Kick Start 2022 - Round D

Analysis: Image Labeler

First, we need to understand how to obtain the median of a set. A median is the value separating the higher and lower halves of a set, so a set of length k will have its median at element $k/2$. For sets with an even number of elements, the median is the average of the two middle elements.

A brute force approach to this problem is described in the example case, where we try each possible combination of assigning regions to categories, and taking the combination with the largest sum of medians.

Time complexity: $O(2^{\mathbf{M}})$ since we need to test each combination of either assigning a region or not to each category. We can reduce the time complexity even further through dynamic programming to avoid recalculating assignmented categories, but there is an even better approach.

A more efficient approach is to maximize the sum by ensuring that the largest regions each form their separate category. One way to do this is with a greedy approach, where we remove the single largest region (say $\mathbf{A_k}$) and assign it to a new category. Then the median of this category is equal to the size of that region. Note that this is the largest median that we can obtain, because all other regions $\mathbf{A_1}, \cdots, \mathbf{A_N} \le \mathbf{A_k}$. If we continue this process for the next $\mathbf{N}-1$ categories, we will have the largest possible medians for all $\mathbf{N}$ regions. The largest regions can be identified by sorting $\mathbf{A}$ in descending order.

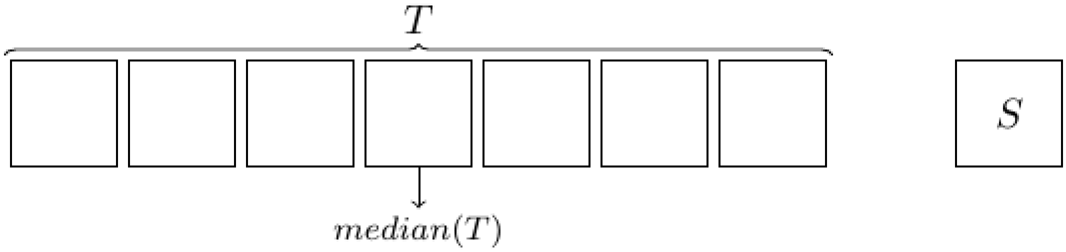
The problem is that we may run out of categories. Precisely, we still have $\mathbf{N}-\mathbf{M}$ regions that we need to assign to some categories. Since we know that for each of these remaining regions that they are smaller or equal to the ones already assigned to categories, then assigning a new region to an existing category may reduce the median of that category. We can decide to decrease the median of only one category — the one that already has the smallest median. Our result (the sum of medians) is then the sum of the largest $\mathbf{M}-1$ regions plus the median of all other regions.

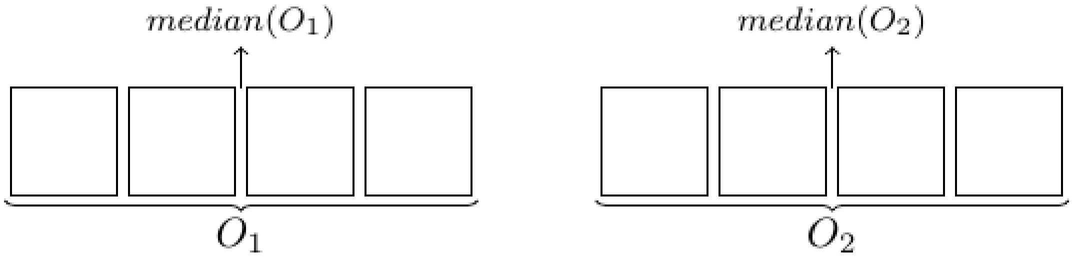
Time complexity: $O(\mathbf{N}\log\mathbf{N}+\mathbf{M})$ since we need to sort $\mathbf{N}$ regions and sum up the top $\mathbf{M}-1$. Calculating the median of the final category is only $O(1)$ because the regions are already in sorted order. We can even solve this problem in linear time, but it was not required for this problem. We only need to find top $\mathbf{M}$-1 regions (and we can do it in $O(\mathbf{N}+\mathbf{M})$ time), and then we can find the median of the remaining regions in linear time as well.

Proof

We can formally prove that the greedy algorithm produces an optimal result.

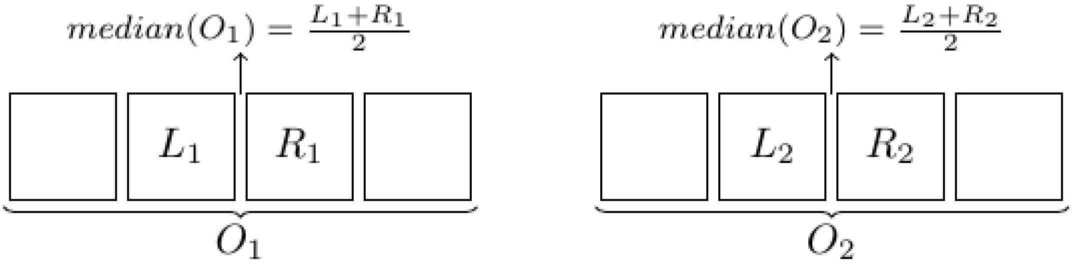
We will prove that given any two arbitrary categories, the sum of medians does not decrease if we reassign the regions within these categories so that the largest region forms one category, and all the other regions forms the other category. Let $\{O_1, O_2\}$ denote these two arbitrary categories. Let S denotes the number of participants in the region with the largest number of participants within these categories and $T = O_1 \cup O_2 \setminus \{S\}$ (S and T are two categories produced by the greedy algorithm). We want to show that $median(O_1) + median(O_2) \le S + median(T)$. Without loss of generality, let us assume that $median(O_1) \le median(O_2)$. In the pictures below, let us assume that the regions within a category are sorted in the non-decreasing order.





In T, there are at least $\left\lceil \frac{|O_1|}{2} \right\rceil + \left\lceil \frac{|O_2|}{2} \right\rceil - 1$ regions larger than or equal to $median(O_1)$ — the regions larger than or equal to the medians of O_1 and O_2, but we need to remove S (hence -1). Now we have to consider two cases:

1. Either in T there are at least $\left\lceil \frac{|O_1|}{2} \right\rceil + \left\lceil \frac{|O_2|}{2} \right\rceil$ regions larger than or equal to $median(O_1)$, which implies $median(T) \geq median(O_1)$, so we can say that $median(O_1) + median(O_2) \leq S + median(T)$, what we wanted to prove.
2. There are exactly $\left\lceil \frac{|O_1|}{2} \right\rceil + \left\lceil \frac{|O_2|}{2} \right\rceil - 1$ regions larger than or equal to $median(O_1)$. Now we will once again consider various cases depending on the parity of $|O_1|$ and $|O_2|$:
    1. $|O_1|$ and $|O_2|$ are even. Let us denote L_1 and R_1 as two "middle" regions of O_1 and L_2 and R_2 as two "middle" regions of O_2.

       
       $median(O_1) = \frac{L_1+R_1}{2}$
       $median(O_2) = \frac{L_2+R_2}{2}$
       L_1
       R_1
       L_2
       R_2
       O_1
       O_2

       Then we can use the similar argument as before and say that in T there are at least $\left\lceil \frac{|O_1|}{2} \right\rceil + \left\lceil \frac{|O_2|}{2} \right\rceil$ regions larger than or equal to $\max(L_1, L_2)$, so $median(T) \geq \min(\max(L_1, L_2), \min(R_1, R_2))$, so $S + median(T) \geq \frac{L_1+R_1}{2} + \frac{L_2+R_2}{2} = median(O_1) + median(O_2)$, since $S \geq \max(R_1, R_2)$.
    2. $|O_1|$ and $|O_2|$ are odd. Then we can see that there are $\left\lceil \frac{|O_1|}{2} \right\rceil + \left\lceil \frac{|O_2|}{2} \right\rceil$ regions with more participants than $median(O_1)$, which is the case that we already described above.
    3. One of $|O_1|$ and $|O_2|$ is even and the other one is odd. The argument here is similar to the arguments above, so we will leave it as an exercise to a reader.

equally well for *everyone, everywhere*, you can download it here.