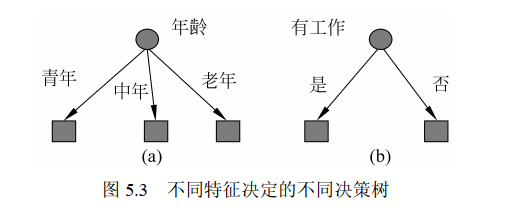
**《决策树模型二》**

**特征选择：**

特征选择在于选取对训练数据具有分类能力的特征。这样可以提高决策树学习的效率。如果利用一个特征进行分类的结果与随机分类的结果没有很大差别，则称这个特征是没有分类能力的。在一些建模的过程中我们会将这个特征删除。

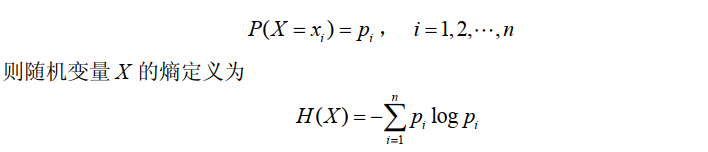
所以又有特征工程这个概念。在决策树中通常特征选择的准则是信息增益或信息增益比。

特征选择是决定用哪个特征来划分特征空间。



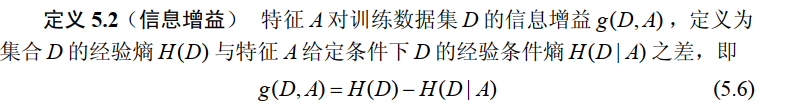
**年龄和工作性质两个特征进行生成不同的决策树，那么什么时候选择年龄什么时候选择有工作进行特征的划分（决策树的生成）。那就是使用信息增益和信息增益比。**

熵是表示随机变量不确定性的度量．设X是一个取有限个值的离散随机变量，其概率分布为：



**熵取值最大，随机变量不确定性最大。**

信息增益表示得知特征 X 的信息而使得类Y 的信息的不确定性减少的程度 。



经验熵 H(D)表示对数据集 D 进行分类的不确定性。

条件熵 H(D|A)表示在特征 A 给定的条件下对数据集 D 进行分类的不确定性。

每一次划分对所有的特征进行信息增益的计算，选择信息增益最大的特征进行划分。