研究生算法课课堂笔记

上课日期：2016 年 11 月 10 日　　第(2)节课

组长学号及姓名：1601111274　常远

组员学号及姓名：1601214730　刘志强

一、 内容概要

本节课所学内容包含以下几点：

1、 Random Forest 算法中树的构造特点讲解

2、 AdaBoost 算法的详细讲解

二、 详细内容

1、 随机森林（Random Forest）中树的构造特点

Q1：随机森林中的树应该是较高还是较矮？

A：较高。

原因：高一点的树是 high variance，low bias 的，即分类能力很强，矮的树则反之。而随机森林方法的优点是比较“稳重”，也就是会降低 variance 提升 bias。如果本来每棵树都已经是 low variance 的话，那么再采用投票的办法效果就会比较差。总之，随机森林中树的构造总的目标是增加多样性，所以应该长得高一些，否则无法刻画样本的细微之处。

2、 AdaBoost 算法

- **AdaBoost 算法简介：**

AdaBoost 的核心思想是利用同一组训练样本的不同加权版本，训练一组弱分类器，然后把这些弱分类器以加权的形式集成起来，形成一个强分类器。所以可见 AdaBoost 与 Random Forest 都是 Ensemble，目标都是获得多样性。相比之下，不同在于获取多样性的方法不同，AdaBoost 算法是通过 re-weighting 来达到目标的。具体而言 AdaBoost 通过提高那些被前一轮弱分类器错误分类样本的权重，而降低那些被正确分类的样本的权值，这样一来，那些没有得到正确分类的数据，由于其权值的加大而受到后一轮的分类器的更大关注。然后 AdaBoost 采用加权多数表决的方法组成最后的强分类器。具体地，加大分类误差率小的弱分类器的权

值，使其在表决中起到较大的作用，减小分类误差率大的弱分类器的权值，使其在表决中起较小的作用。所以 AdaBoost 算法的框架如图 1 所示。

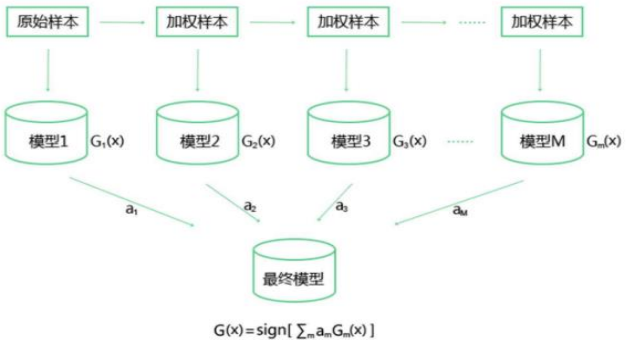


图 1　AdaBoost 算法框架图（图片来源于网络）

- **AdaBoost 算法具体流程如下：**

---

1. 初始化样本权重 $\boldsymbol{w_i = \frac{1}{m}, i = 1, 2, \ldots m}$
2. $\textbf{for t} = \textbf{1 to T}$

（1）使用样本权重$\boldsymbol{w_i}$训练一个分类器，计算当前分类器的误差：

$$\boldsymbol{\varepsilon_t = \sum w_i \ where \ f(x^{(i)}) \neq y^{(i)}}$$

（2）根据当前分类器的误差，计算其权重：

$$\boldsymbol{\alpha_t = \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}}$$

其中每棵树投票的权值为$\boldsymbol{\ln \alpha_t}$

（3）更新样本权重：

$$\textbf{incorrect: } \boldsymbol{w_i = w_i \times \alpha_t}; \ \textbf{correct: } \boldsymbol{w_i = w_i / \alpha_t}$$

（调整权重使得分对和分错的样本权重之和一样大）

（4）归一化处理 **normalization**

Q2：为什么要调整错分样本的权重与划分正确样本的权重一样大？

A：这样可以使上一棵树在当前权重下的分类效果等同于随机猜测，从而保证当前分类器模型与上一分类器不同，进而增加多样性。

- **AdaBoost 两个主要的特点：**

（1） 通过改变样本权重的方式训练新的弱分类器，后一个弱分类器基于前一个分类器的结果来训练。

（2） AdaBoost 能够自动学习多个弱分类器集成时的分类器权重。

● **AdaBoost 算法实例：**

下面用一个简单的例子来描述 AdaBoost 算法的具体过程：

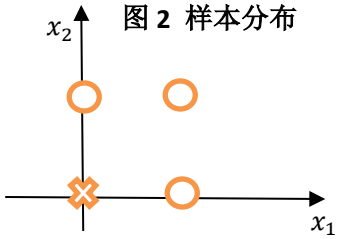
假设一个二类分类样本集 T 如下表 1 所示，

表 1 样本信息

x1	x2	y
0	0	0
0	1	1
1	0	1
1	1	1

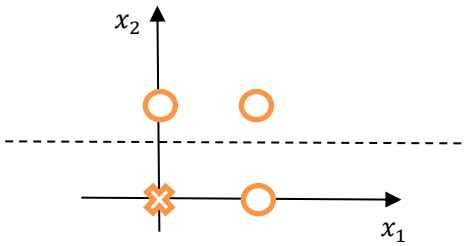
可以用图 2 表示其分布情况为：

图 2 样本分布



用决策树桩对这四个点来分类（只能平行于坐标轴来切分），先以属性x_2来分类，$x_2 = 1$则为正，$x_2 = 0$则为负，如图 3 所示。

图 3 划分情况



此时 AdaBoost 运行结果如表 2 所示。

表 2 运行结果

	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$	ε_t
初始权重 w	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	

分类结果	√	√	×	√	$\varepsilon_t = \frac{1}{4}$
调整权重 w	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{6}$	

然后构造第二个分类器，以属性x_1来分类，$x_1 = 1$则为正，$x_1 = 0$则为负，如图 4 所示。

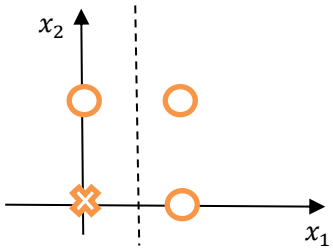


图 4 划分情况

此时 AdaBoost 运行结果如表 3 所示。

表 3 运行结果

	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$	ε_t
初始权重 w	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	
分类结果	√	√	×	√	$\varepsilon_t = \frac{1}{4}$
调整权重 w	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{6}$	
分类结果	√	×	√	√	$\varepsilon_t = \frac{1}{6}$
调整权重 w	$\frac{1}{10}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{10}$	

由上面的例子可以看出，(0,0)和(1,1)两个样本点很容易正确划分，所以其权重变得越来越小。而(0,1)和(1,0)权重变大。

Q3：(0,1)和(1,0)分别被错分一次，为什么权重大小却不一样？

A：(1,0)刚开始被正确划分，导致其权重减小。所以之后被错分时，其权重需要增大的比例相对较大，而(0,1)开始被错分，之后划分正确，所以二者的变化比例不一样，从而导致权重的结果不同。

Q4：如果一个分类器的准确率非常高或非常低，它的权重是怎样的？

A： 1.准确率很低相当于随机猜测，所以$\varepsilon_t = \frac{1}{2}$，$\alpha_t = 1$，所以该分类器权重为 0；

2.准确率很高，则ε_t趋于零，α_t趋于无穷，所以该分类器权重趋于无穷大。

Q5：AdaBoost 抗噪音能力差，若一个样本总被分错，它的权重不断增大怎么办？

A：可以令w_i的上升有个上限，即限制w_i的最大值。

● **Adaboost 算法优缺点补充总结：**

优点：

（1） 能够很大程度上防止过拟合；

（2） 与单个分类器相比，AdaBoost 能够有效提升分类精度；

（3） AdaBoost 提供的是一种框架，弱分类器可以选择决策树或者其他分类算法；

（4） 训练速度相对较快。

缺点：

（1） 由于集成了多个分类器，模型可解释性降低；

（2） 算法抗噪能力较差，对异常值比较敏感。

3、课后题思考

Q6：上面所讲算法为二分类器，若改为多分类器会怎样？

A：每个分类器的权重为$\ln \alpha_t$，为了使权重非负，应该使得每个分类器的准确率不小于 50%，所以在当前分类器的准确率小于 50%时，算法应该停止迭代。在二分类中，即使是随机猜测，准确率也有 50%，所以一般每棵树的准确率都会大于 50%。然而多分类器的随机猜测准确率较低，很有可能出现当前建立的树的分类准确率低于 50%，导致迭代停止，最终的分类准确率降低。

Q7：上面说到，Random Forest 算法中的树较高为好，那么 Adaboost 算法呢？

A：应该矮一点。因为 AdaBoost 是迭代算法，每一步是在弥补上一步的误差，所以算法是降低 bias 的，如果每棵树过高（low bias），则算法容易产生 overfitting。