Comparison of two data sets

Explain how to do the calculations

We have two kinds of dataset that includes a matrix of float numbers from 0 to about 1. in the other point of view. There is a label to find the best gain of the decision tree with these datasets.

The best way to find a better performance of them is by calculating entropy for each data set and then Comparison between the results.

- Step one: Dataset have [17500,1615] data. we have 17500 Row in the data sets, and it is necessary to find unique nods in each row.
- Step two: Each unique node will have entropy result.
- Step three: Now it is necessary to calculate row entropy by using all unique nodes entropy.

$$entropy_{A_i}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times entropy(D_j)$$

- Step four: I calculate entropy of label data to use in final formula.

$$entropy(D) = -\sum_{j=1}^{|C|} \Pr(c_j) \log_2 \Pr(c_j)$$

- Final step: We have to compare the output magnitude of the two-data gain.

$$gain(D, A_i) = entropy(D) - entropy_{A_i}(D)$$

Comparison Between Outputs of two dataset

The entropy(D) inside the report is 0.7763140378900846.

N	Dataset 1 (Entropy Row)	Dataset 1 (Gain)	Dataset 2 (Entropy Row)	Dataset 2 (Gain)
1	0.6550206845307067	0.1212933533593779	0.348028674963614	0.4282853629264706
2	0.604743300436142	0.1715707374539426	0.33893329750137463	0.43738074038870994
3	0.6003920007192063	0.1759220371708783	0.3392022554617249	0.4371117824283597
4	0.6003920007192065	0.17592203717087807	0.3389674907738424	0.43734654711624216
5	0.6550206845307066	0.12129335335937796	0.3264237397864577	0.44989029810362685
6	0.6646154975248125	0.11169854036527205	0.3203047605743426	0.456009277315742
7	0.6003920007192063	0.1759220371708783	0.30818452697445664	0.46812951091562793
8	0.6003920007192064	0.17592203717087818	0.3117403442533627	0.46457369363672185
9	0.6003920007192064	0.17592203717087818	0.31255332813363845	0.4637607097564461
10	0.604743300436142	0.17157073745394258	0.3108762482429228	0.46543778964716176

This is the gain of 10 rows. As information shows that, during all second dataset, we have more amount gain. In the other hand, there is more Zero and One entropy amount in second data set which shows that our branch will be close in this unique nodes.

The following chart shows the amount of gain between rows one through ten.