$$=\sum_{i=1}^{n}\left(y_i-h_\theta\left(x_i\right)\right)x_i$$

将求导结果代入梯度下降迭代公式得：

$$\theta_j=\theta_j-\eta\sum_{i=1}^{n}\left(y_i-h_\theta\left(x_i\right)\right)x_i$$

这样就可以利用迭代公式不断地更新模型参数，直至收敛。常用的梯度下降法包括批量梯度下降法、随机梯度下降法和小批量梯度下降法，这里不进行详细讲解。

逻辑斯蒂回归只能用于解决二分类问题，将它推广为多项逻辑斯蒂回归模型（multi-nominal logistic model，也即softmax函数），用于处理多类分类问题，可以得到softmax回归。

4.3 决策树

决策树是一种通过树形结构来进行分类的方法。在决策树中，树形结构中每个非叶子结点表示对分类目标在某个属性上的一个判断，每个分支代表基于该属性做出的一个判断，每个叶子结点代表一种分类结果，所以决策树可以看作是一系列以叶子结点为输出的决策规则（decision rules）[21]。

4.3.1 决策树分类案例

决策树将分类问题分解为若干基于单个信息的推理任务，采用树状结构来逐步完成决策判断。事实上，人们在逻辑推理过程中经常使用决策树的思想。

下面通过一个例子来解释决策树的分类。银行的数据分析师希望通过历史的贷款记录，包括用户的4种特征（年龄，银行流水，婚姻状况，房产状况）以及最终是否给予贷款，来建立分类模型辅助决策者进行决策。

银行收集整理的数据如表4.6所示。

通过观察，可以画出如图4.7所示的决策树，由图可知以下几点。

表4.6 是否给予贷款与申请人自身状况的关系

序号	年龄	银行流水	是否结婚	拥有房产	是否给予贷款
1	>30	高	否	是	否
2	>30	高	否	否	否
3	20~30	高	否	是	是
4	<20	中	否	是	是
5	<20	低	否	是	是
6	<20	低	是	否	否
7	20~30	低	是	否	是
8	>30	中	否	是	否
9	>30	低	是	是	是
10	<20	中	否	是	是
11	>30	中	是	否	是
12	20~30	中	否	否	是
13	20~30	高	是	是	是
14	<20	中	否	否	否

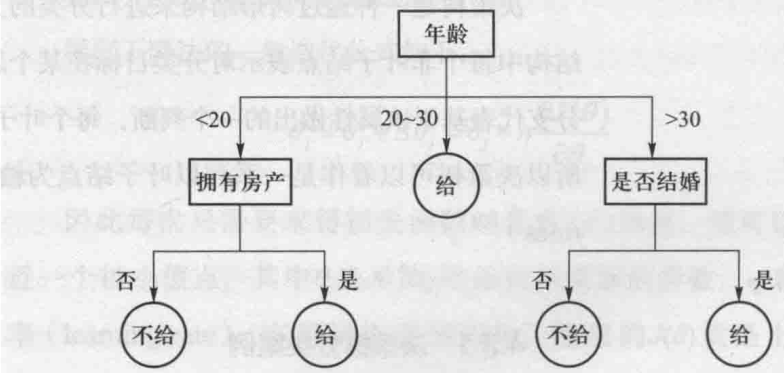

图4.7 银行贷款决策树

• 第一层是年龄状况，分为小于20岁，20岁至30岁，大于30岁三种取值。

• 如果年龄在20岁和30岁之间，样本子集为{3, 7, 12, 13}，这些样本的标签均为“给予贷款”，所以为叶子结点。

• 如果年龄大于30岁，样本子集为{1, 2, 8, 9, 11}，这些样本具有不同的标签，要进一步使用其他属性对这个样本子集进行划分。经观察，通过“是否结婚”这一属性值，可以将该样本子集进一步划分成{1, 2, 8}（未婚）和{9, 11}（已婚）两个样本子集。此时这两个样本子

集内标签一样，不需要再划分。

● 如果年龄小于20岁，样本子集为{4, 5, 6, 10, 14}，这些样本具有不同的标签，同样需要继续划分。通过观察，“拥有房产”这个属性值可将该样本子集进一步划分成{4, 5, 10}（无房产）和{6, 14}（有房产）两个样本子集。此时这两个样本子集内标签一样，不需要再划分。

“银行流水”这一特点及其属性值在这次决策树构造过程中没有使用。

建立决策树的过程，就是不断选择属性值对样本集进行划分，直至每个子样本为同一个类别。上面的案例数据较少，可以通过观察或是穷举的方法来不断选择属性值对样本集进行划分，对于较大的数据集，需要理论和方法来评价不同属性值划分的子样本集的好坏程度，并基于该方法构建决策树。

4.3.2 构建决策树

构建决策树时划分属性的顺序选择是重要的。性能好的决策树随着划分不断进行，决策树分支结点样本集的“纯度”会越来越高，即其所包含样本尽可能属于相同类别。

信息熵（entropy）就是一种衡量样本集合“纯度”的指标，如果计算选择不同属性划分后样本集的“纯度”，那么就可以比较和选择属性。信息熵越大，说明该集合的不确定性越大，“纯度”越低。选择属性划分样本集前后信息熵的减少量被称为信息增益（information gain），也就是说信息增益被用来衡量样本集合复杂度（不确定性）所减少的程度。

假设有K个信息，其组成了集合样本D，记第k个信息发生的概率为$p_k(1 \leqslant k \leqslant K)$。如下定义这$K$个信息的信息熵：

$$E(D) = -\sum_{k=1}^{K} p_k \log_2 p_k$$

$E(D)$值越小，表示D包含的信息越确定，也称D的纯度越高。需要指出，所有p_k累加起来的和为1。

现在应用熵这个度量标准来构建决策树。表4.6中14个样本分属于“给予贷款（9个样本）”和“不给予贷款（5个样本）”两个类别，即$K=2$。

$$Ent(D) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{9}{14} \times \log_2 \frac{9}{14} + \frac{5}{14} \times \log_2 \frac{5}{14}\right) = 0.940$$

表4.6中有年龄、银行流水、是否结婚、拥有房产4个人物相关的属性特征，下面计算这4个特点所对应的信息熵。

以年龄为例，包含“>30”、“20 ~ 30”、“<20”3个属性取值。这3个属性取值对14个样本进行划分，在决策树中产生了3个分支结点，每个分支结点包含一个数据子集，3个数据子集构成了对原数据的划分。如“20 ~ 30”这一属性取值包含4个样本{3, 7, 12, 13}。

对年龄属性划分出的子样本集情况的统计如表4.7所示。这里记属性取值为a_i（即a_0=“>30”，a_1=“20 ~ 30”，a_2=“<20”）。属性取值a_i划分出的子样本集记为D_i，该子样本集包含样本数量记为$|D_i|$。

表4.7 年龄属性划分后子样本集情况统计

年龄属性取值a_i	“>30”	“20~30”	“<20”
对应样本数$\|D_i\|$	5	4	5
正负样本数量	(2+, 3-)	(4+, 0-)	(3+, 2-)

根据表4.7的统计情况，计算每个属性值划分出的子样本集（即每个分支结点）的信息熵：

$$\text{“>30”：}\ Ent(D_0)=-\left(\frac{2}{5}\times\log_2\frac{2}{5}+\frac{3}{5}\times\log_2\frac{3}{5}\right)=0.971$$

$$\text{“20 ~ 30”：}\ Ent(D_1)=-\left(\frac{4}{4}\times\log_2\frac{4}{4}+0\right)=0$$

$$\text{“<20”：}\ Ent(D_2)=-\left(\frac{3}{5}\times\log_2\frac{3}{5}+\frac{2}{5}\times\log_2\frac{2}{5}\right)=0.971$$

得到上述三个的信息熵后，可进一步计算使用年龄属性对原样本集进行划分后的信息增益，计算公式如下：

$$Gain(D, A)=Ent(D)-\sum_{i=1}^{n}\frac{|D_i|}{|D|}Ent(D_i)$$

将A= 年龄 代入。于是选择年龄这一属性划分后的信息增益为：

$$Gain(D, \text{年龄})=0.940-\left(\frac{5}{14}\times0.971+\frac{4}{14}\times0+\frac{5}{14}\times0.971\right)=0.246$$

同理，可以计算银行流水、是否结婚、是否拥有房产3个人物属性的信息增益。通过比较4种属性信息增益的高低来选择最佳属性对原样本集进行划分，得到最大的“纯度”。如果划分后的不同子样本集都只存在同

类样本，那么停止划分。在该案例中，最终可构建如图4.7的决策树。

一般而言，信息增益偏向选择分支多的属性[11]（如上述案例偏向于选择年龄属性），这在一些场合容易导致模型过拟合。为了解决这个问题，一个直接的想法是对分支过多进行惩罚，这就是另外一个“纯度”衡量指标，信息增益率的核心思想。除了计算属性划分后的信息增益为$Gain(D, A)$外，还需要计算划分行为本身带来的信息$info$，$info$和$Gain-ratio$计算公式如下：

$$info = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|}$$

$$Gain - ratio = Gain(D, A) / info$$

另一种计算更简易的度量指标是如下的$Gini$系数：

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2$$

相对于信息熵的计算$E(D) = -\sum_{k=1}^{K} p_k \log_2 p_k$，不用计算对数log，计算更为简易。

4.4 线性判别分析

线性判别分析（linear discriminant analysis，LDA）是一种基于监督学习的降维方法，也称为Fisher线性判别分析（fisher's discriminant analysis，FDA）[12]。对于一组具有标签信息的高维数据样本，LDA利用其类别信息，将其线性投影到一个低维空间上，在低维空间中同一类别样本尽可能靠近，不同类别样本尽可能彼此远离。

LDA与主成分分析（PCA）紧密相关，它们都在寻找最佳解释数据的变量线性组合[18]。

图4.8给出了用矩形和圆圈来表示患有某一疾病或者不患有某一疾病的两类人群。通过调查，这两类人群分别用吸烟频率高低和运动频率高低来描述。当然，在图4.8中，假设这两类人群样本数据符合高斯分布。

为了对这两类人群进行区分，需要将其投影到一个低维空间中。从图中可见，将这些数据向x轴方向和y轴方向投影后，总会存在重叠部分（即若干人群在投影后空间中不可区分）。但是，如果将数据向直线w所