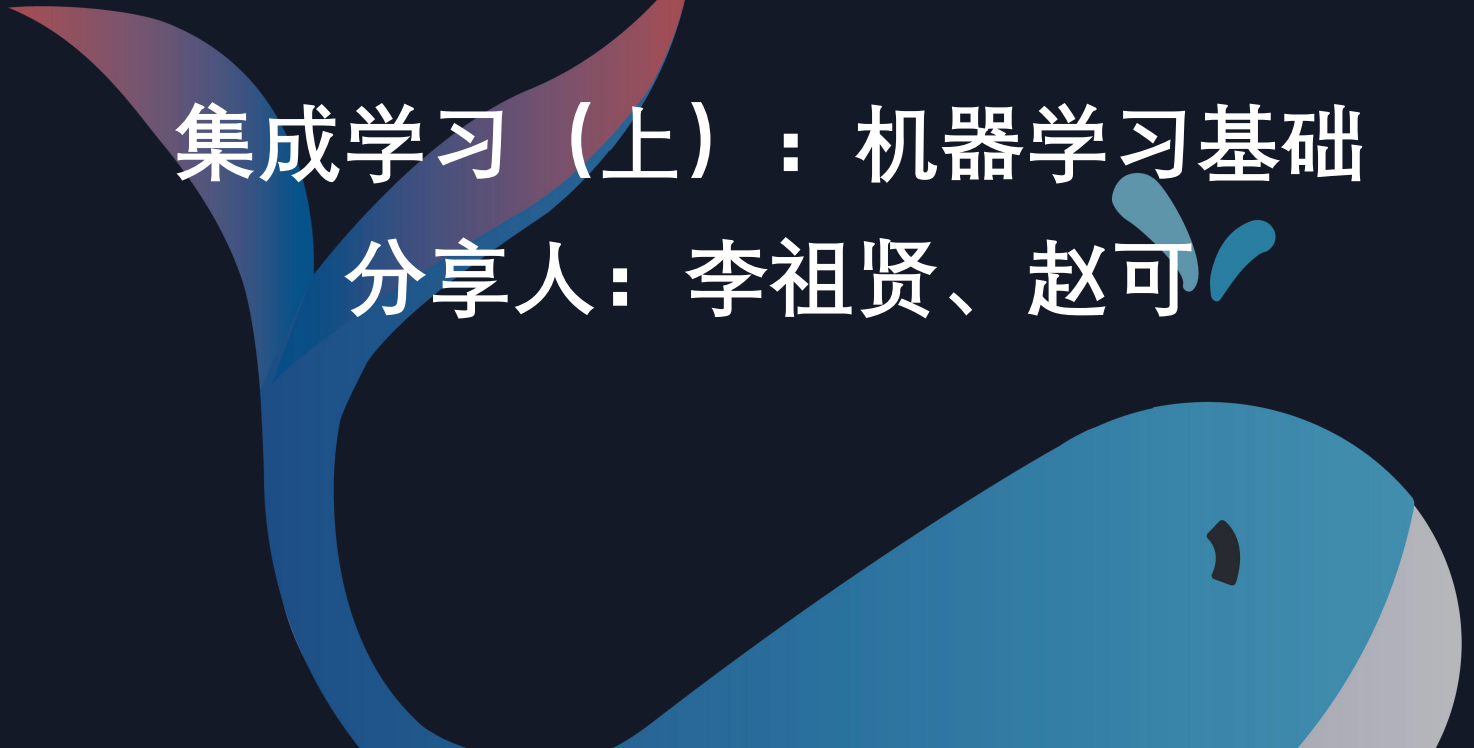
Datawhale开源社区

集成学习（上）：机器学习基础

分享人：李祖贤、赵可



内容贡献者&助教

Datawhale

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

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

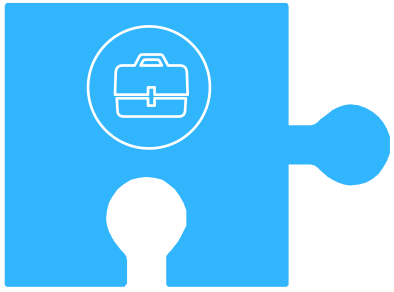
赵可
Datawhale成员，国家电网电气工程师

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

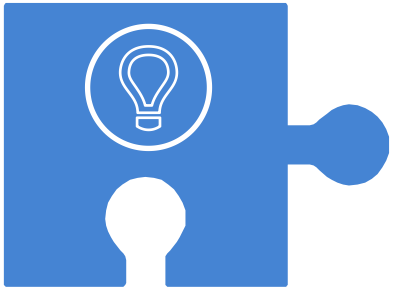
陈琰钰
Datawhale成员，清华大学研究生
https://cyy0214.github.io/

课程设计理念

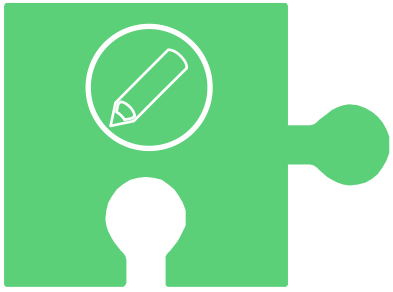
Datawhale



在kaggle等大型数据科学竞赛的高分选手模型中，我们发现一个现象：除了深度学习以外的高分模型，无一例外地见到了集成学习和模型融合的身影。



这个发现迫使我们去学习一些除了基础模型以外的集成学习方法以便在这些比赛上获得更好的成绩。



另外，当我们使用具体的sklearn库的时候，往往因为不懂得模型的一些底层知识而不懂参数的含义。

因此，在本项目中我们会从基础模型的推导以及 sklearn应用过渡到使用集成学习的技术去优化我们的基础模型，使得我们的模型能更好地解决机器学习问题。

学习目标和具体内容

Datawhale

Task01：熟悉机器学习的三大主要任务（1天）	了解传统机器学习领域的三大基本任务——回归、分类、无监督学习。
Task02：掌握基本的回归模型（3天）	掌握基本回归问题中的线性回归以及如何打破线性回归的假设推广至非线性回归，包括多项式回归、广义可加模型、回归树以及支持向量回归，在掌握了这些理论的基础上了解如何使用python及其工具库实现这些算法。
Task03：掌握偏差与方差理论（2天）	在前面的基本回归模型的建模中，我们一直使用最小化训练误差原则，但实际的问题是我们想要最小化未知数据的误差，因此如何权衡训练误差和未知的测试数据误差就是一个急需解决的问题，掌握偏差与方差理论有利于提高模型预测未知数据的能力。偏差与方差的权衡是机器学习基本模型推广至集成学习的关键，也是机器学习面试中必问的一个问题。
Task04：掌握回归模型的评估及超参数调优（3天）	数据科学永恒不变的主题也许就是调参吧，正确的调参姿势也是建立在正确评估模型的基础上的。因此我们要从偏差与方差理论中得到启发，从数学理论和代码上掌握回归模型的评估及超参数调优。
Task05：掌握基本的分类模型（3天）	也许大家并不清楚，分类问题也是从回归问题推广而来的，也正是打破线性回归的基本假设而延伸出多种多样的分类模型。我们需要掌握分类问题中的逻辑回归、基于概率的分类模型（线性判别分析、朴素贝叶斯）、分类决策树、支持向量机以及核函数。
Task06：掌握分类问题的评估及超参数调优（2天）	我们需要像回归问题那样，对分类问题进行正确的评估以及超参数的选择，由于前面回归问题的理论支撑，分类问题的模型评估及超参数选择应该会得心应手！

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机器学习建模和预测，并尝试从偏差-方差理论优化模型等等。

后续课程

Datawhale

两个案例

幸福感预测：一个数据挖掘类型的比赛的baseline。
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与Stacking

Blending会造成信息泄露，所以我们重点学习它的改进版。
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。

Xgboost算法与实例

XGBoost是一个优化的分布式梯度增强库，旨在实现高效，灵活和便携。 它在Gradient Boosting框架下实现机器学习算法。

Boosting、Adaboost与GBDT

Boosting：一组分类器的结果集合得到准确的结果。
Adaboosting：没有先验知识情况下的自适应改进。
GBDT：通过计算梯度来降低错误率。

投票法与Bagging

投票法可以帮助我们提高模型的泛化能力，减少模型的错误率。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。

Datawhale

对学习者最有价值的开源社区