Introduction to Median

Su Jianlin

2026-03-31

Introduction to Median

Recently, I re-learned the concept of median and recorded the key points while the knowledge is still fresh.

When performing outlier rejection or clipping, we often need a "reference point". For example, for a set of non-negative data, we may consider any value greater than 50 times the reference point as an outlier. How should we select such a reference point? A commonly used metric is the mean (average). However, the mean is easily skewed by outliersusing it as a reference may bias the threshold toward outliers, potentially missing some actual anomalies. In such cases, we may consider using the median as the reference.

1 Basic Properties

For one-dimensional data points $x_1, x_2, \cdots, x_n$, their mean is defined as:

$$\text{mean}(x_1, x_2, \cdots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Since all data points directly contribute to the average, if a few values are extremely large, the mean will be pulled upward accordingly, thus interfering with outlier detection.

The idea of the median is to find a "neutral" value in the data as the reference. Specifically, it seeks a partition point such that the number of data points less than or equal to it is equal to the number greater than or equal to it. For this reason, it is also known as the "50% quantile". If n is odd, the median is the $(n+1)/2$-th largest number in the dataset. If n is even, any number between the $n/2$-th and $(n/2+1)$-th largest values can be considered the mediantypically, we take the average of these two values.

From its definition, we can see the median's robustness: as long as the relative ordering between each data point and the median does not change, the median remains unchanged. We can further quantify this robustness using the concept of "breakdown point", defined as "the minimum proportion of outliers required to make the estimate go to infinity". For the mean, this value is nearly zerobecause letting only one data point go to infinity will

drive the mean to infinity. For the median, however, the breakdown point is 50%, because more than half of the data must go to infinity for the median to diverge.

A disadvantage of the median is that it is computationally more complex than the mean, since it requires at least partial sorting of the datamaking it less friendly in distributed computing scenarios. Fortunately, in most situations, a highly precise median is not required, leaving room for approximation. For example, computing local medians and then aggregating them via a global average or median can significantly reduce communication costs.

2 Optimization Perspective

Interestingly, we can unify the mean and median under an optimization framework:

$$\text{mean}(x_1, x_2, \cdots, x_n) = \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{2}$$

$$\text{median}(x_1, x_2, \cdots, x_n) = \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} |x_i - \mu| \tag{3}$$

The derivation for the mean is straightforward and left to the reader. Here, we briefly prove the result for the median. Without loss of generality, assume the x_i are sorted: $x_1 \leq x_2 \leq \cdots \leq x_n$. Define $f(\mu) = \sum_{i=1}^{n} |x_i - \mu|$. Taking the derivative directly yields:

$$f'(\mu) = \sum_{i=1}^{n} \text{sign}(\mu - x_i) = \#\{x_i < \mu\} - \#\{x_i > \mu\} \tag{4}$$

where # denotes the number of elements satisfying the condition. To find the minimum, we want the derivative to be as close to zero as possiblei.e., the number of x_i greater than μ and less than μ should be as equal as possible, which aligns with the idea of the median. A more detailed analysis proceeds case by case:

$$f'(\mu) = \begin{cases} \left.\begin{array}{ll} \underbrace{\#\{x_i < \mu\}}_{\leq k} - \underbrace{\#\{x_i > \mu\}}_{\geq k+1} < 0, & \mu < x_{k+1} \\ \underbrace{\#\{x_i < \mu\}}_{\geq k+1} - \underbrace{\#\{x_i > \mu\}}_{\leq k} > 0, & \mu > x_{k+1} \end{array}\right\} & n = 2k+1 \\ \left.\begin{array}{ll} \underbrace{\#\{x_i < \mu\}}_{\leq k-1} - \underbrace{\#\{x_i > \mu\}}_{\geq k+1} < 0, & \mu < x_k \\ \underbrace{\#\{x_i < \mu\}}_{\geq k+1} - \underbrace{\#\{x_i > \mu\}}_{\leq k-1} > 0, & \mu > x_{k+1} \end{array}\right\} & n = 2k \end{cases} \tag{5}$$

When $n = 2k + 1$ (odd), both $\mu < x_{k+1}$ and $\mu > x_{k+1}$ increase $f(\mu)$, so the minimizer is $\mu^* = x_{k+1}$. When $n = 2k$ (even), both $\mu < x_k$ and $\mu > x_{k+1}$ increase $f(\mu)$, so

$\mu^* \in [x_k, x_{k+1}]$. It can be verified that $f(\mu^*)$ remains constant over this interval, meaning every value in $[x_k, x_{k+1}]$ is a valid minimizer. Thus, μ^* exactly matches the definition of the median.

3 Higher-Dimensional Spaces

From the optimization perspective, we can also understand why the median is more resistant to outliers than the mean: suppose one x_i is extremely large. The loss contributed to the mean is $(x_i - \mu)^2$, while for the median it is $|x_i - \mu|$. Typically, $(x_i - \mu)^2 \gg |x_i - \mu|$, so the mean is more strongly influenced by the outlier in order to minimize the total loss.

Additionally, the optimization perspective allows easy extension to higher-dimensional spaces. The concept of mean naturally generalizes: for a set of vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$, the mean vector is simply $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$. However, the median does not generalize directly, because it relies on sortinga well-defined order is hard to establish for vector data.

But through optimization, the generalization becomes natural:

$$\text{mean}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n) = \underset{\boldsymbol{\mu}}{\text{argmin}} \sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^2 \tag{6}$$

$$\text{median}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n) = \underset{\boldsymbol{\mu}}{\text{argmin}} \sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2 \tag{7}$$

Here, $\|\cdot\|_2$ denotes the Euclidean norm. It is easy to show that the vector minimizing the squared norm sum is exactly $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$, consistent with the empirical definition. The vector minimizing the sum of Euclidean distances is known as the "geometric median". When $n = 3$, this is the classical "Fermat point", so the geometric median is sometimes also referred to as the Fermat point.

Unfortunately, the geometric median has no closed-form solution. It is usually computed using the Weiszfeld iteration:

$$\boldsymbol{\mu}_{t+1} = \frac{\sum_{i=1}^n \boldsymbol{x}_i / \|\boldsymbol{x}_i - \boldsymbol{\mu}_t\|_2}{\sum_{i=1}^n 1 / \|\boldsymbol{x}_i - \boldsymbol{\mu}_t\|_2} \tag{8}$$

4 Further Generalization

Clearly, we can consider a more general formulation:

$$\text{average}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n; \alpha) = \underset{\boldsymbol{\mu}}{\text{argmin}} \sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^\alpha \tag{9}$$

Let $f(\boldsymbol{\mu}) = \sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^\alpha$. Then:

$$\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) = \alpha \sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^{\alpha-2} (\boldsymbol{x}_i - \boldsymbol{\mu}) \tag{10}$$

Setting $\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) = \mathbf{0}$, the equation to solve becomes:

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^{\alpha-2} \boldsymbol{x}_i}{\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_2^{\alpha-2}} \tag{11}$$

Substituting $\boldsymbol{\mu}_t$ on the right-hand side to compute $\boldsymbol{\mu}_{t+1}$ yields a fixed-point iteration. Setting $\alpha = 2$ recovers the mean vector, while $\alpha = 1$ gives the Weiszfeld iteration. Of course, strictly speaking, one must also prove convergence and uniquenessthe details are quite involved and omitted here.

Additionally, a less commonly used form of the high-dimensional median replaces the Euclidean norm with the L1 norm:

$$\operatorname*{argmin}_{\boldsymbol{\mu}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|_1 \tag{12}$$

This is called the "coordinate-wise median", because it essentially computes the univariate median independently for each coordinate. It is computationally simpler, but due to the lack of clear geometric interpretation, its applications are relatively limited.

5 Summary

This article briefly summarizes the concept and properties of the median, as well as its generalizations to higher-dimensional spaces.