

We get the following equation,

$$\sigma_{new}^2 = \frac{(N-1)\sigma^2 + N\bar{x}^2 + z^2 - (N+1)\bar{x}_{new}^2}{N}$$

$\therefore$ We have the value of σ_{new} in terms of σ, $\bar{x}$, $\bar{x}_{new}, z$, and N.