Decision Tree Classification : How to assess the quality of split ?

- Classification Error made by each newly created region.

$$\boxed{\text{Error}(i|j,t_j) = 1 - \max_k p(k|R_i)}$$

where $p(k|R_i)$ is % training pts in R_i that are labeled class k.

Example :

	Class 1	Class 2	$\text{Error}(i\|j,t_j)$
R_1	0	6	$1-\max\{6/6, 0/6\} = 0$
R_2	5	8	$1-\max\{5/13, 8/13\} = 5/13$

We can now try to find predictor j and threshold t_j that minimizes the average classification error over 2 regions, weighted by the population of the regions:

$$\boxed{\min_{j,t_j}\left\{\frac{N_1}{N}\text{Error}(1|j,t_j) + \frac{N_2}{N}\text{Error}(2|j,t_j)\right\}}$$

where N_j is the number of training points inside region R_i.

- Gini Index : impurity of each created region.

$$\boxed{\text{Gini}(i|j,t_j) = 1 - \sum_k p(k|R_i)^2}$$

Example :

	Class 1	Class 2	$\text{Gini}(i\|j,t_j)$
R_1	0	6	$1-[(6/6)^2+(0/6)^2] = 0$
R_2	5	8	$1-[(5/13)^2+(8/13)^2] = 80/169$

We can now try to find predictor j and threshold t_j that minimizes the average Gini Index over 2 regions, weighted by the population of the regions.

$$\boxed{\min_{j,t_j}\left\{\frac{N_1}{N}\text{Gini}(1|j,t_j) + \frac{N_2}{N}\text{Gini}(2|j,t_j)\right\}}$$

- Entropy of the class distribution in each newly created region.

$$\boxed{\text{Entropy}(i|j,k) = -\sum_{k} p(k|R_i)\log_2 p(k|R_i)}$$

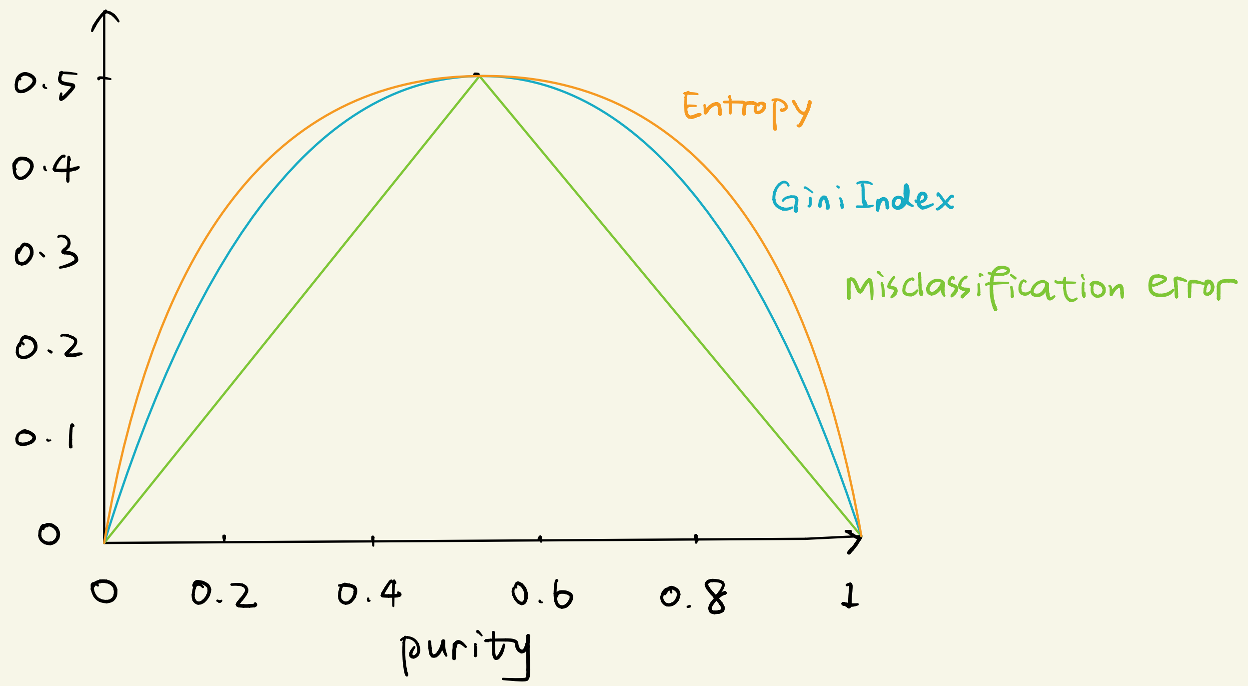
Example :

	Class 1	Class 2	$\text{Entropy}(i|j,t_j)$
R_1	0	6	$-(6/6\log_2 6/6 + 0/6\log_2 0/6) = 0$
R_2	5	8	$-(5/13\log_2 5/13 + 8/13\log_2 8/13) = 1.38$

We can now try to find predictor j and threshold t_j that minimizes the average Entropy over 2 regions, weighted by the population of the regions.

$$\boxed{\min_{j,t_j}\left\{\frac{N_1}{N}\text{Entropy}(1|j,t_j) + \frac{N_2}{N}\text{Entropy}(2|j,t_j)\right\}}$$

Comparison of Criteria :



Entropy penalizes impurity the most.