**第5章 机器学习**

**5.7** 设训练例子集如表5.10所示，请用ID3算法完成其学习过程。

表5.10 训练例子集

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 属性 | | 分类 |
| x1 | x2 |
| 1 | T | T | + |
| 2 | T | T | + |
| 3 | T | F | - |
| 4 | F | F | + |
| 5 | F | T | \_ |
| 6 | F | T | \_ |

**参考解：**

初始化样本集S={S1,S2,…,S6}和属性集X={x1,x2}。

设根节点为(S,X)，尽管它包含了所有的训练例子，但却没有包含任何分类信息，因此具有最大的信息熵。即：

E(S,X)= -PS(+)log PS(+) - PS(-)log PS(-)

式中

PS(+)=3/6，PS(-)=3/6，log 2(3/6)= -1

即有

E(S,X)= -(3/6)\*log 2(3/6) - (3/6)\*log 2(3/6)

= 0.5+0.5 = 1

按照ID3算法，再计算根节点(S, X)关于每个属性的加权信息熵。

先考虑属性x1，对x1的不同取值：

当x1=T时，有ST={1，2，3}

当x1=F时，有SF={4，5，6}

其中，ST 和SF中的数字均为例子集S中的各个例子的序号，|S|、| ST|和|SF|分别为例子集S、ST和SF 的大小且有|S|=6，| ST |=| SF |=3。

由ST可知：

　　　　PST(+)=2/3，PST(-)=1/3，log2(2/3)= -0.5850，log2(1/3) =- 1.5850

则有：

E(ST, X)= - PST(+)log2 PST(+) - PST(-)log2 PST(- )

= - (2/3)\*(-0.5850)- (1/3)\*(-1.5850)

=0.9183

再由SF可知：

　　　　PSF(+)=1/3，PSF(-)=2/3

则有：

E(SF,X)= -PSF(+)log2 PST(+)-PSF(-)log2 PSF(- )

= - (1/3)\*(-1.5850)- (2/3)\*(-0.5850)

=0.9183

E(S,X), xi)= ( |ST| / |S|)\* E(ST, X) + ( |SF| / |S|)\* H(SF, X)

= (3/6)\*0.9183+(3/6)\*0.9183

= 0.9183

再考虑属性x2，对x2的不同取值：

当x2=T时，有S’T={1，2，5，6}

当x2=F时，有S’F={3，4}

其中，S’T 和S’F中的数字均为例子集S中的各个例子的序号，|S|、| S’T|和|S’F|分别为例子集S、S’T和S’F 的大小且有|S|=6，| S’T |=4 ，| S’F |=2。

由S’T可知：

　　　　P’ST(+)=2/4，P’ST(-)=2/4，log2(2/4)= -1

则有：

E(S’T, X)= - P’ST(+)log2 P’ST(+) – P’ST(-)log2 P’ST(- )

= - (2/4)\*(-1) - (2/4)\*(-1)

=1

再由S’F可知：

　　　　P’SF(+)=1/2，P’SF(-)=1/2，log2(1/2)= -1

则有：

E(S’F,X)= - (P’SF(+)log2 P’ST(+)-P’SF(-)log2 P’SF(- ))

= - (1/2)\*(-1)- (1/2)\*(-1)

= 1

E(S,X), x2)= ( |S’T| / |S|)\* E(S’T, X) + ( |S’F| / |S|)\* H(S’F, X)

= (4/6)\*1+(2/6)\*1 = 1

据此，可得到各属性的信息增益分别为

G((S,X) x1)= E(S, X) - E(S, X), xi) = 1- 0.9183=0.0817

G((S,X) x2)= E(S, X) - E(S, X), x2) = 1-1=0

显然，x1的信息熵最大，因此应该对x1进行扩展。扩展x1后生成的部分决策树如下

S

(+,+,-)

(+,-,-)

x1=T

x1=F

扩展x1后得到的部分决策树

从属性集中删去已扩展的属性x1，得新的属性集集X1={x2}。由于现在属性集中的属性仅剩x1，故扩展x1，得到最终的完整决策树，如下图所示。

S

(+,+,-)

(+,-,-)

x1=T

x1=F

(-)

x2=T

(+,+)

x2=F

(-,-)

(+)

x2=F

x2=T

x2=F

扩展x1、x2后得到的完整决策树

5.15 假设给出的训练样本如表5.11所示

表5.11 训练例子集

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 样本序号  输入输出 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 输入X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 输出Y | 1 | 1 | -1 | -1 | -1 | 1 | 1 | -1 |

其中，X为输入向量，Y为输出向量，请用AdaBoost算法训练出一个强分类器。

**参考解：**

问题的实质是对训练样本集

D={ (0,1), (1,1), (2,-1), (3,-1), (4,-1), (5,1), (6,1), (7,-1) }

如何使用AdaBoost算法把0、1、2、3、4、5、6、7这8个数分成两类，其中一类输出为“1”，另一类输出为“-1”。从训练集可以看出

对输入0、1，对应的输出为1

对输入2、3、4，对应的输出为-1

对输入5、6，对应的输出为1

对输入7，对应的输出为-1

因此，对应的分界点应该在x为1和2之间，或x为4和5之间，或x为6和7之间。故可取1.5、4.5、或6.5为分界点。

在本例中，样本总个数N=8，初始化迭代轮数t=1，初始化训练样本的权重分布

P1(xi) = 1/N = 0.125 i=1,2,…,8

(1) 迭代过程1

对分界点1.5，若x<1.5时y取1，x>1.5时y取-1，则x为5、6时被错分，错分率为2\*0.125=0.25。

对分界点4.5，若x<4.5时y取-1，x>4.5时y取1，则x为0、1、7时被错分，错分率为3\*0.125=0.375。

对分界点6.5，若x<6.5时y取1，x>6.5时y取-1，则x为2、3、4时被错分，错分率为3\*0.125=0.375。

可见，对分界点x=1.5，其错分率=0.25，在3个分界点中为最低，故取x=1.5构造第1轮迭代的弱分类器1，即

此时数据0、1、2、3、4、7分类正确，而5和6被错分，其错分率为

下面根据计算弱分类器h1(x)在合成过程中的系数

它代表h1(x)最终强分类器中所占的比重。

接着按如下公式

更新训练样本的权值分布，用于下一轮迭代。为计算

计算归一化因子。

+

故有

即

P2(x)=(0.1250, 0.1250, 0.1250, 0.1250, 0.1250, 0.3750, 0.3750, 0.1250)

此时，可得到带权重的基本分类器

由以上分析可知，该分类器不能实现对样本数据正确分类，因此求解过程还需要再继续迭代。

(2) 迭代过程2

此时，迭代轮数t=2，权值分布为P2(x)，训练例子集还是原来例子集。分析可知

对分界点1.5，若x<1.5时y取1，x>1.5时y取-1，则x为5、6时被错分，错分率为2\*0.3750=0.75。大于0.5，不可取。

对分界点4.5，若x<4.5时y取-1，x>4.5时y取1，则x为0、1、7时被错分，错分率为3\*0.125=0.375。

对分界点6.5，若x<6.5时y取1，x>6.5时y取-1，则x为2、3、4时被错分，错分率为3\*0.125=0.375。

可见，对分界点x=4.5和6.5，其错分率均为0.375，假设取x=4.5构造第2轮迭代的弱分类器2，即

此时数据2、3、4、5、6分类正确，而0、1和7被错分，其错分率为

下面根据计算弱分类器h2(x)在合成过程中的系数

它代表h2(x)最终强分类器中所占的比重。

接着按如下公式

更新训练样本的权值分布，用于下一轮迭代。为计算

计算归一化因子。

+

故有

即

P3(x)=(0.0858, 0.0858, 0.0858, 0.0858, 0.0858, 0.1388, 0.1388, 0.0858)

此时，可得到带权重的基本分类器

由以上分析可知，该分类器不能实现对样本数据正确分类，因此求解过程还需要再继续迭代。

(3) 迭代过程3

此时，迭代轮数t=3，权值分布为P3(x)，训练例子集还是原来例子集。分析可知

对分界点1.5，若x<1.5时y取1，x>1.5时y取-1，则x为5、6时被错分，错分率为2\*0.1388=0.2776。

对分界点4.5，若x<4.5时y取-1，x>4.5时y取1，则x为0、1、7时被错分，错分率为3\*0.0858=0.2574。

对分界点6.5，若x<6.5时y取1，x>6.5时y取-1，则x为2、3、4时被错分，错分率为3\*0.0858=0.2574。

可见，对分界点x=4.5和6.5，其错分率=0.2574，在3个分界点中为最低，假设取x=6.5构造第3轮迭代的弱分类器3，即

此时数据0、1、5、6、7分类正确，而2、3和4被错分，其错分率为

下面根据计算弱分类器h3(x)在合成过程中的系数

它代表h3(x)最终强分类器中所占的比重。

此时，可得到带权重的基本分类器

实际上，到目前为止，AdaBoost算法的迭代过程已经完成。将上述3个基分类器集成到一起，就可得到一个能够解决所给问题的强分类器。即

对该强分类器的功能可验证如下：

可见，该集成分类器完全可以实现对所有样本数据的正确分类。