机器学习导论
习题六

学号：141130077
作者姓名：邱梓豪
邮箱：2957606241@qq.com

2017 年 6 月 5 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)
（1）Boosting的核心思想是将从弱学习算法出发，反复学习，得到一系列弱分类器（基分类器），然后组合这些弱分类器，得到一个强分类器。
增强分类器多样性的办法有如下几种：（1）数据样本扰动，如从初始的数据集中产生不同的数据子集，再利用不同的数据子集训练出不同的个体学习器。（2）输入属性扰动，如从不同的属性子空间训练出个体学习器。（3）输出表示扰动，对输出表示进行操纵以增强多样性。（4）算法参数扰动，随机设置某些学习器的初始参数，往往能产生差别较大的学习器。
（2）因为在个体决策树的构建过程中，Bagging使用的是“确定型”决策树，在选择划分属性时要对结点的所有属性进行考察，而随机森林使用的“随机型”决策树则只考察属性集合的一个子集。所以随机森林为何比决策树Bagging集成的训练速度更快。

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. 此处用于写证明(中英文均可)
（1）证明：
将（2.4）带入（2.5）可得：

$$E_{bag} = \mathbb{E}_{\mathbf{x}}[\epsilon_{bag}(\mathbf{x})^2] = \mathbb{E}_{\mathbf{x}}[\{\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\}^2] \tag{2.7}$$

因为假设各学习器产生的误差互不相关（$\forall\ m \neq l$，$\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] = 0$）并且各学习器产生的误差期望和为0（$\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})] = 0$），所以（2.7）可以写为：

$$\begin{aligned}
E_{bag} &= \mathbb{E}_{\mathbf{x}}[\{\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\}^2] && (2.8)\\
&= \mathbb{E}_{\mathbf{x}}[\{\frac{1}{M}\sum_{i=1}^{M}\epsilon_i(\mathbf{x})\}\{\frac{1}{M}\sum_{j=1}^{M}\epsilon_j(\mathbf{x})\}] && (2.9)\\
&= \mathbb{E}_{\mathbf{x}}[\frac{1}{M^2}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})^2] && (2.10)\\
&= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2] && (2.11)\\
&= \frac{1}{M}E_{av} && (2.12)
\end{aligned}$$

（2.10）到（2.11）的原因是期望的线性性质。
（2）证明：

对应凸函数f，由Jensen不等式，有：

$$f(\sum_{m=1}^{M}\omega_m x_m) \leq \sum_{m=1}^{M}\omega_m f(x_m) \tag{2.13}$$

其中$\omega_i \geq 0$，$\sum_i \omega_i = 1$。
由（1）中可知：

$$E_{bag} = \mathbb{E}_{\mathbf{x}}[\{\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\}^2] = \mathbb{E}_{\mathbf{x}}[\{\sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})\}^2] \tag{2.14}$$

$$E_{av} = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2] = \mathbb{E}_{\mathbf{x}}[\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})^2] = \mathbb{E}_{\mathbf{x}}[\sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})^2] \tag{2.15}$$

所以，将$\frac{1}{M}$看成Jensen不等式中的ω_i，将$\epsilon_m(\mathbf{x})$看成Jensen不等式中的x_i，令$f(x) = x^2$，则有：

$$\{\sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})\}^2 \leq \sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})^2 \tag{2.16}$$

因为上式对所以x成立，因此也对x的期望成立，因此有：$E_{bag} \leq E_{av}$成立。

□

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. 此处用于写解答(中英文均可)
这次编程实践让我对AdaBoost又有了进一步的认识。先说编程中遇到的问题在一开始编程时，我没能理解“基于分布D_t从数据集D中训练出分类器h_t”这句话的含义，导致我的程序运行结果是无论有多少个基学习器，精度都不变。后来我发现了这处问题，在模型训练时为每个样本加上了相应的权重，解决了这个问题。
从这次编程的结果来看，基分类器的个数越多，整体精度越高，但相应的也带来了运行时间增长的问题。我觉得AdaBoost这个算法也可以做类似随机森林算法那样的随机化操作，即每次选一些属性进行训练，这样一来可以提高基训练器的多样性，也可以缩短训练时间。