

# 机器学习导论

## 习题六

学号, 作者姓名, 邮箱

2017 年 5 月 24 日

### 1 [20pts] Ensemble Methods

- (1) [10pts] 试说明Boosting的核心思想是什么, Boosting中什么操作使得基分类器具备多样性?
- (2) [10pts] 试析随机森林为何比决策树Bagging集成的训练速度更快。

**Solution.** 此处用于写解答(中英文均可)

### 2 [20pts] Bagging

考虑一个回归学习任务 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 。假设我们已经学得 $M$ 个学习器 $\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x}), \dots, \hat{f}_M(\mathbf{x})$ 。我们可以将学习器的预测值看作真实值项加上误差项

$$\hat{f}_m(\mathbf{x}) = f(\mathbf{x}) + \epsilon_m(\mathbf{x}) \quad (2.1)$$

每个学习器的期望平方误差为 $\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2]$ 。所有的学习器的期望平方误差的平均值为

$$E_{av} = \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2] \quad (2.2)$$

$M$ 个学习器得到的Bagging模型为

$$\hat{f}_{bag}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \hat{f}_m(\mathbf{x}) \quad (2.3)$$

Bagging模型的误差为

$$\epsilon_{bag}(\mathbf{x}) = \hat{f}_{bag}(\mathbf{x}) - f(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \epsilon_m(\mathbf{x}) \quad (2.4)$$

其期望平均误差为

$$E_{bag} = \mathbb{E}_{\mathbf{x}}[\epsilon_{bag}(\mathbf{x})^2] \quad (2.5)$$

- (1) [10pts] 假设 $\forall m \neq l, \mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})] = 0, \mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] = 0$ 。证明

$$E_{bag} = \frac{1}{M} E_{av} \quad (2.6)$$

- (2) [10pts] 试证明不需对 $\epsilon_m(\mathbf{x})$ 做任何假设,  $E_{bag} \leq E_{av}$ 始终成立。(提示: 使用Jensen's inequality)

**Proof.** 此处用于写证明(中英文均可)

□

### 3 [30pts] AdaBoost in Practice

- (1) [25pts] 请实现以Logistic Regression为基分类器的AdaBoost, 观察不同数量的ensemble带来的影响。详细编程题指南请参见链接: [http://lamda.nju.edu.cn/ml2017/PS6/ML6\\_programming.html](http://lamda.nju.edu.cn/ml2017/PS6/ML6_programming.html)
- (2) [5pts] 在完成上述实践任务之后, 你对AdaBoost算法有什么新的认识吗? 请简要谈谈。

**Solution.** 此处用于写解答(中英文均可)