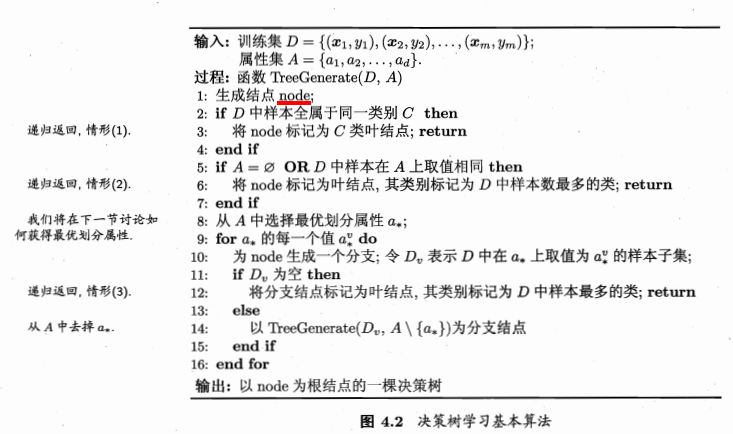
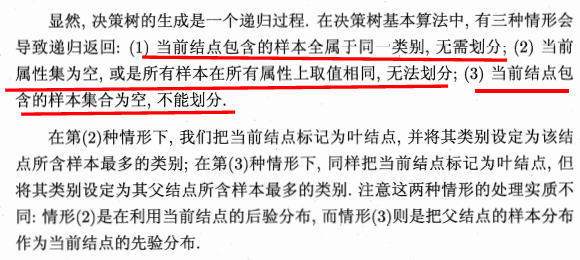
Decision Tree





Using the decision algorithm, we start at the tree root and split the data on the feature that results in the **largest information gain (IG).**In an iterative process, we can then repeat this splitting procedure at each child node until the leaves are pure. **We typically prune the tree by setting a limit for the maximal depth of the tree to prevent overfitting.**

**Maximizing information gain – getting the most bang for the buck （ID3算法）**

Our objective function is to maximize the information gain at each split, which we define as follows:



f is the feature to perform the split, Dp and Dj are dataset of the parent and jth chile node, and Nj is the number of samples in the jth child node. As we can see, the information gain is simply the difference between the impurity of the parent node and the sum of the child node impurities – the lower the impurity of the child nodes, the larger the information gain. However, for simplicity and to reduce the combinatorial search space, most libraries implement binary decsiton trees. This means that each parent node is split into two child nodes, Dleft and Dright:



The three impurity measures or splitting criteria that are commonly used in binary decision decision tree are Gini index (Ig),entropy (Ih) and the classification error (Ie).

增益率：

实际上，信息增益准则对可取值数目较多的属性有所偏好，为减少这种偏好可能带来的不利影响，著名的C4.5 决策树算法不直接使用信息增益，而是使用“增益率”（gain ratio）来选择最优划分属性。增益率定义为：



其中，



称为属性a的固有值。属性a的可能取值数目越多（即V越大），则IV(a)的值通常会越大。需注意的是，增益率准则对可取值数目较少的属性有所偏好，因此，C4.5算法并不是直接选择增益率最大的候选划分属性，而是使用了一个启发式：先从候选划分中找出信息增益高于平均水平的属性，再从中选择增益率最高的。

1. Entroy:



Here ,  is the proportion of the samples that belongs to class c for a particular node t. The entropy is therefore 0 if all samples at a node belong to the same class, and the entropy is maximal if we have a unfiormal class distribution.

For example, in a binary class setting, the entropy is 0 if p (i=1|t ) = 1 or p (i=0|t ) = 0 . If the classes are distributed uniformly with p(i=1|t ) = 0.5 and p (i=0|t )= 0.5, the entropy is 1. Therefore, we can say that the entropy criterion attempts to maximize the mutual information  
in the tree.

1. Gini index （CART算法，Classification and Regression Tree）: can be understood as criterion to minimize the probability of misclassification:

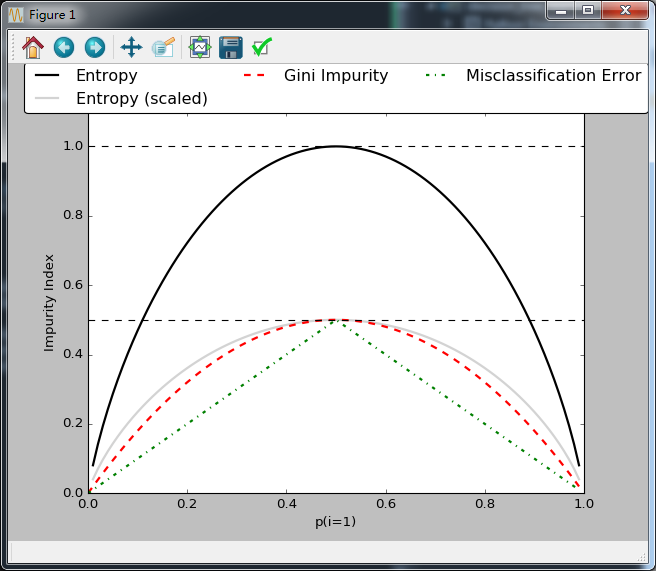


Similar to entropy, the Gini index is maximal if the classes are perfectly mixed.

1. Classification error:

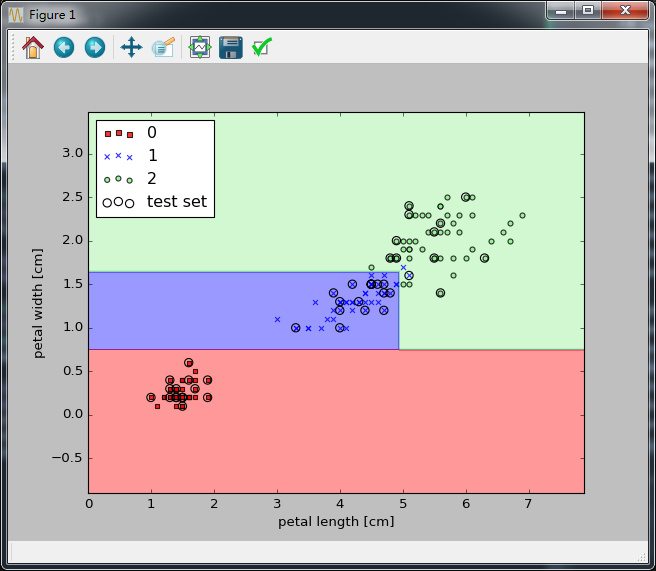


This is a useful criterion for pruning but not recommended for growing a decision three, since it is less sensitive to changes in the class probabilities of the nodes.



**Building a decision tree**

The deeper the decision tree, the more complex the decision boundary becomes, which can easily result in overfitting.



Reference:

1. Python machine learning
2. 机器学习\_周志华