# AdaBoost步骤及算例

关于Adaboost的算例，网上的例子都比较统一，基本都来源于《统计方法学》这本书，这里也是对使用该例子，只是详细说明了计算过程，便于理解公式。

## AdaBoost算法的步骤

1. 初始化训练数据的权值分布，权值最终影响的是分类误差率的计算，进而影响弱分类器的系数。
2. 使用弱分类器Gm对训练数据进行分类，统计其分类误差率：



公式这么理解：m表示当前弱分类器的序号，i表示样本的序号，N表示有N个样本，Wmi表示统计第m个弱分类器时，样本i所具有的权重，Gm(xi)表示样本xi在Gm分类器上分类的结果，yi是xi的标签，+1或-1，公式的表示的是分类误差率为所有分类错误的样本所对应的权重之和。

1. 计算若分类器Gm的系数,公式表达为：



1. 更新训练数据的权值分布



其中：



这里的Zm只是为了将Wmi归一化，使得所有数据的权值和为1。公式中当分类器分类结果与对应标签相同，即分类正确时，yiGm(xi) = 1, 否则为-1，所以上面的公式可以这样表示：



当所有的弱分类器的系数计算完成后，构建基本分类器的线性组合：



得到最终的分类器：



## 算例

给定以下训练数据：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Y | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 |

弱分类器十分简单，为x<v或x>v产生，其中v为阈值，每个弱分类器的v不同。

试用Adaboost算法学习一个强分类器。

首先假设这里有三个弱分类器

  

这些分类器在测试集上的正确率为70%，60%，70%，都比随机猜测高，因此使用这3个分类器构成强分类器。由于v = 2.5和v = 8.5时分类器误差最低，首先随便选择其中一个开始计算，这里选v=2.5

首先，对第v=2.5分类器，初始化权值，这里共有10个样本，每个样本权值为1/10=0.1：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Y | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 |
| W | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |

当v=2.5时，7，8，9号样本分类错误，故误差率根据公式为:



计算分类器的系数：



注意这里的log是以e为底的求对数，而不是以10为底。

再对v=8.5的分类器，更样权值如下：

先计算归一化参数Z：



对于正确分类的样本，新的权重为：



以第一个样本为例，得：



对于错误的分类样本，新的权重为：



以第七个样本为例，得：



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Y | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 |
| W | 0.0714 | 0.0714 | 0.0714 | 0.0714 | 0.0714 | 0.0714 | 0.1666 | 0.1666 | 0.1666 | 0.0714 |

重复上面的步骤，可以得到，，于是最终的分类器为：

