机器学习导论
习题六

Boardwell, Nanjing University

2017 年 7 月 10 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 core idea of Boosting algorithms are continuely adjusting each sample's weight by performance of basic learner on them and pay more attention to those wrongly classified in the last turn. At last use their combination to make the final prediction (in AdaBoost the combination is weighted addition).
The different weights assigned to each sample in each turn enable those basic learners to have diversity.
(2): Bagging: Train the complete Decision Tree model in each part of basic learners. In each layer of Decision Tree the model need to caculate each feature's information gain or Gini index to decide use which feature as a node.
Random Forest: In each layer of Decision Tree the model only need to caculate the selected k features' index and use a optimal one, which makes the caculation less than Bagging.

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):

$$\begin{aligned} E_{bag} &= \mathbb{E}_{\mathbf{x}}[(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x}))^2] \\ &= \mathbb{E}_{\mathbf{x}}[\frac{1}{M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] \end{aligned} \tag{2.7}$$

Because of $\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] = 0$ so that:

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

(2):
At first prove the following inequality:
Because of the expectation of random variable's square will always be positive, we can prove that (in this proof a and b are random variables)

$$\begin{aligned} &\mathbb{E}[(a-b)^2] \geq 0 \\ &\Rightarrow \mathbb{E}[a^2 + b^2 - 2ab] \geq 0 \\ &\Rightarrow \mathbb{E}[a^2] + \mathbb{E}[b^2] \geq \mathbb{E}[2ab] \\ &\Rightarrow \mathbb{E}[ab] \leq \frac{1}{2}(\mathbb{E}[a^2] + \mathbb{E}[b^2]) \end{aligned} \tag{2.8}$$

And back to this problem:

$$\begin{aligned} E_{bag} &= \mathbb{E}_{\mathbf{x}}[(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x}))^2] \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] \\ &\leq \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}(\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2] + \mathbb{E}_{\mathbf{x}}[\epsilon_l(\mathbf{x})^2]) \\ &= \frac{1}{M^2}\sum_{m=1}^{M}M\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2] \\ &= E_{av} \end{aligned} \tag{2.9}$$

So we have proved the original problem. □

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. (2):
Effect of parameter setting of basic learner:
Because of the regularzation term changes the loss function to minimize, the final parameter of logistic regression will differ a lot, so that basic learner's current accuracy will affect the precedure of AdaBoost a lot. Moreover, setting of parameters will also affect this model's generalization ability, which contributes to the result. If don't add the optimal regularzation term, because the logistic regression is relatively simple, basic learner may get high accuracy in training data but perform bad in test data. So we should use good regularzation term to reduce training accuracy and use AdaBoost to improve its generalization ability later.

AdaBoost：
1) AdaBoost are effective in combining many weak learners but if the origin model's accuracy is relatively high, the precedure of AdaBoost is unnecessary and hard to continue in some circumstances. (For example, if don't add the regularzation term, many classmates meet the problem of $accuracy = 1$ and the next turns of basic learner will never be updated.)
2) AdaBoost can hardly improve strong learners further more and often keep the same

accuracy although you may combine many learners together.
3) The number of learners combined need to be discovered for specific conditions. In my programming, the test accuracy in condition of T=5 is worse than T=1 (origin Logistic Regression). The accuracy of T=10 and T=100 are better than T=1 but their difference are relatively small, which means it is unnecessary to train 100 learners and only 10 is enough.