Pattern Recognition

第七次作业

牛李金梁
201928014629008

2020 年 1 月 1 日

Question 1

(a)

训练数据集为 D，特征集 A_1, A_2, A_3, A_4 分别表示年龄、有工作、有自己的房子和信贷情况 4 个特征。

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

- 年龄 (A_1) 对数据集 D 的经验条件熵：$H(D|A_1) = \frac{5}{15}H(D_1) + \frac{5}{15}H(D_2) + \frac{5}{15}H(D_3) = 0.888$，年龄 (A_1) 的信息增益为：$g(D, A_1) = H(D) - H(D|A_1) = 0.083$
- 有工作 (A_2) 对数据集 D 的经验条件熵：$H(D|A_2) = \frac{5}{15}H(D_2) + \frac{10}{15}H(D_2) = 0.647$，有工作 (A_2) 的信息增益为：$g(D, A_2) = H(D) - H(D|A_2) = 0.324$
- 有自己的房子 (A_3) 对数据集 D 的经验条件熵：$H(D|A_3) = \frac{6}{15}H(D_1) + \frac{9}{15}H(D_2) = 0.551$，有自己的房子 (A_3) 的信息增益为：$g(D, A_3) = H(D) - H(D|A_3) = 0.420$
- 信贷情况 (A_4) 对数据集 D 的经验条件熵：$H(D|A_4) = \frac{5}{15}H(D_1) + \frac{6}{15}H(D_2) + \frac{4}{15}H(D_3) = 0.608$，信贷情况 (A_4) 的信息增益为：$g(D, A_4) = H(D) - H(D|A_4) = 0.363$

(b)

由（a）得，在数据集 D 中，特征 A_3(有自己的房子) 的信息增益最大，所以选取特征 A_3 作为根节点的特征。于是数据集 D 被划分为两个子集 D_1(A_3 取值为“是”) 和 D_2(A_3 取值为“否”)。

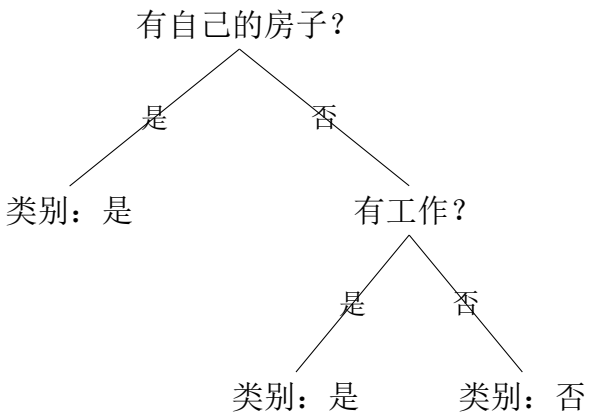
由于 D_1 只含有同一类样本点，所以它成为一个叶子节点，节点类型标记为“是”。接下来求剩余特征 (A_1, A_2, A_4) 在子集 D_2 的信息增益。

D_2 的经验熵为 $H(D_2) = -\frac{2}{9}\log_2\frac{2}{9} - \frac{7}{9}\log_2\frac{7}{9} = 0.918$

- 年龄 (A_1) 的信息增益：$g(D_2, A_1) = H(D_2) - H(D_2|A_1) = 0.918 - 0.667 = 0.083$
- 有工作 (A_2) 的信息增益：$g(D_2, A_2) = H(D_2) - H(D_2|A_2) = 0.918 - 0 = 0.918$
- 信贷情况 (A_4) 的信息增益：$g(D_2, A_4) = H(D_2) - H(D_2|A_4) = 0.918 - 0.444 = 0.474$

选择信息增益最大的特征 A_2(有工作) 作为节点特征。A_2 有两个可能取值，从该节点引出两个子节点：一个对应“是”，包含三个样本，属于同一类，所以构成一个叶子节点，类别标记为“是”。另一个对应“否”的子节点，包含 6 个样本，它们也属于同一类，所以也构成一个叶子节点，类别标记为“否”。

最终使用两个特征 A_2, A_3 构成了一个决策树：



Question 2

比较划分前后的泛化性来决定是否应该进行此划分。若划分后泛化性有提升，则保留划分，否则删除此划分。

```
Pre-Pruning(root)
P' = the classification accuracy on test set using rule of root;
for node = root->first child to root->last child do
// traverse root's children
    P = the classification accuracy on test set using rule of node;
    if P > P' then
        remain the node root;
        Pre-Pruning(node);
    else then
        discard the node root;
    end if
end for
return root;
```

Question 3

对于 N 个 D 维样本 $x_1, x_2, \ldots, x_N$，得到一个数据空间 $X \in \mathbb{R}^{N\times D}$。主成分分析是将每个样本从 D 维投影到 K 维，其中 $K < D$，此时的数据空间为 $X \in \mathbb{R}^{N\times K}$。计算步骤如下：

1. 计算数据集均值：$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$，其中 $\bar{x} \in \mathbb{R}^D$。
2. 计算样本的协方差矩阵：$S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T$，其中 $S \in \mathbb{R}^{D\times D}$。
3. 对协方差矩阵 S 做特征值分解，得到 D 个特征值。取前 K 个最大的特征值 $\lambda_1, \lambda_2, \ldots, \lambda_K$，以及每个特征值对应的特征向量 $u_1, u_2, \ldots, u_K$。
4. 构造映射矩阵 $U = [u_1, u_2, \ldots, u_k]$，其中 $U \in \mathbb{R}^{D\times K}$。将每个样本投影到新的数据空间 $z_n = U^T x_n$，其中 $z_n \in \mathbb{R}^K$。