决策树

决策树基本原理

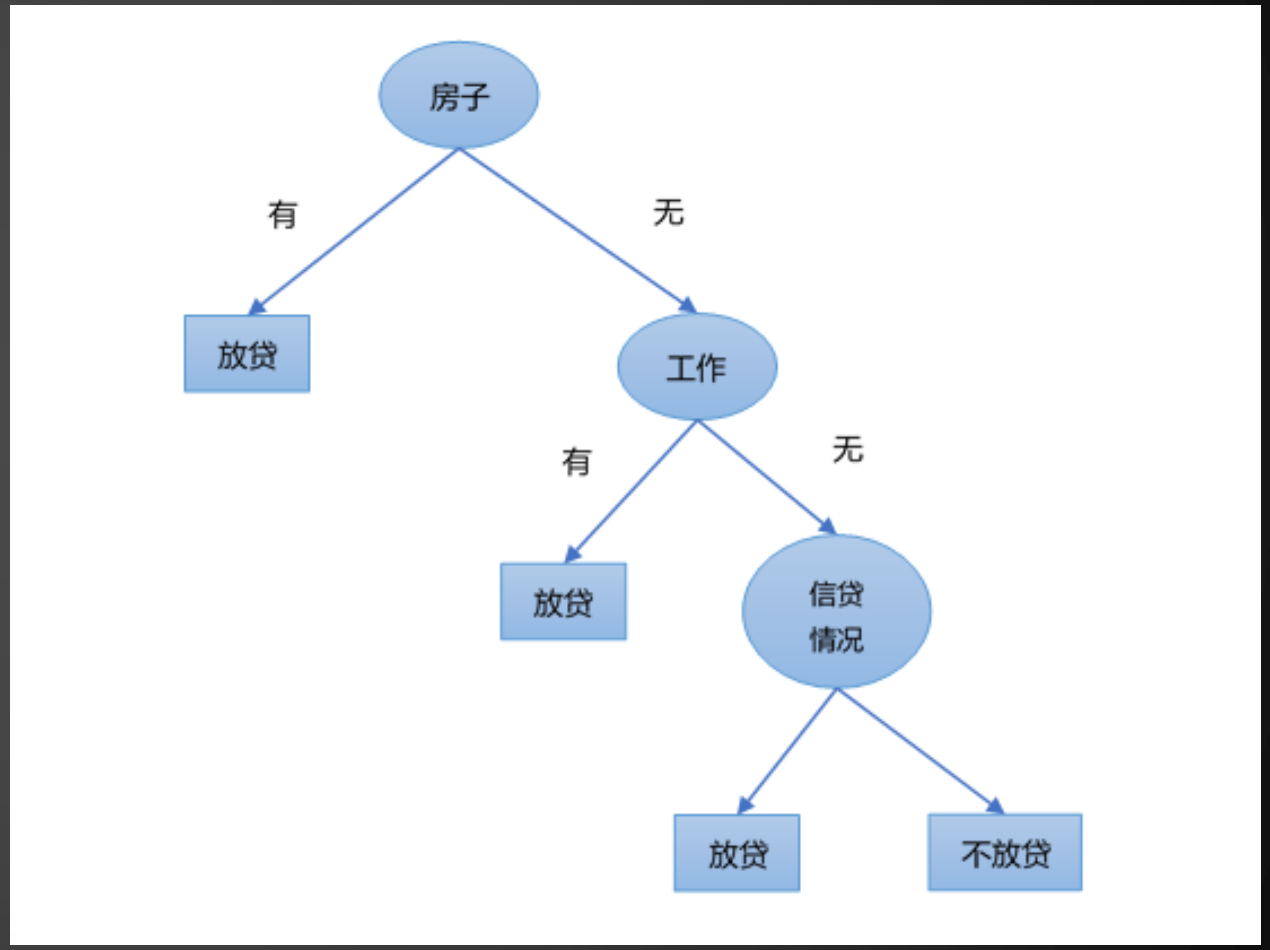
- 决策树：用树状图来模拟人的决策行为的一种算法，本质是一系列if/else问题。
- 例如，金融机构判断是否考虑给一个人放贷，可能考虑的因素有：房子、工作、信贷情况及年龄等。其决策流程大致为：

1. 房子：是否有房？如果有房，则考虑放贷。如果没房，则继续考虑工作情况。
2. 工作：在没有房子的情况下，如果有工作，则考虑放贷，否则继续考虑信贷情况。
3. 信贷情况：在没有工作的情况下，如果信贷情况非常好，考虑放贷，否则不考虑放贷。

决策树基本原理

图中，涉及以下概念：

- 决策树：用来描述决策过程的树状图
- 根节点：决策树的第一个节点
- 内部节点：用于描述中间过程的节点
- 叶子节点：代表最终决策结果的节点。



熵

- 例如，有两个数据集D1和D2。
- D1={男，男，女，男，女，女}
- D2={男，男，女，男，男，男}

- 数据集D1对我们说来是混乱的，因为不同性别标签的样本都混在一起，数据集D2相对不那么混乱的，因为其中的大部分样本都是一种性别，男，尽管有一个性别标签为女的样本。

- 那么，如何衡量一个数据集的混乱程度呢?
- 答案是熵，熵在信息论中很常见，可以用来衡量数据的混乱程度，表示信息的期望值。

D_1={男，男，女，男，女，女}，D_2={男，男，女，男，男，男}

对于数据集D1来说，包含两个类别，男和女。
对于男这个类别来说，其出现的概率记为$p_1 = \frac{3}{6}$，接着，取一个底为2对数并加一个负号，表示该类别的信息值，即$l(\text{男}) = -\log_2 p_1 = 1$

按照同样的方式，可以计算出女这个类别的信息值，即$l(\text{女}) = -\log_2 p_2 = 1$

接着，计算整个数据集的所有类别的信息期望值，就是数据集D1的熵，记为

$$H(D_1) = -p_1 \times \log_2 p_1 - p_2 \times \log_2 p_2 = 1$$

对于数据集D2来说，也可以按照类似的方式计算出它的熵，即

$$H(D_2) = -p_1 \times \log_2 p_1 - p_2 \times \log_2 p_2 = -\frac{5}{6} \times \log_2 \frac{5}{6} - \frac{1}{6} \times \log_2 \frac{1}{6} = 0.65$$

从上面可以看到，数据集D1的熵明显大于数据集D2的熵，即数据集D1的混乱程度大于数据集D2。

熵

以上计算过程涉及两个概念：某个分类的信息和熵。

1. 某个类别的信息

   定义为：$l(x_i) = -\log_2 p(x_i)$，其中，$p(x_i)$表示选择该类别的概率。

2. 熵，数据集中所有类别的信息期望值

   用公式表示为：$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$，其中，$n$是类别数目。

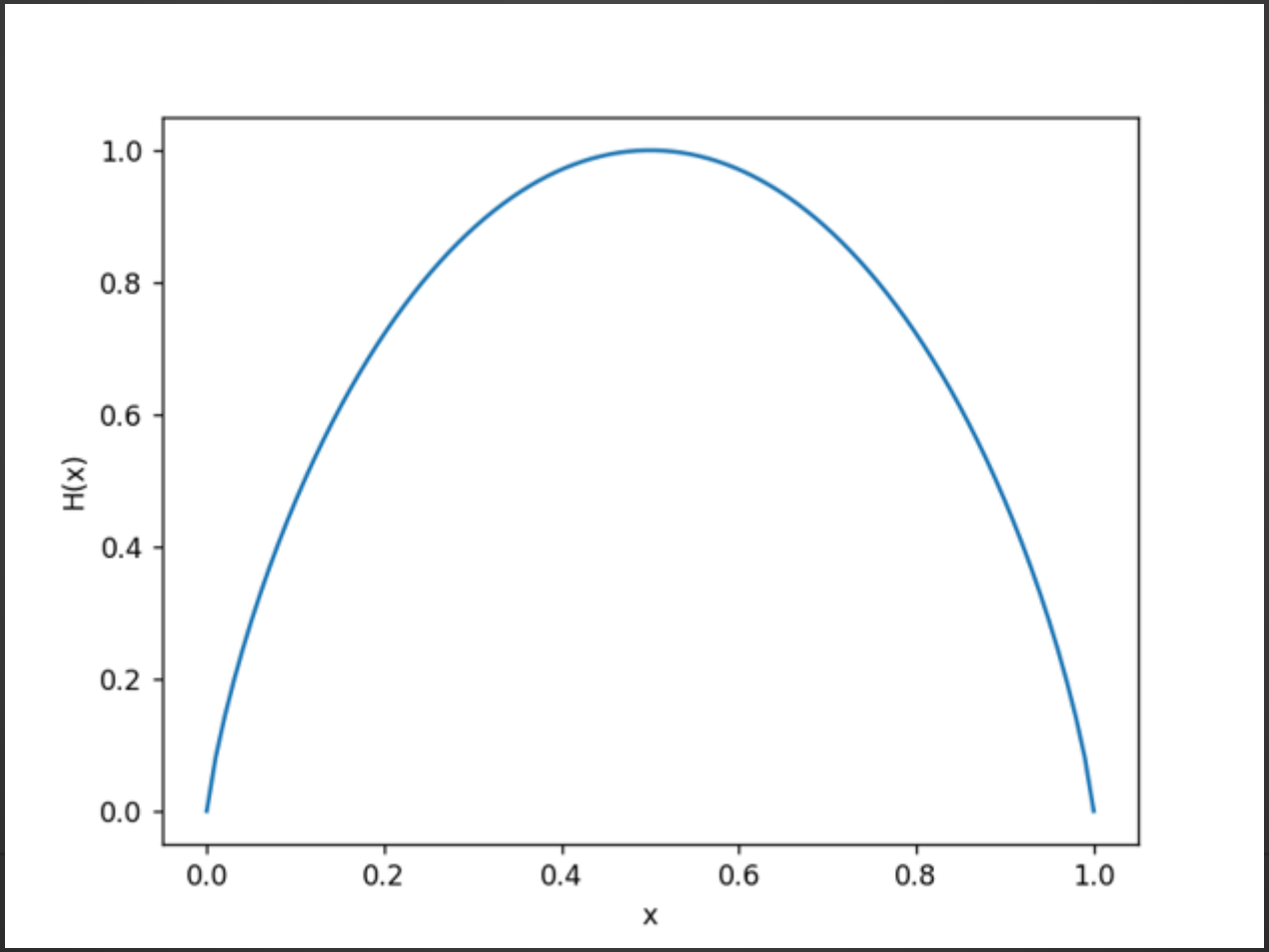
   对于常见的二分类问题，$n = 2$。

对于一个二分类问题来说，只有两个类别，整个数据集的熵的计算公式为：

$$H(X) = -p \log_2(p) - (1-p) \log_2(1-p)$$

熵

- 将H(X)看成是一个关于p的函数，$H(X) = -p\log_2(p) - (1-p)\log_2(1-p)$



经验条件熵和信息增益

- 为了说明经验条件熵和信息增益，我们给数据集D加上一个特征头发，分为短发和长发。

编号	头发	性别
1	短发	男
2	长发	男
3	长发	女
4	短发	男
5	长发	女
6	短发	女

首先，计算出数据集D的**总体熵**，计算公式如下。

$$H(D) = -\frac{3}{6} \times \log_2 \frac{3}{6} - \frac{3}{6} \times \log_2 \frac{3}{6} = 1$$

将头发这个特征称为A，根据这个特征A，将数据集分为：D1={1:男，4:男，6:女}和D2={2:男，3:女，5:女}
这两部分的熵分别为：

$$H(D_1) = -\frac{2}{3} \times \log_2 \frac{2}{3} - \frac{1}{3} \times \log_2 \frac{1}{3} = 0.92$$

$$H(D_2) = -\frac{1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3} = 0.92$$

接着，计算这两个数据集的信息期望值，将这个信息期望值叫作**经验条件熵**，记为$H(D|A)$

$$H(D|A) = \frac{3}{6} \times H(D_1) + \frac{3}{6} \times H(D_2) = 0.92$$

可以看到，数据集划分前后，信息发生了变化，我们将其称为**信息增益**，记为$g(D, A)$。

$$g(D, A) = H(D) - H(D|A) = 1 - 0.92 = 0.08$$

构造决策树主要步骤

1. 计算整个数据集的总体熵
2. 计算各个特征的信息增益（总体熵-经验条件熵）
3. 选择信息增益最大的特征为分类节点
4. 以此类推，构造出整个决策树

决策树构造实例：金融机构借贷数据

- 样本数：15个
- 特征：年龄、工作、房子、信贷情况
- 标签：是/否

- 任务：构造决策树

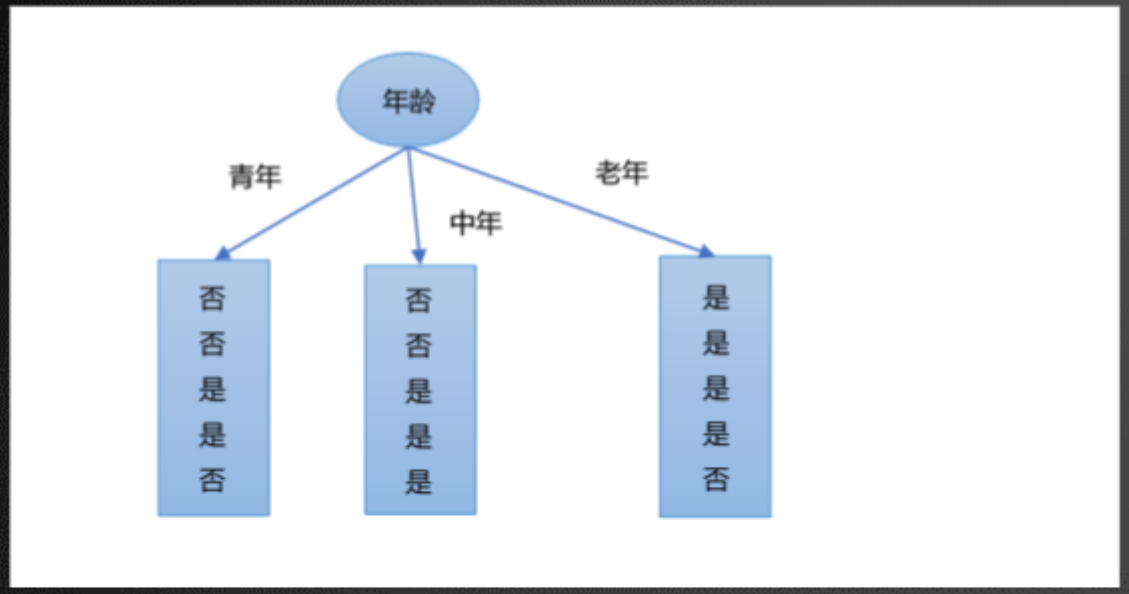
ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

- 计算整个数据集D的熵
- 类别为是：9个
- 类别为否：6个

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

$$H(D) = -\frac{9}{15}\times\log_2\frac{9}{15} - \frac{6}{15}\times\log_2\frac{6}{15} = 0.971$$

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否



根据特征年龄A_1，将数据集D划分为三部分D1,D2,D3，这三部分的熵分别为：

$$H(D_1) = -\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5}) = 0.971$$

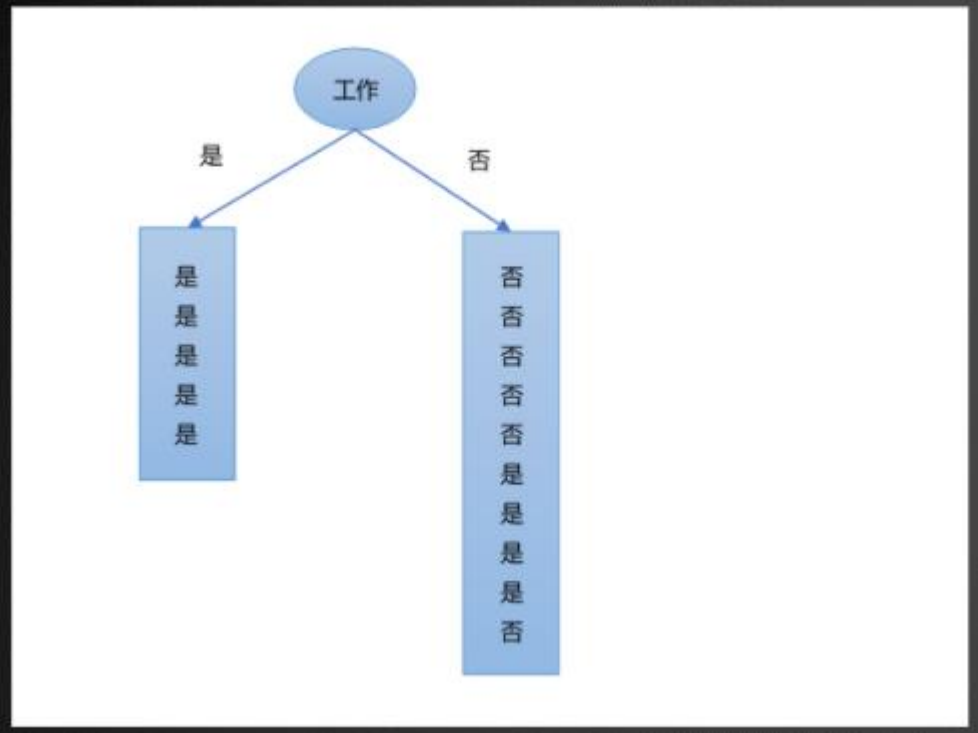
$$H(D_2) = -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}) = 0.971$$

$$H(D_3) = -\frac{1}{5}\log_2(\frac{1}{5}) - \frac{4}{5}\log_2(\frac{4}{5}) = 0.723$$

根据数据集D，年龄分别取青年、中年、老年的概率为：$\frac{5}{15},\frac{5}{15},\frac{5}{15}$

经验条件熵：$H(D|A_1) = \frac{5}{15}H(D_1) + \frac{5}{15}H(D_2) + \frac{5}{15}H(D_3) = 0.888$

信息增益：$g(D,A_1) = H(D) - H(D|A_1) = 0.971 - 0.888 = 0.083$



根据特征工作A_2，将数据集D划分为三部分D1,D2，这两部分的熵分别为：

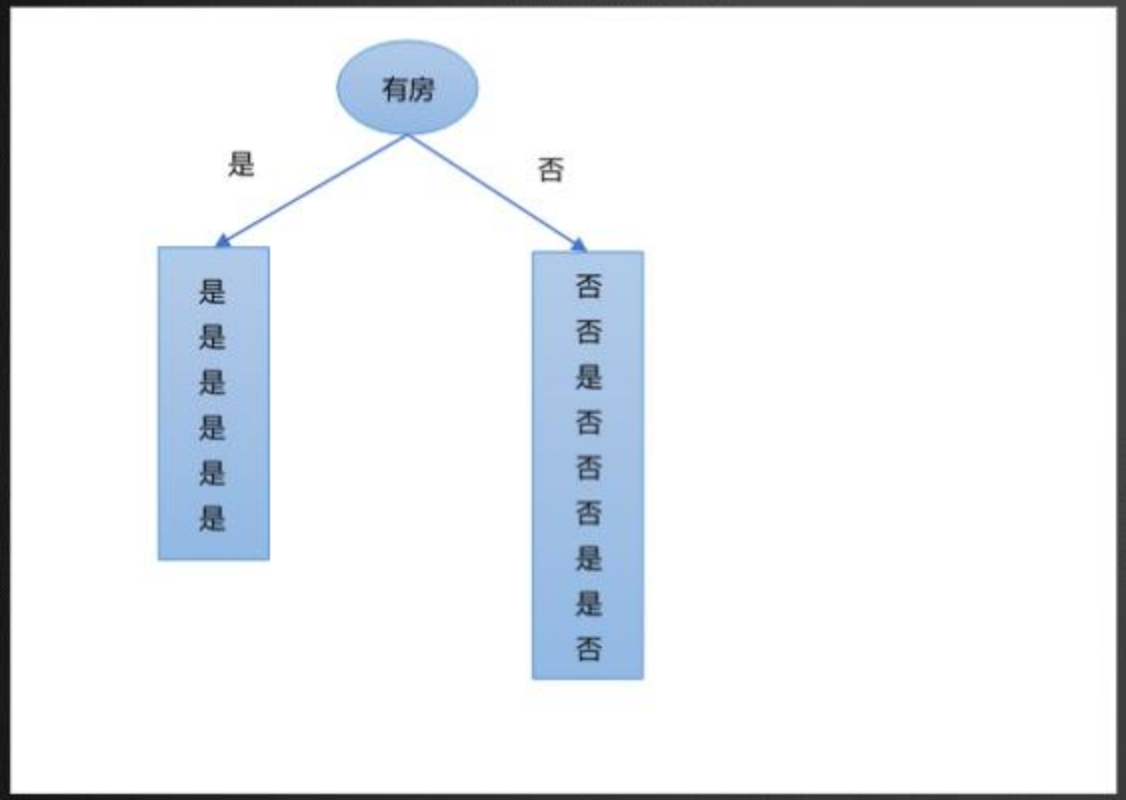
$$H(D_1) = -\frac{5}{5}\log_2(\frac{5}{5}) = 0$$

$$H(D_2) = -\frac{4}{10}\log_2(\frac{4}{10}) - \frac{6}{10}\log_2(\frac{6}{10}) = 0.971$$

根据数据集D，工作分别取是、否的概率为：$\frac{5}{15}, \frac{10}{15}$

经验条件熵：$H(D|A_2) = \frac{5}{15}H(D_1) + \frac{10}{15}H(D_2) = 0.647$

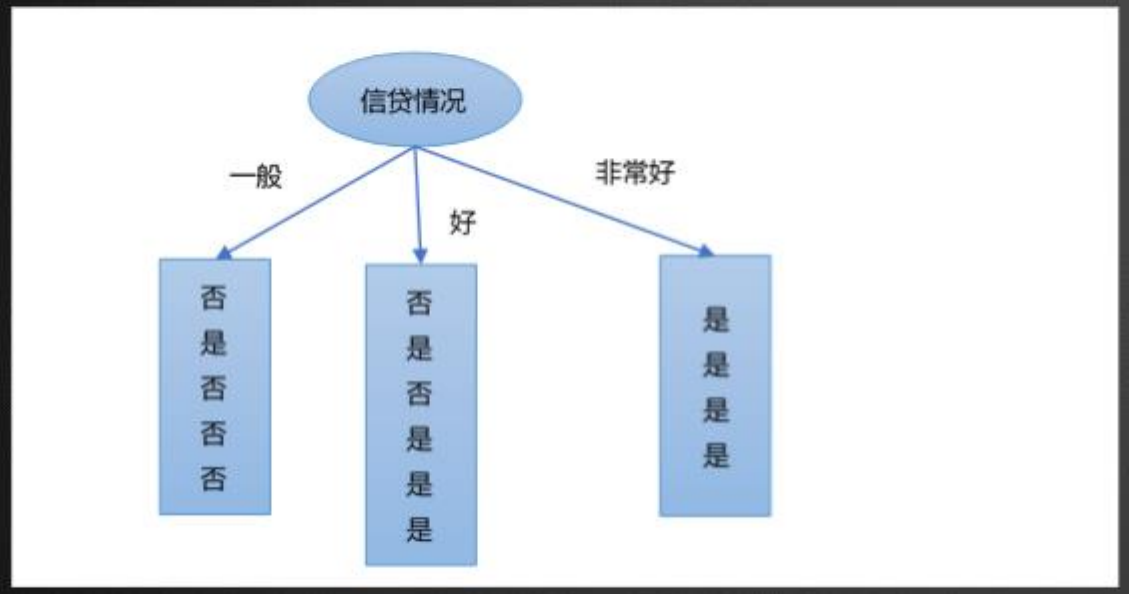
信息增益：$g(D, A_2) = H(D) - H(D|A_2) = 0.971 - 0.647 = 0.324$



对于特征有房，其信息增益为：

$$g(D, A_3) = H(D) - H(D|A_3) = 0.971 - [\frac{6}{15}(-\frac{6}{6}\log_2\frac{6}{6}) + \frac{9}{15}(-\frac{3}{9}\log_2\frac{3}{9} - \frac{6}{9}\log_2\frac{6}{9})] = 0.971 - 0.551 = 0.420$$

特征信贷情况的划分



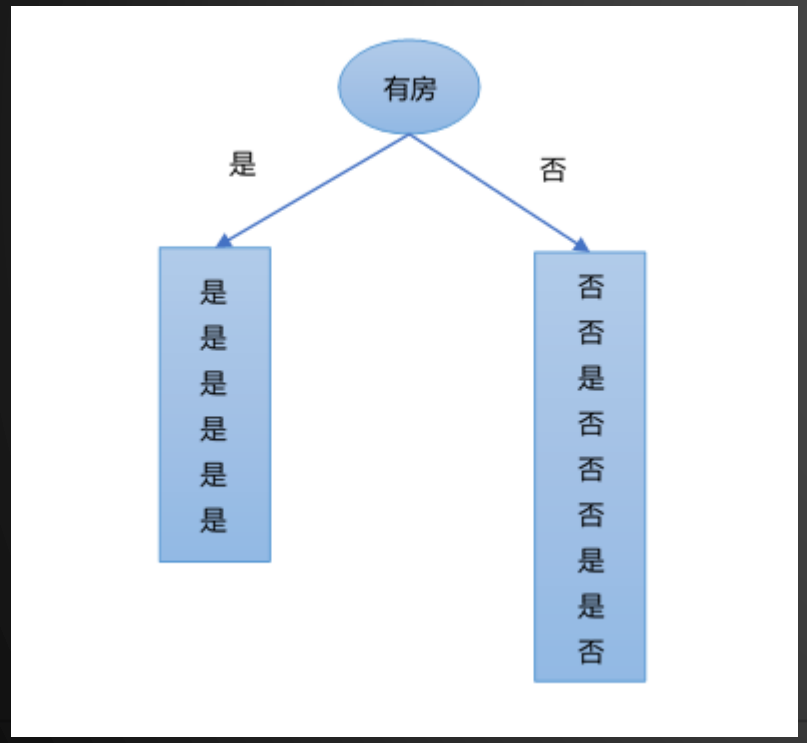
对于特征信贷情况，其信息增益为：

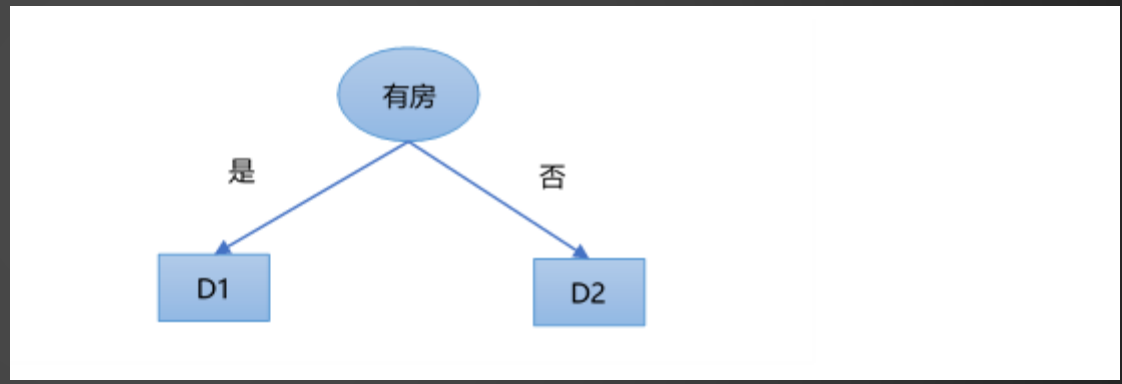
$$g(D, A_4) = H(D) - H(D|A_4) = 0.971 - 0.608 = 0.363$$

比较各特征的信息增益值：

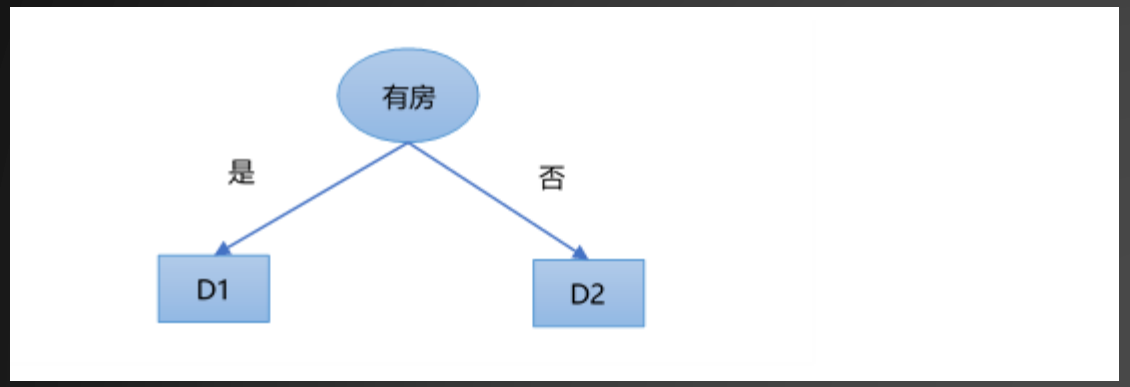
年龄：0.083，是否有工作：0.324，是否有房：0.420，信贷情况：0.363。

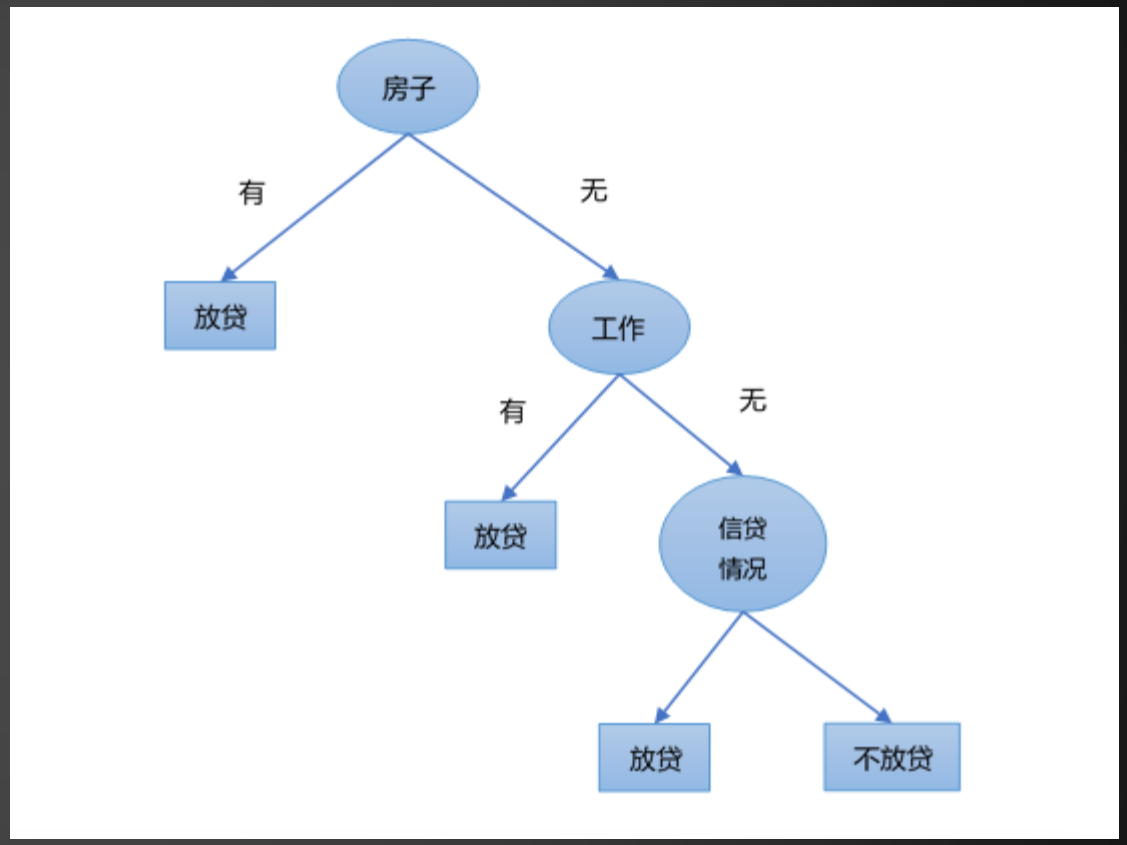
特征A3(是否有房)的信息增益值最大，所以选择特征A3作为最有特征，构造决策树如下。





对于根据特征"有房"得到的数据集D2继续使用如上方法，可以构造出如下决策树。





常用的决策树生成算法

- **ID3：**在决策树各节点上应用信息增益准则选择特征，递归构建决策树。

$$g(D,A) = H(D) - H(D|A)$$

问题：存在偏向于选择取值较多的特征的问题。

- **C4.5：**对ID3算法进行了改进，用信息增益比来选择特征。

$$g\_r_{(D,A)} = \frac{g(D,A)}{H(D|A)}$$

- **CART：**Classifation and Regression Tree，分类与回归树，用Gini指数来选择特征。

$$Gini(D) = 1 - \sum_{i=1}^{n} (\frac{|C_i|}{|D|})^2$$

其中，C_i是D中属于第i类的样本子集，n是类别的个数。

树的剪枝（pruning）

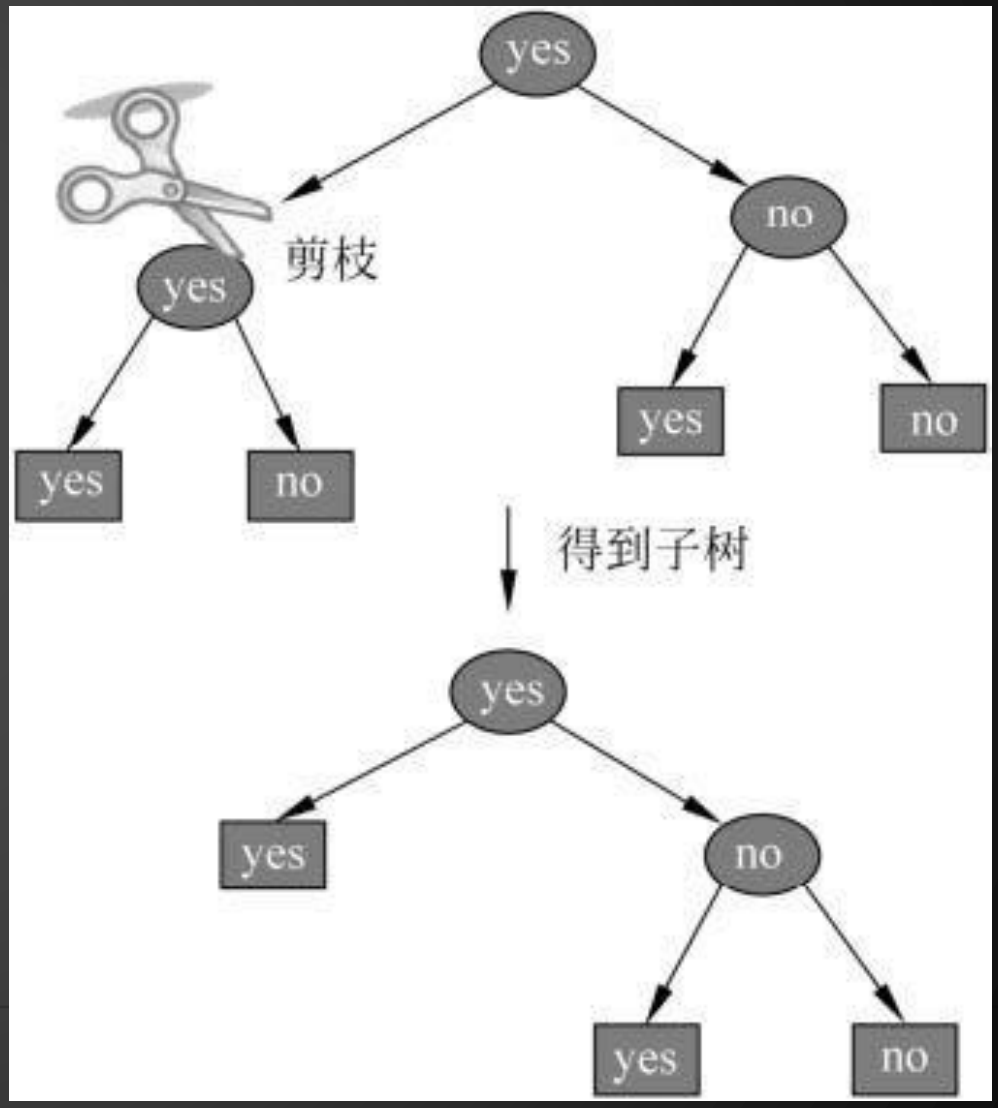
将已生成的树进行简化的过程称为剪枝（pruning）

剪枝通常有两种策略：

- **预剪枝：**提前控制树的生长，当熵减少的数量达到某一个阈值时，就停止树的分支的生成。
- **后剪枝**：先构造一棵树，然后删除信息较少的分支。

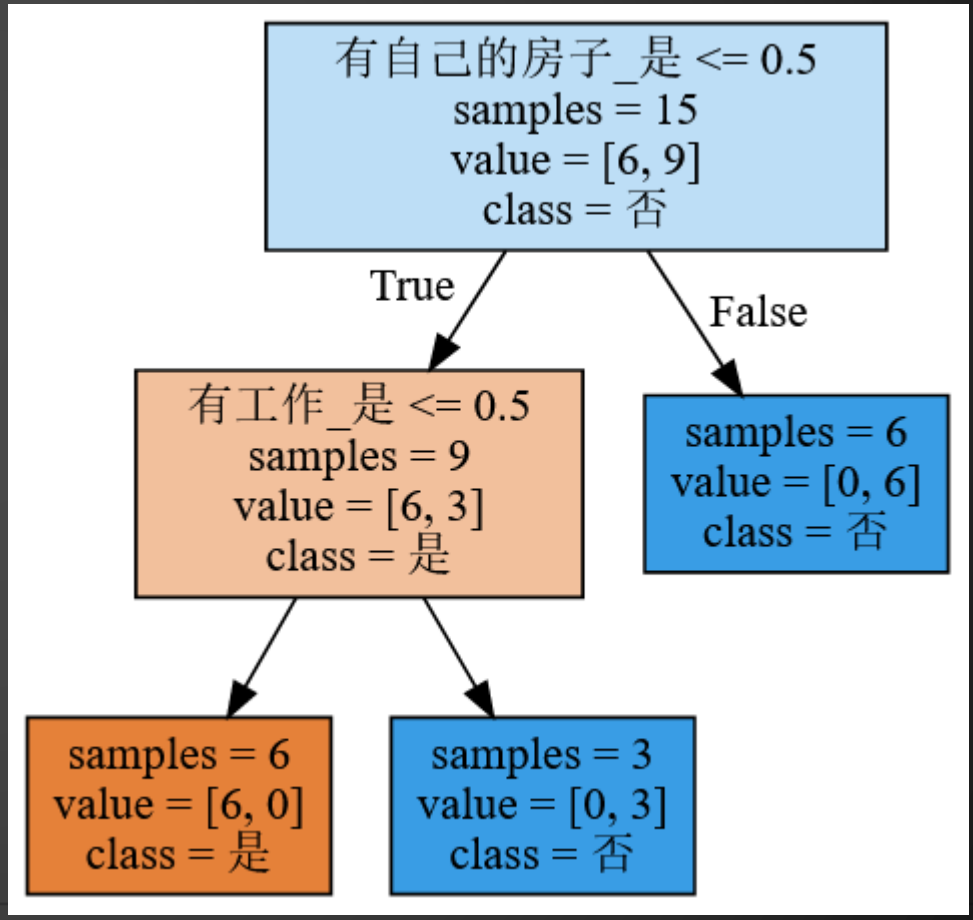
本质：防止过拟合，控制决策树的复杂度

sklearn中需要关注的参数：max_depth、max_leaf_nodes等。



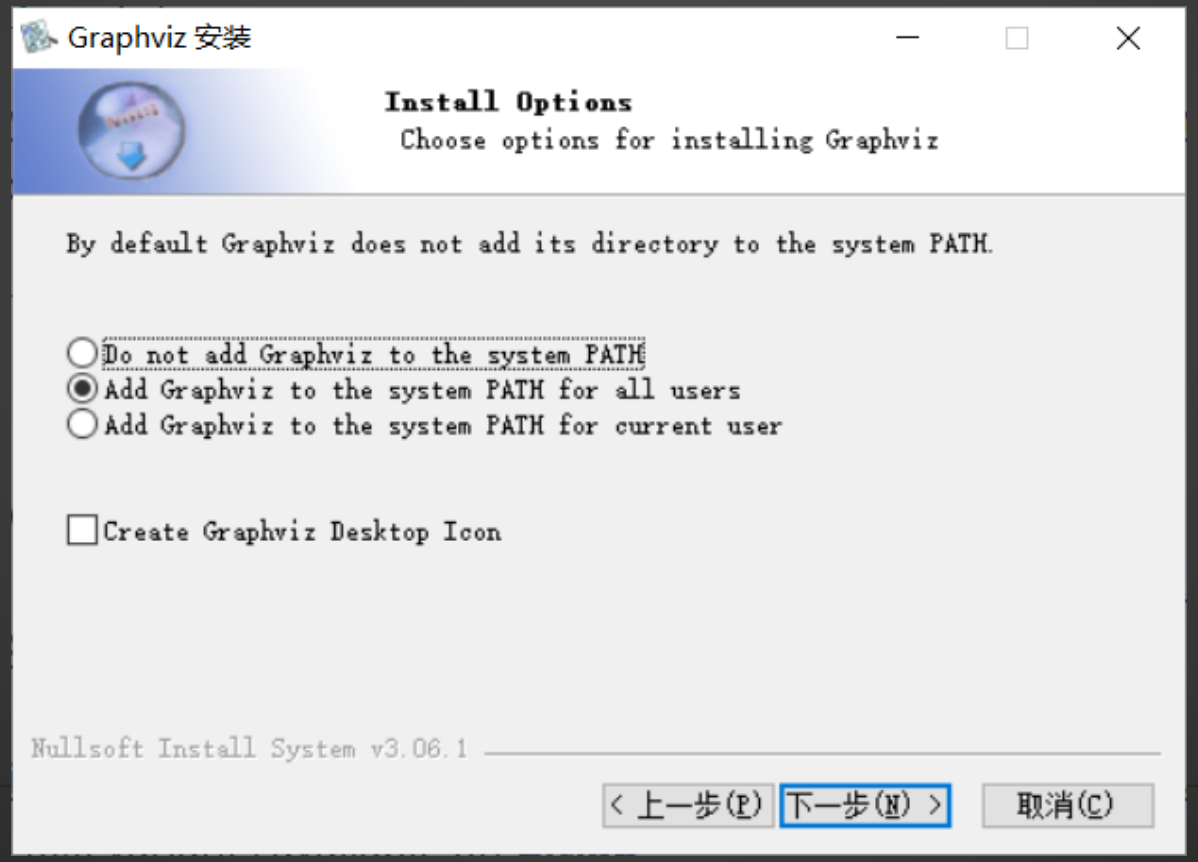
决策树的可视化

- 决策树可视化工具：graphviz
- 需要安装Graphviz软件和graphviz库



安装Graphviz软件

1. 下载并安装Graphviz
2. 下载地址：http://www.graphviz.org/



安装graphviz库

pip安装graphviz库

案例：利用决策树预测员工是否离职

案例及数据介绍：见逻辑回归最后部分讲解

代码实现主要步骤：

1. 读取数据、认识数据
2. 数据探索及预处理
3. 利用sklearn进行决策树建模
4. 使用交叉验证评估模型
5. 利用graphviz进行决策树可视化
6. 树的特征重要性