（十三）通俗易懂理解——Adaboost算法原理

看了一堆的博客，都很难找到能够综合起来完全理解这个原理，直到这位博主文章的出现……不多说了，直接上。

本博客将会详细介绍AdaBoost算法过程，并给出了一个Adaboost例子的详细求解过程，当然也给出了Matlab代码求解过程。碍于太多复杂公式，文章是在电脑Word文档写好再复制上博客的，为了排版好看，有些地方给出了截图。

下面给出几个我认为不错的博客资料：

【1】*http://blog.csdn.net/v_july_v/article/details/40718799*感谢这位博主给出了 Adaboost算法的原理与推导，本文章很多地方都参考了他的内容

【2】*http://blog.csdn.net/m0_37407756/article/details/67637400*该博客有一个Adaboost算法的例子，但其过程简略太多，初学者很难看懂。本文章的Adaboost 算法例子也是与之相对应的，但本人给出了详细的步骤和分析过程。话说，图都是我一个一个画上去，心疼我用了两天时间！！

一、AdaBoost简介

Boosting, 也称为增强学习或提升法，是一种重要的集成学习技术， 能够将预测精度仅比随机猜度略高的弱学习器增强为预测精度高的强学习器，这在直接构造强学习器非常困难的情况下，为学习算法的设计提供了一种有效的新思路和新方法。其中最为成功应用的是，Yoav Freund和Robert Schapire在1995年提出的AdaBoost算法。
AdaBoost是英文"Adaptive Boosting"（自适应增强）的缩写，它的自适应在于：前一个基本分类器被错误分类的样本的权值会增大，而正确分类的样本的权值会减小，并再次用来训练下一个基本分类器。同时，在每一轮迭代中，加入一个新的弱分类器，直到达到某个预定的足够小的错误率或达到预先指定的最大迭代次数才确定最终的强分类器。
Adaboost算法可以简述为三个步骤：
（1）首先，是初始化训练数据的权值分布D1。假设有N个训练样本数据，则每一个训练样本最开始时，都被赋予相同的权值：w1=1/N。
（2）然后，训练弱分类器hi。具体训练过程中是：如果某个训练样本点，被弱分类器hi准确地分类，那么在构造下一个训练集中，它对应的权值要减小；相反，如果某个训练样本点被错误分类，那么它的权值就应该增大。权值更新过的样本集被用于训练下一个分类器，整个训练过程如此迭代地进行下去。
（3）最后，将各个训练得到的弱分类器组合成一个强分类器。各个弱分类器的训练过程结束后，加大分类误差率小的弱分类器的权重，使其在最终的分类函数中起着较大的决定作用，而降低分类误差率大的弱分类器的权重，使其在最终的分类函数中起着较小的决定作用。
换而言之，误差率低的弱分类器在最终分类器中占的权重较大，否则较小。

二、AdaBoost算法过程

给定训练数据集：（x1,y1）,(x2,y2)···(xn,yn)，其中yi属于{1，-1}用于表示训练样本的类别标签，*i=1,...,N*。Adaboost的目的就是从训练数据中学习一系列弱分类器或基本分类器，然后将这些弱分类器组合成一个强分类器。

相关符号定义：


$D_t(i)$：训练样本集的权值分布；
w_i：每个训练样本的权值大小；
h：弱分类器，
H：基本分类器；
H_{final}=最终的强分类器
e：误差率；
α_t：弱分类器的权重；

Adaboost的算法流程如下：

（1）首先，初始化训练数据的权值分布。每一个训练样本最开始时都被赋予相同的权值：$w_i=1/N$，这样训练样本集的初始权值分布 $D_1(i)$：

$$D_1(i)=(w_1,w_2,...w_N)=(\frac{1}{N},....,\frac{1}{N})$$

（2）进行迭代 $t=1,...,T$

（a）选取一个当前误差率最低的弱分类器 h 作为第 t 个基本分类器 Ht，并计算弱分类器 $h_t:X\rightarrow\{-1,1\}$，该弱分类器在分布 D_t 上的误差为：

$$e_t=P(H_t(x_i)\neq y_i)=\sum_{i=1}^{N}w_{ti}I(H_t(x_i)\neq y_i)$$

PS：由上述式子可知，$H_t(x)$在训练数据集上的误差率 e_t就是被 $H_t(x)$误分类样本的权值之和。

（b）计算该弱分类器在最终分类器中所占的权重（弱分类器权重用α表示）：

$$\alpha_t=\frac{1}{2}\ln\left(\frac{1-e_t}{e_t}\right)$$

（c）更新训练样本的权值分布 D_{t+1}：

$$D_{t+1}=\frac{D_t(i)\exp(-\alpha_t y_i H_t(x_i))}{Z_t}$$

其中 Z_t为归一化常数 $Z_t=2\sqrt{e_t(1-e_t)}$

（3）最后，按弱分类器权重α_t组合各个弱分类器，即

$$f(x)=\sum_{t=1}^{T}\alpha_t H_t(x)$$

通过符号函数 $sign$ 的作用，得到一个强分类器为：

$$H_{final}=sign(f(x))=sign\left(\sum_{t=1}^{T}\alpha_t H_t(x)\right).$$

相关说明：

因为权重更新依赖于α,而α又依赖于误差率e，所以我们可以直接将权重更新公式用e表示。样本权重更新公式：$D_{t+1}(i)=\dfrac{D_t(i)\exp(-\alpha_t y_i H_t(x_i))}{Z_t}$，其中

$Z_t=2\sqrt{e_t(1-e_t)}$

（1）当样本分错时，$y_i H_t(x_i)=-1$

$$D_{t+1}(i)=\frac{D_t(i)}{Z_t}\exp(\alpha_1)=\frac{D_t(i)}{Z_t}\exp(\frac{1}{2}\ln(\frac{1-e_1}{e_1}))=\frac{D_t(i)}{Z_t}\sqrt{\frac{1-e_t}{e_t}}=\frac{D_t(i)}{2\sqrt{e_t(1-e_t)}}\sqrt{\frac{1-e_t}{e_t}}=\frac{D_t(i)}{2e_t}$$

（2）当样本分对时，$y_i H_t(x_i)=1$

$$D_{t+1}(i)=\frac{D_t(i)}{Z_t}\exp(-\alpha_1)=\frac{D_t(i)}{Z_t}\exp(-\frac{1}{2}\ln(\frac{1-e_1}{e_1}))=\frac{D_t(i)}{Z_t}\sqrt{\frac{e_t}{1-e_t}}=\frac{D_t(i)}{2\sqrt{e_t(1-e_t)}}\sqrt{\frac{e_t}{1-e_t}}=\frac{D_t(i)}{2(1-e_t)}$$

综合上面的推导，可得样本分错与分对时，其权值更新的公式为：

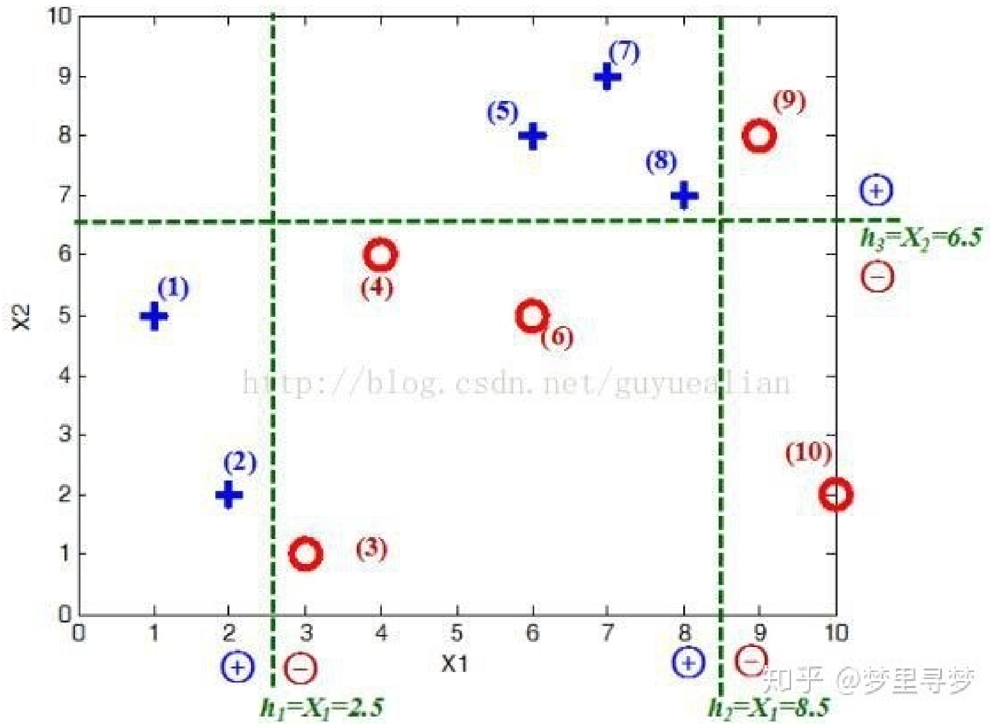
错误分类样本，权值更新： $D_{t+1}(i)=\dfrac{D_t(i)}{2\varepsilon_t}$

正确分类样本，权值更新： $D_{t+1}(i)=\dfrac{D_t(i)}{2(1-\varepsilon_t)}$

三、**AdaBoost**实例讲解

例：给定如图所示的训练样本，弱分类器采用平行于坐标轴的直线，用**Adaboost**算法的实

现强分类过程。



数据分析：

将这10个样本作为训练数据，根据 X 和Y 的对应关系，可把这10个数据分为两类，图中用“+”表示类别1，用“O”表示类别-1。本例使用水平或者垂直的直线作为分类器，图中已经给出了三个弱分类器，即：

$$h_1=\begin{cases}1\ ,X_1<2.5\\-1,X_1>2.5\end{cases},\quad h_2=\begin{cases}1\ ,X_1<8.5\\-1,X_1>8.5\end{cases},\quad h_3=\begin{cases}1\ ,X_2>6.5\\-1,X_2<6.5\end{cases}$$

初始化：

首先需要初始化训练样本数据的权值分布，每一个训练样本最开始时都被赋予相同的权值：$wi=1/N$，这样训练样本集的初始权值分布$D1(i)$：

令每个权值$w1i$ = $1/N$ = 0.1，其中，N = 10，i = 1,2, ..., 10，然后分别对于t= 1,2,3, ...等值进行迭代（t表示迭代次数，表示第t轮），下表已经给出训练样本的权值分布情况：

样本序号	1	2	3	4	5	6	7	8	9	10
样本点 X	(1,5)	(2,2)	(3,1)	(4,6)	(6,8)	(6,5)	(7,9)	(8,7)	(9,8)	(10,2)
类别 Y	1	1	-1	-1	1	-1	1	1	-1	-1
权值分布 D_1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

第1次迭代t=1：

初试的权值分布$D1$为1/N（10个数据，每个数据的权值皆初始化为0.1），

D1=[0.1, 0.1, 0.1, 0.1, 0.1, 0.1,0.1, 0.1, 0.1, 0.1]

在权值分布$D1$的情况下，取已知的三个弱分类器$h1$、$h2$和$h3$中误差率最小的分类器作为第1个基本分类器$H1(x)$（三个弱分类器的误差率都是0.3，那就取第1个吧）

$$H_1=\begin{cases}1\ ,X_1<2.5\\-1,X_1>2.5\end{cases}$$

PS：某个分类器的误差率等于该分类器的被**错分类样本的权值之和**

在分类器$H1(x)$=$h1$情况下，样本点“5 7 8”被错分，因此基本分类器$H1(x)$的误差率为：

误差率为：

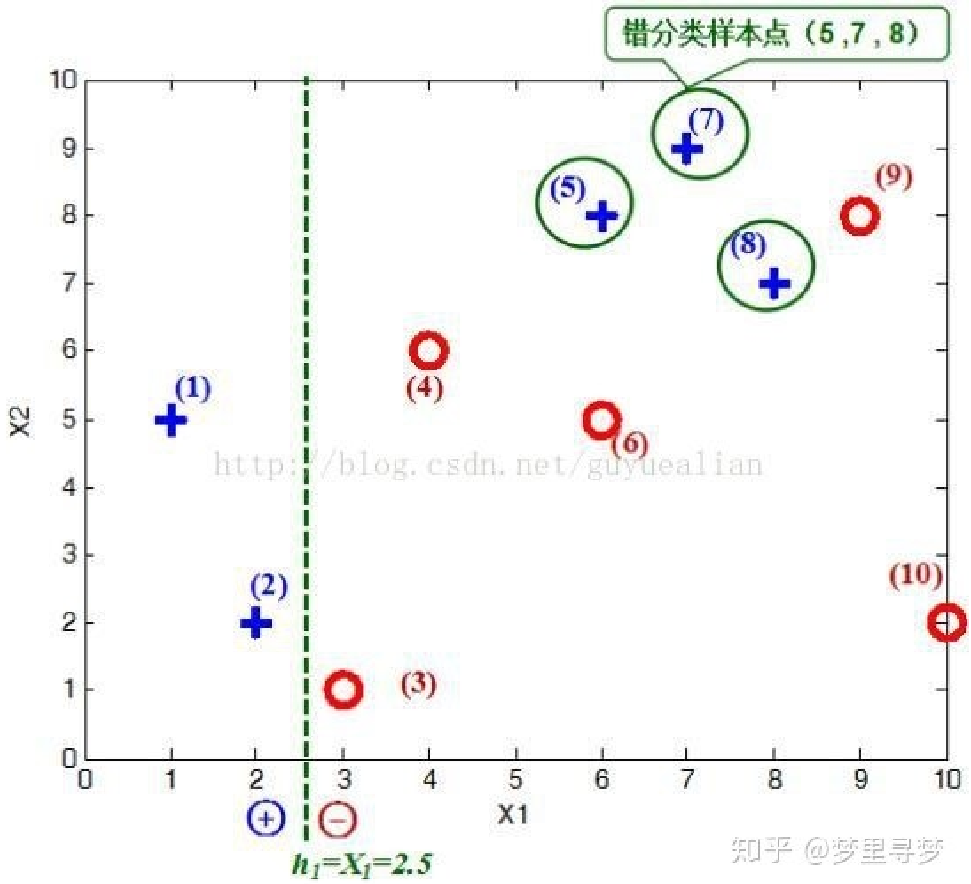
$$e_1=(0.1+0.1+0.1)=0.3$$

根据误差率 e_1 计算 H_1 的权重：

$$\alpha_1=\frac{1}{2}\ln\left(\frac{1-e_1}{e_1}\right)=\frac{1}{2}\ln\left(\frac{1-0.3}{0.3}\right)=0.4236$$

PS：这个 α_1 代表 $H_1(x)$ 在最终的分类函数中所占的权重为 0.4236

可见，被误分类样本的权值之和影响误差率e，误差率e影响基本分类器在最终分类器中所占的权重α。



然后，更新训练样本数据的权值分布，用于下一轮迭代，对于正确分类的训练样本“1 2 3 4 6 9 10”（共7个）的权值更新为：

$$D_2=\frac{D_1}{2(1-\varepsilon_1)}=\frac{1}{10}\times\frac{1}{2\times(1-0.3)}=\frac{1}{14}$$

PS：可见，**正确分类**的样本的权值由原来 1/10 减小到 1/14。

对于所有错误分类的训练样本“5 7 8”（共 3 个）的权值更新为：

$$D_2(i)=\frac{D_1(i)}{2e_1}=\frac{1}{10}\times\frac{1}{2\times 03}=\frac{1}{6}$$

PS：可见，**错误分类**的样本的权值由原来 1/10 增大到 1/6。

这样，第1轮迭代后，最后得到各个样本数据新的权值分布：

D2=[1/14,1/14,1/14,1/14,1/6,1/14,1/6,1/6,1/14,1/14]

由于样本数据“5 7 8”被H_1(x)分错了，所以它们的权值由之前的0.1增大到1/6；反之，其它数据皆被分正确，所以它们的权值皆由之前的0.1减小到1/14，下表给出了权值分布的变换情况：

样本序号	1	2	3	4	5	6	7	8	9	10
样本点 X	(1,5)	(2,2)	(3,1)	(4,6)	(6,8)	(6,5)	(7,9)	(8,7)	(9,8)	(10,2)
类别 Y	1	1	-1	-1	1	-1	1	1	-1	-1
权值分布 $D1$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
权值分布 $D2$	1/14	1/14	1/14	1/14	1/6	1/14	1/6	1/6	1/14	1/14
$sign(f_1(x))$	1	1	-1	-1	-1	-1	-1	-1	-1	-1

被 H_1(x)分错的样本

PS：用浅绿色底纹标记的表格，是被 H_1(x)分错的样本“5 7 8”；没有底纹（白色的）是正确分类的样本

可得分类函数：$f_1(x) = \alpha_1 H_1(x) = 0.4236 H_1(x)$。此时，组合一个基本分类器$sign(f_1(x))$作为强分类器在训练数据集上有3个误分类点（即5 7 8），此时强分类器的训练错误为：0.3

第二次迭代t=2：

在权值分布D2的情况下，再取三个弱分类器h1、h2和h3中误差率最小的分类器作为第2个基本分类器H2(x)：

① 当取弱分类器h1=X1=2.5时，此时被错分的样本点为“5 7 8”：

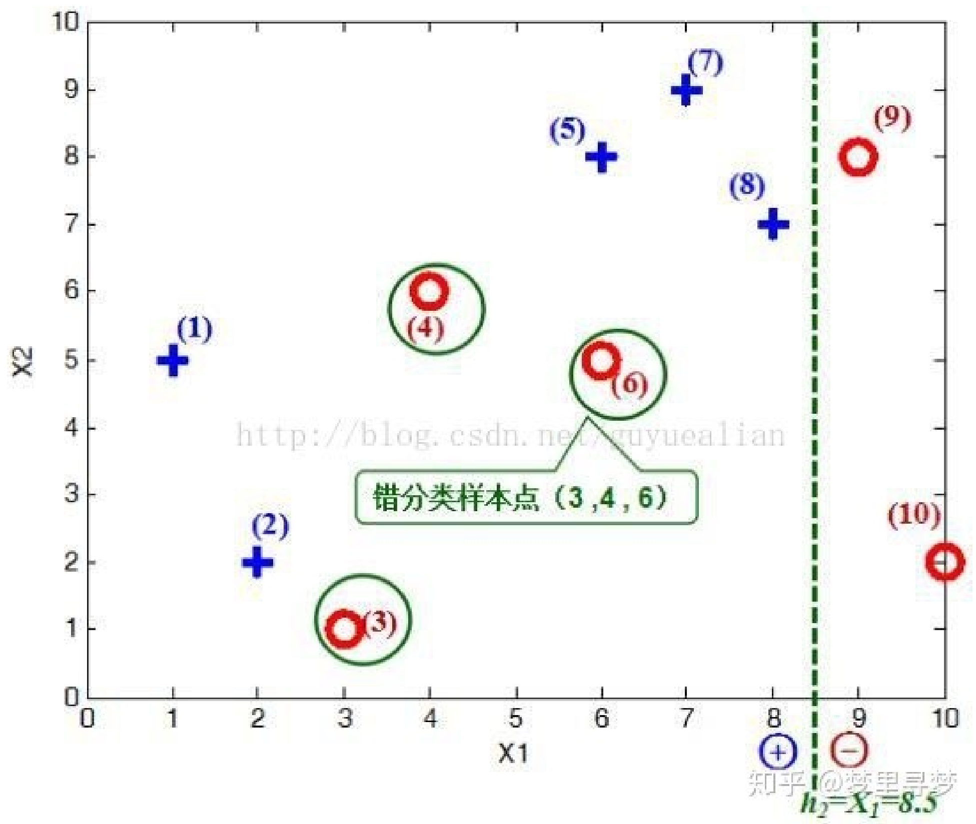
误差率e=1/6+1/6+1/6=3/6=1/2；

② 当取弱分类器h2=X1=8.5时，此时被错分的样本点为“3 4 6”：

误差率e=1/14+1/14+1/14=3/14；

③ 当取弱分类器h3=X2=6.5时，此时被错分的样本点为“1 2 9”：

误差率e=1/14+1/14+1/14=3/14；



因此，取当前最小的分类器h2作为第2个基本分类器H2(x)

$$H_2 = \begin{cases} 1\ , X_1 < 8.5 \\ -1, X_1 > 8.5 \end{cases}$$

显然，$H2(x)$把样本“3 4 6”分错了，根据$D2$可知它们的权值为$D2(3)=1/14$，$D2(4)=1/14$，$D2(6)=1/14$，所以$H2(x)$在训练数据集上的误差率：

$$e_2 = P(H_2(x_i) \neq y_i) = 3 \times \frac{1}{14} = \frac{3}{14} \text{（即权值之和）}$$

根据误差率 e_2 计算 H_2 的权重：

$$\alpha_2 = \frac{1}{2}\ln(\frac{1-e_2}{e_2}) = 0.6496$$

更新训练样本数据的权值分布，对于**正确分类**的样本权值更新为：

$$D_3(i) = \frac{D_2(i)}{2(1-e_2)} = \frac{7}{11} D_2(i),$$

对于**错误分类**的权值更新为：

$$D_3(i) = \frac{D_2(i)}{2e_2} = \frac{7}{3} D_2(i)$$

这样，第2轮迭代后，最后得到各个样本数据新的权值分布：

D3=[1/22,1/22,1/6,1/6,7/66,1/6,7/66,7/66,1/22,1/22]

下表给出了权值分布的变换情况：

序号	1	2	3	4	5	6	7	8	9	10
样本点 X	(1,5)	(2,2)	(3,1)	(4,6)	(6,8)	(6,5)	(7,9)	(8,7)	(9,8)	(10,2)
类别 Y	1	1	-1	-1	1	-1	1	1	-1	-1
权值分布 $D1$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
权值分布 $D2$	1/14	1/14	1/14	1/14	1/6	1/14	1/6	1/6	1/14	1/14
$sign(f_1(x))$	1	1	-1	-1	-1	-1	-1	-1	-1	-1
权值分布 $D3$	1/22	1/22	1/6	1/6	7/66	1/6	7/66	7/66	1/22	1/22
$sign(f_2(x))$	1	1	1	1	1	1	1	1	-1	-1

被 H_2(x)分错的样本

可得**分类函数**：$f2(x)=0.4236H1(x) + 0.6496H2(x)$。此时，组合两个基本分类器$sign(f2(x))$作为强分类器在训练数据集上有3个误分类点（即3 4 6），此时强分类器的训练错误为：0.3

第三次迭代t=3:

在权值分布D3的情况下，再取三个弱分类器h1、h2和h3中误差率最小的分类器作为第3个基本分类器H3(x)：

① 当取弱分类器h1=X1=2.5时，此时被错分的样本点为“5 7 8”：

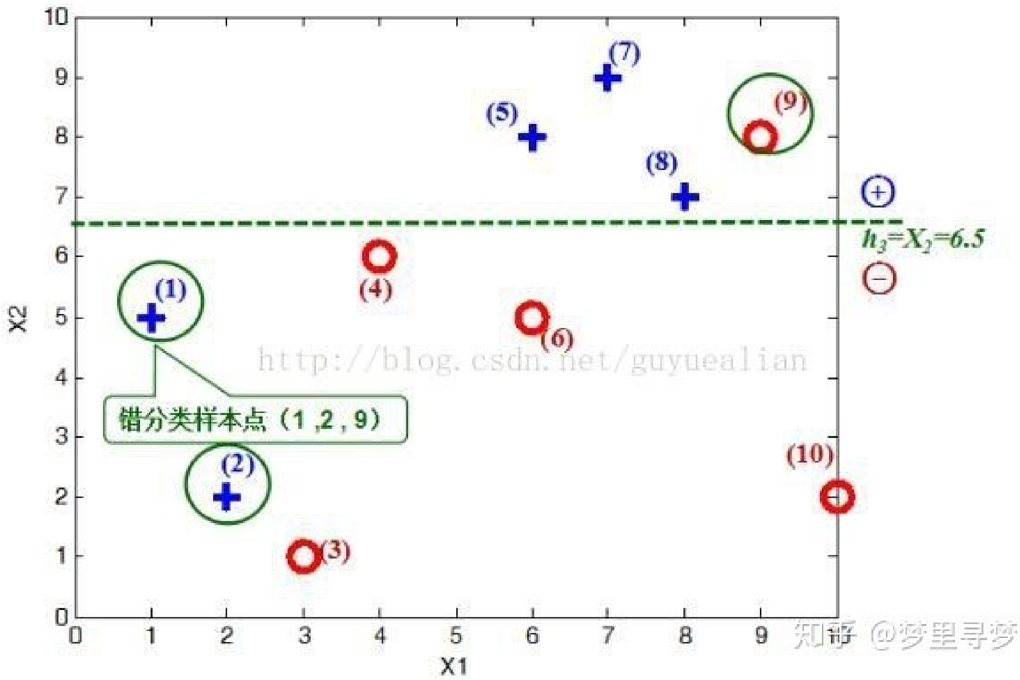
误差率e=7/66+7/66+7/66=7/22；

② 当取弱分类器h2=X1=8.5时，此时被错分的样本点为“3 4 6”：

误差率e=1/6+1/6+1/6=1/2=0.5；

③ 当取弱分类器h3=X2=6.5时，此时被错分的样本点为“1 2 9”：

误差率e=1/22+1/22+1/22=3/22；



因此，取当前最小的分类器$h3$作为第3个基本分类器$H3(x)$：

$$H_3(x)=\begin{cases}1\ ,X_2>6.5\\-1,X_2<6.5\end{cases}$$

此时被 $H_3(x)$误分类的样本是"1 2 9"，根据 D_3 可知它们的权值为 $D_3(1)=1/22$，$D_3(2)=1/22$， $D_3(9)=1/22$，所以 $H_3(x)$在训练数据集上的误差率：

$$e_3=P(H_3(x_i)\neq y_i)=3\times\frac{1}{22}=\frac{3}{22}$$ （即权值之和）

根据误差率 e_3 计算 H_3 的权重：

$$\alpha_3=\frac{1}{2}\ln(\frac{1-e_3}{e_3})=0.9229$$

更新训练样本数据的权值分布，对于**正确分类**的样本权值更新为：

$$D_4(i)=\frac{D_3(i)}{2(1-e_3)}=\frac{11}{19}D_3(i)$$

对于**错误分类**的权值更新为：

$$D_4(i)=\frac{D_3(i)}{2e_3}=\frac{11}{3}D_3(i)$$

这样，第3轮迭代后，得到各个样本数据新的权值分布为：

D4=[1/6,1/6,11/114,11/114,7/114,11/114,7/114,7/114,1/6,1/38]

下表给出了权值分布的变换情况：

序号	1	2	3	4	5	6	7	8	9	10
样本点 X	(1,5)	(2,2)	(3,1)	(4,6)	(6,8)	(6,5)	(7,9)	(8,7)	(9,8)	(10,2)
类别 Y	1	1	-1	-1	1	-1	1	1	-1	-1
权值分布 $D1$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
权值分布 $D2$	1/14	1/14	1/14	1/14	1/6	1/14	1/6	1/6	1/14	1/14
$sign(f_1(x))$	1	1	-1	-1	-1	-1	-1	-1	-1	-1
权值分布 $D3$	1/22	1/22	1/6	1/6	7/66	1/6	7/66	7/66	1/22	1/22
$sign(f_2(x))$	1	1	1	1	1	1	1	1	-1	-1
权值分布 $D4$	1/6	1/6	11/114	11/114	7/114	11/114	7/114	7/114	1/6	1/38
$sign(f_3(x))$	1	1	-1	-1	1	-1	1	1	-1	-1

可得分类函数：$f3(x)=0.4236H1(x)+0.6496H2(x)+0.9229H3(x)$。此时，组合三个基本分类

器*sign*(*f*3(x))作为强分类器，在训练数据集上有0个误分类点。至此，整个训练过程结束。

整合所有分类器，可得最终的强分类器为：

$$H_{final} = sign\left(\sum_{t=1}^{T} \alpha_t H_t(x)\right) = sign\left(0.4236H_1(x) + 0.6496H_2(x) + 0.9229H_3(x)\right)$$

这个强分类器***Hfinal***对训练样本的错误率为**0**！

三、**AdaBoost**的优点和缺点

优点

（1）Adaboost提供一种框架，在框架内可以使用各种方法构建子分类器。可以使用简单的弱分类器，不用对特征进行筛选，也不存在过拟合的现象。

（2）Adaboost算法不需要弱分类器的先验知识，最后得到的强分类器的分类精度依赖于所有弱分类器。无论是应用于人造数据还是真实数据，Adaboost都能显著的提高学习精度。

（3）Adaboost算法不需要预先知道弱分类器的错误率上限，且最后得到的强分类器的分类精度依赖于所有弱分类器的分类精度，可以深挖分类器的能力。Adaboost可以根据弱分类器的反馈，自适应地调整假定的错误率，执行的效率高。

（4）Adaboost对同一个训练样本集训练不同的弱分类器，按照一定的方法把这些弱分类器集合起来，构造一个分类能力很强的强分类器，即“三个臭皮匠赛过一个诸葛亮”。

缺点：

在Adaboost训练过程中，Adaboost会使得难于分类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。此外，Adaboost依赖于弱分类器，而弱分类器的训练时间往往很长。

Adaboost算法的某些特性是非常好的，这里主要介绍Adaboost的两个特性。（1）是训练的错误率上界，随着迭代次数的增加，会逐渐下降；（2）是Adaboost算法即使训练次数很多，也不会出现过拟合的问题。

如果有不懂的地方，欢迎咨询：

原文博主文章：https://blog.csdn.net/guyuealian/article/details/70995333

优秀博文：http://www.cnblogs.com/willnote/p/6801496.html

优秀博文：https://blog.csdn.net/v_july_v/article/details/40718799