Questions and Answers for Decision Trees

Bingzhang Zhu

October 04, 2021

1 What other criteria are used to measure the quality of a split besides Entropy(ID3)?

Gini Impurity(CART).

Used by the CART (classification and regression tree) algorithm for classification trees, Gini impurity (named after Italian mathematician Corrado Gini) is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The Gini impurity can be computed by summing the probability p_i of an item with label i being chosen times the probability $\sum_{k\neq i} p_k = 1 - p_i$ of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

The Gini impurity is also an information theoretic measure and corresponds to Tsallis Entropy with deformation coefficient $q = 2$, which in physics is associated with the lack of information in out-of-equilibrium, non-extensive, dissipative and quantum systems. For the limit $q \to 1$ one recovers the usual Boltzmann-Gibbs or Shannon entropy. In this sense, the Gini impurity is but a variation of the usual entropy measure for decision trees.

To compute Gini impurity for a set of items with J classes, suppose $i \in \{1, 2, ..., J\}$, and let p_i be the fraction of items labeled with class i in the set.

$$I_G(p) = \sum_{i=1}^{J} \left(p_i \sum_{k\neq i} p_k \right) = \sum_{i=1}^{J} p_i(1 - p_i) = \sum_{i=1}^{J} (p_i - p_i^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=1}^{J} p_i^2$$

Reference: https://en.wikipedia.org

2 How to prune the Decision Trees?

Pre-prune:

Pre-pruning, also known as Early Stopping Rule, is the method where the subtree construction is halted at a particular node after evaluation of some measure. These measures can be the Gini Impurity or the Information Gain. In pre-pruning, we evaluate the pruning condition based on the above measures at each node. Examples of pruning conditions include seting the minimal gain and the max depth of the tree. If the condition is satisfied, we prune the subtree. That means we replace the decision node with a leaf node. Otherwise, we continue building the tree using our decision tree algorithm.

Post-prune:

As the name suggests, post-pruning means to prune after the tree is built. You grow the tree entirely using your decision tree algorithm and then you prune the subtrees in the tree in a bottom-up fashion. You start from the bottom decision node and, based on measures such as Gini Impurity or Information Gain, you decide whether to keep this decision node or replace it with a leaf node. For example, say we want to prune out subtrees that result in least information gain. When deciding the leaf node, we want to know what leaf our decision tree algorithm would have created if it didn't create this decision node.

Reference: https://www.educative.io