机器学习导论
习题六

141242006, 袁帅, 141242006@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) The key idea in Boosting is to distinguish sample weights, attaching more importance to mistaken samples and less importance to correctly classified samples. It combines all learners' results as the output. The diversity in Boosting is provided by the different datasets or weights used in the training session.

(2) In every iteration, Random Forests randomly pick a small feature set and select an optimal feature out of that set to split, so the cost of splitting would be decreased dramatically. In addition, in Random Forest, we don't require every decision tree to be completely precise, so we can set arguments such as maximum samples per leaf or maximum depth to restrain the complexity of each decision tree, so the training process will be faster.

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)

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

(2) By Cauchy's Inequality, we obtain $\forall \boldsymbol{x}$,

$$\left(\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x})\right)^2 \leq \left(\sum_{m=1}^{M}1^2\right)\left(\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x})^2\right) = M\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x})^2,$$

which gives $\mathbb{E}_{\boldsymbol{x}}[\left(\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x})\right)^2] \leq M \cdot \mathbb{E}_{\boldsymbol{x}}[\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x})^2]$. Therefore, we have

$$E_{\text{bag}} = \mathbb{E}_{\boldsymbol{x}}[\epsilon_{\text{bag}}(\boldsymbol{x})^2] = \frac{1}{M^2}\mathbb{E}_{\boldsymbol{x}}[\left(\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x})\right)^2] \leq \frac{1}{M}\mathbb{E}_{\boldsymbol{x}}[\sum_{m=1}^{M}\epsilon_m(\boldsymbol{x})^2] = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\boldsymbol{x}}[\epsilon_m(\boldsymbol{x})^2] = E_{\text{av}}.$$

.

□

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. I wrote two versions of AdaBoost, namely *AdaBoost_resample()* and *AdaBoost_reweight()*, in regards to the different ways of utilizing sample weights. In both cases, I use the default L2-regularized Logistic Regression as base classifier. The results are listed below:

	Total running time	Accuracy			
		1 base	5 bases	10 bases	100 bases
Re-weight(C=1)	$<$ 1 second	57.52%	72.70%	73.97%	73.97%
Re-sample(C=1)	6∼8 seconds	80.68%	82.58%	84.46%	84.05%
Re-sample(C=100)	50 seconds	81.95%	84.88%	85.08%	87.60%

There is one pitfall in this assignment: the labels are 0s and 1s; however, the AdaBoost Algorithm on textbook only applies for labels {-1, +1} because there is a weighted summation process as the output. Generally speaking, the accuracy is increasing when we use more base learners, so Ensemble Learning does help! Changing parameters could make a significant difference because all base learners' performance are altered and the degree of diversity is also changed.