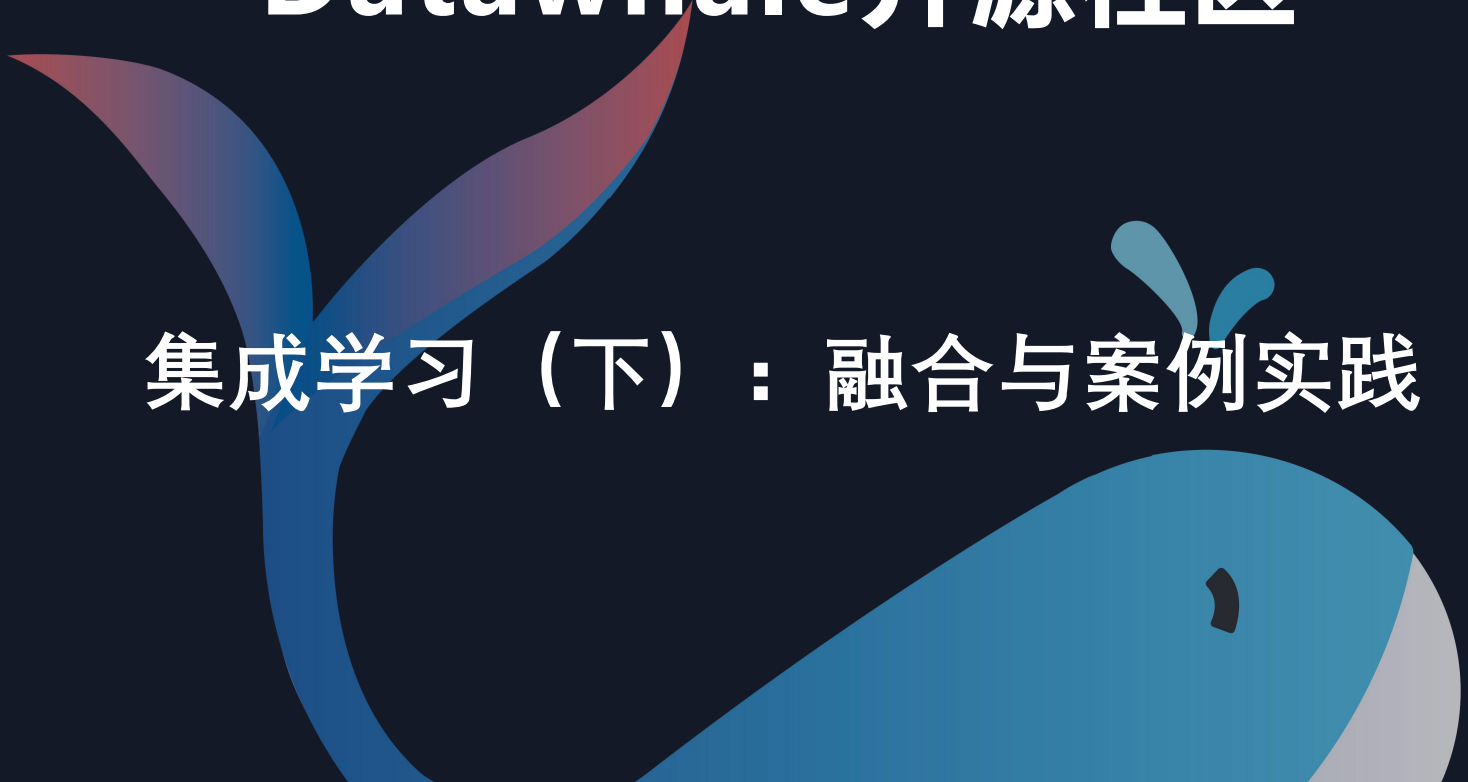
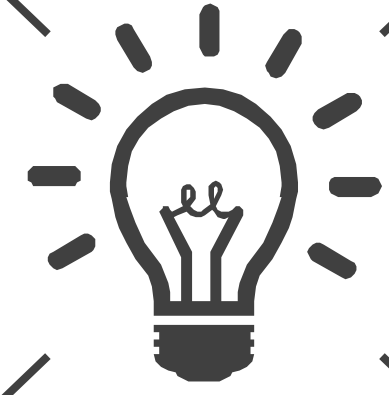
Datawhale开源社区

集成学习（下）：融合与案例实践



内容贡献者&助教

Datawhale



李祖贤
Datawhale成员，深圳大学，项目负责人
https://www.zhihu.com/people/meng-di-76-92/posts

赵可
Datawhale成员，国家电网电气工程师

薛传雨
Datawhale成员，康涅狄格大学在读博士
http://chuanyuxue.com/

杨毅远
Datawhale成员，清华大学研究生
https://github.com/yyysjz1997

陈琰钰
Datawhale成员，清华大学研究生
https://cyy0214.github.io/

李嘉骐
清华大学在读博士
https://www.zhihu.com/people/li-jia-qi-16-9/posts

课程设计理念

Datawhale

学习说明

之前大家系统学习了机器学习的经典的算法、bagging/boosting等基本集成方法的思路、理论推导和库文件调用。本次我们将学习使用多个模型的训练结果进行最终的融合。在模型训练和融合全部完成后，我们还将使用两个较大的真实数据来进行完整的调参融合练习。

任务路线1

掌握多模型集成融合方法

任务路线2

综合三期所学知识，从单模型调参到多模型融合完成完整的数据实践项目

定位人群

已完成集成学习（上）、（中）课程内容的学习，具备《高等数学》、《线性代数》、《概率论与数理统计》基础，了解机器学习常用模型、集成模型的理论知识，能够调用相应模型库解决的学习者。

学习目标和具体内容

Datawhale

Task12：Blending算法分析与案例调参实例（2天）

Blending是学习Stacking算法的基础，不知道大家小时候有没有过这种经历：老师上课提问到你，那时候你因为开小差而无法立刻得知问题的答案。就在你彷徨的时候，由于你平时人缘比较好，因此周围的同学向你伸出援手告诉了你他们脑中的正确答案，因此你对他们的答案加以总结和分析最终的得出正确答案。相信大家都有过这样的经历，这就是Blending算法的核心。

Task13：Stacking算法分析与案例调参实例（2天）

Blending在集成的过程中只会用到验证集的数据，对数据实际上是一个很大的浪费。为了解决这个问题，如果能将交叉验证的思想附加到集成算法中，将能顺利解决这个问题，Stacking算法就是这么诞生的。Stacking的思路是先使用交叉验证训练多个不同模型，然后使用Blending方法将交叉验证的结果堆叠融合，以获得更高的预测准确率。

Task14：集成学习案例一（幸福感预测）（5天）

通过前13个task的学习，我们已经掌握了集成学习的基础知识和多种基本算法，那集成学习中的算法在实践中是如何使用的呢？“幸福感预测”这一案例就是以分类为目标的集成学习。此案例是一个数据挖掘类型的比赛——幸福感预测的baseline。比赛的数据使用的是官方的《中国综合社会调查（CGSS）》文件中的调查结果中的数据，其共包含有139个维度的特征，包括个体变量（性别、年龄、地域、职业、健康、婚姻与政治面貌等等）、家庭变量（父母、配偶、子女、家庭资本等等）、社会态度（公平、信用、公共服务）等特征。

Task15：集成学习案例二（蒸汽量预测）（5天）

不同于task14的“幸福感预测”的分类问题，本案例中的“蒸汽量预测”是以回归为目标的集成学习。此案例的数据产生于实际的工业大数据的生产中，具有十分重要的现实意义，经脱敏后的锅炉传感器采集的数据（采集频率是分钟级别），根据锅炉的工况，预测产生的蒸汽量。与上面的案例相同，本案例展示了一个完整的集成学习解决方案供大家参考和改进。通过以上较为全面的案例分析，希望帮助大家更加直观、深入地学会使用集成学习的思想来解决自己所面对的问题。

如何更好地学习本课程

Datawhale

01

理解Blending&Stacking的集成思路理念，特别是堆叠融合的思路。在熟练使用的基础上，可以尝试将不同的模型和集成方法堆叠使用，组合、创造出新的集成方法。

02

了解并掌握几种常见的数据预处理方法及可视化方法，能够针对业务特性和数据特性，对竞赛数据进行常规的预处理及可视化，满足后续建模调参的需要。

03

知晓各种常用模型、各种多模型集成方法的特性和优缺点，能够根据需要和数据情况选择合适的模型并使用合适的方法进行多模型集成，最终得出优于单模型的结果。

组队学习模式

Datawhale

集成学习开源内容

https://github.com/datawhalechina/team-learning-data-mining/tree/master/EnsembleLearning

学习与交流

学习输出

如何提问

Datawhale

相互尊重

减少重复提问

请大家注意关注群公告，有关学习内容、学习方法和组队学习流程等，都会在公告中提示。提问前请先翻阅公告，公告已有内容不再重复回答。

准确描述问题

提问请尽量准确地描述问题，并附上完整截图。若没有得到解答请再次提问并@答疑助教萌弟（李祖贤）。

围绕学习内容

组队学习期间请尽量围绕本次学习内容进行提问，与本次学习关联不大的提问若被助教忽略请谅解。

如何完成作业

Datawhale

01 选择一个分享平台如CSDN、Github（经常打不开，不建议使用）、简书、B站等

02 将自己的学习体会，输出成学习笔记或学习视频

03 将分享的链接，填至小程序的“打卡链接”中相应的位置

内　容

打卡包括但不限于对理论知识的理解、扩展、代码实现、公式推导等等，也可直播分享自己的学习过程。不需要复制粘贴教程原文。

提　示

如果笔记中需要引用教程内容或其它资料，希望注明出处，并附上来源链接，避免版权纠纷。

- 字数少于50
- 教程复制粘贴
- 教程大纲复制粘贴
- 与本任务内容无关
- 错误链接

以上情况将被视同未打卡，由助教抱出群并关闭后续打卡！

Datawhale

对学习者最有价值的开源社区