Gini Robustness

Chris Coffee

May 11, 2023

Definition of Gini Robustness

Gini Robustness measures how a partitioning's Gini Coefficient would be affected by new data. Two scenarios are considered for clusters comprising S discrete elements:

1. What would happen to the Gini Coefficient if we added a new cluster of S^2 elements?
2. What would happen to the Gini Coefficient if we added $\sqrt{S}$ singleton clusters?

The two scenarios above are designed to test the robustness of the Gini Coefficient, which must first be normalized with respect to S. The normalized Gini Coefficient is defined as the Gini Coefficient divided by the maximum possible Gini Coefficient given S elements. A normalized Gini Coefficient of 1 indicates that the cluster sizes are maximally uneven for the given number of elements. Cluster sizes can be said to be maximally uneven if they comprise $\left\lfloor \sqrt{S+\frac{1}{4}}-\frac{1}{2} \right\rfloor$ singletons and one cluster of size $S-\left\lfloor \sqrt{S+\frac{1}{4}}-\frac{1}{2} \right\rfloor$ (although other configurations are sometimes equivalent).

The two values chosen for the number of elements in the new cluster are S^2 and $\sqrt{S}$ because they are the smallest values that could maximize the Gini Coefficient.

To calculate Gini Robustness, let x_k be the size of the k^{th} cluster in a partition of S elements into K clusters. Then (unnormalized) Gini Robustness is defined as:

$$\text{Robustness}(\mathbf{x}_K) = \left(1-\widehat{\text{Gini}}\left((\mathbf{x}_K, S^2)\right)\right)\left(1-\widehat{\text{Gini}}\left(\left(\mathbf{1}_{\sqrt{S}}, \mathbf{x}_K\right)\right)\right)$$

where $\text{Gini}(\mathbf{x}_K)$ is the Gini Coefficient of the (sorted) cluster sizes:

$$\text{Gini}(\mathbf{x}_K) = \frac{\sum_{k=1}^{K}\sum_{l=1}^{K} x_k x_l |k-l|}{2K\sum_{k=1}^{K} x_k}$$

and $\widehat{\text{Gini}}(\mathbf{x}_K)$ is the Gini Coefficient of the cluster sizes divided by the largest possible Gini Coefficient for S elements, which is $\text{Gini}\left(\left(\mathbf{1}_{\left\lfloor \sqrt{S+\frac{1}{4}}-\frac{1}{2} \right\rfloor}, S-\left\lfloor \sqrt{S+\frac{1}{4}}-\frac{1}{2} \right\rfloor\right)\right)$.

Finally, the result is divided by the maximum possible result given S.

These definitions are convenient in that singletons and single clusters can be considered minimally Gini robust.

Values range from 0 to 1, with higher values indicating higher robustness.

A perhaps more compelling way to think of Gini Robustness is as a measure of the (lack of) "explosiveness" of $\frac{K^{\text{new}}}{K}$ and $\frac{\mu^{\text{new}}}{\mu}$ where "new" denotes the new value after additional elements are discovered and $\mu = \frac{S}{K}$. One would generally prefer the addition of new data to not drastically change the existing number of clusters or the average cluster size. This is because we expect the data to be i.i.d. and thus these values to be stable.