第六章 集成学习

一、填空题

1. 如果将多项式看做一个集成学习系统，其个体学习器是（幂/多项式）函数，个体学习器的参数是（多项式系数及常数项）。
2. 当所有个体学习器类型相同，称它们是（同质的）的，此时的个体学习器又称为（基学习器）。
3. Boosting 方法采用（并行）方式学习个体分类器，Bagging 方法采用（串行）方式学习个体分类器。
4. Adaboost 算法中集成学习器$H(x)$应最小化（指数）损失函数。
5. 在标准 Adaboost 算法中个体学习器$h_t(x)$输出的结果为（+1）或（-1）
6. 在标准 Adaboost 算法中个体学习器$h_t(x)$的正确率为 75%，则α_t的值为$(\ln\sqrt{3})$
7. 在标准 Adaboost 算法中，当训练集有 N 个样本时，每个样本的初始化权重为（$\frac{1}{N}$）。

二、判断题

8. 使用相对多数投票时，若没有任何一类得票数高于 50%，则拒绝分类。(×)
9. 采用前向分步优化策略得到的最优解就是集成学习问题的全局最优解。(×)
10. Adaboost 算法中，被个体学习器h_t错误分类的样本的权重在h_{t+1}的学习中一定会增加。(×)

三、选择题

11. 设样本集为$X=\{1,3,5,7\}, Y=\{+1,-1,-1,+1\}$，权重为$\mathcal{D}_t=\{0.1,0.5,0.2,0.2\}$，个体分类器函数形式为$h_t(x)=\text{sign}(x-v)$。则当前$v$的最优取值为：（D）

A. 0　　B. 2　　C. 4　　D. 6

12. 当样本数趋近于无穷大时，Bootstrap 采样时样本被抽中的概率约为：（B）

A. 0.368　　B. 0.612　　C. 1　　D. 0.5

四、简答题

13. 设训练集为$\{\boldsymbol{x}_i, y_i | i=1,\ldots,N\}$,请将加法模型的学习问题转化为参数$\boldsymbol{\theta}_t$与系数$\alpha_t$的优化问题，并写出目标函数表达式，并加以说明
14. 简述集成学习中个体学习器的设计与学习要满足什么样的规则，并加以解释。

15. 试说明 boosting 方法与 bagging 方法的主要异同之处。

16. AdaBoost 的算法采用什么措施使个体分类器“不同”。

17. 随机森林算法采用哪些措施使个体分类器“不同”。

五、计算（画图）题

18. 给定如表所示训练数据集。假设弱分类器由$x<v$或$x>v$产生，试使用 AdaBoost 算法求解。

序号	1	2	3	4	5	6
x	0	1	2	3	4	5
y	1	1	-1	-1	1	1

1） 给出前两个个体分类器的训练过程和结果

（1）初始化样本权重：

$$\omega_i = \frac{1}{N} = \frac{1}{6}$$

故上述训练集 D 所对应的权重集合：

$$D_1 = \{\omega_{1,1}, \omega_{1,2}, \omega_{1,3}, \omega_{1,4}, \omega_{1,5}, \omega_{1,6}\} = \left\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right\}$$

由题可知，以序号为标准，阈值v可取-0.5，0.5，1.5，2.5，3.5，4.5，5.5。分类误差率为样本 $\boldsymbol{x}$ 的预测分类 $\mathrm{h}_t(\boldsymbol{x})$ 和实际分类 $f(\boldsymbol{x})$ 不相同的概率：$\epsilon_1 = P_{\boldsymbol{x}\sim D_t}(\mathrm{h}_t(\boldsymbol{x}) \neq f(\boldsymbol{x}))$

$\boldsymbol{v}$	$\boldsymbol{x}<\boldsymbol{v}$错分样本	$\boldsymbol{x}>\boldsymbol{v}$错分样本	最小分类误差$\boldsymbol{e}$
-0.5	1,2,5,6	3,4	2/6
0.5	2,5,6	1,3,4	3/6
1.5	5,6	1,2,3,4	2/6
2.5	3,5,6	1,2,4	3/6
3.5	3,4,5,6	1,2	2/6
4.5	3,4,6	1,2,5	3/6
5.5	3,4	1,2,5,6	2/6

故选择分类器$x > v$，阈值 $v = -0.5$，对应的分类误差 $e = 2/6 < 0.5$

此时的分类器权重：

$$\alpha_1 = \frac{1}{2}\ln\frac{1-\epsilon_1}{\epsilon_1} = \frac{1}{2}\ln\frac{1-\frac{1}{3}}{\frac{1}{3}} = \ln\sqrt{2}$$

规范化因子Z_i的计算如下，它使各个值的新权重值之和为 1:

$$Z_1 = \sum_{i=1}^{6} \omega_{1,i} e^{-\alpha_1 f(x) h_1(x)} = \frac{1}{6} \mathrm{e}^{-\ln\sqrt{2}*1*(-1)} \times 2 + \frac{1}{6} \mathrm{e}^{-\ln\sqrt{2}*1*1} \times 4 = \frac{2\sqrt{2}}{3}$$

被错误分类的样本的权重(i = 3,4):

$$\omega_{2,i} = \frac{\omega_{1i}}{Z_1} e^{-\alpha_1 f(x) h_1(x)} = \frac{\frac{1}{6}}{\frac{2\sqrt{2}}{3}} e^{-\ln\sqrt{2}*1*(-1)} = \frac{1}{4}$$

被正确分类的样本的权重(i =1,2,5,6):

$$\omega_{2,i} = \frac{\omega_{1i}}{Z_1} e^{-\alpha_1 f(x) h_1(x)} = \frac{\frac{1}{6}}{\frac{2\sqrt{2}}{3}} e^{-\ln\sqrt{2}*1*1} = \frac{1}{8}$$

此时更新的权重集合:

$$D_2 = \{\omega_{2,1}, \omega_{2,2}, \omega_{2,3}, \omega_{2,4}, \omega_{2,5}\} = \left\{\frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right\}$$

(2) 权重更新之后，错分样本和加权误差情况如下

v	$x < v$错分样本	$x > v$错分样本	最小加权误差
-0.5	1,2,5,6	3,4	1/2
0.5	2,5,6	1,3,4	3/8
1.5	5,6	1,2,3,4	1/4
2.5	3,5,6	1,2,4	3/8
3.5	3,4,5,6	1,2	1/4
4.5	3,4,6	1,2,5	3/8
5.5	3,4	1,2,5,6	1/2

根据加权分类误差，选择分类器$x < v$，阈值$v = 1.5$

$$\alpha_2 = \frac{1}{2} ln \frac{1-\epsilon_1}{\epsilon_1} = \frac{1}{2} ln \frac{1-1/4}{1/4} = ln\sqrt{3}$$

$$Z_2 = \sum_{j=1}^{6} \omega_{2,j} e^{-\alpha_2 f(x) G_2(x_i)} = 2 * \frac{1}{8} * \sqrt{3} + 2 * \frac{1}{4} * \frac{\sqrt{3}}{3} + 2 * \frac{1}{8} * \frac{\sqrt{3}}{3} = \frac{\sqrt{3}}{2}$$

被错误分类的样本的权重(i = 5, 6):

$$\omega_{3,i} = \frac{\omega_{2,i}}{Z_1} e^{-\alpha_2 f(x) h_2(x)} = \frac{\frac{\sqrt{3}}{8}}{\frac{\sqrt{3}}{2}} = \frac{1}{4}$$

上次被错误分类，此次被正确分类的样本的权重(i = 3,4)：

$$\omega_{3,4} = \frac{\omega_{2,1}}{Z_1} e^{-\alpha_1 f(x) h_2(x)} = \frac{\frac{\sqrt{3}}{12}}{\frac{\sqrt{3}}{2}} = \frac{1}{6}$$

上次被正确分类，此次仍被正确分类的样本的权重(i = 1,2)：

$$\omega_{3,i} = \frac{\omega_{2,i}}{Z_1} e^{-\alpha_i f(x) h_2(x)} = \frac{\frac{\sqrt{3}}{24}}{\frac{\sqrt{3}}{2}} = \frac{1}{12}$$

此时更新的权重集合：

$$D_3 = \{\omega_{3,1}, \omega_{3,2}, \omega_{3,3}, \omega_{3,4}, \omega_{3,5}, \omega_{3,6}\} = \left\{\frac{1}{12}, \frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{4}, \frac{1}{4}\right\}$$

故两轮训练后，弱分类器输出：

$$H(\boldsymbol{x}) = sign(ln\sqrt{2} * \text{sign}(x + 0.5) + ln\sqrt{3} * \text{sign}(1.5 - \text{x}))$$

2) 采用 Bootstrap 方法对上述训练集进行采样，从统计效果上看，单个样本在某一轮采样中未被选中的概率是多少？

答：共 6 个样本。每个样本单次未被选中的概率为：$1 - \frac{1}{6} = \frac{5}{6}$;

从统计效果上看，每一轮次的 Bootstrap 包含 6 次采样，则单个样本未被选中的概率为：

$$P_{miss} = \left(1 - \frac{1}{6}\right)^6 \approx 0.335$$

19. 某集成学习框架下，5 个个体分类器对于一个二分类问题各自给出的结果及相应的系数如下表所示。请分别根据绝对多数投票，相对多数投票和加权投票的策略给出集成学习框架的最终识别结果。

分类器序号	1	2	3	4	5
权重	0.1	0.2	0.3	0.3	0.1
分类结果	是	否	是	否	是

答：根据绝对多数投票，3 个是，2 个否，结果为是；

根据相对多数投票，3 个是，2 个否，结果为是；

根据加权投票：

$$0.1 * (+1) + 0.2 * (-1) + 0.3 * (+1) + 0.3 * (-1) + 0.1 * (+1) = 0$$

结果为拒绝识别。