# Entropy /Gini Gain

Tuesday, September 11, 2018 4:22 PM

`H(Y) = - ∑ (p(yj) * log2(p(yj)))`
In words, select an attribute and for each value check target attribute value ... so p(yj) is the fraction of patterns at Node N are in category yj - one for true in target value and one one for false

From <https://stackoverflow.com/questions/14363689/calculating-entropy-in-decision-tree-machine-learning>

**Gini** is to minimize misclassification
**Entropy** is for exploratory analysis

From <https://datascience.stackexchange.com/questions/10228/gini-impurity-vs-entropy>

Gini impurity and Information Gain Entropy are pretty much the same. And people do use the values interchangeably. Below are the formulae of both:

1. $Gini: Gini(E) = 1 - \sum_{j=1} p_j^2$ Gini:Gini(E)=1−∑j=1cpj2
2. $Entropy: H(E) = -\sum_{j=1} p_j \log p_j$

From <https://datascience.stackexchange.com/questions/10228/gini-impurity-vs-entropy>

**Tree models are suscpetible to overfit if the tree grows too long**

- Gini of a Node

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

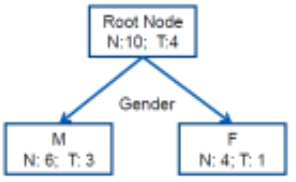
- Gini of Split Node is computed as Weighted Avg Gini of each Node at Split Node level

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

$n_i$ = number of records at child i,
n = Total number of records in parent node

- Gini Gain = Gini(t) – Gini(split)

| Cust_ID | Gender | Occupation | Age | Target |
|---|---|---|---|---|
| 1 | M | Sal | 22 | 1 |
| 2 | M | Sal | 22 | 0 |
| 3 | M | Self-Emp | 23 | 1 |
| 4 | M | Self-Emp | 23 | 0 |
| 5 | M | Self-Emp | 24 | 1 |
| 6 | M | Self-Emp | 24 | 0 |
| 7 | F | Sal | 25 | 1 |
| 8 | F | Sal | 25 | 0 |
| 9 | F | Sal | 26 | 0 |
| 10 | F | Self-Emp | 26 | 0 |




| Node | Gini Computation Formula | Gini Index |
|---|---|---|
| Overall | = 1 - ( (4/10)^2 + (6/10)^2 ) | 0.48 |
| Gender = M | = 1 - ( (3/6)^2 + (3/6)^2) | 0.50 |
| Gender = F | = 1 - ( (1/4)^2 + (3/4)^2) | 0.375 |
| Gender | = (6/10) * 0.5 + (4/10) * 0.375 | 0.45 |
| Gini Gain | = Gini (Overall) – Gini (Gender) | 0.03 |

Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure. We can calculate Gini as following:

- Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
- Calculate Gini for split using weighted Gini score of each node of that split

Entropy is the measure of impurity as given by (for binary class):

$$\text{Entropy} = -p \log_2 p - q \log_2 q$$

Here p and q is probability of success and failure respectively in that node. Entropy is zero when a node is homogeneous. It is maximum when a both the classes are present in a node at 50% – 50%. Lower entropy is desirable.