Question 7 : Updating Mean, Median, and Standard Deviation in MATLAB

Kavya Gupta

August 21, 2023

1 Instructions

All the MATLAB functions `UpdateMean, UpdateMedian, UpdateStd` are present in `Q7.m` file.
Also note that `IMG_4045.jpeg` and `IMG_4046.jpeg` in the zip are the solutions to UpdateStd derivation which is also attached in this pdf.

2 Introduction

In this report, we will derive and implement MATLAB functions to update the mean, median, and standard deviation of a dataset without recalculating them from scratch. This will allow us to efficiently incorporate new data points into our existing statistics.

3 Derivation: Updating Mean

Let's consider a dataset with n data points: $x_1, x_2, \ldots, x_n$.

The mean of the dataset is calculated as:

$$mean = \frac{1}{n} \sum_{i=1}^{n} x_i$$

We want to update the mean when adding a new data point x_{new} to the dataset.

3.1 Step 1: Original Mean

The original mean of the dataset with n data points is represented as:

$$oldMean = \frac{1}{n} \sum_{i=1}^{n} x_i$$

We can also express the sum of all data points as:

$$oldSum = \sum_{i=1}^{n} x_i$$

Since $mean = \frac{oldSum}{n}$, we have:

$$oldSum = oldMean \cdot n$$

3.2 Step 2: Updating with New Data

When we add a new data point x_{new} to the dataset, we need to update the mean. The updated sum can be calculated as:

$$newSum = oldSum + x_{new}$$

The updated mean, denoted as $newMean$, is then given by:

$$newMean = \frac{newSum}{n+1}$$

3.3 Step 3: Final Formula

Substituting the expression for $oldSum$ and $newSum$, we get:

$$newMean = \frac{(oldMean \cdot n) + x_{new}}{n+1}$$

This is the formula to update the mean when a new data point is added to the dataset.

4 Derivation: Updating Median

Let's consider a sorted dataset with n unique data points: $A = [x_1, x_2, \ldots, x_n]$.

We want to update the median when adding a new data point x_{new} to the dataset.

4.1 Step 1: Original Median

The original median of the dataset is the middle value of the sorted array A. For an odd number of data points (n is odd):

$$oldMedian = A\left(\frac{n+1}{2}\right)$$

For an even number of data points (n is even):

$$oldMedian = \frac{A\left(\frac{n}{2}\right) + A\left(\frac{n}{2}+1\right)}{2}$$

4.2 Step 2: Updating with New Data

When we add a new data point x_{new} to the sorted array A, we need to update the median.

We insert x_{new} into the array A to get a new sorted array $sortedA$.

4.3 Step 3: Calculating the Updated Median

After inserting x_{new} into the array, the total number of data points becomes $n+1$.

If $n+1$ is odd, the new median is the middle value of the updated array:

$$newMedian = sortedA\left(\frac{n+2}{2}\right)$$

If $n+1$ is even, the new median is the average of the two middle values in the updated array:

$$newMedian = \frac{sortedA\left(\frac{n+1}{2}\right) + sortedA\left(\frac{n+1}{2}+1\right)}{2}$$

5 Derivation : Update Standard Deviation

Handwritten Solutions given :-
`IMG_4045.jpeg` and `IMG_4046.jpeg` in the zip are same photos of my handwritten solution, which are also attached in this pdf.
Final Formula :-

$$newStd = \sqrt{\frac{(n-1)OldStd^2 + nOldMean^2 + NewDataValue^2 - (n+1)NewMean^2}{n}} \tag{1}$$

6 Histogram

We will check in which bin (interval) the `NewDataValue` lies in and then increase the frequency of that bin by 1.
To check in which bin the value will be in, we will compare it with lower and upper value of the bin.

Updating Standard deviation

$$\sigma = \text{OldStd} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

$$\text{So } \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \text{oldMean})^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n} \left(x_i^2 + (\bar{x})^2 - 2(x_i \cdot \bar{x}) \right)}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 + n \times \bar{x}^2 - 2 \times \sum_{i=1}^{n} x_i \cdot \bar{x}}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2n\bar{x}\bar{x}}{n-1}$$

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

$$\text{So } \sum_{i=1}^{n} x_i^2 = (n-1)(\text{OldStd})^2 + n(\text{OldMean})^2$$

Now restart newStd, same formula as above

$$\text{newStd}^2 = \frac{\sum_{i=1}^{n+1} x_i^2 - (n+1)(\text{newMean})^2}{n}$$

Now $\sum_{i=1}^{n+1} x_i^2 = \sum_{i=1}^{n} x_i^2 + (\text{newdata})^2$

$= (n-1)(\text{OldStd})^2 + n(\text{OldMean})^2 + (\text{newdata})^2$

so $(\text{newStd})^2 = \dfrac{(n-1)(\text{OldStd})^2 + n(\text{OldMean})^2 + (\text{newdata})^2 - (n+1)(\text{newMean})^2}{n}$

so, Final Formula.

$$\text{newStd} = \sqrt{\frac{(n-1)(\text{OldStd})^2 + n(\text{OldMean})^2 + (\text{newdata})^2 - (n+1)(\text{newMean})^2}{n}}$$