UCAS

PATTERN RECOGNITION

Assignment 7

Yuxun Qu

October 2019

Question 1

请简述 Adaboost 算法的设计思想。

下表是一个由 15 个样本组成的贷款申请训练数据，包括四个特征（年龄，有无工作，有无房屋，信贷情况），最后一列是类别，表示是否同意其贷款。问题如下：

（1）计算所有特征对上表中数据集的信息增益

（2）用 ID3 算法建立决策树

表 5.1　贷款申请样本数据表

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

图 1: 信息表

Solution:

(1) 首先计算经验熵 $H(D)$

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

然后计算各个特征的信息增益，分别以 A_1, A_2, A_3, A_4 代表年龄，有工作，有房子和信贷四个特征

$$\begin{aligned} g(D, A_1) =& H(D) - \left[\frac{5}{15}H(D_{A_1=\text{青年}}) + \frac{5}{15}H(D_{A_1=\text{中年}}) + \frac{5}{15}H(D_{A_1=\text{老年}})\right] \\ =& 0.971 - \left[\frac{5}{15}(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) + \frac{5}{15}(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}) + \frac{5}{15}(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5})\right] \\ =& 0.971 - 0.888 = 0.083 \end{aligned}$$

$$\begin{aligned} g(D, A_2) =& H(D) - \left[\frac{5}{15}H(D_{A_2=\text{有工作}}) + \frac{10}{15}H(D_{A_1=\text{无工作}})\right) \\ =& 0.971 - \left[\frac{5}{15}(-0\log_2 0 - \log_2 1) + \frac{10}{15}(-\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10})\right] \\ =& 0.971 - 0.647 = 0.324 \end{aligned}$$

$$
\begin{aligned}
g(D, A_3) =& H(D) - \left[\frac{6}{15}H(D_{A_2=\text{有房子}}) + \frac{9}{15}H(D_{A_1=\text{无房子}})\right] \\
=& 0.971 - \left[\frac{6}{15}(-0\log_2 0 - \log_2 1) + \frac{9}{15}(-\frac{6}{9}\log_2\frac{6}{9} - \frac{3}{9}\log_2\frac{3}{9})\right] \\
=& 0.971 - 0.551 = 0.42
\end{aligned}
$$

$$
\begin{aligned}
g(D, A_4) =& H(D) - \left[\frac{4}{15}H(D_{A_2=\text{非常好}}) + \frac{6}{15}H(D_{A_1=\text{好}})) + \frac{5}{15}H(D_{A_1=\text{一般}}))\right] \\
=& 0.971 - \left[\frac{4}{15}(-0\log_2 0 - \log_2 1) + \frac{6}{15}(-\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}) + \frac{5}{15}(-\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5})\right] \\
=& 0.971 - 0.608 = 0.363
\end{aligned}
$$

(2) 第一步取出信息增益最大的一项，即是否有房 (A_3) 作为分类的第一个节点，将数据集分为 D_1（有房）与 D_2（无房）两个部分，其中有房已经完全分类，对无房的 D_2 分类。

对 A_1, A_2, A_4 进行分类

计算经验熵 $H(D)$

$$
H(D_2) = -\frac{6}{9}\log_2\frac{6}{9} - \frac{3}{9}\log_2\frac{3}{9} = 0.918
$$

$$
\begin{aligned}
g(D_2, A_1) =& H(D) - \left[\frac{5}{15}H(D_{A_1=\text{青年}}) + \frac{5}{15}H(D_{A_1=\text{中年}}) + \frac{5}{15}H(D_{A_1=\text{老年}})\right] \\
=& 0.971 - \left[\frac{3}{9}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) + \frac{3}{9}(-0\log_2 0 - \log_2 1) + \frac{3}{9}(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3})\right] \\
=& 0.918 - 0.612 = 0.306
\end{aligned}
$$

$$
\begin{aligned}
g(D_2, A_2) =& H(D) - \left[\frac{3}{9}H(D_{A_2=\text{有工作}}) + \frac{6}{9}H(D_{A_1=\text{无工作}}))\right] \\
=& 0.971 - \left[\frac{3}{9}(-0\log_2 0 - \log_2 1) + \frac{6}{9}(-0\log_2 0 - \log_2 1)\right] \\
=& 0.918 - 0 = 0.918
\end{aligned}
$$

易知，A_4 分类的经验条件熵大于 0，所以最大信息增益的特征为 A_2。A_2 的两个分支都能完全分类，决策树停止。

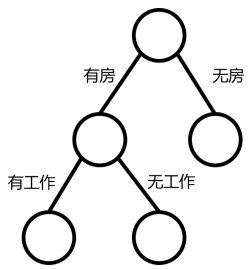

图 2: 决策树

Question 2

用伪代码描述一种决策树剪枝的方法

Solution:

以下算法为预剪枝算法，该算法运行的位置为决策树产生时，选择完最优分割属性之后，产生子节点之前。

Algorithm 1 决策树剪枝（预剪枝）

Require: 当前节点上的训练集 $\mathbf{D}$，当前节点上的验证集 $\mathbf{V}$，属性集 $A = \{a_1, a_2, ...a_d\}$，最优分属性 a_*，已生成的决策树 T，当前节点 $Node$.

Ensure: 剪枝后的决策树 T

function PrePruning($\mathbf{D}$,$\mathbf{V}$,$A = \{a_1, a_2, ...a_d\}$,a_*,T,$Node$)
    在 $Node$ 节点上选取计数最多的类别 C_{Node}，并取得数据集中属于该类别的数据样本个数 T_{Node}. 初始化计数 T_a
    for a_* 的每一种取值 a_*^v **do**
        取 $a_* = a_*^v$ 的训练子集 $\mathbf{D}_v$ 与验证子集 $\mathbf{V}_v$.
        if D_v 是空集 **then**
            a_*^v 上的标记类别 $C_{a^v} \leftarrow C_{Node}$
        else
            a_*^v 上的标记类别 C_{a^v} 为 $\mathbf{D}_v$ 中最多的类别.
        end if
        计算验证子集 $\mathbf{V}_v$ 中被分为 C_{a^v} 的个数 T_{a^v}
        $T_a = T_a + T_{a^v}$
    end for
    if then$T_a \geq T_a$
        分支节点 $Node$ 不再产生儿子节点，将其标记为叶子节点.
    end if
end function

Question 3

有 N 个样本 $x_1, ..., x_N$，每个样本维数 D，希望将样本维数降低到 K，请给出 PCA 算法的计算过程

Solution:

Algorithm 2 PCA 算法

Require: $\mathbf{X}_{D\times N} = \{x_1, ...x_N\}$ 数组，D 降维前的维度，K 降维后的样本维度.

Ensure: 降维后的数据 $\mathbf{Y}$

function PCA($\mathbf{X} = \{x_1, ...x_N\}$)
    计算协方差矩阵 $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x})$
    计算 $\mathbf{S}$ 的特征值与特征向量 $\{\lambda_1, \lambda_2...\lambda_N\}$,$\{u_1, u_2...u_N\}$, 其中特征值按绝对值从大到小排列，即 $\lambda_i < \lambda_j, i < j$.
    取前 K 个特征值构成投影矩阵 $\mathbf{U}_{D\times K}$.
    降维后的数据为 $\mathbf{Y}_{K\times N} = \mathbf{U}^T\mathbf{X}$
end function