Decision Trees: Measuring Impurity for Splits

As outlined in the learning steps, the Decision Tree algorithm needs a way to quantify how "good" a potential split is. The goal is to make splits that result in child nodes that are more "pure" (i.e., containing samples predominantly from a single class) than the parent node. This is achieved by measuring the **impurity** of a node. Two common measures of impurity used are the **Gini Index** and **Entropy**.

Gini Index (or Gini Impurity)

* **Definition:** The Gini Index measures the impurity or disorder of a set of items (e.g., the samples within a node). Specifically, it calculates the probability of misclassifying a randomly chosen element from the set if it were randomly labeled according to the distribution of labels in the subset.
* **Range:** The values of the Gini Index range from **0 to 1**.
  + **Gini = 0:** Represents **absolute purity**. All elements in the node belong to a single class. There is no impurity.
  + **Gini = 0.5 (for binary classification):** Represents the **maximum impurity**. The elements are equally distributed among the classes (e.g., 50% Class A, 50% Class B).
  + **Gini = 1 - (1/n) (for n classes):** Represents the maximum impurity when elements are equally distributed among n classes. (Note: The slide mentions a max value of 1, which is often used conceptually, but the theoretical max for equal distribution is slightly less unless n=2 where it's 0.5).
* **Calculation:** The Gini Index for a specific node (i) is calculated as 1 minus the sum of the squared probabilities of each class (k) present in that node.
* Gini(i) = 1 - Σ [ (p\_{i,k})² ] (sum over all classes k=1 to n)

Where:

* + p\_{i,k} is the proportion (or probability) of samples belonging to class k in node i. (Calculated as: Number of samples of class k in node i / Total number of samples in node i).
* **Use in Splitting:** When evaluating a potential split, the algorithm calculates a *weighted average* of the Gini Indices of the resulting child nodes. The feature and threshold combination that results in the **minimum weighted Gini Index** (i.e., the lowest impurity after the split) is chosen as the best split for the current node.

Entropy

Entropy is another common measure of impurity or disorder, originating from information theory.

* **Definition:** Entropy measures the level of uncertainty or "chaos" in a set of items. Higher entropy means more disorder or mixed classes; lower entropy means less disorder or more purity.
* **Range:**
  + **Entropy = 0:** Represents **absolute purity**. All elements in the node belong to a single class. There is zero uncertainty.
  + **Maximum Entropy:** The maximum value depends on the number of classes (n). It occurs when samples are equally distributed among all classes.
    - For 2 classes (binary): Max entropy = log₂(2) = 1
    - For 4 classes: Max entropy = log₂(4) = 2
    - For 8 classes: Max entropy = log₂(8) = 3
    - For n classes: Max entropy = log₂(n)
* **Calculation:** Entropy for a specific node (i) is calculated as the negative sum of the probability of each class (k) multiplied by the logarithm (usually base 2) of that probability.
* Entropy(i) = - Σ [ p\_{i,k} \* log₂(p\_{i,k}) ] (sum over all classes k=1 to n, where p\_{i,k} > 0)

Where:

* + p\_{i,k} is the proportion (or probability) of samples belonging to class k in node i.
  + The term 0 \* log₂(0) is defined as 0.
* **Use in Splitting (Information Gain):** While Entropy measures impurity, Decision Tree algorithms like ID3 and C4.5 typically use **Information Gain** as the splitting criterion.
  + **Information Gain = Entropy(parent) - Weighted Average Entropy(children)**
  + The algorithm chooses the feature and threshold combination that **maximizes the Information Gain**, which is equivalent to minimizing the resulting weighted average entropy of the child nodes.

**Gini vs. Entropy:**

* Both are common and often lead to similar trees.
* Gini Index is slightly faster to compute as it doesn't involve logarithmic calculations.
* Entropy (via Information Gain) might lead to slightly more balanced trees in some cases.
* The default criterion in scikit-learn's DecisionTreeClassifier is Gini impurity.

The choice between Gini and Entropy typically has a small impact on the final performance, but understanding both helps in comprehending how the tree decides where and how to split the data.