MATHEMATICS

The Median Minimizes the Sum of Absolute Deviations (The L_1 Norm)

Asked 7 years, 9 months ago Active 7 months ago Viewed 49k times

92

Suppose we have a set S of real numbers. Show that

$$\sum_{s \in S} |s - x|$$

is minimal if x is equal to the median.

76

This is a sample exam question of one of the exams that I need to take and I don't know how to proceed.

optimization convex-optimization absolute-value median

edited Apr 6 at 10:45
Rodrigo de Azevedo 14.2k 4 23 69

asked Feb 25 '12 at 16:48
hattenn 1,269 1 10 11

18 Please replace *THE median* by *ANY median*. – Did Feb 25 '12 at 18:47

8 Answers

71

Introduction: The solution below is essentially the same as the solution given by Brian M. Scott, but it will take a lot longer. You are expected to assume that S is a finite set. with say k elements. Line them up in order, as $s_1 < s_2 < \cdots < s_k$.

The situation is a little different when k is odd than when k is even. In particular, if k is even there are (depending on the exact definition of median) many medians. We tell the story first for k odd.
Recall that $|x - s_i|$ is the **distance** between x and s_i, so we are trying to minimize the sum of the distances. For example, we have k people who live at various points on the x-axis. We want to find the point(s) x such that the **sum** of the travel distances of the k people to x is a minimum.

The story: Imagine that the s_i are points on the x-axis. For clarity, take $k = 7$. Start from well to the left of all the s_i, and take a tiny step, say of length ϵ, to the right. Then you have gotten ϵ closer to every one of the s_i, so the sum of the distances has decreased by 7ϵ.

Keep taking tiny steps to the right, each time getting a decrease of 7ϵ. This continues until you hit s_1. If you now take a tiny step to the right, then your distance from s_1 *increases* by ϵ, and your distance from each of the remaining s_i decreases by ϵ. What has happened to the sum of the distances? There is a decrease of 6ϵ, and an increase of ϵ, for a net decrease of 5ϵ in the sum.

This continues until you hit s_2. Now, when you take a tiny step to the right, your distance from each of s_1 and s_2 increases by ϵ, and your distance from each of the five others decreases by ϵ, for a
net decrease of 3ϵ.

This continues until you hit s_3. The next tiny step gives an increase of 3ϵ, and a decrease of 4ϵ, for a net decrease of ϵ.

This continues until you hit s_4. The next little step brings a total increase of 4ϵ, and a total decrease of 3ϵ, for an *increase* of ϵ. Things get even worse when you travel further to the right. So the minimum sum of distances is reached at s_4, the median.

The situation is quite similar if k is even, say $k = 6$. As you travel to the right, there is a net decrease at every step, until you hit s_3. When you are between s_3 and s_4, a tiny step of ϵ increases your distance from each of s_1, s_2, and s_3 by ϵ. But it decreases your distance from each of the three others, for no net gain. Thus any x in the interval from s_3 to s_4, including the endpoints, minimizes the sum of the distances. In the even case, I prefer to say that **any** point between the two "middle" points is **a** median. So the conclusion is that the points that minimize the sum are the medians. But some people prefer to define **the** median in the even case to be the average of the two "middle" points. Then the median does minimize the sum of the distances, but some other points also do.

edited Feb 25 '12 at 22:32
Michael Hardy
225k 24 212 495

answered Feb 25 '12 at 20:37
André Nicolas
463k 38 451 847

54

We're basically after:

$$\arg\min_{x} \sum_{i=1}^{N} \left| s_i - x \right|$$

One should notice that $\frac{\mathrm{d}\left|x\right|}{\mathrm{d}x} = \operatorname{sign}(x)$ (Being more rigorous would say it is a Sub Gradient of the non smooth L_1 Norm function).

Hence, deriving the sum above yields $\sum_{i=1}^{N} \operatorname{sign}\left(s_i - x\right)$.

This equals to zero only when the number of positive items equals the number of negative which happens when $x = \operatorname{median}\left\{s_1, s_2, \cdots, s_N\right\}$.

One should notice that the `median` of a discrete group is not uniquely defined.
Moreover, it is not necessarily an item within the group.

edited Jul 31 '18 at 5:49

answered Nov 16 '14 at 16:45
Royi
4,708 1 28 56

4 Using derivatives here is overkill; the problem can be done by more elementary methods. – Michael Hardy Oct 28 '16 at 17:06

9 @MichaelHardy actually among all answers I find this one to be the simplest, rather than the "walk to the left and then to the right of the real line". – gented Feb 21 at 23:21

To the point. :) – dksahuji Apr 12 at 8:26

The derivative of |x| is x/|x| as proven here : math.stackexchange.com/questions/83861/... – clyton dantis Apr 15 at 3:07

@clytondantis, The function $\frac{x}{|x|}$ is one of the definitions of $\operatorname{sign}(\cdot)$. – Royi Apr 15 at 4:04

35

Suppose that the set S has n elements, $s_1 < s_2 < \cdots < s_n$. If $x < s_1$, then

$$f(x) = \sum_{s \in S} |s - x| = \sum_{s \in S} (s - x) = \sum_{k=1}^{n} (s_k - x) \ .$$

As x increases, each term of (1) decreases until x reaches s_1, therefore $f(s_1) < f(x)$ for all $x < s_1$.

Now suppose that $s_k \le x \le x + d \le s_{k+1}$. Then

$$\begin{aligned} f(x+d) &= \sum_{i=1}^{k} \Big(x + d - s_i\Big) + \sum_{i=k+1}^{n} \Big(s_i - (x+d)\Big) \\ &= dk + \sum_{i=1}^{k} (x - s_i) - d(n-k) + \sum_{i=k+1}^{n} (s_i - x) \\ &= d(2k-n) + \sum_{i=1}^{k} (x - s_i) + \sum_{i=k+1}^{n} (s_i - x) \\ &= d(2k-n) + f(x) \ , \end{aligned}$$

so $f(x+d) - f(x) = d(2k-n)$. This is negative if $2k < n$, zero if $2k = n$, and positive if $2k > n$. Thus, on the interval $[s_k, s_{k+1}]$

$$f(x) \text{ is } \begin{cases} \text{decreasing}, & \text{if } 2k < n \\ \text{constant}, & \text{if } 2k = n \\ \text{increasing}, & \text{if } 2k > n \ . \end{cases}$$

From here it shouldn't be too hard to show that $f(x)$ is minimal when x is the median of S.

edited Nov 21 '18 at 12:01
pgmank
103 4

answered Feb 25 '12 at 17:22
Brian M. Scott
472k 42 558 965

3 But a small typo. In "As x increases, each term of (1) decreases until x reaches x_1, so" You intended to say s_1 instead of x_1. – Neo M Hacker Sep 28 '17 at 3:23

13

You want the median of n numbers. Say x is bigger than 12 of them and smaller than 8 of them (so $n = 20$). If x increases, it's getting closer to 8 of the numbers and farther from 12 of them, so the sum of the distances gets greater. And if x decreases, it's getting closer to 12 of them and farther from 8 of them, so the sum of the distances gets smaller.

A similar thing happens if x is smaller than more of the n numbers than x is bigger than.

But if x is smaller than 10 of them and bigger than 10 of them, then when x moves, it's getting farther from 10 of them and closer to just as many of them, so the sum of the distances is not changing.

So the sum of the distances is smallest when the number of data points less than x is the same as the number of data points bigger than x.

edited Apr 12 '15 at 6:45
Lord_Farin
16.4k 6 40 111

answered Feb 25 '12 at 22:37
Michael Hardy
225k 24 212 495

9

Starting with

$$f(x) = \sum_{i=1}^{n} |s_i - x|$$

Assume we rearranged our terms such that $s_1 < s_2 < \cdots < s_n$

We first proceed by making the following observation

$$\sum_{i=1}^{n} |s_i - x| = \sum_{i=2}^{n-1} |s_i - x| + (s_n - s_1) \quad \text{when} \quad x \in [s_1, s_n]$$

Now suppose that n is odd, then by applying the above identity repeatedly we get

$$f(x) = \sum_{i=1}^{n} |s_i - x| = |s_{\frac{n+1}{2}} - x| + (s_n - s_1) + (s_{n-1} - s_2) + \cdots + (s_{\frac{n+3}{2}} - s_{\frac{n-1}{2}})$$

or in other words

$$f(x) = |s_{\frac{n+1}{2}} - x| + \text{constant}$$

This is just the absolute value function with its vertex being at $(s_{\frac{n+1}{2}}, \text{constant})$, the minimum of the absolute value function occurs at its vertex, therefore $s_{\frac{n+1}{2}}$(median) minmizes $f(x)$.

Now suppose n is even, again by using our identity, we have

$$f(x) = \sum_{i=1}^{n} |s_i - x| = |s_{\frac{n}{2}} - x| + |s_{\frac{n+2}{2}} - x| + \text{constant}$$

Where the minimum occurs at $f'(x) = 0$(or when not defined), therefore by differentiating and setting $f'(x)$ to zero we get

$$\frac{|s_{\frac{n}{2}} - x|}{s_{\frac{n}{2}} - x} + \frac{|s_{\frac{n+2}{2}} - x|}{s_{\frac{n+2}{2}} - x} = 0$$

Observe that $s := \dfrac{s_{\frac{n+2}{2}} + s_{\frac{n}{2}}}{2}$(median) satisfies the above equation, since s is halfway between $s_{\frac{n}{2}}$ and $s_{\frac{n+2}{2}}$

$$s_{\frac{n}{2}} - s = -(s_{\frac{n+2}{2}} - s)$$

that is by setting $x = s$ we get

$$\frac{|s_{\frac{n}{2}} - s|}{s_{\frac{n}{2}} - s} + \frac{|s_{\frac{n}{2}} - s|}{-(s_{\frac{n}{2}} - s)} = 0$$

Therefore s is a minimum.

edited Mar 25 '17 at 1:16 Michael Hardy 225k 24 212 495

answered Jul 17 '16 at 22:01 Omar Nagib 1,038 7 16

I think some theory about minimum being where $f'(x) = 0$ for non differentiable functions is needed here – Guerlando OCs Aug 27 '18 at 0:56

5

Consider two x_i's x_1 and x_2,

For $x_1 \le a \le x_2$, $\sum_{i=1}^{2} |x_i - a| = |x_1 - a| + |x_2 - a| = a - x_1 + x_2 - a = x_2 - x_1$

For $a < x_1$, $\sum_{i=1}^{2} |x_i - a| = x_1 - a + x_2 - a = x_1 + x_2 - 2a > x_1 + x_2 - 2x_1 = x_2 - x_1$

For $a > x_2$, $\sum_{i=1}^{2} |x_i - a| = -x_1 + a - x_2 + a = -x_1 - x_2 + 2a > -x_1 - x_2 + 2x_2 = x_2 - x_1$

$\implies$ for any two x_i's the sum of the absolute values of the deviations is minimum when $x_1 \le a \le x_2$ or $a \in [x_1, x_2]$.

When n is odd,

$$\sum_{i=1}^{n} |x_i - a| = |x_1 - a| + |x_2 - a| + \cdots + \left|x_{\frac{n-1}{2}} - a\right| + \left|x_{\frac{n+1}{2}} - a\right| + \left|x_{\frac{n+3}{2}} - a\right| + \cdots + \left|x_{n-1} - a\right| + |x_n - a|$$

consider the intervals $[x_1, x_n], [x_2, x_{n-1}], [x_3, x_{n-2}], \ldots, \left[x_{\frac{n-1}{2}}, x_{\frac{n+3}{2}}\right]$. If a is a member of all these intervals. i.e, $\left[x_{\frac{n-1}{2}}, x_{\frac{n+3}{2}}\right]$,

using the above theorem, we can say that all the terms in the sum except $\left|x_{\frac{n+1}{2}} - a\right|$ are minimized. So

$$\sum_{i=1}^{n} |x_i - a| = (x_n - x_1) + (x_{n-1} - x_2) + (x_{n-2} - x_3) + \cdots + \left(x_{\frac{n+3}{2}} - x_{\frac{n-1}{2}}\right) + \left|x_{\frac{n+1}{2}} - a\right| = \left|x_{\frac{n+1}{2}} - a\right| + \text{costant}$$

Now since the derivative of modulus function is signum function, $f'(a) = \operatorname{sgn}\left(x_{\frac{n+1}{2}} - a\right) = 0$ for $a = x_{\frac{n+1}{2}} = \text{Median}$

$\implies$ When n is odd,the median minimizes the sum of absolute values of the deviations.

When n is even,

$$\sum_{i=1}^{n} |x_i - a| = |x_1 - a| + |x_2 - a| + \cdots + |x_{\frac{n}{2}} - a| + |x_{\frac{n}{2}+1} - a| + \cdots + |x_{n-1} - a| + |x_n - a|$$

If a is a member of all the intervals $[x_1, x_n], [x_2, x_{n-1}], [x_3, x_{n-2}], \ldots, \left[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}\right]$, i.e, $a \in \left[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}\right]$,

$$\sum_{i=1}^{n} |x_i - a| = (x_n - x_1) + (x_{n-1} - x_2) + (x_{n-2} - x_3) + \cdots + \left(x_{\frac{n}{2}+1} - x_{\frac{n}{2}}\right)$$

$\implies$ When n is even, any number in the interval $[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}]$, i.e, including the median, minimizes the sum of absolute values of the deviations. For example consider the series:2, 4, 5, 10, median, $M = 4.5$.

$$\sum_{i=1}^{4} |x_i - M| = 2.5 + 0.5 + 0.5 + 5.5 = 9$$

If you take any other value in the interval $\left[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}\right] = [4, 5]$, say 4.1

$$\sum_{i=1}^{4} |x_i - 4.1| = 2.1 + 0.1 + 0.9 + 5.9 = 9$$

For any value outside the interval $\left[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}\right] = [4, 5]$, say 5.2

$$\sum_{i=1}^{4} |x_i - 5.2| = 3.2 + 1.2 + 0.2 + 4.8 = 9.4$$

4

Consider two real numbers $a < b$. Then the objective becomes

$$dist(a, b) = |x - a| + |x - b|$$

This expression is minimum when $a \le x \le b$. It can be proved by calculating the objective on 3 cases ($x < a, a \le x \le b, x > b$).

Now consider the general case where S has n elements. Sort them in increasing order as $S_1, S_2, \ldots, S_n$.

Pair the smallest and the largest numbers. As explained above, $dist(S_1, S_n)$ is minimum when $S_1 \le x \le S_n$. Remove these two elements from the list and continue this procedure until there is at most one element left in the set.

If there is an element S_i left, then $x = S_i$ minimizes $dist(x - S_i)$. It also lies between all the pairs.

In the case of even elements, finally the sequence will be empty. As in the case above, median lies between all the pairs.

1

Suppose S is finite (with cardinal s), without repetitions, and ordered. Then the sum of absolute values is continuous (sum of continuous functions), and piecewise linear (hence differentiable), with left-most slope $-s$. By induction, the slope increases by 2 for each interval from left to right, with right-most slope $+s$. Hence the piece-wise slope first reaches either -1 or 0 at index $\left\lfloor\frac{s+1}{2}\right\rfloor$, and 0 or $+1$ at index $\left\lceil\frac{s+1}{2}\right\rceil$.

Hence the function attains its minima in the interval $\left[\left\lfloor\frac{s+1}{2}\right\rfloor, \left\lceil\frac{s+1}{2}\right\rceil\right]$, which reduces to a singleton when s is odd.

The notion of median for continuous functions is detailed in Sunny Garlang Noah, The Median of a Continuous Function, Real Anal. Exchange, 2007

edited Dec 1 '15 at 13:16

answered Dec 1 '15 at 11:42

Laurent Duval

5,631 1 13 42