AdaBoost 算法

By Xian2207, 13689903575, wszhangxian@126.com

1 介绍

AdaBoost 是 Adaptive Boosting 算法的简称，即自适应提升法。该算法特点是通过训练若干弱分类器，然后将弱分类器组合成强分类器进行分类。训练时，各个弱分类器之间是串行训练的，当前弱分类器的训练依赖于上一轮弱分类器的训练结果。各个弱分类器的权重是不同的，效果好的弱分类器的权重大，效果差的弱分类器的权重小。值得注意的是，AdaBoost 不止适用于分类模型，也可以用来训练回归模型。这需要将弱分类器替换成回归模型，并改动损失函数。

2 缺点

该算法多解决二分类问题。遇到多分类的情况，需要借助 one-versus-rest 的思想来训练多分类模型。

3 原理

Step 1: 训练当前最优的弱分类器

最优弱分类器是错误率最小的那个弱分类器。错误率的计算公式是

$$e_m = \sum_{i=1}^{N} w_{mi} I(G_m(x_i) \neq y_i)$$

e_m是第m次迭代的错误率，w_{mi}代表第i个样本在第m次迭代时的权值是w，I是指示函数，其值为 1 或 0. 当I括号中的表达式为真时，I的结果为 1；当I括号中的表达式为假时，I结果为 0. 取错误率最低的弱分类器为当前迭代的最优弱分类器。注意，第一轮迭代时，每个样本的权重初始化为总样本数分之一，即$\frac{1}{N}$

Step 2: 计算最优弱分类器的权重

$$\alpha_m = \frac{1}{2} \cdot \log \frac{1 - e_m}{e_m}$$

α_m代表在第m次迭代时最优弱分类器的权重。错误率越小，该弱分类器权值越大；错误率越大，权值越小；

Step 3: 根据错误率更新样本权重

样本权重的更新与当前样本权重和弱分类器的权重有关

$$w_{m+1,i} = \frac{w_{m,i}}{z_m} e^{-\alpha_m y_i G_m(x_i)}$$

$$z_m = \sum_{i}^{N} w_{m,i} \cdot e^{-\alpha_m y_i G_m(x_i)}$$

当样本被正确分类时，y_i和G_m取值一致，则样本在$m+1$次迭代时，权重变小；当样本被错误分类时，y_i和G_m取值不一致，则样本在$m+1$次迭代时，权重变大。这样处理可以使被错误分类的样本权重变大，从而在下一轮迭代中得到重视。

Step 4: 迭代终止条件

不断重复 1,2,3 步骤，直到达到终止条件为止。终止条件是强分类器的错误率低于最低错误率阈值或达到最大迭代次数。

4 实例

实例用到的数据集如下表所示。为方便说明，本文所用弱分类器为形如 x<1.5,则 y=1，否则 y=-1 的简单分类算法

Table 1 数据集

x	0	1	2	3	4	5
Y	1	1	-1	-1	1	-1

第一次迭代$m=1$，设置分类器切分点为 0.5, 1.5, 2.5, 3.5, 4.5; 样本总数 N=6，每个样本的权值初始值为$w_{1,i}=\frac{1}{6}=0.167$;

[1]计算错误率

按照 0.5 分割，得弱分类器$x<0.5$时，$y=1$；$x>0.5$时，$y=-1$，根据表 1，$x=1$ 和 $x=4$ 时$y=1$分类错误，故根据错误率公式

$$e_m=\sum_{i=1}^{N} w_{mi} I(G_m(x_i)\neq y_i)$$

可知，错误率为$2*0.167=0.334$;

按照 1.5 分割，得弱分类器$x<1.5$时，$y=1$；$x>1.5$时，$y=-1$，根据表 1，$x=4$时$y=1$分类错误，故错误率为$1*0.167=0.167$;

按照 2.5 分割，得弱分类器$x<2.5$时，$y=1$；$x>2.5$时，$y=-1$，根据表 1，$x=2$ 和 $x=4$ 时$y=1$分类错误，故错误率为$2*0.167=0.334$;

按照 3.5 分割，得弱分类器$x<3.5$时，$y=1$；$x>3.5$时，$y=-1$，根据表 1，$x=2$，$x=3$ 和 $x=4$时，y分类错误，故错误率为$3*0.167=0.501$;

按照 4.5 分割，得弱分类器$x<4.5$时，$y=1$；$x>4.5$时，$y=-1$，根据表 1，$x=2$ 和 $x=3$ 时$y=1$分类错误，故错误率为$2*0.167=0.334$;

[2]计算最优分类器权重

$$\alpha_m=\frac{1}{2}\cdot\log\frac{1-e_m}{e_m}$$

根据错误率越小，该最优分类器权重越大的推论，这里选择错误率 0.167 计算，

$$\alpha_1=\frac{1}{2}\cdot\log\frac{1-0.167}{0.167}=0.8047$$

[3]更新样本的权值

$$z_m=\sum_{i}^{N} w_{m,i}\cdot e^{-\alpha_m y_i G_m(x_i)}$$

$$w_{m+1,i}=\frac{w_{m,i}}{z_m}e^{-\alpha_m y_i G_m(x_i)}$$

当取 0.167 这一错误率时，$x=0,1,2,3,5$时，y分类正确，根据上述公式，$i=1,2,3,4,6$时

$$w_{1,i}\cdot e^{-\alpha_1 y_i G_1(x_i)}=\frac{1}{6}e^{-0.8047\cdot 1\cdot 1}=0.167*\exp(-0.8047)=0.075$$

$x=4$时，y分类错误，$i=5$时

$$w_{1,5}\cdot e^{-\alpha_1 y_4 G_4(x_4)}=\frac{1}{6}e^{-0.8047\cdot 1\cdot -1}=0.167*\exp(0.8047)=0.373$$

新样本权重总和为

$$z_1 = \sum_i^N w_{1,i} \cdot e^{-\alpha_1 y_i G_1(x_i)} = 0.075 * 5 + 0.373 = 0.748$$

进而可知

$$w_{1+1,1} = \frac{w_{1,1}}{z_1} e^{-\alpha_1 y_1 G_1(x_1)} = \frac{0.075}{0.748} = 0.1$$

$$w_{1+1,2} = \frac{w_{1,2}}{z_1} e^{-\alpha_1 y_2 G_1(x_2)} = \frac{0.075}{0.748} = 0.1$$

$$w_{1+1,3} = \frac{w_{1,3}}{z_1} e^{-\alpha_1 y_3 G_1(x_3)} = \frac{0.075}{0.748} = 0.1$$

$$w_{1+1,4} = \frac{w_{1,4}}{z_1} e^{-\alpha_1 y_4 G_1(x_4)} = \frac{0.075}{0.748} = 0.1$$

$$w_{1+1,5} = \frac{w_{1,5}}{z_1} e^{-\alpha_1 y_5 G_1(x_5)} = \frac{0.373}{0.748} = 0.5$$

$$w_{1+1,6} = \frac{w_{1,6}}{z_1} e^{-\alpha_1 y_6 G_1(x_6)} = \frac{0.075}{0.748} = 0.1$$

此时强分类器为 G(x) = 0.8047 * G1(x)。G1(x)为 x < 1.5,则 y = 1; x > 1.5, 则 y = -1。则强分类器的错误率为 1 / 6 = 0.167

第二次迭代$m = 2$

[1]计算错误率

若按 0.5 切分数据，得弱分类器 x > 0.5,则 y = 1; x < 0.5, 则 y = -1。此时错误率为 0.1 * 4 = 0.4

若按 1.5 切分数据，得弱分类器 x < 1.5,则 y = 1; x > 1.5, 则 y = -1。此时错误率为 1 * 0.5 = 0.5

若按 2.5 切分数据，得弱分类器 x > 2.5,则 y = 1; x < 2.5, 则 y = -1。此时错误率为 0.1 * 4 = 0.4

若按 3.5 切分数据，得弱分类器 x > 3.5,则 y = 1; x < 3.5, 则 y = -1。此时错误率为 0.1 * 3 = 0.3

若按 4.5 切分数据，得弱分类器 x < 4.5,则 y = 1; x > 4.5, 则 y = -1。此时错误率为 2 * 0.1 = 0.2

[2]计算最弱分类器权重

$$\alpha_2 = \frac{1}{2} \cdot \log \frac{1 - 0.2}{0.2} = 0.6931$$

[3]更新样本的权值

x = 0, 1, 4, 5 时，y 分类正确

$$w_{1,1} \cdot e^{-\alpha_2 y_0 G_2(x_0)} = 0.1 * e^{-.6931 \cdot 1 \cdot 1} = 0.1 * \exp(-0.6931) = 0.05$$

$$w_{1,2} \cdot e^{-\alpha_2 y_1 G_2(x_1)} = 0.1 * e^{-.6931 \cdot 1 \cdot 1} = 0.1 * \exp(-0.6931) = 0.05$$

$$w_{1,5} \cdot e^{-\alpha_2 y_4 G_2(x_4)} = 0.5 * e^{-.6931 \cdot 1 \cdot 1} = 0.5 * \exp(-0.6931) = 0.25$$

$$w_{1,6} \cdot e^{-\alpha_2 y_5 G_2(x_5)} = 0.1 * e^{-.6931 \cdot 1 \cdot 1} = 0.1 * \exp(-0.6931) = 0.05$$

x = 2,3 时，y 分类错误

$$w_{1,3} \cdot e^{-\alpha_2 y_2 G_2(x_2)} = 0.1 * e^{-.6931 \cdot 1 \cdot -1} = 0.1 * \exp(0.6931) = 0.2$$

$$w_{1,4} \cdot e^{-\alpha_2 y_2 G_2(x_2)} = 0.1 * e^{-.6931 \cdot 1 \cdot -1} = 0.1 * \exp(0.6931) = 0.2$$

新样本权重总和为

$$z_2 = \sum_i^N w_{1,i} \cdot e^{-\alpha_1 y_i G_1(x_i)} = 0.05 * 3 + 0.25 + 0.2 * 2 = 0.8$$

进而可知

$$w_{2+1,1} = \frac{w_{1,1}}{z_1} e^{-\alpha_2 y_1 G_2(x_1)} = \frac{0.05}{0.8} = 0.0625$$

$$w_{2+1,2} = \frac{w_{1,2}}{z_1} e^{-\alpha_2 y_2 G_2(x_2)} = \frac{0.05}{0.8} = 0.0625$$

$$w_{2+1,3} = \frac{w_{1,3}}{z_1} e^{-\alpha_2 y_3 G_2(x_3)} = \frac{0.2}{0.8} = 0.25$$

$$w_{2+1,4} = \frac{w_{1,4}}{z_1} e^{-\alpha_2 y_4 G_2(x_4)} = \frac{0.2}{0.8} = 0.25$$

$$w_{2+1,5} = \frac{w_{1,5}}{z_1} e^{-\alpha_2 y_5 G_2(x_5)} = \frac{0.25}{0.8} = 0.3125$$

$$w_{2+1,6} = \frac{w_{1,6}}{z_1} e^{-\alpha_2 y_6 G_2(x_6)} = \frac{0.05}{0.8} = 0.0625$$

此时强分类器为 G(x) = 0.8047 * G1(x) + 0.6931 * G2(x)。G1(x)为 x < 1.5,则 y = 1; x > 1.5, 则 y = -1。G2(x)为 x < 4.5,则 y = 1; x > 4.5, 则 y = -1。按 G(x)分类会使 x=4 分类错误，则强分类器的错误率为 1 / 6 = 0.167；

第三次迭代，

[1] 计算错误率

按 0.5 切分，得弱分类器 x < 0.5,则 y = 1; x > 0.5, 则 y = -1。错误率为 0.0625 + 0.3125 = 0.375

按 1.5 切分，得弱分类器 x < 1.5,则 y = 1; x > 1.5, 则 y = -1。错误率为 1 * 0.3125 = 0.3125

按 2.5 切分，得弱分类器 x > 2.5,则 y = 1; x < 2.5, 则 y = -1。错误率为 0.0625 * 2 + 0.250 + 0.0625 = 0.4375

按 3.5 切分，得弱分类器 x > 3.5,则 y = 1; x < 3.5, 则 y = -1。错误率为 0.0625 * 3 = 0.1875

按 4.5 切分，得弱分类器 x < 4.5,则 y = 1; x > 4.5, 则 y = -1。错误率为 2 * 0.25 = 0.5

[2] 计算最优弱分类器的权重

由于按 3.5 划分数据时错误率最小为 0.1875，故

α_2 = 0.5 * log((1 –0.1875) / 0.1875) = 0.7332

[3] 更新样本权重

x = 2, 3 时，y 分类正确，则样本权重为：

0.25 * exp(-0.7332) = 0.1201

x = 4 时，y 分类正确，则样本权重为：

0.3125 * exp(-0.7332) = 0.1501

x = 0, 1, 5 时，y 分类错误，则样本权重为：

0.0625 * exp(0.7332) = 0.1301

新样本权重总和为 0.1201 * 2 + 0.1501 + 0.1301 * 3 = 0.7806

规范化后，

x = 2, 3 时，样本权重更新为：

0.1201 / 0.7806 = 0.1539

x = 4 时，样本权重更新为：

0.1501 / 0.7806 = 0.1923

x = 0, 1, 5 时，样本权重更新为：

0.1301 / 0.7806 = 0.1667

综上，新的样本权重为(0.1667, 0.1667, 0.1539, 0.1539, 0.1923, 0.1667)。

此时强分类器为 G(x) = 0.8047 * G1(x) + 0.6931 * G2(x) + 0.7332 * G3(x)。G1(x)为 x < 1.5,则 y = 1; x > 1.5, 则 y = -1。G2(x)为 x < 4.5,则 y = 1; x > 4.5, 则 y = -1。G3(x)为 x > 3.5,则 y = 1; x <

3.5, 则 y = -1。按 G(x)分类所有样本均分类正确，则强分类器的错误率为 0 / 6 = 0。则停止迭代，最终强分类器为 G(x) = 0.8047 * G1(x) + 0.6931 * G2(x) + 0.7332 * G3(x).