DSC 40A - Discussion 01

January 14, 2020

1 Summation Notation

You can often verify for yourself if something is true about summation notation by "expanding" the summation symbol and seeing if the property holds. For instance, suppose we want to see if it is true that

$$\sum_{i=1}^{n} c \cdot x_i = c \sum_{i=1}^{n} x_i$$

We start by "expanding" $\sum_{i=1}^{n} c \cdot x_i$:

$$\sum_{i=1}^{n} c \cdot x_i = cx_1 + cx_2 + cx_3 + \ldots + cx_n$$

Now we see that the c can be factored out:

$$\begin{aligned} &= c(x_1 + x_2 + x_3 + \ldots + x_n) \\ &= c \sum_{i=1}^{n} x_i. \end{aligned}$$

This is a simple proof that the property is true. On the other hand, we can prove that a property doesn't hold in the same way: by expanding both sides and showing that they are not equal.

Problem 1.

Show that $\sum_{i=1}^{n}(x_i + y_i) = \left(\sum_{i=1}^{n} x_i\right) + \left(\sum_{i=1}^{n} y_i\right)$.

Solution:

$$\begin{aligned} \sum_{i=1}^{n}(x_i + y_i) &= (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_n + y_n) \\ &= (x_1 + x_2 + \ldots x_n) + (y_1 + y_2 + \ldots y_n) \\ &= \left(\sum_{i=1}^{n} x_i\right) + \left(\sum_{i=1}^{n} y_i\right) \end{aligned}$$

2 Chaining Inequalities

Suppose we have collected a bunch of numbers, $y_1, \ldots, y_n$. Let's assume, too, that these numbers are in sorted order, so that $y_1 \leq y_2 \leq \ldots \leq y_n$.

The *midpoint* of $y_1, \ldots, y_n$ is the average of the smallest and largest number:

$$\text{midpoint} = \frac{y_1 + y_n}{2}.$$

Intuitively, the midpoint is at most y_n and is at least y_1; it lies somewhere in the middle of these two numbers. We can easily prove this with a *chain* of inequalities.

First, we show that the midpoint is at most y_n. We start with the definition:

$$\text{midpoint} = \frac{y_1 + y_n}{2}$$

We can do anything to the right hand side that makes it bigger, keeping in mind that we're trying to get it to look like y_n. Right now there is y_1 hanging out; can we simply change it to a y_n? Yes! Remember that $y_n \geq y_1$, so this would make the right hand side bigger. Therefore, we have to write $\leq$:

$$\leq \frac{y_n + y_n}{2}$$

We can simplify this:

$$= \frac{2y_n}{2}$$

Notice that we wrote $=$ on the last line, not $\leq$. This is because the line is indeed equal to the one before it.

$$= y_n$$

We have made a chain of inequalities and equalities; this one looks like $=, \leq, =, =$. Since $\leq$ is the "weakest link" in the chain, the strongest statement we can make is that the midpoint is $\leq y_n$, but this is what we wanted to say.

Problem 2.

Prove that the midpoint is $\geq y_1$.

Solution:

$$\begin{aligned} y_1 &\leq y_n \\ y_1 + y_1 &\leq y_n + y_1 \\ \frac{2y_1}{2} &\leq \frac{y_n + y_1}{2} \\ y_1 &\leq \mathit{midpoint} \end{aligned}$$

Problem 3.

Suppose $y_1, \ldots, y_n$ are all positive numbers. The *geometric mean* of $y_1, \ldots, y_n$ is defined to be:

$$\left(y_1 \cdot y_2 \cdots y_n\right)^{1/n}.$$

Prove that the geometric mean is less than or equal to y_n and greater than or equal to y_1 using a chain of inequalities.

Solution: Assuming the numbers are ordered, let's first show that the geometric mean $\geq y_1$. We know that the below inequalities hold by definition .

$$\begin{aligned} y_1 &\leq y_1 \\ y_1 &\leq y_2 \\ y_1 &\leq y_3 \\ y_1 &\leq y_4 \\ \ldots & \\ y_1 &\leq y_n \end{aligned}$$

Since $y_i > 0 \ \forall i$, we can multiply the n inequalities to get

$$y_1 y_1 \dots y_1 \leq y_1 y_2 \dots y_n$$

So,

$$\begin{aligned} y_1^n &\leq y_1 y_2 \dots y_n \\ (y_1^n)^{1/n} &\leq (y_1 y_2 \dots y_n)^{1/n} \\ y_1 &\leq \textit{geometric mean} \end{aligned}$$

You can similarly show that *geometric mean* $\leq y_n$ by using the fact that $y_i \leq y_n$ for $i = 1, 2, \dots n$.

Problem 4.

The *standard deviation* of $y_1, \dots, y_n$ is

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2}$$

where $\mu = \text{Mean}(y_1, \dots, y_n)$. Suppose δ is the biggest distance between μ and any data point. Is it true that the standard deviation is $\leq \delta$?

Solution: Define the distance between μ and y_i to be $\delta_i = |\mu - y_i|$. δ is defined as the maximum of these distances. Therefore, we know that $\delta_i \leq \delta$ for $i = 1, 2, \dots n$. Also, because distances can only be non-negative, we can square both sides (or multiply the inequality with itself) to get $\delta_i^2 \leq \delta^2$ for $i = 1, 2, \dots n$. Hence,

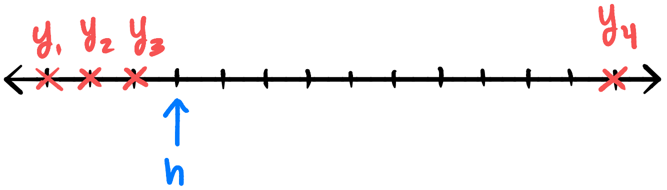
$$\begin{aligned} \delta_1^2 &\leq \delta^2 \\ \delta_2^2 &\leq \delta^2 \\ \dots & \\ \delta_n^2 &\leq \delta^2 \end{aligned}$$

Then, summing all equations, we get

$$\begin{aligned} \sum_{i=1}^{n} \delta_i^2 &\leq n\delta^2 \\ \sum_{i=1}^{n} |\mu - y_i|^2 &\leq n\delta^2 \\ \sum_{i=1}^{n} (y_i - \mu)^2 &\leq n\delta^2 \\ \frac{1}{n}\sum_{i=1}^{n} (y_i - \mu)^2 &\leq \delta^2 \\ \left|\sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \mu)^2}\right| &\leq \delta \\ \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \mu)^2} &\leq \delta \\ \textit{standard deviation} &\leq \delta \end{aligned}$$

3 The Mean and Median

In class we saw that the mean's sensitivity to outliers is due to its role as a minimizer of the mean squared error. Now we'll make this more clear. Suppose we have the data $y_1, \ldots, y_n$ drawn below:



For this problem you do not need to know the exact position of the data points, but, if you like, you can assume that the space between each tick mark is one unit and that $y_1 = 1$, $y_2 = 2$, $y_3 = 3$, $y_4 = 14$ and $h = 4$.

Suppose we start out with the prediction h as shown above. There is a tug-of-war going on in the picture above: y_1, y_2, y_3 want h to move closer to them, while y_4 wants h to move to the right to be closer to it. Who wins depends on the loss function that is used.

Problem 5.

Suppose that the absolute loss is used. Suppose that h is moved one unit to the left. This increases the error for y_4, but decreases the error for y_1, y_2, y_3. Show that the decrease in $|y_1 - h| + |y_2 - h| + y_3 - h|$ makes up for the increase in $|y_4 - h|$ so that moving h to the left decreases the overall error.

Solution: Recall that absolute loss is defined as: $L_A(h) = \sum_{i=1}^{n} |h - y_i|$. We want to show that $L_A(h) \geq L_A(h')$, where h' is the new position of h.

$L_A(h) = L_A(4) = |4 - 1| + |4 - 2| + |4 - 3| + |4 - 14| = 3 + 2 + 1 + 10 = 16$
$L_A(h') = L_A(3) = |3 - 1| + |3 - 2| + |3 - 3| + |3 - 14| = 2 + 1 + 0 + 11 = 14.$

As seen above, moving h to the left decreases the overall absolute error.

Problem 6.

Now suppose that the square loss is used. Again suppose that h is moved one unit to the left. Show that the increase in $(y_4 - h)^2$ is larger than the decrease in $(y_1 - h)^2 + (y_2 - h)^2 + (y_3 - h)^2$, so that moving h to the left increases the overall squared error.

Solution: Recall that square loss is defined as: $L_S(h) = \sum_{i=1}^{n} (h - y_i)^2$. We want to show that $L_S(h) \leq L_S(h')$, where h' is the new position of h.

$L_S(h) = L_S(4) = (4 - 1)^2 + (4 - 2)^2 + (4 - 3)^2 + (4 - 14)^2 = 9 + 4 + 1 + 100 = 114$
$L_S(h') = L_S(3) = (3 - 1)^2 + (3 - 2^2 + (3 - 3)^2 + (3 - 14)^2 = 4 + 1 + 0 + 121 = 126.$

As seen above, moving h to the left increases the overall square error.

Informally, moving h to the left always increases the loss associated with y_4, whether the absolute loss or square loss is used. That is, y_4 always protests against moving h to the left. This protest isn't strong enough in the case of the absolute loss, but if the square loss is used, y_4's voice is amplified, and it wins.