Exercise

Compare the version of decision trees used by the sklearn software with the theory seen in class.

Solution

In class, we discussed using ID3 algorithm as part of the training process to build Decision Trees. However, as per scikit-learn's document, scikit-learn uses an optimised version of the CART algorithm. ID3 uses Entropy and the greatest information gain to select and build nodes. CART uses Gini Impurity, which is a measure of the homogeneity of the nodes. scikit-learn splits the tress at a threshold where the sum of Gini Impurity Index is minimised across the groups being split with the said threshold. Gini Impurity Index is calculated as follows:

Suppose we have a categorical variable that takes values $C_i,\ i = 1, \dots, K$

and that the probability of category j arising is $\mathbb{P}(C_i) = p_i$, where $p = (p_1, \dots, p_K)$

The Gini index is defined as

$$G(p) = \sum_{i \neq j} p_i p_j = 1 - \sum_{i=1}^{K} p_i^2$$