第九次作业

阅读机器学习实战 第7章 7.5节

Exercise 9.1

以下那种算法不是集成学习算法( )

A. AdaBoost
B. GBDT
C. 决策树
D.随机森林

解析：单棵决策树不属于集成学习算法，因为决策树并没有综合多颗树的结果

Exercise 9.2

[多选]下面关于随机森林和梯度提升集成方法的说法哪个是正确的？ ( )

A. 这两种方法都可以用来做分类
B. 随机森林用来做分类，梯度提升用来做回归
C. 两种方法都可以用来做回归
D. 随机森林用来做回归，梯度提升用来做分类

Exercise 9.3

AdaBoost中核心参数α的取值为(e为模型错误率)：
A. $\frac{1}{2}\ln(\frac{1-e}{e})$
B. $\ln(\frac{1-e}{e})$
C. $\frac{1}{2}\ln(\frac{e}{1-e})$
D.$\ln(\frac{e}{1-e})$

(c) 根据分类错误率ϵ_t计算当前弱分类器的权重系数α_t:

解析:

$$\alpha_t = \ln \lambda_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$$

Exercise 9.4

According to $\alpha_t = ln(\lambda_t)$, and $\lambda_t = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$, when would $\alpha_t > 0$?
A、 $\epsilon_t < \frac{1}{2}$
B、 $\epsilon_t > \frac{1}{2}$
C、 $\epsilon_t \neq 1$
D、 $\epsilon_t \neq 0$

Exercise 9.5

问题:GBDT算法的描述，不正确的是( )

A、决策树+Boosting=GBDT
B、GBDT算法主要是用了Boosting方法
C、GBDT与AdaBoost 的对比，都是 Boosting 家族成员，使用弱分类器；都使用前向分步算法
D、梯度提升算法通过迭代地选择一个梯度方向上的基函数来逐渐逼近局部极小值

解析：梯度提升算法通过迭代地选择一个负梯度方向上的基函数来逐渐逼近局部极小值

Exercise 9.6

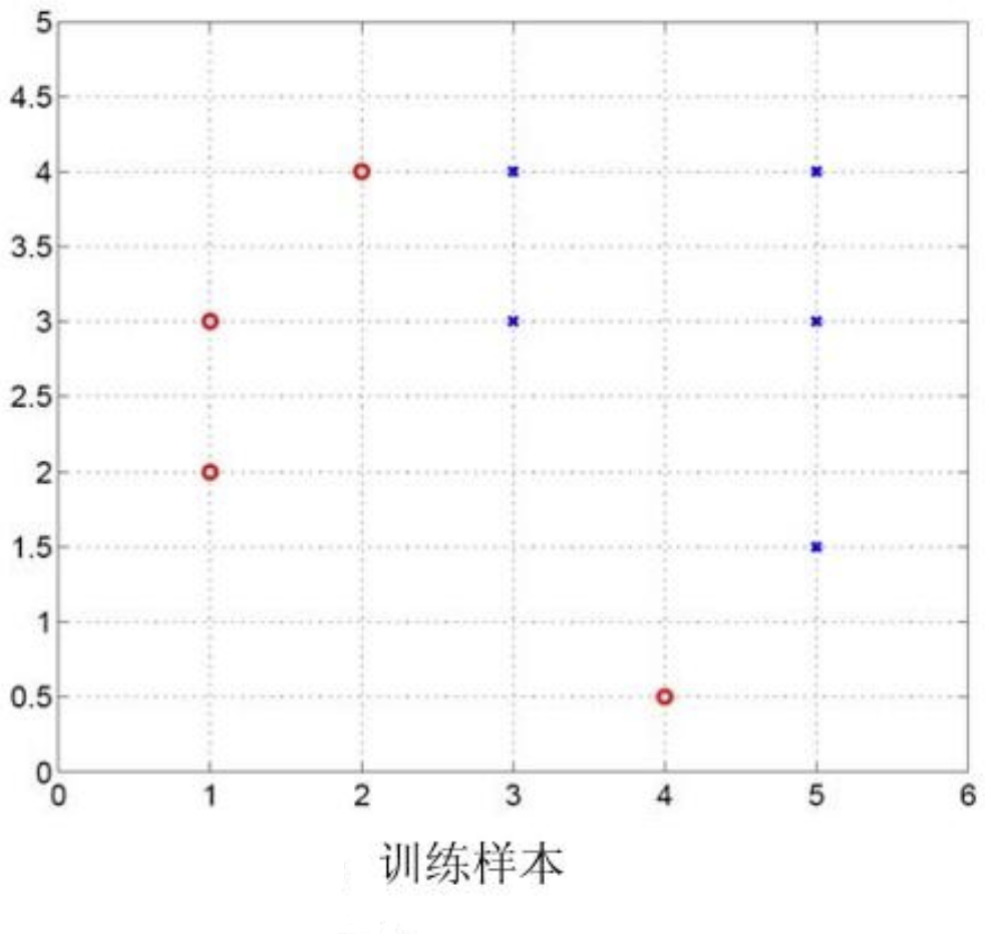
在AdaBoosting迭代中,从第 t 轮到第 $t+1$ 轮,某个被错误分类样本的惩罚增加了,可能因为该样本

A. 被第 t 轮训练的弱分类器错误分类
B. 被第 t 轮后的集成分类器（强分类器）错误分类
C. 被到第 t 轮为止训练的大多数弱分类器错误分类
D. 以上都有可能

解析:

Exercise 9.7

考虑如下图所示的训练样本，其中"x"和"o"分别表示负样本和正样本，我们采用Adaboost算法对上述样本进行分类。在Boosting的每次迭代中，我们选择加权错误率最小的弱分类器。假设采用的弱分类器为平行于两个坐标轴的线性分类器。



(1) 请在图中标出第一次迭代选择的弱分类器(L_1), 并给出决策面的"+"和"-"面?

(2) 请在图中用圆圈标出在第一次迭代后权重最大的样本，其权重是多少?

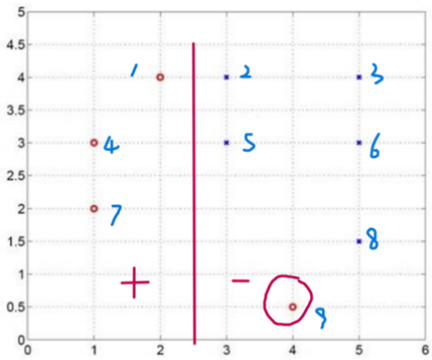
(3) 第一次迭代后权重最大的样本在经过第二次迭代后权重变为多少?

(4) 强分类器为弱分类器的加权组合。则在这些点中，存在被经过第二次迭代后的强分类器错分的样本吗？给出简短理由

解析:

1. 一共有9个样本，初始化每个样本的权重u_i = 1/9,i=1,2.....9
   第一次迭代选择分类器在x=2.5时，错误率最小
   决策面L_1 (x)为

$$L_1(x) = \begin{cases} 1, x <= 2.5 \\ -1, x > 2.5 \end{cases}$$



训练样本

2. 第一次迭代后权重最大的样本如图所示
   此时错误率为$\epsilon_1 = \frac{1}{9} \quad \lambda_1 = \sqrt{8}$
   所以它的权重为 $u_2 = u_1 \times \lambda_1 = \frac{\sqrt{8}}{9}$
3. 第一轮迭代后:
   错误分类的样本权重为$\frac{\sqrt{8}}{9}$， 正确分类的样本权重为$\frac{\sqrt{2}}{36}$
   得到弱分类器$(L2)$如图中蓝线所示
   此时错误率为$\epsilon_2 = \frac{2\sqrt{2}}{36} \approx 0.0786 \quad \lambda_2 \approx 3.42$
   所以，第二次迭代后，它的权重变成了$u_3 \approx 0.091$

4. $L1$的权重 $\alpha_1 = ln\lambda_1 = 1.039$
   $L2$的权重 $\alpha_2 = ln\lambda_2 = 1.231$
   所以强分类器可表示为
   $G(x) = sign(1.039 \cdot L1 + 1.231 \cdot L2)$
   所以对于$L1$和$L2$之间的两个点，其分类结果为
   $G = sign(-1.039 + 1.231) = 1$ 为正类，而实际这两个点都是负类。
   所以存在经过第二次迭代后被强分类器错误分类的点。

Exercise 9.8

如果你的AdaBoost集成对训练数据欠拟合，你应该调整哪些超参数？怎么调整?

解析：如果你的AdaBoost集成欠拟合训练集，可以尝试提升估算器的数量或是降低基础估算器的正则化超参数。你也可以尝试略微提升学习率

Exercise 9.9

如果你的梯度提升集成对训练集过拟合，你是应该提升还是降低学习率?

解析：如果你的梯度提升集成过拟合训练集，你应该试着降低学习率，也可以通过提前停止法来寻找合适的预测器数量（可能是因为预测器太多）。

Exercise 9.10

现有某个公司四位员工的考评信息即月薪如下表所示，试根据该数据集和 GBDT 学习算法构造集成模型，其中树的深度为3（不包括初始提升树），迭代次数M=2，并使用该集成模型预测工龄为 25 年，绩效得分为 65 分的员工的月薪

表 员工信息数据集

编号	工龄（年）	绩效得分	月薪（万元）
1	5	20	1.1
2	7	30	1.3
3	21	70	1.7
4	30	60	1.8

解析:

（1）依据以下目标初始化第一个个体学习器$T_0(\mathbf{x})$:

$T_0(\mathbf{x}) = \arg\min_c \sum_{i=1}^4 L(y_i, c)$

假设采用平方损失函数$L(y, \mathbf{x}) = \frac{1}{2}(y - \mathbf{x})^2$进行求解，易求得$c = \bar{y}$, 由于此时只包含根节点，该节点对应整个数据集，因此取:

$c = \frac{1}{4}\sum_{i=1}^4 y_i = 1.475$

故求得初始化个体学习器为:

$H_0(\mathbf{x}) = T_0(\mathbf{x}) = 1.475$

（2）计算 D 中每个样本所对应的负梯度:

$$r_{1i} = -\frac{\partial L(y_i, H(\mathbf{x}_i))}{\partial H(\mathbf{x}_i)}|_{H_0(\mathbf{x}_i)=H(\mathbf{x}_i)}, \quad i = 1, 2, 3, 4$$

可构建如下表所示的数据集D_1。

编号	工龄(年)	绩效得分	r_{1i}
1	5	20	-0.375
2	7	30	-0.175
3	21	70	0.025
4	30	60	0.325

通过数据集 D_1 和方差最小化原则构造回归树，具体做法为遍历计算所有可能划分点所对应的方差值，计算结果如下表所示。

划分点	小于划分点的数据集	大于划分点的数据集	方差
工龄=5	$\emptyset$	{2,3,4}	0.082
工龄=7	{1}	{2,3,4}	0.047
工龄=21	{1,2}	{3,4}	0.0125
工龄=30	{1,2,3}	{4}	0.062
得分=20	$\emptyset$	{2,3,4}	0.082
得分=30	{1}	{2,3,4}	0.047
得分=60	{1,2}	{3,4}	0.0125
得分=70	{1,2,3}	{4}	0.867

由上表可知最优划分点有两个：工龄 $=21$和绩效得分 $=60$，此处随机选择工龄 $=21$作为划分点，左右子树的节点集合为$R_{11}=\{\mathbf{x}_1,\mathbf{x}_2\}, R_{12}=\{\mathbf{x}_3,\mathbf{x}_4\}$

我们设置的参数中树的深度max_depth=3，现在树的深度只有2，需要再进行一次划分，这次划分要对左右两个节点分别进行划分。

对于**左节点**，只含有1,2两个样本，根据下表我们选择**工龄7**划分：

划分点	小于划分点的数据集	大于划分点的数据集	方差
工龄=5	$\emptyset$	{1,2}	0.01
工龄=7	{1}	{2}	0
得分=20	$\emptyset$	{1,2}	0.01
得分=30	{1}	{2}	0

对于**右节点**，只含有3,4两个样本，根据下表我们选择**工龄30**划分

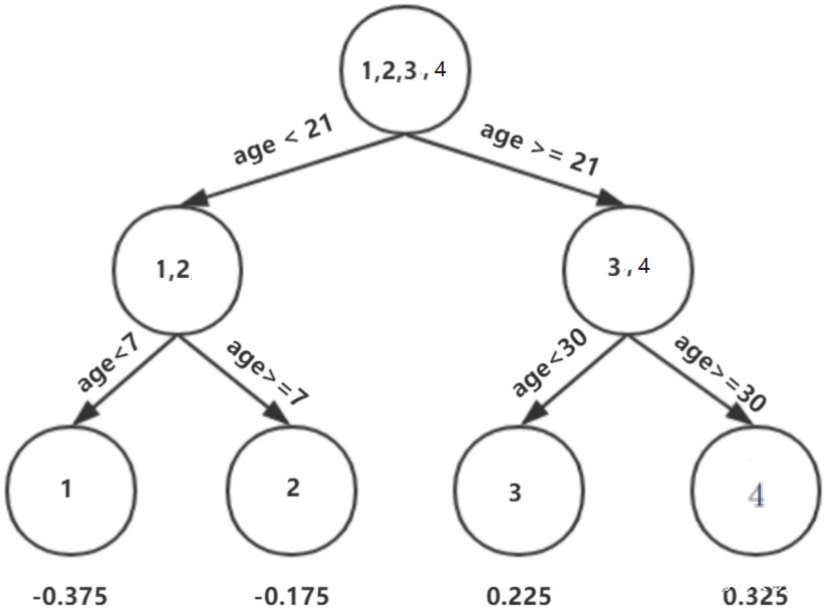
划分点	小于划分点的数据集	大于划分点的数据集	方差
工龄=21	$\emptyset$	{3,4}	0.0025
工龄=30	{3}	{4}	0
得分=60	$\emptyset$	{3,4}	0.025
得分=70	{3}	{4}	0

计算叶子节点输出值，

$c_{1j} = \arg\min_c \sum_{\mathbf{x}_i \in R_{1j}} L(y_i, H_0(\mathbf{x}_i) + c), \quad j = 1, 2, 3, 4$

$c_{11} = -0.375,\ c_{12} = -0.175,\ c_{13} = -0.225,\ c_{14} = 0.325,$

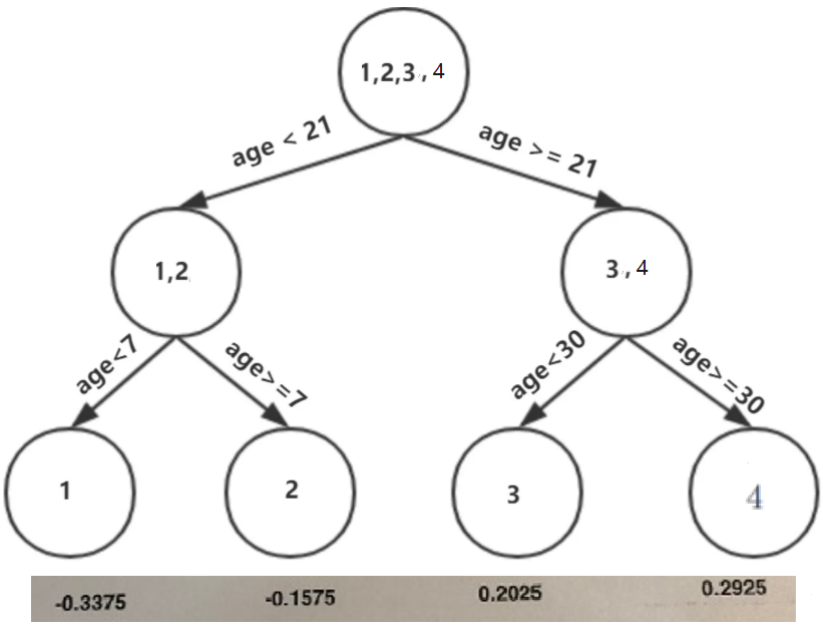
此时决策树如下：



考虑参数学习率，假设 $lr = 0.1$

$H_1(\mathbf{x}) = H_0(\mathbf{x}) + 0.1 * \sum_{j=1}^{4} c_{1j} I(\mathbf{x} \in R_{1j})$

继续进行第二次迭代，并得到如何的决策树，



则 $f(\mathbf{x}) = H_0(\mathbf{x}) + 0.1 * \sum_{j=1}^{4} c_{1j} I(\mathbf{x} \in R_{1j}) + 0.1 * \sum_{j=1}^{4} c_{2j} I(\mathbf{x} \in R_{2j})$

计算得 $f(x_{test}) = 1.51775$