



清华大学

Tsinghua University

实验三 集成学习



实验3：集成学习

- 实现不同集成学习算法
- 对比它们与不同基分类器结合时的效果
- 共4种组合
 - 两个集成学习算法: Bagging, AdaBoost
 - 四个基分类器任选两个: Bayes, KNN, SVM, Decision Tree
- 基分类器可以调用已有工具包
- 但集成学习算法需要自己动手实现



- 基于评论的评分预测任务



Arnulfo Castillero



Problems

Reviewed in the United States on October 12, 2019

Color: Black | Size: 64GB | **Verified Purchase**

Good morning, I bought this article through Amazon and brought it to Panama, I have problems with the phone to unlock it. What should I do to solve this problem, please

- exp3-reviews.csv:220,000
- 请勿使用其他资源的数据，包括预训练的词向量



- *overall*: 标签 (label) 列, 表示用户对物品的评分 (从1到5)
- *reviewerID*: 每个评论者的唯一标识
- *asin*: 每个物品的唯一标识
- *unixReviewTime*: 评论文本的时间戳
- *summary*: 评论摘要内容, 以英文表示, 未经预处理
- *reviewText*: 评论内容, 以英文表示, 未经预处理



- 自己**实现** Bagging、AdaBoost (60%)
- 比较和分析 4 种组合 (40%)
 - 指标：至少包括 MAE、RMSE
 - 分类和回归任务均可
 - 将数据按照9:1的比例划分为训练集和测试集
 - 如果数据过大，可以对其进行采样
- **请注意**：评分不是基于模型的性能 (i.e.,准确率)，而是基于你实现算法的方式、评估其性能以及分析结果的方式



包含以下内容的**一个压缩文件**:

- Source Code:
 - 包含必要的注释
 - 确保助教能够理解并运行代码复现结果 (注意**设置随机种子**)
- README
 - 文本文件 (utf8编码), 含姓名、学号、联系方式, 及代码运行指南
- Report
 - PDF文件: 实验设计/结果/分析/讨论 (**请不要直接复制代码**)
- **请不要上传数据集**



截止时间以及其他信息

- 截止时间: **2025.04.20 Sunday 23:59:00**
 - 上传压缩文件到网络学堂, 文件名为 姓名_学号
 - 迟交作业会有扣分
 - 迟交 \leq 一周: 0.8; 迟交 $>$ 一周: 0.6
 - 如果有特殊情况, 请提前告知助教
 - 不允许抄袭代码和报告
 - 学校提供了作业查重系统, 被确认为抄袭的作业会严重扣分
- 如有任何问题, 请联系助教:
 - 王亦凡, 何祉瑜, 江楚萌
 - {yf-wang21, hezy22, jcm24}@mails.tsinghua.edu.cn



- SVM
 - Sklearn: <https://scikit-learn.org/stable/modules/svm.html>
 - LibSVM: <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>
 - SVM-light: https://www.cs.cornell.edu/people/tj/svm_light/
- 决策树
 - Sklearn: <https://scikit-learn.org/stable/modules/tree.html>
 - C4.5: <http://www.rulequest.com/Personal/>
 - C5.0: <http://www.rulequest.com/see5-info.html>
- 请注意，不能使用包内提供的集成学习工具