**集成学习**（ensemble）思想是为了解决单个模型或者某一组参数的模型所固有的缺陷，从而整合起更多的模型，取长补短，避免局限性。随机森林就是集成学习思想下的产物，将许多棵决策树整合成森林，并合起来用来预测最终结果。

**bootstrap**：即从样本自身中再生成很多可用的同等规模的新样本，从自己中产生和自己类似的，所以叫做自助，即不借助其他样本数据。自助法的具体含义如下：

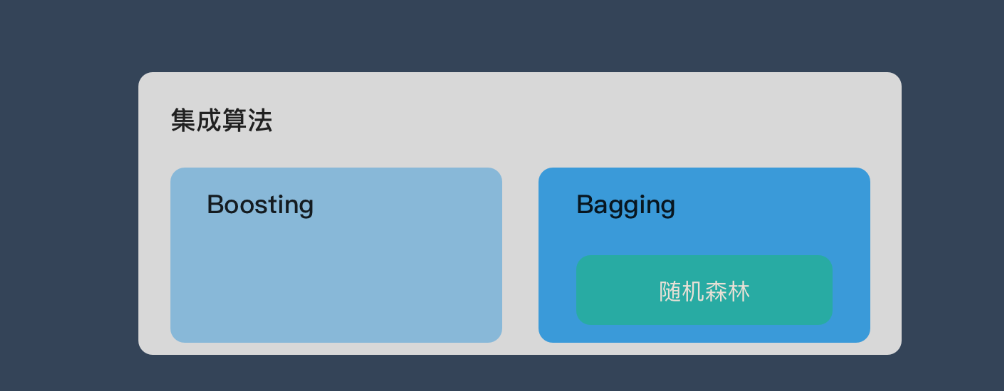
如果我们有个大小为N的样本，我们希望从中得到m个大小为N的样本用来训练。那么我们可以这样做：首先，在N个样本里随机抽出一个样本x1，然后记下来，放回去，再抽出一个x2，… ，这样重复N次，即可得到N的新样本，这个新样本里可能有重复的。重复m次，就得到了m个这样的样本。实际上就是一个有放回的随机抽样问题。每一个样本在每一次抽的时候有同样的概率（1/N）被抽中。

这个方法在样本比较小的时候很有用，自助法不会降低训练样本的规模，又能留出验证集（因为训练集有重复的，但是这种重复又是随机的），因此有一定的优势。

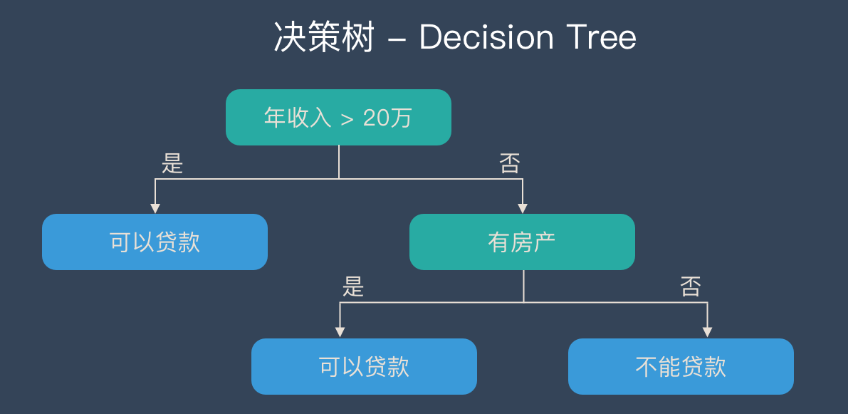
**bagging**意思是自助抽样集成，这种方法将训练集分成m个新的训练集，然后在每个新训练集上构建一个模型，各自不相干，最后预测时我们将这个m个模型的结果进行整合，得到最终结果。整合方式就是：分类问题用majority voting，回归用均值。

**什么是随机森林？**

随机森林属于 集成学习 中的 Bagging（Bootstrap AGgregation 的简称） 方法。如果用图来表示他们之间的关系如下：

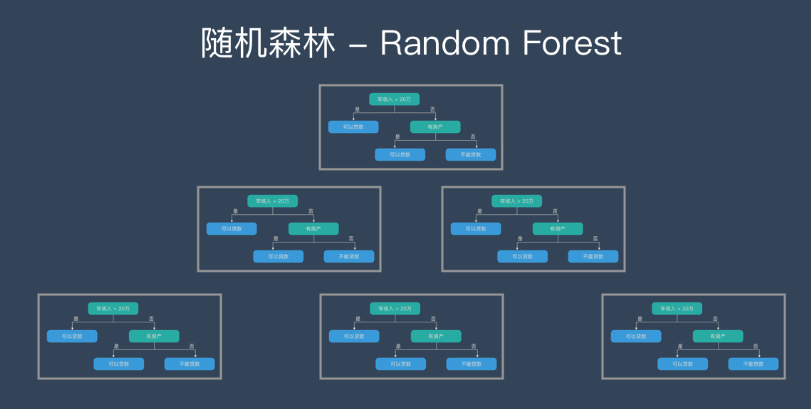


**决策树 - Decision Tree**



**随机森林 - Random Forest | RF**

随机森林是由很多决策树构成的，不同决策树之间没有关联。



当我们进行分类任务时，新的输入样本进入，就让森林中的每一棵决策树分别进行判断和分类，每个决策树会得到一个自己的分类结果，决策树的分类结果中哪一个分类最多，那么随机森林就会把这个结果当做最终的结果。对于回归问题则是取它们的平均值。

**构造随机森林的 4 个步骤**

1.假如有N个样本，则有放回的随机选择N个样本(每次随机选择一个样本，然后返回继续选择)。这选择好了的N个样本用来训练一个决策树，作为决策树根节点处的样本。

2.假设每个样本有M个属性，在决策树的每个节点进行分裂时，随机从这M个属性中选取出m个属性，满足条件m << M。然后从这m个属性中采用某种策略（比如说信息增益）来选择1个属性作为该节点的分裂属性。

3.决策树形成过程中每个节点都要按照步骤2来分裂（很容易理解，如果下一次该节点选出来的那一个属性是刚刚其父节点分裂时用过的属性，则该节点已经达到了叶子节点，无须继续分裂了）。一直到不能够再分裂为止。注意整个决策树形成过程中没有进行剪枝。

4.按照步骤1~3建立大量的决策树，这样就构成了随机森林了。