Adaboost

Boosting（提升方法）是一族可以将弱分类器提升为强分类器的方法。这族算法的工作原理是：先从初始训练集训练出一个基分类器，再根据基分类器的表现对训练样本权值进行调整，使得被基分类器分错的样本得到更多的关注，然后根据调整权值后的样本来训练下一个基分类器，如此重复，直至基分类器的数目达到预先设定的阈值T，最终将这T个基分类器进行加权组合。
Boosting方法的理论基础是强可学习与弱可学习是等价的。强可学习是指存在一个多项式的学习算法能够得到一个正确率很高的分类器；弱可学习是指存在一个多项式的学习算法能够得到一个正确率仅比随机猜测略好的分类器。可以证明，强可学习与弱可学习是等价的。也就是说，对于一个数据集如果存在弱可学习算法，则可以通过一些办法将它提升为强可学习算法。
Adaboost算法是Boosting算法族最常用的算法，其有多种推导方式，比较容易理解的是基于”加性模型“，即基分类器的加权线性组合。Adaboost算法如下：

Input: Data set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$;

Base learning algorithm $\mathcal{L}$;

Number of learning rounds T.

Process:

$D_1(i) = 1/m.$ % Initialize the weight distribution

for $t = 1, \cdots, T$:

$h_t = \mathcal{L}(\mathcal{D}, D_t)$; % Train a weak learner h_t from $\mathcal{D}$ using distribution D_t

$\epsilon_t = \Pr_{i \sim D_i}[h_t(\boldsymbol{x}_i \neq y_i)]$; % Measure the error of h_t

$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$; % Determine the weight of h_t

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \exp(-\alpha_t) & \text{if } h_t(\boldsymbol{x}_i) = y_i \\ \exp(\alpha_t) & \text{if } h_t(\boldsymbol{x}_i) \neq y_i \end{cases}$$

$$= \frac{D_t(i) \exp(-\alpha_t y_i h_t(\boldsymbol{x}_i))}{Z_t}$$ % Update the distribution, where Z_t is

% a normalization factor which enables D_{t+1} be a distribution

end.

Output: $H(\boldsymbol{x}) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x})\right)$

图1 The AdaBoost Algorithm

简要描述其工作步骤：

1. 先从初始训练集训练出T个基分类器；

2. 再根据基分类器的表现选择错误率最小的基分类器作为当前轮迭代的基分类器，根据错误率对训练样本权重进行调整，使得被先前的基分类器误判的训练样本在后续受到更多的关注，同时需要计算出该轮所用基分类器的权重；
3. 然后基于调整后的样本权重来训练下一个基分类器；
4. 如此重复，直到基分类器数量达到给定的值T为止；
5. 最终将这T个基分类器进行加权组合得到最终的强分类器。

在Adaboost中分类器的错误率ϵ计算方法是被该分类器错分的样本权值之和。
Adaboost算法的两个核心步骤：

1. 权值调整：Adaboost算法提高那些被前一轮基分类器错误分类样本的权值，降低那些被前一轮基分类器正确分类样本的权值，从而使得那些没有正确分类的样本由于权值的加大而受到后一轮基分类器的更大关注。
2. 基分类器组合，Adaboost采用加权组合的方法：
    - 加大分类误差较小的弱分类器的权值，使得它在表决中起较大作用；
    - 减小分类误差较大的弱分类器的权值，使得它在表决中起较小作用。
      分类器权重计算方法：

$$\alpha = \frac{1}{2}\ln\frac{1-\epsilon}{\epsilon} = \ln\sqrt{\frac{1-\epsilon}{\epsilon}}$$

每次权值调整幅度应为多大？Boosting的思想是总权值的一半赋予被错误分类的样本，总权值的一半赋予剩余的样本。由于初始权值均匀分布在所有样本上且和为1，这样分配给被错误分类的样本权值等于错误率ϵ，后续将被错误分类的样本的权重更新为被错误分类的样本权重乘以$\frac{1}{2\epsilon}$，将被正确分类的样本的权重更新为被正确分类样本的权重乘以$\frac{1}{2(1-\epsilon)}$。
下面通过实例来说明Adaboost算法的整个过程，所用样本如下：

样本序号	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
样本	(1,5)	(2,2)	(3,1)	(4,6)	(6,8)	(6,5)	(7,9)	(8,7)	(9,8)	(10,2)
标签	1	1	-1	-1	1	-1	1	1	-1	-1

Base learning algorithm
本示例使用垂直坐标系的直线作为分类器，有如下三个弱分类器：

$$h_1 = \begin{cases} 1, & X < 2.5 \\ -1, & X \geq 2.5 \end{cases}$$

$$h_2 = \begin{cases} 1, & X < 8.5 \\ -1, & X \geq 8.5 \end{cases}$$

$$h_3 = \begin{cases} 1, & Y > 6.5 \\ -1, & Y \leq 6.5 \end{cases}$$

直观化的表示如图2所示：

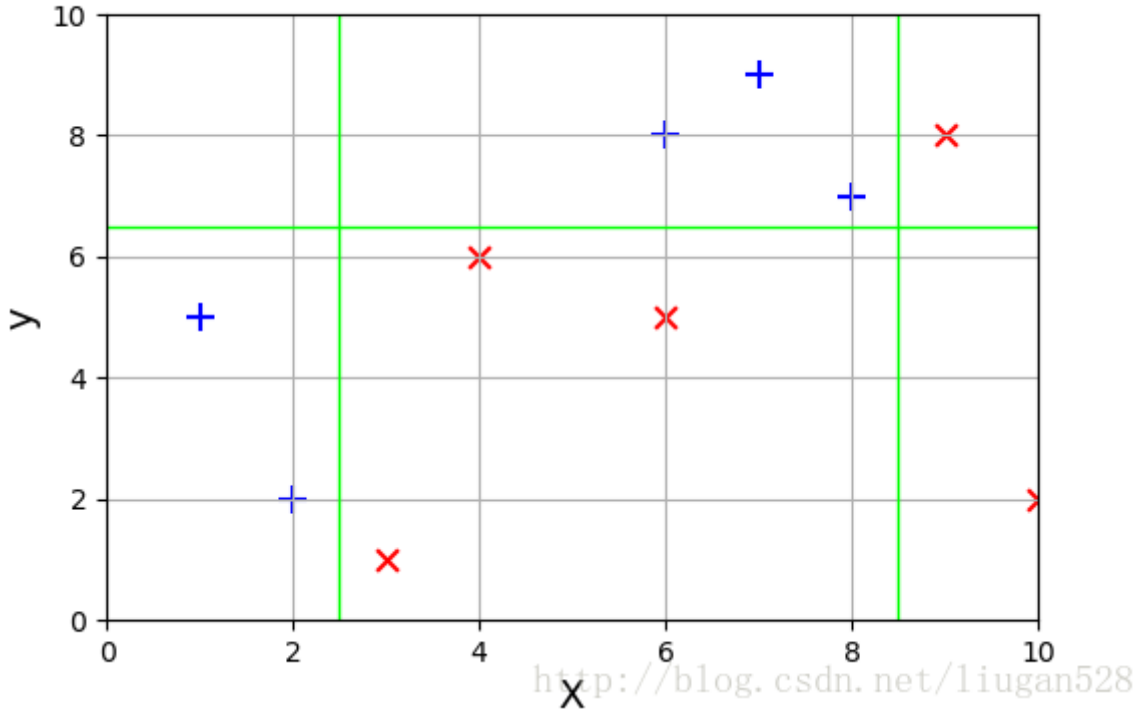

图2 样本&基分类器

本示例的总迭代次数$T = 3$

初始化：

首先给训练集的10个样本分配初始权值$\frac{1}{m}$，其中m为样本个数，在这里是10。每个样本的权值分布D_1如下：

样本序号	1	2	3	4	5	6	7	8	9	10
样本	(1,5)	(2,2)	(3,1)	(4,6)	(6,8)	(6,5)	(7,9)	(8,7)	(9,8)	(10,2)
标签	1	1	-1	-1	1	-1	1	1	-1	-1
权值D_1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

第一轮迭代$t = 1$得到第一个强分类器：

在样本权值分布为D_1的情况下，选择h_1、h_2、h_3三个中错误率最小的分类器作为第一个基

分类器H_1，由图2可知三个分类器分类错误率都是0.3，且0.3<0.5，因此可取三个中的任何一个，这里选择h_1作为H_1，即：

$$H_1 = \begin{cases} 1, & X < 2.5 \\ -1, & X \geq 2.5 \end{cases}$$

使用H_1进行分类时样本点(6,8)、(7,9)、(8,7)会被分错，H_1的错误率为：

$$\epsilon_1 = 0.1 + 0.1 + 0.1 = 0.3$$

根据ϵ_1计算H_1的权重

$$\alpha_1 = \ln \sqrt{\frac{1-\epsilon_1}{\epsilon_1}} = \ln \sqrt{\frac{1-0.3}{0.3}} = 0.4236$$

此时得到的强分类器为$sign(\alpha_1 H_1)$，即$sign(0.4236 H_1)$，此时的强分类器的训练错误率为0.3。
根据H_1的表现更新样本的权值，错误分类的样本有三个，因此这三个错误样本的权值之和是0.5，剩下的七个样本权值之和为0.5。因此三个被错误分类的样本(6,8)、(7,9)、(8,7)的权值均为$\frac{1}{2} \times \frac{1}{3}$，即$\frac{1}{6}$，剩余的七个样本权值均为$\frac{1}{2} \times \frac{1}{7}$，即$\frac{1}{14}$。
经过第一轮迭代之后样本的权值分布D_2为：

样本序号	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
样本	(1,5)	(2,2)	(3,1)	(4,6)	(6,8)	(6,5)	(7,9)	(8,7)	(9,8)	(10,2)
标签	1	1	-1	-1	1	-1	1	1	-1	-1
权值D_2	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{6}$	$\frac{1}{14}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{14}$	$\frac{1}{14}$

第二轮迭代$t = 2$得到第二个强分类器：
在样本权值分布为D_2的情况下，选择h_1、h_2、h_3三个中错误率最小的分类器作为第二个基分类器H_2：

- 选取h_1作为第二个基分类器时(6,8)、(7,9)、(8,7)会被分错，错误率为：

$$\epsilon = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

- 选取h_2作为第二个基分类器时(3,1)、(4,6)、(6,5)会被分错，错误率为：

$$\epsilon = \frac{1}{14} + \frac{1}{14} + \frac{1}{14} = \frac{3}{14}$$

- 选取h_3作为第二个基分类器时(9,8)、(1,5)、(2,2)会被分错，错误率为：

$$\epsilon = \frac{1}{14} + \frac{1}{14} + \frac{1}{14} = \frac{3}{14}$$

当选择h_2或h_3作为第二个基分类器时错误率均为$\frac{3}{14}$且$\frac{3}{14}$<0.5，可以选择h_2作为第二个基分类器H_2，即

$$H_2 = \begin{cases} 1, & X < 8.5 \\ -1, & X \geq 8.5 \end{cases}$$

由图2可知使用H_2进行分类时样本点(3,1)、(4,6)、(6,5)会被分错，H_2的错误率为：

$$\epsilon_2 = \frac{1}{14} + \frac{1}{14} + \frac{1}{14} = \frac{3}{14} < 0.5$$

根据ϵ_2计算H_2的权重

$$\alpha_2 = \ln \sqrt{\frac{1-\epsilon_2}{\epsilon_2}} = \ln \sqrt{\frac{1-\frac{3}{14}}{\frac{3}{14}}} = 0.6496$$

此时得到的强分类器为$sign(\alpha_1 H_1 + \alpha_2 H_2)$，即$sign(0.4236 H_1 + 0.6496 H_2)$，此时的强分类器的训练错误率仍为0.3。
对于被H_2分错的样本点(3,1)、(4,6)、(6,5)权重均更新为：$\frac{1}{2} \times \frac{1}{3}$，即$\frac{1}{6}$。
对于被H_2正确分类的样本点(1,5)、(2,2)、(9,8)、(10,2)，其上一轮的权重均为$\frac{1}{14}$，本轮将它们权重更新为$\frac{1}{14} \times \frac{1}{2\times(1-\epsilon_2)} = \frac{1}{14} \times \frac{1}{2\times(1-\frac{3}{14})} = \frac{1}{22}$。
对于被H_2正确分类的样本点(6,8)、(7,9)、(8,7)，其上一轮的权重均为$\frac{1}{6}$，本轮将它们权重更新为$\frac{1}{6} \times \frac{1}{2\times(1-\epsilon_2)} = \frac{1}{6} \times \frac{1}{2\times(1-\frac{3}{14})} = \frac{7}{66}$。
经过第二轮迭代之后样本的权值分布D_3为：

样本序号	1	2	3	4	5	6	7	8	9	10
样本	(1,5)	(2,2)	(3,1)	(4,6)	(6,8)	(6,5)	(7,9)	(8,7)	(9,8)	(10,2)
标签	1	1	-1	-1	1	-1	1	1	-1	-1
权值D_3	$\frac{1}{22}$	$\frac{1}{22}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{7}{66}$	$\frac{1}{6}$	$\frac{7}{66}$	$\frac{7}{66}$	$\frac{1}{22}$	$\frac{1}{22}$

第三轮迭代$t=3$得到第三个强分类器：

在样本权值分布为D_3的情况下，选择h_1、h_2、h_3三个中错误率最小的分类器作为第三个基分类器H_3：

- 选取h_1作为第三个基分类器时(6,8)、(7,9)、(8,7)会被分错，错误率为：

$$\epsilon=\frac{7}{66}+\frac{7}{66}+\frac{7}{66}=\frac{7}{22}$$

- 选取h_2作为第三个基分类器时(3,1)、(4,6)、(6,5)会被分错，错误率为：

$$\epsilon=\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{1}{2}$$

- 选取h_3作为第三个基分类器时(9,8)、(1,5)、(2,2)会被分错，错误率为：

$$\epsilon=\frac{1}{22}+\frac{1}{22}+\frac{1}{22}=\frac{3}{22}$$

当选择h_3作为第三个基分类器时错误率均为$\frac{3}{22}$最小且$\frac{3}{22}$<0.5，所以选择h_3作为第三个基分类器H_3，即

$$H_3=\begin{cases}1, & Y>6.5\\ -1, & Y\leq 6.5\end{cases}$$

由图2可知使用H_3进行分类时样本点(9,8)、(1,5)、(2,2)会被分错，H_3的错误率为：

$$\epsilon_3=\frac{1}{22}+\frac{1}{22}+\frac{1}{22}=\frac{3}{22}<0.5$$

根据ϵ_3计算H_3的权重

$$\alpha_3=\ln\sqrt{\frac{1-\epsilon_3}{\epsilon_3}}=\ln\sqrt{\frac{1-\frac{3}{22}}{\frac{3}{22}}}=0.9229$$

此时得到的强分类器为$sign(\alpha_1H_1+\alpha_2H_2+\alpha_3H_3)$，即$sign(0.4236H_1+0.6496H_2+0.9229H_3)$，此时的强分类器的训练错误率为0。

使用H_3进行分类时样本点(9,8)、(1,5)、(2,2)会被分错，它们的权重均更新为：$\frac{1}{2}\times\frac{1}{3}$，即$\frac{1}{6}$。

对于被H_3正确分类的样本点(3,1)、(4,6)、(6,5)，其上一轮的权重均为$\frac{1}{6}$，本轮将它们权重更新为$\frac{1}{6}\times\frac{1}{2\times(1-\epsilon_3)}=\frac{1}{6}\times\frac{1}{2\times(1-\frac{3}{22})}=\frac{11}{114}$。

对于被H_3正确分类的样本点(6,8)、(7,9)、(8,7)，其上一轮的权重均为$\frac{7}{66}$，本轮将它们权

重更新为$\frac{7}{66} \times \frac{1}{2\times(1-\epsilon_3)} = \frac{7}{66} \times \frac{1}{2\times(1-\frac{3}{22})} = \frac{7}{114}$。
对于被H_3正确分类的样本点(10,2)，其上一轮的权重为$\frac{1}{22}$，本轮将它们权重更新为$\frac{1}{22} \times \frac{1}{2\times(1-\epsilon_3)} = \frac{1}{22} \times \frac{1}{2\times(1-\frac{3}{22})} = \frac{1}{38}$。
经过第三轮迭代之后样本的权值分布D_4为：

样本序号	1	2	3	4	5	6	7	8	9	10
样本	(1,5)	(2,2)	(3,1)	(4,6)	(6,8)	(6,5)	(7,9)	(8,7)	(9,8)	(10,2)
标签	1	1	-1	-1	1	-1	1	1	-1	-1
权值D_4	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{11}{114}$	$\frac{11}{114}$	$\frac{7}{114}$	$\frac{11}{114}$	$\frac{7}{114}$	$\frac{7}{114}$	$\frac{1}{6}$	$\frac{1}{38}$

由于三个基分类器都用完了，完成了三次迭代之后算法终止。
因此最终得到的强分类器为：

$$H = sign(\sum_{t=1}^{T} \alpha_t H_t) = sign(0.4236H_1 + 0.6496H_2 + 0.9229H_3)$$

可以验证最终的强分类器H的错误率为0，例如对于点(9,8)：

$$H = sign(0.4236 * (-1) + 0.6496 * (-1) + 0.9229 * 1) = sign(-0.1503) = -1$$

所以判定点(9,8)属于-1的类别，判断正确。

Adaboost算法优缺点
优点：可解释性强；不容易过拟合；不用做特征筛选；可以使用很多方法（决策树、SVM等）构建基分类器。
缺点：对于离群值比较敏感，在Adaboost训练过程中，Adaboost会使得离群样本的权值呈指数增长，训练将会过于偏向这类困难的样本，导致Adaboost算法易受噪声干扰。

参考文献

周志华. 机器学习 [D]. 清华大学出版社，2016.
华校专、王正林. Python大战机器学习 [D]. 电子工业出版社，2017.
Peter Flach著、段菲译. 机器学习 [D]. 人民邮电出版社，2016.

Xindong Wu, Vipin Kumar, et al. Top 10 algorithms in data mining[J]. Knowl Inf Syst, 2008, 14:1–37.