机器学习导论
习题六

141220120, 徐世坚, xsj13260906215@gmail.com

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. 此处用于写解答(中英文均可)
(1)Boosting的核心思想是先基于原始数据集训练出一个基学习器，然后根据这个学习器的表现对训练集样本分布进行调整，使得先前做错的训练样本在后续得到更高的关注，然后基于调整后的数据集训练下一个基学习器，迭代进行下去。
Boosting中对训练集样本分布的调整使得基分类器具备多样性。
(2)在决策树Bagging集成中，每次选择属性划分需要考察结点所有的属性，而随机森林在每个结点上，只需要随机考察一个属性子集。所以随机森林的训练速度更快。

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) $E_{bag} = \mathbb{E}_{\mathbf{x}}[\epsilon_{bag}(\mathbf{x})^2]$
$= \mathbb{E}_{\mathbf{x}}[(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x}))^2]$
$= \mathbb{E}_{\mathbf{x}}[\frac{1}{M^2}(\sum_{m=1}^{M}\epsilon_m(\mathbf{x}))^2]$
$= \frac{1}{M^2}\mathbb{E}_{\mathbf{x}}[\sum_{m=1}^{M}\sum_{n=1}^{M}\epsilon_m(\mathbf{x})\epsilon_n(\mathbf{x})]$
$= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{n=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_n(\mathbf{x})]$
$\because \forall\ m \neq l$，$\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})] = 0$，$\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] = 0$
$\therefore E_{bag} = \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2] = \frac{1}{M}E_{av}$
(2) 由Jensen's inequality可知
$(\frac{1}{M}\sum_{m=1}^{M}\epsilon_{\mathbf{m}}(\mathbf{x}))^2 \leq \frac{1}{M}\sum_{m=1}^{M}\epsilon_{\mathbf{m}}(\mathbf{x})^2$
两边同取期望，可得
$\mathbb{E}_{\mathbf{x}}[(\frac{1}{M}\sum_{m=1}^{M}\epsilon_{\mathbf{m}}(\mathbf{x}))^2] \leq \mathbb{E}_{\mathbf{x}}[\frac{1}{M}\sum_{m=1}^{M}\epsilon_{\mathbf{m}}(\mathbf{x})^2]$
$\mathbb{E}_{\mathbf{x}}[\epsilon_{bag}(\mathbf{x})^2] \leq \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2]$
$\therefore E_{bag} \leq E_{av}$ □

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)
直观感受是，adaboost随着集成数目的增加，精度会提高很多。
另外就是，由于在sklearn中的logistic regression 是按如下方式实现的：
$\underset{w,c}{min}\ \frac{1}{2}w^Tw + C\sum_{i=1}^{n}\log(\exp(-y_i(X_i^Tw + c)) + 1)$
其中的C的作用和SVM中的具有相同作用。所以，当适当调高C的值时，单个logistics regression得到的分类器的精度会提高很多，从而导致整体的精度也会提高。

至于样本权重的归一化问题，一个现象是，不归一化的权重得到的结果反而比归一化之后的结果更好，这我不是很理解，不知道实现上是怎么处理的。

最后就是从未碰到过的坑了——Python的深浅拷贝问题。因为我在类中定义了一个基分类器成员，所以我每次都是用同一个基分类器来训练的，然后将fit得到的模型通过list.append()加进去。但是list.append()函数是浅拷贝，导致最终只有一个模型是有效的。这个错误实在是太难发现了。