5.26 《机器学习基础》作业9

王宇哲 18000011828

T1

解：(回归问题提升树算法框架)

输入：训练样本集 $D=\{(x_i,y_i)\}_{i=1}^{N}$；
基学习器的个数 M，损失函数
(回归树)
$L(y,f(x))$

输出：提升树 $f_M(x)$

1) 初始化 $f_0(x)=0$

2) 对 $m=1,2,\cdots,M$

i) 计算残差

$$r_{mi}=y_i-f_{m-1}(x_i),\ i=1,\cdots,N$$

ii) 拟合残差 r_{mi}，得到回归树 $T(x;\theta_m)$

$$\theta_m=\arg\min_{\theta}\sum_{i=1}^{N}L(y_i,f_{m-1}(x_i)+T(x;\theta))$$

iii) 更新

$$f_m(x)=f_{m-1}(x)+T(x;\theta_m)$$

3) 得到回归问题提升树

$$f_M(x)=\sum_{m=1}^{M}T(x;\theta_m)$$

T2

解：(随机森林算法框架)

输入：训练样本集 $D=\{(x_i,y_i)\}_{i=1}^{N}$；
基分类器个数 T，决策树特征选择算法 $\mathbb{L}$

输出：集成分类器 $f(x)$ 最优划分

1) 对 $t=1,2,\cdots,T$

i) 从 D 中利用自助采样法随机抽取 N 个样本，得到训练集 D_t

ii) 生成基决策树 $f_t(x)$

a) 生成结点 node

b) 若 D_t 中样本均属同一类别，将 node 标记为该类别叶结点，返回；
否则转 c)

c) 从当前节点对应的 d 个特征中先随机选择 k 个特征，从这 k 个特征中依 $\mathbb{L}$ 生成分支；重复递归进行该过程，直至分支节点被标记为叶结点；递归返回

d) 返回决策树 $f_t(x)$

2) 依多数占优的简单投票法生成集成分类器 $f(x)$

$$f(x)=\arg\max_{y}\sum_{t=1}^{T}\mathbb{I}(f_t(x)=y)$$

对随机森林和 Bagging 方法进行比较：

i) 随机森林以 Bagging 方法为基础，以决策树算法作为基学习器的学习算法；相比 Bagging，在决策树学习算法中引入了随机属性选择

ii) 随机森林增加了关于划分特征的随机扰动，因此相比 Bagging 增加了基学习器的多样性差异，增加了泛化能力，减小泛化误差

iii) 随机森林相比 Bagging 在相同任务下基分类器训练效率高，训练速度相对较快