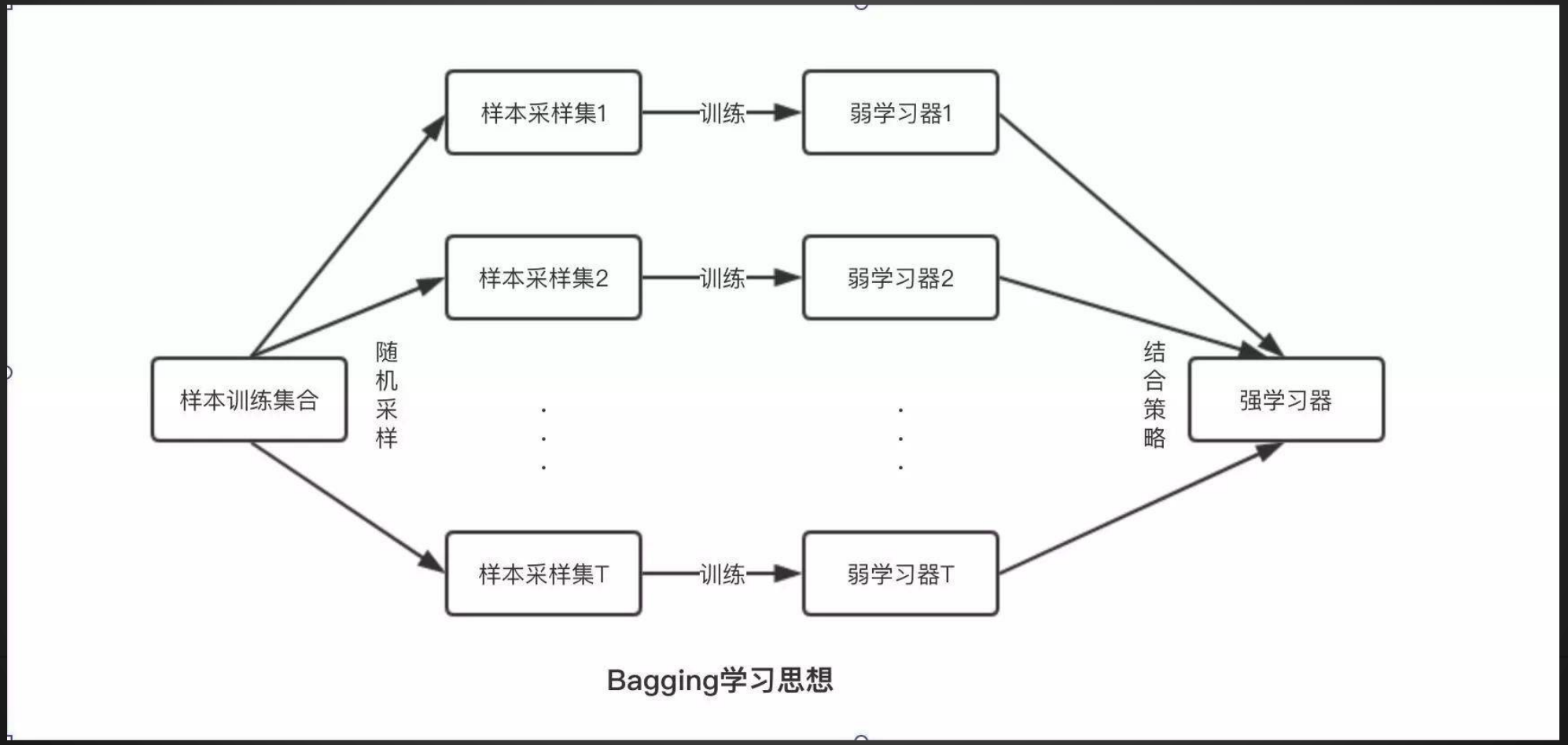
集成学习（Ensemble Learning）

集成学习

- 集成：通过集成多个机器学习模型来构建更强大的模型。
- 三个臭皮匠顶个诸葛亮
- 两个流派：Bagging流派和Boosting流派

Bagging流派

- Bagging是bootstrap aggregating的缩写，bootstrap表示自助抽样法，是一种有放回的随机抽样，aggregating表示合计、聚合的意思。



Bagging学习思想

Bagging算法典型代表：随机森林

- 随机森林由多个决策树组成，其构造过程分为以下三步。

1. 通过对样本数据进行随机抽样，对数据的行（样本）和列（特征）都进行抽样。
2. 用第1步得到的多个样本数据和对应的特征，训练出多棵不同的决策树。
3. 预测，对于一个样本来说，每一棵树都会给出一个预测结果，对每一棵树的预测结果进行投票，哪个类别多，这个样本就属于哪个类别。

- 调参涉及的重要参数：

1. 决策树的个数：n_estimators
2. 特征抽样数量：max_features
3. 树的剪枝：max_depth

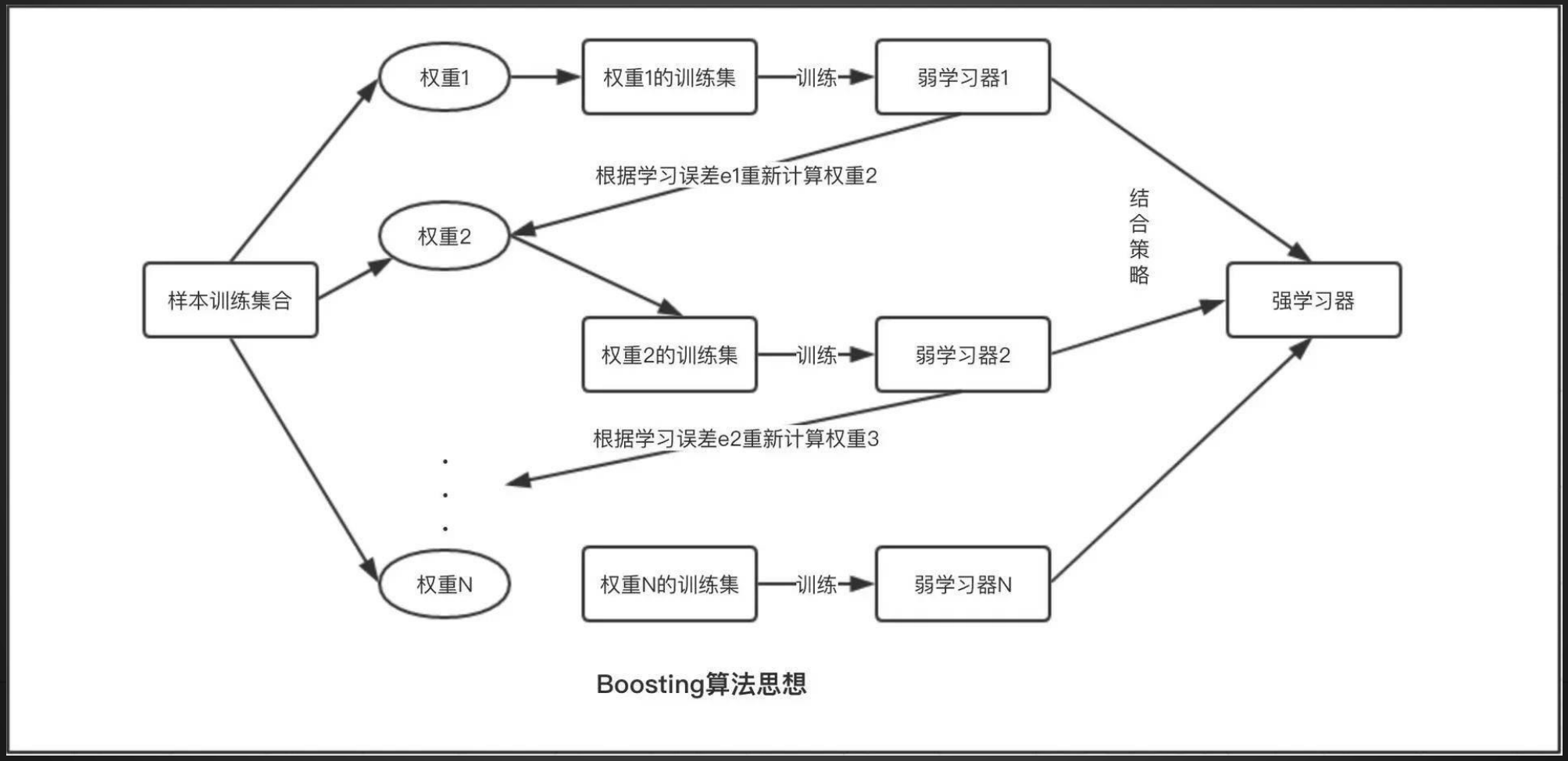
案例：利用随机森林预测企业员工是否离职

1. 读取数据、认识数据
2. 数据探索及预处理
3. 随机森林建模
4. 模型评估及调参
5. 查看特征重要性

关于案例介绍、数据读取、数据探索及预处理请参考逻辑回归中的讲解。

Boosting流派

- 英文单词Boosting表示提升的意思，也叫提升方法。



Boosting算法典型代表

- Boosting算法中的典型代表：梯度提升决策树、AdaBoost及XGBoost等。
- 梯度提升决策树，英文名称是Gradient Boosting Decision Tree，简称GBDT，是一种基于决策树的集成学习方法。
- 梯度提升决策树通过连续不断地构造决策树，每一棵树都试图纠正前一棵树的错误，串行地构造一个强学习器。

- 调参涉及的重要参数：

1. 决策树的个数：n_estimators
2. 学习率：learning_rate

案例：利用梯度提升回归树预测企业员工是否离职

1. 读取数据、认识数据
2. 数据探索及预处理
3. 梯度提升决策树建模
4. 模型评估及调参
5. 查看特征重要性

关于案例介绍、数据读取、数据探索及预处理请参考逻辑回归中的讲解。