建立简单的决策树：

ID3（Iterative Dichotomier 3）：越强大的属性越靠近根节点（最基础）

如何选择属性？--> 属性的度量

- **Ross Quinlan**: http://www.rulequest.com/
- One of the most influential Decision Trees models
- Top-down, greedy search through the space of possible decision trees
- Since we want to construct short trees ...
- It is better to put certain attributes higher up the tree.
- Some attributes split the data more purely than others.
- Their values correspond more consistently with the class labels.
- Need to have some sort of measure to compare candidate attributes.

$$Gain(S, District) = Entropy(S) - \frac{5}{14}Entropy(S_{District=Suburban}) - \frac{5}{14}Entropy(S_{District=Urban}) - \frac{4}{14}Entropy(S_{District=Rural})$$

$$= 0.940 - \frac{5}{14}\cdot 0.971 - \frac{5}{14}\cdot 0.971 - \frac{4}{14}\cdot 0 = 0.247$$

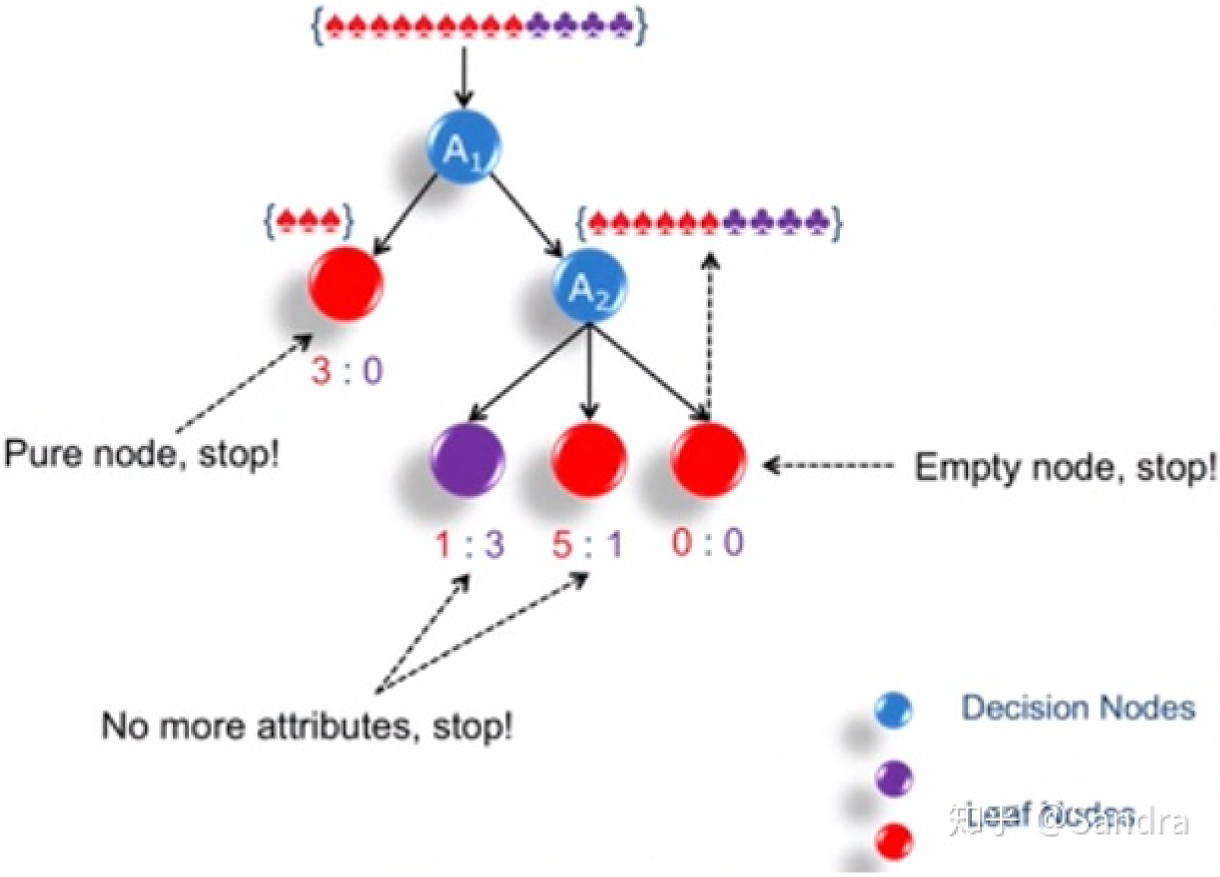
$$Gain(S, Income) = Entropy(S) - \frac{7}{14}Entropy(S_{Income=High}) - \frac{7}{14}Entropy(S_{Income=Low})$$

$$= 0.940 - \frac{7}{14}\cdot 0.9852 - \frac{7}{14}\cdot 0.5917 = 0.152$$

算法是如何工作的：递归建树

挑选对当前分类效果最好的属性，赋值 --> 然后在该节点下有不同的分支，分支所指向的子集若是“纯”的，无需再分类，打标签为正的；否则，在剩余的属性中挑选最好的放置在该分支的节点上，查看新增属性的分类子集。

但有可能，用完所有的属性，分至最后都无法做到全“纯”，那么少数服从多数



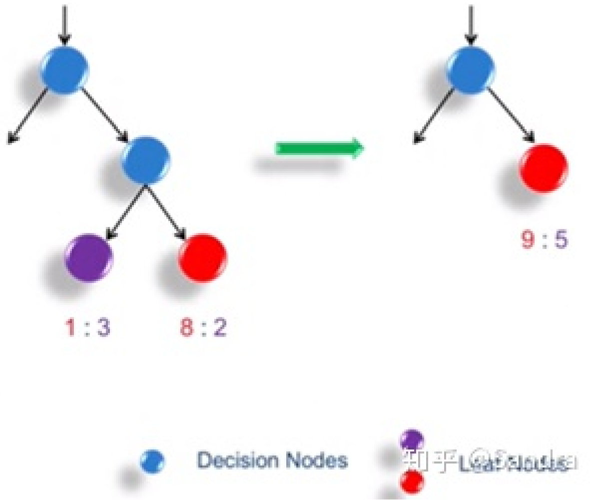
图中的空节点（Empty node）服从A2的“少数服从多数”

过学习：训练集中A比B好，测试集中B比A好 --> A过学习

- It is possible to create a separate rule for each training sample.
  - Perfect Training Accuracy vs. Overfitting
  - Random Noise, Insufficient Samples
- We want to capture the general underlying functions or trends.
- Definition
  - Given a hypothesis space H, a hypothesis $h \in H$ is said to overfit the training data if there exists some alternative hypothesis $h' \in H$, such as h has smaller error than h' over the training samples, but h' has a smaller error than h over the entire distribution of instances.
- Solutions
  - Stop growing the tree earlier.
  - Allow the tree to overfit the data and then post-prune the tree.

1. 要控制树的规模（深度有范围控制）
2. 剪枝（想让其自由生长，最后进行修剪），提高泛化能力

剪枝（其实是合并）：



生成模式时需要三种数据集

- Training Set：让树自由生长
- Validation Set：训练集的一种，剪枝的时候要观看在校验集的准确性曲线（误差大小曲线），在其拐点的位置收手
- Test Set

若教室中男生和女生的生日没有重复的，那么生日属性来进行分类，会让每个分支下只有一个纯子集，看起来效果很好，但样本分的过于琐碎，极易过学习 --> 引入Entropy Bias（切分越细，数值越大）：

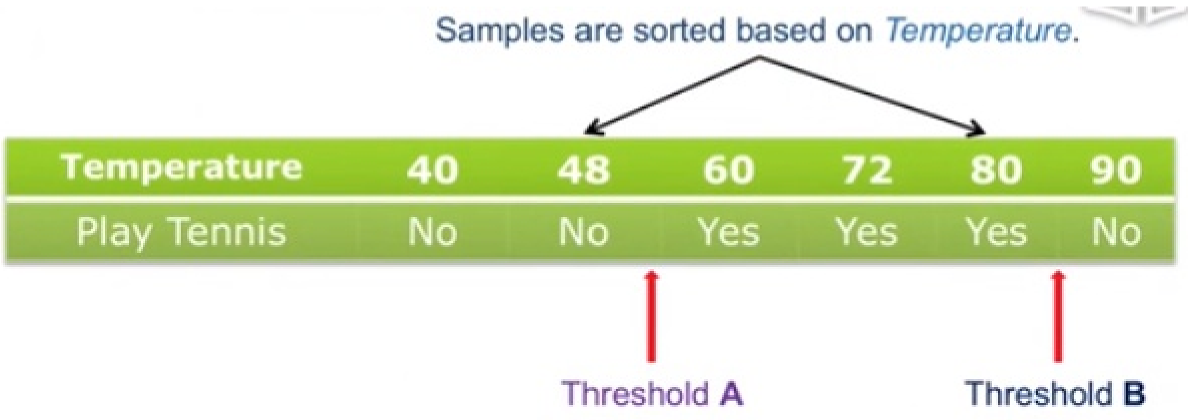
- The entropy measure guides the entire tree building process.
- There is a natural bias that favours attributes with many values.
- Consider the attribute "Birth Date"
  - Separate the training data into very small subsets.
  - Very high information gain
  - A very poor predicator of the target function over unseen instances.
- Such attributes need to be penalized!

$$SplitInformation(S,A) = -\sum_{i=1}^{C}\frac{|S_i|}{|S|}\log_2\frac{|S_i|}{|S|}$$

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}$$

连续型：（可以进行离散化）

离散切分界限，要根据熵进行判断好坏：



$$Gain(S,A) = Entropy(S) - \frac{1}{3}\cdot 0 - \frac{2}{3}\cdot(-\frac{3}{4}\cdot\log_2\frac{3}{4} - \frac{1}{4}\cdot\log_2\frac{1}{4}) = 1 - 0.541 = 0.459$$

$$Gain(S,B) = Entropy(S) - \frac{1}{6}\cdot 0 - \frac{5}{6}\cdot(-\frac{3}{5}\cdot\log_2\frac{3}{5} - \frac{2}{5}\cdot\log_2\frac{2}{5}) = 1 - 0.809 = 0.191$$

好的切分点要出现在label发生变化的时候

提供一些自学补充材料：

- Online Tutorial
  - http://www.decisiontrees.net/node/21 (with interactive demos)
  - http://www.autonlab.org/tutorials/dtree18.pdf
  - http://people.revoledu.com/kardi/tutorial/DecisionTree/index.html
  - http://www.public.asu.edu/~kirkwood/DAStuff/decisiontrees/index.html
- Tom Mitchell, *Machine Learning*, Chapters 3&6, McGraw-Hill.
- Additional reading about Naïve Bayes Classifier
  - http://www-2.cs.cmu.edu/~tom/NewChapters.html
- Software for text classification using Naïve Bayes Classifier
  - http://www-2.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html

补充知识：

熵：（计算不确定性的程度，最大值为1）

$$Entropy(S) = -\sum_{i=1}^{C} p_i \log(p_i)$$

p_i: the proportion of instances in the dataset that take the i^{th} target value

$$S = [9/14\,(responses), 5/14\,(no\ responses)]$$

原始数据的不确定性

增加了属性 -->不同的subset，计算每个subset的熵，计算时前面要有添加它的权重（每个属性所对应的子集的大小）

$$Entropy(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in A}\frac{|S_v|}{|S|}Entropy(S_v)$$

S_v: the subset of S where attribute A takes the value v.

熵为0.94，很不确定，引入新的属性-->获得了信息的增益，越大越好