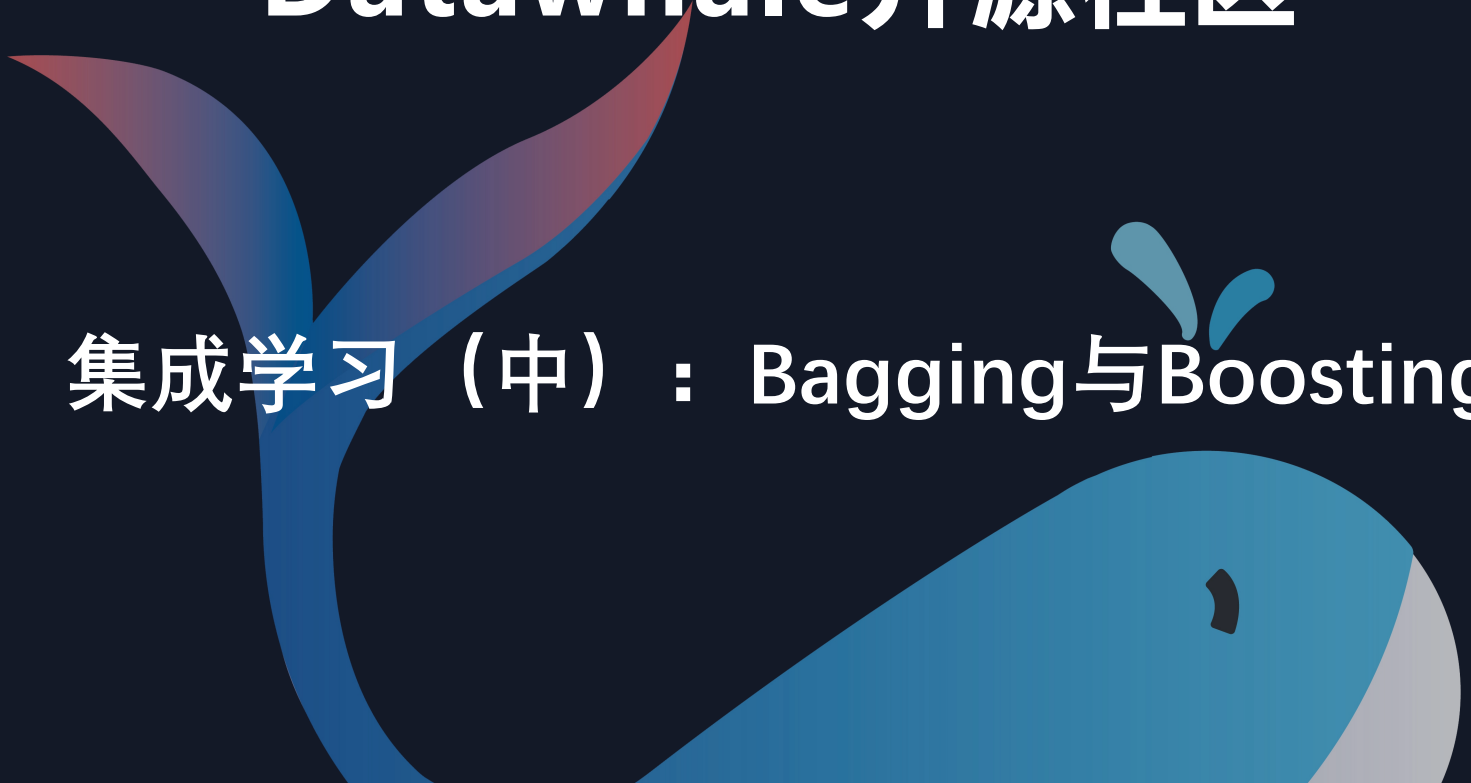
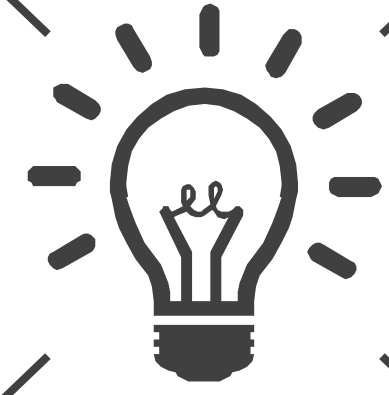
Datawhale开源社区

集成学习（中）：Bagging与Boosting



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

任务路线

掌握基本的模型集成方法和常见组合集成模型的理论及模型调用调参

定位人群

已完成集成学习（上）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，了解机器学习经典模型的理论知识，能够调用相应模型库解决的学习者。

学习说明

上期学习大家系统了解了机器学习的经典的算法数学推导和代码调用，本期我们将进行bagging/boosting等基本集成方法以及常见的集成方法组合的学习。我们依然对于每个算法都进行了细致的理论分析以及必要的代码演示，希望大家的理论知识水平和代码实践能力两个方面都能够获得均衡的提高。在案例的代码中，我们给出了详细的代码注释，尽量让学习者们不会因为看不懂代码而感到烦恼。

如何更好地学习本课程

Datawhale

01

算法工程师

机器学习研究的话，就得从头到尾每个公式自己手推，推完公式后查看sklearn官方api查看各个参数意义。

跨学科使用算法

只需要理解算法的建模思路，理清公式脉络后查看sklearn文档使用api建模即可。

02

理解 jupyter notebook 中的文字注释和代码（若发现有任何问题，大家可以及时与助教反馈），重点是理解代码思路和代码逻辑，什么时候该干什么内容。

03

在上面的基础上，大家可以举一反三，例如尝试自己使用其他数据集进行建模和预测，甚至自己尝试对各种集成方法进行组合等等。

组队学习模式

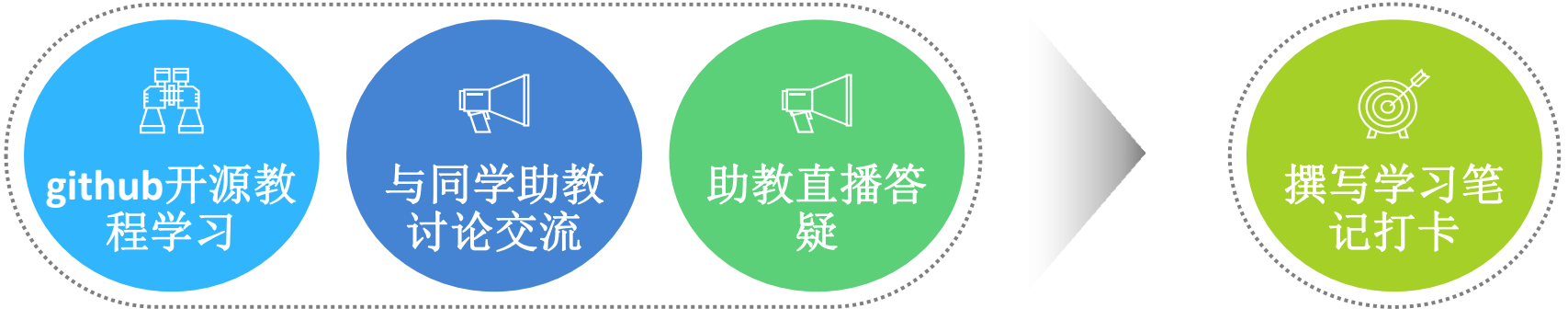
Datawhale

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



学习目标和具体内容

Datawhale

Task07：投票法的原理和案例分析（3天）	投票法是集成学习中常用的技巧，可以帮助我们提高模型的泛化能力，减少模型的错误率。它的数学原理相对简单，算法实现相对容易且快速。主要需要掌握回归预测、硬投票软投票等思路。
Task08：Bagging的原理和案例分析（3天）	Bagging不仅仅集成模型最后的预测结果，同时采用一 定策略来影响基模型训练，保证基模型可以服从一定的假设。Bagging的核心在于自助采样(bootstrap)这一概念，即有放回的从数据集中进行采样。
Task09：Boosting的思路与Adaboost算法（2天）	Boosting是与Bagging截然不同的思想，Boosting方法是使用同一组数据集进行反复学习，得到一系列简单模型，然后组合这些模型构成一个预测性能十分强大的机器学习模型。显然，Boosting思想提高最终的预测效果是通过不断减少偏差的形式，与Bagging有着本质的不同。
Task10：前向分步算法与梯度提升决策树（2天）	GBDT是回归树而不是分类树，它使用加法模型+前向分步算法的框架实现回归问题。和AdaBoost的主要区别就在于AdaBoost是在每一次迭代中修改样本权重来使得后一次的树模型更加关注被分错的样本，而GBDT则是后一次树模型直接去拟合残差。
Task11：XGBoost算法分析与案例调参实例（2天）	XGBoost是一个优化的分布式梯度增强库。它是陈天奇等人开发的一个开源机器学习项目，高效地实现了GBDT算法并进行了算法和工程上的许多改进，被广泛应用在Kaggle竞赛及其他许多机器学习竞赛中并取得了不错的成绩。近年机器学习方面的竞赛高分思路，绝大部分都使用了xgboost以及它的优化改进模型。

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！

后续课程

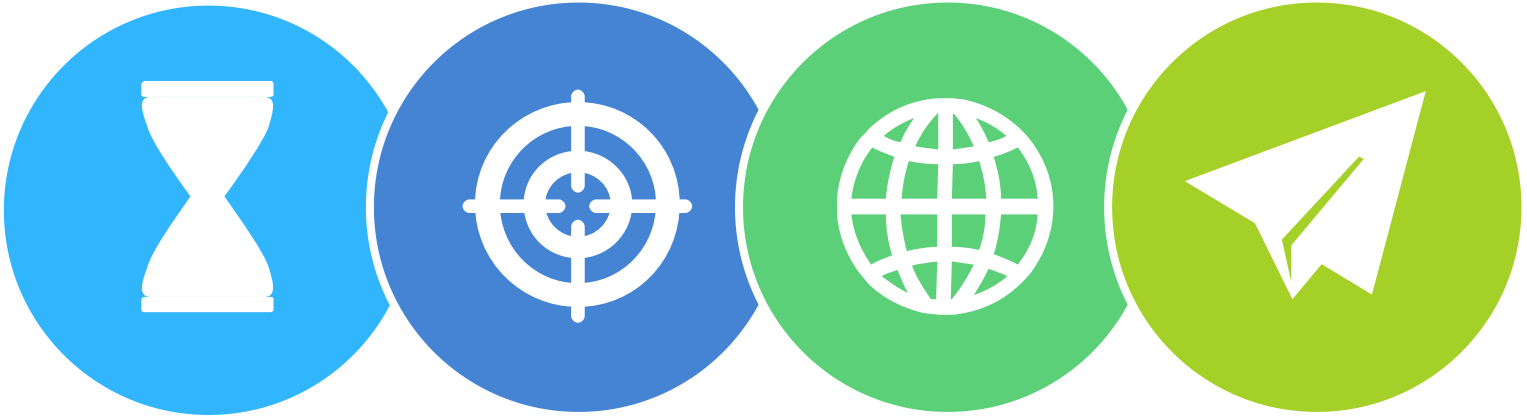
Datawhale

Stacking

Stacking称为“懒人”算法，因为它不需要理解太多的理论，只需要在实际中加以运用即可。

蒸汽量预测

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。



Blending

Blending会造成信息泄露，所以我们重点学习它的改进版。

幸福感预测

一个数据挖掘类型的比赛的baseline。

Datawhale

对学习者最有价值的开源社区