随机森林

1、定义

随机森林：是由多棵决策树构建而成的，多棵决策树一起运算整合---->**集成算法**

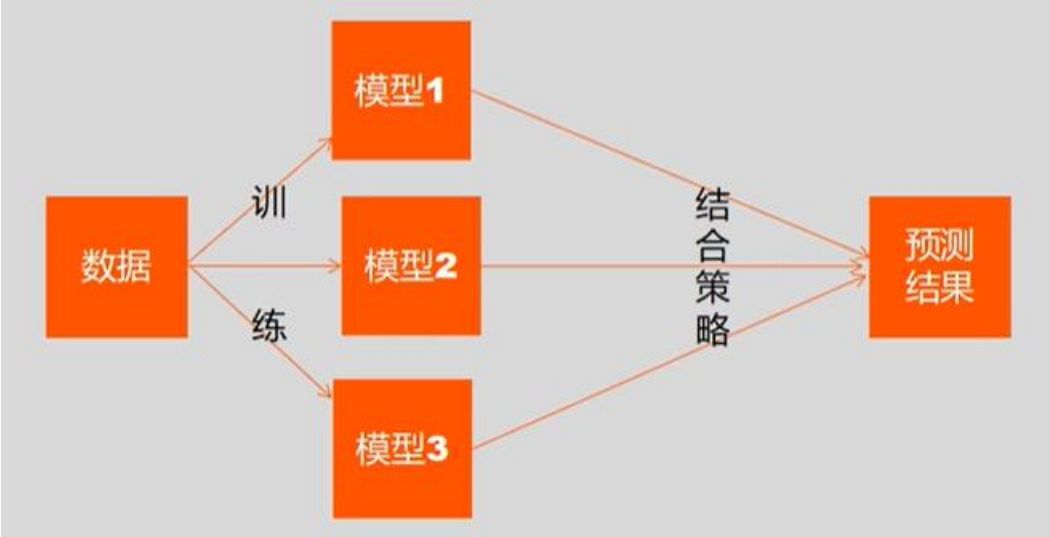


图 1 集成学习算法结构图

集成学习的范例 1（多数表决原则）：

	A	B
模型1	1	0
模型2	0	1
模型3	0	1
模型4	1	0
模型5	0	1

A:2
B:3
3>2
结果为**B**

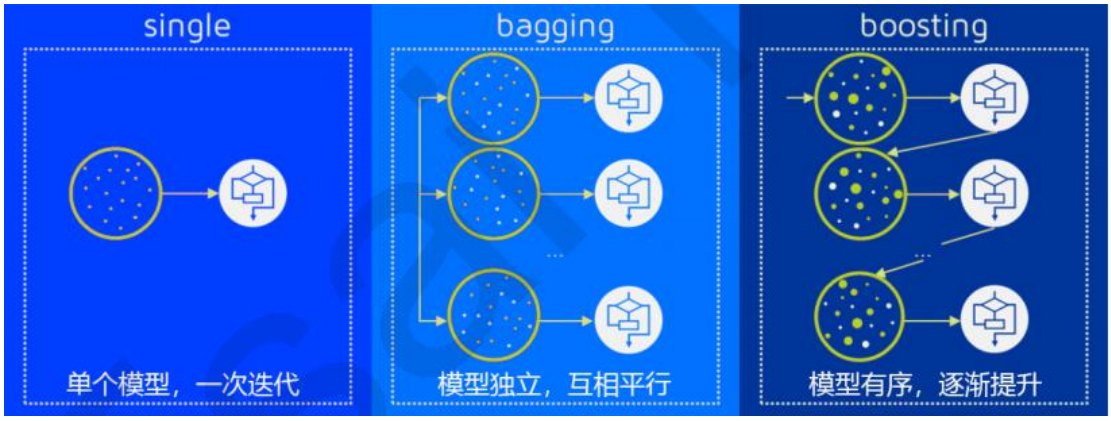
有 5 个模型，第一个模型预测结果是 A，第二个模型预测结果是 B，第三个模型预测结果是 B，第四个模型预测结果是 A，第五个模型预测结果是 B；

集成学习的范例 2（平均原则）：

	A	B
模型1	90%	10%
模型2	40%	60%
模型3	30%	70%
模型4	80%	20%
模型5	20%	80%

A：(0.9+0.4+0.3+0.8+0.2)/5=0.52
B：(0.1+0.6+0.7+0.2+0.8)/5=0.48
0.52>0.48
结果为A

多个模型集成成为的模型叫做集成评估器（ensemble estimator），组成集成评估器的每个模型都叫做基评估器（base estimator）。通常来说，有三类集成算法：装袋法（Bagging），提升法（Boosting）和 stacking。



装袋法（Bagging）	随机森林
提升法（Boosting）	Adaboost 和梯度提升树

2、随机性

图 2 随机森林定义形象比喻

随机森林的随机性体现在：

- 每棵决策树的训练样本是随机的
- 决策树中每个节点的分裂属性集合也是随机选择确定的

有了这 2 个随机的保证，随机森林就不会产生过拟合的现象了。

随机森林的本质是一种**装袋集成算法**（bagging），装袋集成算法是对基评估器的预测结果进行**平均**或用**多数表决原则**来决定集成评估器的结果。

3、Bagging

3.1 bootstrap & oob_score

3.1.1 bootstrap

- 要让基分类器尽量都不一样，一种很容易理解的方法是使用不同的训练集来进行训练，而袋装法正是通过**有放回的随机抽样技术**来形成不同的训练数据，bootstrap 就是用来控制抽样技术的参数。
- bootstrap 参数默认 True，代表采用这种有放回的随机抽样技术。

3.1.2 (out of bag data，简写为 oob)

- 在使用随机森林时，我们可以不划分测试集和训练集，只需要用袋外数据来测试我们的模型即可。
- 如果希望用袋外数据来测试，则需要在实例化时就将 oob_score 这个参数调整为 True，训练完毕之后，我们可以用随机森林的另一个重要属性：oob_score_来查看我们的在袋外数据上测试的结果。