Disadvantages of entropy:

1. It requires a log of probability.
2. The log is computationally expensive
3. We have to calculate the entropy for each feature at each node
4. This becomes time-consuming.

Gini Impurity

The Gini Impurity of the random variable Y is given by:

- $GI(Y) = 1-\sum^{k}_{i=1}p(y_i)^2$

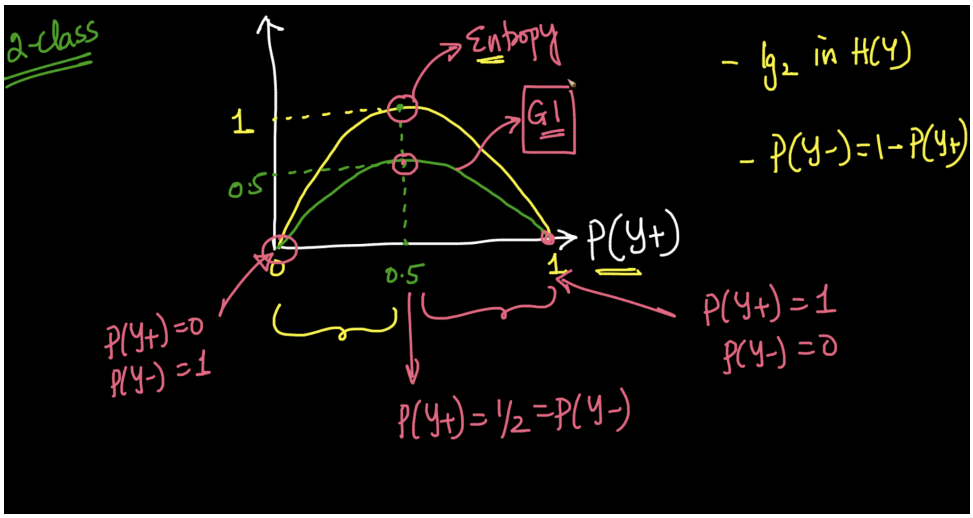
For a binary classification:

$GI(y) = 1-[p(y_i=1)^2+p(y_i=0)^2]$

What is the difference in behaviors of Gini-Impurity and entropy?

- Gini-Impurity is high when entropy is high
- It is low when the entropy is low

Let's plot the graph of Entropy as well as Gini Impurity:



Notice that,

1. When the nodes are pure i.e for
   - if $p(y_+) = 1$ and $p(y_-) = 0$, or
   - if $p(y_+) = 0$ and $p(y_-) = 1$

   The entropy and Gini-Impurity are zero

2. when the probability of $p(y_+) = 1/2$ and $p(y_-) = 1/2$
   the entropy and Gini- Impurity are the maximum

So, we can conclude that Gini-Impurity has the same behavior as entropy.

Information Gain and Gini - Reduction

When a node is split into multiple children, each child has a separate impurity

How can we compare these multiple impurity values obtained for each feature ?
- We combine these impurity values

Information Gain: Entropy of parent - Weighted average of entropy of children

Gini Reduction: G(Parent Node) - weighted average of gini impurity of children

Calculating Feature Importance

Feature importance is calculated based on Information Gain.
- Feature with more information gain is considered more important.

Handling numerical features

Typically, we will have a threshold and
- We compare each value of the feature with a threshold
- and split them based on the threshold.

How to choose the threshold?

1. Arrange values of the feature in increasing order
2. Set each value of the feature as a threshold
3. Next, we calculate the IG of that split.
4. The threshold giving the maximum IG is selected as IG obtained

Bias/Variance in Decision Trees

- Overfits when the Depth of the tree is too high as deeper trees can capture more complex relationships in the data, allowing the model for not only capture finer distinctions between classes or values but also capture noise/outliers
    - This leads DT to have low bias but high variance
- Underfits when the Depth of the tree is too low as the model oversimplify the data, resulting in poorer performance on both training and test data.
    - This leads DT to have high bias and low variance

Miscellaneous Info:

A decision Stump is a Decision Tree with depth = 1

A shallow tree has a small depth value

Deep tree is when the tree's depth is very large
Outliers: Outliers impact the Decision Tree when the depth is very large

Standardization: Standardization does not impact entropy, Gini Impurity, and information gain

Train Complexity: o(nlog(n) d)

Run Time Complexity: o(d)

- n: Nodes in the tree
- d: Depth of the tree

Feature Importance can be measured using DT

$$FI(f_i) = IG_1 \times \frac{n_1}{n} + IG_2 \times \frac{n_2}{n}$$

- Where IG_1 is the information gain from 1st split of feature f_i having n_1 samples
- And IG_2 is the information gain from the 2nd split of feature f_i having n_2 samples

- DT can be used for Regression task
    - use MSE for splitting nodes
    - $MSE = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2$
        - Where n is the sample for a node
        - $\bar{y}$ is the mean of all values in the node