CO395 Group 57: Dongxiao Huang, Zheng Xun Phang, Yufeng Zhang

Implementation

How do we select the best attribute at each node? For all attributes, we compute the information gain if the dataset is split on that attribute:

$$\text{Gain(Attribute)} = \mathcal{I}(p,n) - \left[\frac{p_0+n_0}{p+n}\mathcal{I}(p_0,n_0) + \frac{p_1+n_1}{p+n}\mathcal{I}(p_1,n_1)\right]$$

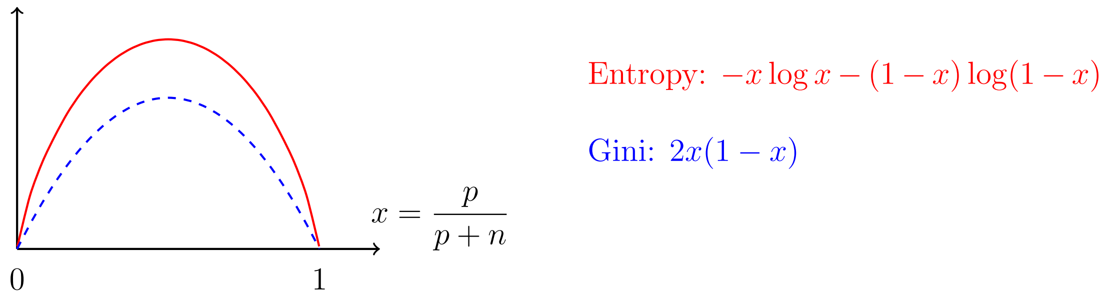
$$\begin{aligned} p_k &= \text{Number of positive examples with attribute} = k \\ n_k &= \text{Number of negative examples with attribute} = k \\ p = p_0 + p_1 &= \text{Number of positive examples before split} \\ n = n_0 + n_1 &= \text{Number of negative examples before split} \end{aligned}$$

There are two common ways to measure information.

Entropy: $$\mathcal{I}(p,n) = -\frac{p}{p+n}\log\frac{p}{p+n} - \frac{n}{p+n}\log\frac{n}{p+n} \qquad \text{if } p,n \neq 0$$
$$\mathcal{I}(p,0) = \mathcal{I}(0,n) = 0$$

Gini impurity: $$\mathcal{I}(p,n) = \frac{p}{p+n}\left(1-\frac{p}{p+n}\right) + \frac{n}{p+n}\left(1-\frac{n}{p+n}\right)$$

We used entropy since it's stated in the specification, but Gini impurity is faster to compute. Both metrics should give similar results since their graphs have a similar shape:



When comparing their graphs, their relative heights do not matter because minimizing a function is equivalent to minimizing any positive multiple of that function.

To evaluate our decision tree, we performed cross validation as follows:

1. Shuffle the dataset and split it into $K = 10$ parts
2. For each $k \in \{1, \cdots, K\}$ we train the decision tree on the dataset *excluding* part k and then test the decision tree on part k. During testing, the relevant cells in the confusion matrix are incremented.

Evaluation

Each cell of the confusion matrix is a total, not an average, over *all* folds of cross validation.

Predicted	Actual					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger						
Disgust						
Fear						
Happiness						
Sadness						
Surprise						

From the confusion matrix above, we can compute these summary statistics:

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Precision						
Recall						
F_1 score						

Miscellaneous

Noisy-Clean Datasets Question

The noisy dataset has lower performance.

Ambiguity Question

In case our 6 trees predict that an image depicts more than 1 emotion, we considered the following methods of selecting 1 emotion:

1. Pick the first emotion in alphabetical order
This is effectively selecting an emotion at random.

2. Disable each active unit in turn, and take a majority vote
Example:

Pruning Question