Online Weighted Mean

by Joshua Burkholder

Given the following set of inputs and their associated weights:

$$\{(x_1, w_1), (x_2, w_2), \ldots, (x_{n-1}, w_{n-1}), (x_n, w_n)\}$$

Let n be the number of inputs and their associated weights, $\overline{x}_n^{\text{weighted}}$ is the weighted sample mean for the first n inputs and their associated weights, $\overline{x}_{n-1}^{\text{weighted}}$ be the weighted sample mean of the first $n-1$ inputs and their associated weights, w_n be the n-th weight associated with input x_n, and x_n be the n-th input associated with w_n. Then, the recurrence equation for the weighted sample mean (a.k.a. online weighted mean) is:

$$\overline{x}_n^{\text{weighted}} = \overline{x}_{n-1}^{\text{weighted}} - \frac{w_n\left(\overline{x}_{n-1}^{\text{weighted}} - x_n\right)}{\sum_{i=1}^{n} w_i}$$

where $\sum_{i=1}^{n} w_i \neq 0$. Preferably, all the weights are positive such that $\sum_{i=1}^{n} w_i > 0$.

Proof:

The definition of the sample mean is:

$$\overline{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$$

The definition of the weighted sample mean is:

$$\overline{x}_n^{\text{weighted}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

If we expand this definition, we have:

$$\overline{x}_n^{\text{weighted}} = \frac{\sum_{i=1}^{n-1} w_i x_i + w_n x_n}{\sum_{i=1}^{n-1} w_i + w_n}$$

From algebra, we know that for arbitrary a, b, c, and d:

$$
\begin{aligned}
\frac{a+b}{c+d} &= \frac{a+b}{c+d} + \frac{a}{c} - \frac{a}{c} \\
&= \frac{a}{c} + \frac{a+b}{c+d} - \frac{a}{c} \\
&= \frac{a}{c} + \left(\frac{a+b}{c+d}\right)\left(\frac{c}{c}\right) - \left(\frac{a}{c}\right)\left(\frac{c+d}{c+d}\right) \\
&= \frac{a}{c} + \frac{ac+bc-ac-ad}{c(c+d)} \\
&= \frac{a}{c} + \frac{\cancel{ac}+bc-\cancel{ac}-ad}{c(c+d)} \\
&= \frac{a}{c} + \frac{bc-ad}{c(c+d)} \\
&= \frac{a}{c} + \frac{bc}{c(c+d)} + \frac{-ad}{c(c+d)} \\
&= \frac{a}{c} + \frac{b\cancel{c}}{\cancel{c}(c+d)} + \frac{-ad}{c(c+d)} \\
&= \frac{a}{c} + \frac{-ad}{c(c+d)} + \frac{b}{(c+d)}
\end{aligned}
$$

Hence, we have:

$$
\begin{aligned}
\overline{x}_n^{\text{weighted}} &= \frac{\overbrace{\sum_{i=1}^{n-1} w_i x_i}^{a} + \overset{b}{w_n x_n}}{\underbrace{\sum_{i=1}^{n-1} w_i}_{c} + \underset{d}{w_n}} \\
&= \frac{\left(\sum_{i=1}^{n-1} w_i x_i\right)}{\left(\sum_{i=1}^{n-1} w_i\right)} + \frac{-\left(\sum_{i=1}^{n-1} w_i x_i\right)(w_n)}{\left(\sum_{i=1}^{n-1} w_i\right)\left(\left(\sum_{i=1}^{n-1} w_i\right)+(w_n)\right)} + \frac{(w_n x_n)}{\left(\left(\sum_{i=1}^{n-1} w_i\right)+(w_n)\right)} \\
&= \left(\frac{\sum_{i=1}^{n-1} w_i x_i}{\sum_{i=1}^{n-1} w_i}\right) - \left(\frac{\sum_{i=1}^{n-1} w_i x_i}{\sum_{i=1}^{n-1} w_i}\right)\left(\frac{w_n}{\sum_{i=1}^{n} w_i}\right) + \frac{(w_n x_n)}{\left(\sum_{i=1}^{n} w_i\right)}
\end{aligned}
$$

Since the weighted sample mean for the first $n-1$ inputs and their associated weights is defined as $\overline{x}_{n-1}^{\text{weighted}} = \frac{\sum_{i=1}^{n-1} w_i x_i}{\sum_{i=1}^{n-1} w_i}$, we have:

$$\overline{x}_n^{\text{weighted}} = \left(\overline{x}_{n-1}^{\text{weighted}}\right) - \left(\overline{x}_{n-1}^{\text{weighted}}\right)\left(\frac{w_n}{\sum_{i=1}^{n} w_i}\right) + \frac{\left(w_n x_n\right)}{\left(\sum_{i=1}^{n} w_i\right)}$$

Factoring out the -1, we have:

$$\overline{x}_n^{\text{weighted}} = \left(\overline{x}_{n-1}^{\text{weighted}}\right) - \left(\left(\overline{x}_{n-1}^{\text{weighted}}\right)\left(\frac{w_n}{\sum_{i=1}^{n} w_i}\right) - \frac{\left(w_n x_n\right)}{\left(\sum_{i=1}^{n} w_i\right)}\right)$$

Combining the fractions and factoring out the w_n, we have:

$$\begin{aligned}\overline{x}_n^{\text{weighted}} &= \overline{x}_{n-1}^{\text{weighted}} - \left(\frac{\overline{x}_{n-1}^{\text{weighted}} w_n - w_n x_n}{\sum_{i=1}^{n} w_i}\right) \\ &= \overline{x}_{n-1}^{\text{weighted}} - \frac{w_n\left(\overline{x}_{n-1}^{\text{weighted}} - x_n\right)}{\sum_{i=1}^{n} w_i}\end{aligned}$$

Therefore, the recurrence equation for the weighted sample mean (a.k.a. online weighted mean) is:

$$\overline{x}_n^{\text{weighted}} = \overline{x}_{n-1}^{\text{weighted}} - \frac{w_n\left(\overline{x}_{n-1}^{\text{weighted}} - x_n\right)}{\sum_{i=1}^{n} w_i}$$

where $\sum_{i=1}^{n} w_i \neq 0$.

Note: If all the weights are the same constant value c (i.e. $w_i = c$ for $i = 1, \ldots, n$), the weighted sample mean would be:

$$\begin{aligned}
\bar{x}^{\text{weighted}} &= \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \\
&= \frac{\sum_{i=1}^{n} c x_i}{\sum_{i=1}^{n} c} \\
&= \frac{c\left(\sum_{i=1}^{n} x_i\right)}{c\left(\sum_{i=1}^{n} 1\right)} \\
&= \frac{\not{c}\left(\sum_{i=1}^{n} x_i\right)}{\not{c}\,(n)} \\
&= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \bar{x}
\end{aligned}$$

For instance, if all the weights are 1, then the weighted sample mean is the sample mean:

$$\begin{aligned}
\bar{x}^{\text{weighted}} &= \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \\
&= \frac{\sum_{i=1}^{n} (1) x_i}{\sum_{i=1}^{n} (1)} \\
&= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \bar{x}
\end{aligned}$$

Similarly, the online weighted mean with weights of the same constant value c would be:

$$\begin{aligned}
\bar{x}_n^{\text{weighted}} &= \bar{x}_{n-1}^{\text{weighted}} - \frac{w_n\left(\bar{x}_{n-1}^{\text{weighted}} - x_n\right)}{\sum_{i=1}^{n} w_i} \\
&= \bar{x}_{n-1}^{\text{weighted}} - \frac{c\left(\bar{x}_{n-1}^{\text{weighted}} - x_n\right)}{\sum_{i=1}^{n} c} \\
&= \bar{x}_{n-1}^{\text{weighted}} - \frac{c\left(\bar{x}_{n-1}^{\text{weighted}} - x_n\right)}{c\left(\sum_{i=1}^{n} 1\right)} \\
&= \bar{x}_{n-1}^{\text{weighted}} - \frac{\not{c}\left(\bar{x}_{n-1}^{\text{weighted}} - x_n\right)}{\not{c}\left(n\right)} \\
&= \bar{x}_{n-1}^{\text{weighted}} - \frac{\left(\bar{x}_{n-1}^{\text{weighted}} - x_n\right)}{n} \\
&= \bar{x}_{n-1} - \frac{\left(\bar{x}_{n-1} - x_n\right)}{n} \\
&= \bar{x}_n
\end{aligned}$$

Therefore, if all the weights are the same constant value c, the online weighted mean is the same as the online mean.

Example of C++ code that computes the online weighted mean:

```
#include <iostream>
#include <iomanip>

int main () {

    double x;
    double weight;
    double sum_of_weights = 0;
    double weighted_mean = 0;
    double prev_weighted_mean;

    if ( std::cin >> x && std::cin >> weight ) {
        sum_of_weights += weight;
        weighted_mean = x;
        while ( std::cin >> x && std::cin >> weight ) {
            prev_weighted_mean = weighted_mean;
            sum_of_weights += weight;
            weighted_mean = (
                prev_weighted_mean - weight * ( prev_weighted_mean - x ) / sum_of_weights
            );
        }
    }

    std::cout << "sum_of_weights: " << std::setprecision( 17 ) << sum_of_weights << '\n';
    std::cout << "weighted_mean:  " << std::setprecision( 17 ) << weighted_mean  << '\n';
}
```

Example of data.txt:

```
-19.313117172629575    2.718281828459045
-34.14656787734913     7.38905609893065
-14.117521595690334    20.085536923187668
.                      .
.                      .
.                      .
```

Command line:

```
g++ -o main.exe main.cpp -std=c++11 -march=native -O3 -Wall -Wextra -Werror -static
./main.exe < data.txt
```

Sample Output:

```
sum_of_weights: 34843.773845331321
weighted_mean:  -28.368899576339764
```