17. AdaBoost常见问题

1．Adaboost 步骤概览

a．初始化训练样本的权值分布，每个训练样本的权值应该相等（如果一共有N个样本，则每个样本的权值为1/N)

b．依次构造训练集并训练弱分类器。如果一个样本被准确分类，那么它的权值在下一个训练集中就会降低；相反，如果它被分类错误，那么它在下个训练集中的权值就会提高。权值更新过后的训练集会用于训练下一个分类器。

c．将训练好的弱分类器集成为一个强分类器，误差率小的弱分类器会在最终的强分类器里占据更大的权重，否则较小。

2．Adaboost 算法流程

给定一个样本数量为m的数据集T= $(x_1, y_1), \ldots, (x_m, y_m)$ ，yi 属于标记集合{-1,+1}。

训练集的在第k个弱学习器的输出权重为

$$D(k) = (w_{k1}, w_{k2}, \ldots w_{km}); \quad w_{1i} = \frac{1}{m}; i = 1, 2 \ldots m$$

① 初始化训练样本的权值分布，每个训练样本的权值相同：

$$D(1) = (w_{11}, w_{12}, \ldots w_{1m}); \quad w_{1i} = \frac{1}{m}; i = 1, 2 \ldots m$$

② 进行多轮迭代，产生T个弱分类器。

for t = 1, ... , T :

a. 使用权值分布 D(t) 的训练集进行训练，得到一个弱分类器 $G_t(x)$: $\chi \to -1, +1$ b. 计算 Gt(x) 在训练数据集上的分类误差率（其实就是被 Gt(x) 误分类样本的权值之和）：

$$e_t = P\left(G_t\left(x_i\right) \neq y_i\right) = \sum_{i=1}^{m} w_{ti} I\left(G_t\left(x_i\right) \neq y_i\right)$$

c. 计算弱分类器 Gt(x) 在最终分类器中的系数(即所占权重) $\alpha_t = \frac{1}{2}\ln\frac{1-e_t}{e_t}$

d. 更新训练数据集的权值分布，用于下一轮（t+1）迭代

$$D(t+1) = (w_{t+1,1}, w_{t+1,2}, \cdots w_{t+1,i} \cdots, w_{t+1,m})$$

for i = 1, ... , m:

$$w_{t+1,i} = \frac{w_{t,i}}{Z_t} \times \begin{cases} e^{-\alpha_t} & (if G_t(x_i) = y_i) \\ e^{\alpha_t} & (if G_t(x_i) \neq y_i) \end{cases} = \frac{w_{t,i}}{Z_t}\exp(-\alpha_t y_i G_t(x_i))$$

其中 Zt是规范化因子，使得D(t+1)成为一个概率分布（和为1）：

$$Z_t = \sum_{j=1}^{m} w_{t,i}\exp(-\alpha_t y_i G_t(x_i))$$

③ 集成 T 个弱分类器为1个最终的强分类器：

$$G(x) = \text{sign}\left(\sum_{t=1}^{T}\alpha_t G_t(x)\right)$$

3．权值更新过程举例说明

给定一个数据集T，由10个训练样本组成：x1，x2，...，x10，整个训练集样本总数 m=10。初始权重设置为 $w_{1i} = \frac{1}{m} = 0.1$

$p_1(\mathbf{x}_1)$	$p_1(\mathbf{x}_2)$	$p_1(\mathbf{x}_3)$	$p_1(\mathbf{x}_4)$	$p_1(\mathbf{x}_5)$	$p_1(\mathbf{x}_6)$	$p_1(\mathbf{x}_7)$	$p_1(\mathbf{x}_8)$	$p_1(\mathbf{x}_9)$	$p_1(\mathbf{x}_{10})$
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

根据权值分布，我们训练出第一个弱分类器G1（对于无法接受带权样本的基学习算法，可以通过重采样resampling来处理，后面会举例介绍一下）。假设分类器G1在数据集T上的效果为：正确分类出样本 x1 - x7，将样本 x8 - x10 错误分类。我们可以计算出赋权后的误差率：

$$e_1 = \sum_{i=1}^{10} w_{1i} I(G_1(x_i) \neq y_i) = 0.3$$

可以求出系数 α1 = 0.424，

根据上面 ②-d 步骤我们可以得出新的权值（还未规范化）：

$p_2(x_1)$	$p_2(x_2)$	$p_2(x_3)$	$p_2(x_4)$	$p_2(x_5)$	$p_2(x_6)$	$p_2(x_7)$	$p_2(x_8)$	$p_2(x_9)$	$p_2(x_{10})$
0.0656	0.0656	0.0656	0.0656	0.0656	0.0656	0.0656	0.152	0.152	0.152

经过规范化因子规范化后的权值分布（和为1）：

$p_2(x_1)$	$p_2(x_2)$	$p_2(x_3)$	$p_2(x_4)$	$p_2(x_5)$	$p_2(x_6)$	$p_2(x_7)$	$p_2(x_8)$	$p_2(x_9)$	$p_2(x_{10})$
0.0717	0.0717	0.0717	0.0717	0.0717	0.0717	0.0717	0.166	0.166	0.166

下一个分类器从此分布中产生。

4． Adaboost 的优缺点?

优点：

- 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。
- 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。

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。
- Adaboost依赖于弱分类器，而弱分类器的训练时间往往很长。

5．AdaBoost 需要归一化吗

Adaboost是指的一类集成的方法，他可以使用各种不同的分类器，你的预处理要根据你的分类器具体去定，如果你使用决策树做分类器，也是不用做归一化的。

6．AdaBoost 与 GBDT 对比有什么不同?

区别在于两者boosting的策略：Adaboost通过不断修改权重、不断加入弱分类器进行boosting；GBDT通过不断在负梯度方向上加入新的树进行boosting。

7．为什么能快速收敛?

因为每轮训练后，都会增大上一轮训练错误的样本的权重，下一轮的分类器为了达到较低的分类误差，会把权重高的样本分类正确，这样导致的结果是虽然每个弱分类器都有可能分错，但是能保证权重大的样本分正确。

8．Adaboost对噪声敏感吗?

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。

9．Adaboost和随机森林算法的异同点

相同点：

- 随机森林和Adaboost算法都可以用来分类

- 它们都是优秀的基于决策树的组合算法
- 二者都是Bootsrap自助法选取样本。

不同点：

- Adaboost是基于Boosting的算法，随机森林是基于Bagging的算法
- Adaboost减少的是偏差，随机森林减少的是方差
- 随机森林在训练每一棵树的时候，随机挑选了部分特征作为拆分特征，而不是所有的特征都去作为拆分特征。
- Adaboost后面树的训练，其在变量抽样选取的时候，对于上一棵树分错的样本，抽中的概率会加大。
- 在预测新数据时，Adaboost中所有的树**加权投票**来决定因变量的预测值，每棵树的权重和错误率有关；随机森林按照所有树中**少数服从多数**树的分类值来决定因变量的预测值（或者求取树预测的平均值）。

- [ ] https://zhuanlan.zhihu.com/p/62106410 <https://zhuanlan.zhihu.com/p/62106410>
- [ ] https://blog.csdn.net/Heitao5200/article/details/103758643?utm_medium=distribute.pc_aggpage_search_result.none-task-blog-2~aggregatepage~first_rank_ecpm_v1~rank_aggregation-13-103758643.pc_agg_rank_aggregation&utm_term=adaboost%E9%9D%A2%E8%AF%95&spm=1000.2123.3001.4430 <https://blog.csdn.net/Heitao5200/article/details/103758643?utm_medium=distribute.pc_aggpage_search_result.none-task-blog-2~aggregatepage~first_rank_ecpm_v1~rank_aggregation-13-103758643.pc_agg_rank_aggregation&utm_term=adaboost%E9%9D%A2%E8%AF%95&spm=1000.2123.3001.4430>
- [ ] https://www.cnblogs.com/pacino12134/p/11340106.html <https://www.cnblogs.com/pacino12134/p/11340106.html>