Assignment 7

姓名：费政聪
学号：201928013229003

2019 年 12 月 26 日

1. (1) 计算所有特征对表中数据集的信息增益。

计算经验熵H(D)。

$$H(D) = -\frac{9}{15}log_2\frac{9}{15} - \frac{6}{15}log_2\frac{6}{15} = 0.971 \tag{1}$$

分别以A_1,A_2,A_3,A_4表示年龄，有工作，有自己的房子和信贷情况四个特征。

$$g(D, A_1) = H(D) - [\frac{5}{15}H(D_1) + \frac{5}{15}H(D_2) + \frac{5}{15}H(D_3)] = 0.083 \tag{2}$$

$$g(D, A_2) = H(D) - [\frac{5}{15}H(D_1) + \frac{10}{15}H(D_2)] = 0.324 \tag{3}$$

$$g(D, A_3) = H(D) - [\frac{6}{15}H(D_1) + \frac{9}{15}H(D_2)] = 0.420 \tag{4}$$

$$g(D, A_4) = H(D) - [\frac{5}{15}H(D_1) + \frac{6}{15}H(D_2) + \frac{4}{15}H(D_3)] = 0.363 \tag{5}$$

(2) 用ID3算法建立决策树。

根据上题，选择A_3有自己的房子作为根节点特征，划分为子集D_1(是)和D_2(否)。由于子集D_1只有同一类样本点，所以它是叶节点，标记为“是”。

对子集D_2从特征A_1,A_2,A_4中选择。

$$g(D_2, A_1) = H(D_2) - H(D_2|A_1) = 0.251 \tag{6}$$

$$g(D_2, A_2) = H(D_2) - H(D_2|A_2) = 0.918 \tag{7}$$

$$g(D_2, A_4) = H(D_2) - H(D_2|A_4) = 0.474 \tag{8}$$

选择A_2特征作为节点特征，以此分出的两个子集均为同一类，叶节点，分别标记为“是”和“否”。得到的决策树结构如图1所示。

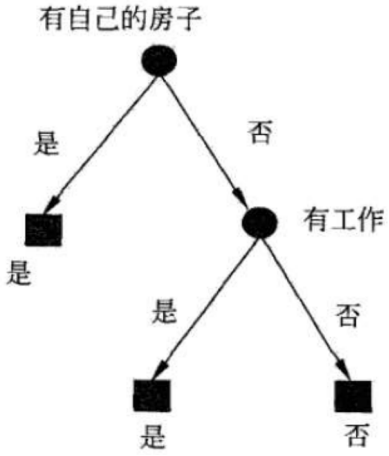

图 1: 决策树结构。

2. 用伪代码描述一种决策树剪枝的方法。

Algorithm 1 Post-Pruning

Input: features, data set;

Output: decision tree;

```
Grow decision tree to its entirety;
Split data into training and validation set;
Do until further pruning is harmful:
    Evaluation on validation set of pruning each possible node;
    Greedily remove that most improves validation set;
return smallest version of most accurate subtree;
```

3. PCA算法的计算过程。

(1) 计算数据的均值。

$$\hat{x} = \frac{1}{N}\sum_{i=1}^{N} x_n \tag{9}$$

(2) 计算数据的协方差矩阵。

$$S = \frac{1}{N}\sum_{i=1}^{N}(x_n - \hat{x})(x_n - \hat{x})^T \tag{10}$$

(3) 对$D \times D$协方差矩阵做特征值分解。
(4) 取前K大的特征值对应的特征向量构成投影矩阵$U = [u_1, \ldots, u_K]$。
(5) 降维后的数据为 $z_n = U^T x_n$。