

决策树与 Boosting 方法

数 41 李博扬 2014012118

1、

(科目: 模式识别) **数 学 作 业 纸**

编号: 2014012118 班级: 数41 姓名: 李博扬 第 1 页

1. (1) 证明:

$$\begin{aligned} E_{h_B} &= \frac{1}{n} \sum_{j=1}^n (h_B(x_j) - y(x_j))^2 \\ &= \frac{1}{n} \sum_{j=1}^n \left(\frac{1}{m} \sum_{i=1}^m h_i(x_j) - y(x_j) \right)^2 \\ &= \frac{1}{n} \sum_{j=1}^n \left(\frac{1}{m} \sum_{i=1}^m (h_i(x_j) - y(x_j)) \right)^2 \\ &= \frac{1}{n} \sum_{j=1}^n \frac{1}{m^2} \left(\sum_{i=1}^m \varepsilon_i(x_j) \right)^2 \\ &= \frac{1}{m^2 n} \sum_{j=1}^n \left(\sum_{i=1}^m \varepsilon_i^2(x_j) + 2 \sum_{i < k} \varepsilon_i(x_j) \varepsilon_k(x_j) \right) \\ &= \frac{1}{m^2 n} \sum_{j=1}^n \left(\sum_{i=1}^m \varepsilon_i^2(x_j) \right) + \frac{2}{m^2 n} \sum_{j=1}^n \sum_{i < k} \varepsilon_i(x_j) \varepsilon_k(x_j) \\ &= \frac{1}{m} \bar{E}_h \quad \quad \quad = 0 \text{ (由 6)} \end{aligned}$$

(2) $E_{h_B} = \frac{1}{m^2 n} \sum_{j=1}^n \left(\sum_{i=1}^m \varepsilon_i^2(x_j) \right) \leq \frac{1}{mn} \sum_{j=1}^n \left(\sum_{i=1}^m \varepsilon_i^2(x_j) \right) = \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{j=1}^n \varepsilon_i^2(x_j) = \bar{E}_h$ □

Cauchy 不等式

2. 1 训练基于决策树桩的 adaboost (训练、测试错误率见 mat 文件)

2. 2 训练决策树

最大分裂数	训练错误率	测试错误率
4	9.0%	10.15%

20	8.7%	9.49%
100	8.6%	9.49%

2. 3 训练随机森林

分类器数	训练错误率	测试错误率
30	5.3%	8.12%
100	3.4%	4.58%
300	4.7%	4.98%

2. 4

由实验结果可见，**adaboost** 算法随着迭代次数的增加，在足够大时，训练错误率可以达到 0，测试集错误率一直降低。而决策树与随机森林没有这么明显的效果。且在训练与测试错误率方面，随机森林要优于决策树，但不如 **adaboost**。建立决策树的关键在于当前状态下选择哪个属性作为分类依据。随机森林在 **bagging** 基础上做了修改，建立多棵 **CART** 决策树形成随机森林，通过投票决定分类。**Adaboost** 通过增加训练被前一分类器分错的样本的权重，降低错误率，但对噪声数据和异常数据很敏感。相对于其他分类方法，**Adaboost** 不容易出现过拟合。