机器学习导论
习题六

141220132, 银琦, 141220132@smail.nju.edu.cn

2017 年 6 月 8 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.

(1)Boosting的核心思想是：强分类器算法比较难以获得，而弱分类器较易获得，所有基于易得到的弱分类器，达到强分类器的识别效果。Boosting产生一系列的分类器，然后对所有的分类器的结果进行加权融合。

Boosting中的重采样使得基分类器具备多样性。

(2)因为随机森林在以决策树为基学习器构建Bagging集成的基础上，进一步在决策树的训练过程中引入了随机属性选择，即随机森林不仅会随机样本，还会在所有样本属性中随机几种出来计算。这样每次生成分类器时都是对部分属性计算最优属性，而Bagging是对全部属性计算最优属性，所以随机森林速度会比Bagging要快。

2 [20pts] Bagging

考虑一个回归学习任务$f:\mathbb{R}^d \to \mathbb{R}$。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

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

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

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

M个学习器得到的Bagging模型为

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

Bagging模型的误差为

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

其期望平均误差为

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

(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

$$E_{bag} = \frac{1}{M}E_{av} \tag{2.6}$$

(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)

(1)由题：

$$\begin{aligned}E_{bag} &= \mathbb{E}_{\mathbf{x}}[\epsilon_{bag}(\mathbf{x})^2]\\ &= \mathbb{E}_{\mathbf{x}}[(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x}))^2]\\ &= \mathbb{E}_{\mathbf{x}}[\frac{1}{M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})]\end{aligned}$$

由于$\forall\ m \neq l$，$\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] = 0$，所以代入消去后只剩m=l的部分，即：

$$\begin{aligned}E_{bag} &= \mathbb{E}_{\mathbf{x}}[\frac{1}{M^2}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})^2]\\ &= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2]\\ &= \frac{1}{M}E_{av}\end{aligned}$$

得证。

(2)令$t_m = \epsilon_m(\mathbf{x})$，则：

$$\begin{aligned}E_{bag} &= \mathbb{E}_{\mathbf{x}}[(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x}))^2]\\ &= \mathbb{E}_{\mathbf{x}}[(\frac{1}{M}\sum_{m=1}^{M}t_m)^2]\end{aligned}$$

显然，$y = t^2$为凸函数，由Jensen's inequality不等式，

$$\begin{aligned}E_{bag} &\leq \mathbb{E}_{\mathbf{x}}[\sum_{m=1}^{M}\frac{t_m^2}{M}]\\ &= \mathbb{E}_{\mathbf{x}}[\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})^2]\\ &= E_{av}\end{aligned}$$

得证。 □

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`

(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.

(2)AdaBoost算法通过每次更新权重，利用权重更新数据集，进行下一轮迭代，多轮迭代后产生多个弱分类器，将其组合后得出结果，理论上迭代轮数增加，精确度逐渐震荡直至收敛，但是若错误率过低，则会使得权重计算出错，影响精度，所以需要进行其它参数的调整；对率回归使用的是作业二中的代码，所以没有正则化，并且训练时间十分长。

不同的基分类器参数设置会使得数据的值有所侧重，使得分类效果更佳。