机器学习导论
习题六

151220023, 段建辉, djhbarca@163.com

2017 年 6 月 9 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的核心思想为：根据前一轮学习器的表现对训练样本分布进行调整，使得前一个学习器做错的样本受到更大关注然后进行下一轮的学习器训练，如此重复多次，通过这样的序列化方法最后将顺序生成的学习器进行加权结合。相当于“知错就改”，利用残差不断逼近最优解。
基分类器多样性原因：在训练过程的每一轮中根据样本分布为每一个训练样本重新赋予新的权重。如果基学习器不接受权重则根据重新调整过的样本分布对训练集重新采样然后再训练。这样使得基分类器具有多样性。
(2) Bagging使用的是“确定型”决策树，在选择划分属性的时候要对节点的所有属性进行考察，而随机森林使用的“随机型”决策树只需要考察一个属性子集，所以随机森林每次计算的属性都是部分属性而bagging计算全属性，所以随机森林更快。

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$\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x}), \ldots, \hat{f}_M(\mathbf{x})$。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.
(1)由于：

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

那么带入即可到到：

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

因为$\forall\, m \neq l$，$\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})] = 0$，$\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] = 0$，那么对于$(\sum_{m=1}^{M}\epsilon_m(\mathbf{x}))^2$而言，我们有:

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

将该式带入E_{bag}等式之后可以得到：

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

证毕。
(2)
由(1)我们有：

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

由于$f(x) = x^2$是一个下凸函数，根据琴生不等式我们有：

$$f(\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}) \leqslant \frac{f(x_1) + f(x_2) + f(x_3) + \cdots + f(x_n)}{n}$$

那么由于$\varphi(x) = (\frac{1}{M^2}\sum_{m=1}^{M}\epsilon_m(\mathbf{x}))^2$是一个类二次函数，那么也满足上述性质。因此：

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

证毕。

□

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. 一开始，训练总是出错，因为对象创建在了一开始，所以循环里面训练一次$T = 1$之后就停止了，导致5/10/100均没有训练，在循环里面创建对象成功。
后来训练，发现迭代多少轮正确率都是0.5585。但是迭代过程是没有问题的。