机器学习导论
习题六

学号, 作者姓名, 邮箱

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）核心思想是通过调整训练样本的分布保证基分类器多样性，串行生成多个分类器组成强分类器。
Boosting的基分类器按顺序训练，训练每个基分类器时所使用的训练集是加权重的，而训练集中的每个样本的权重系数取决于前一个基分类器的性能。如果前一个基分类器错误分类地样本点，那么这个样本点在下一个基分类器训练时会有一个更大的权重，这样就能确保基分类器的多样性。
（2）因为随机森林在训练时没有使用全部特征，只是选取了特征的一个子集进行训练。

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

$$\begin{aligned}
E_{bag} &= \mathbb{E}_{\mathbf{x}}[\epsilon_{bag}(\mathbf{x})^2] \\
&= \mathbb{E}_{\mathbf{x}}[(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x}))^2] \\
&= \frac{1}{M^2}\mathbb{E}_{\mathbf{x}}[(\epsilon_1(\mathbf{x}) + ... + \epsilon_M(\mathbf{x}))^2] \\
&= \frac{1}{M^2}\mathbb{E}_{\mathbf{x}}[\sum_{m=1}^{M}\epsilon_m(\mathbf{x})^2] \\
E_{av} &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2]
\end{aligned}$$

$E_{bag} = \frac{1}{M}E_{av}$得证
(2)由前面推导可得，

$$\begin{aligned}
E_{bag} &= \frac{1}{M^2}\mathbb{E}_{\mathbf{x}}[(\epsilon_1(\mathbf{x}) + ... + \epsilon_M(\mathbf{x}))^2] \\
E_{av} &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2]
\end{aligned}$$

要证$E_{baf} \leq E_{av}$，即证$E_x[(\frac{(\epsilon_1+...+\epsilon_M)^2}{M}] \leq E_x[\epsilon_1^2 + ... + \epsilon_M^2]$，即证$(\frac{(\epsilon_1+...+\epsilon_M)^2}{M} \leq \epsilon_1^2 + ... + \epsilon_M^2$
根据Jessen不等式$f(\frac{1}{M}\sum_{i=1}^{M}\epsilon_i) \leq \frac{1}{M}\sum_{i=1}^{M}f(\epsilon_i)(f{:}p)$
可得$\frac{(\epsilon_1+...+\epsilon_M)}{M} \leq \frac{1}{M}(\epsilon_1 + ... + \epsilon_M)^2 \leq \epsilon_1^2 + ... + \epsilon_M^2$
命题得证 □

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.
尝试了正则化参数C=1,10,100三种，基分类器的精度依次上升，然而adboost的提升效果依次下降。由误差分歧分解可知adboost的精度是由个体学习器准确性和多样性共同决定，单个学习器的学习精度太高达到0.9时，多样性明显下降，导致adboost优化效果不明显。