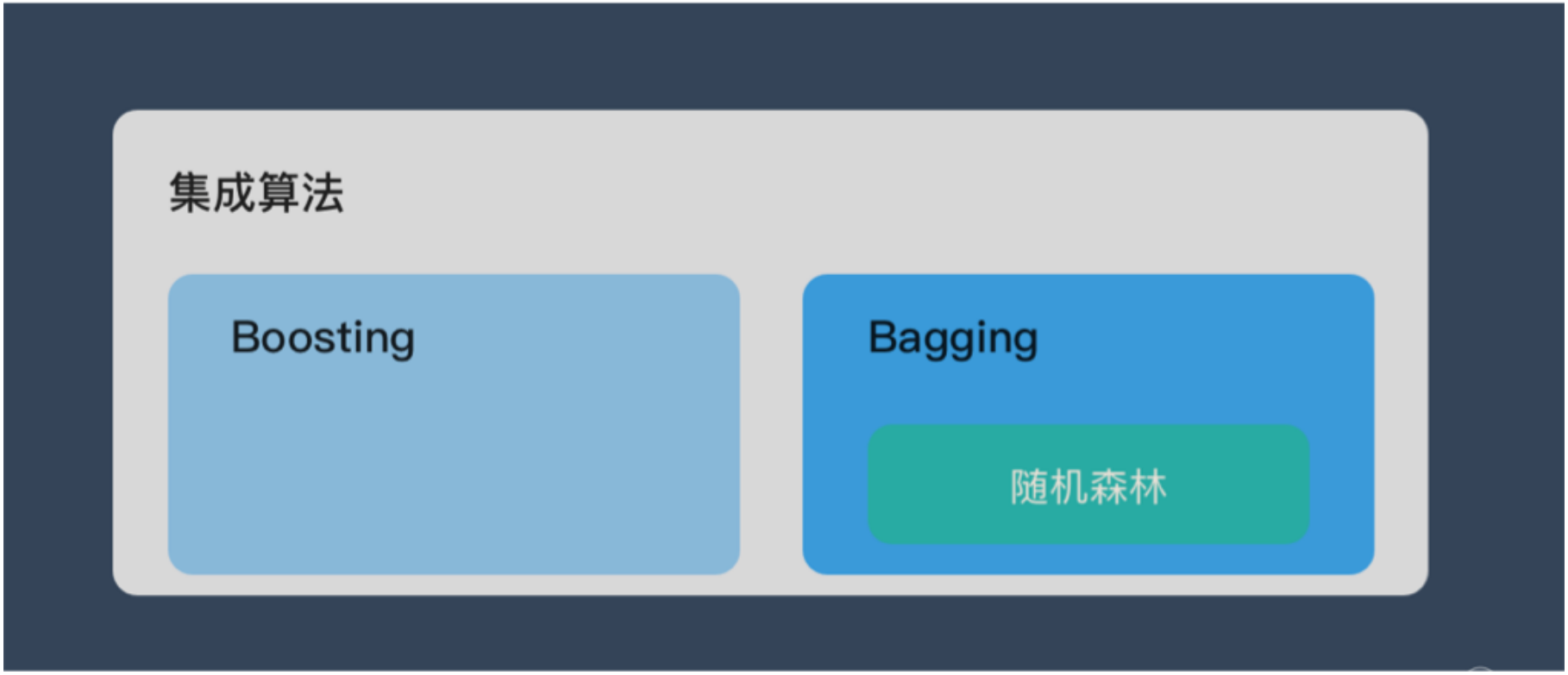
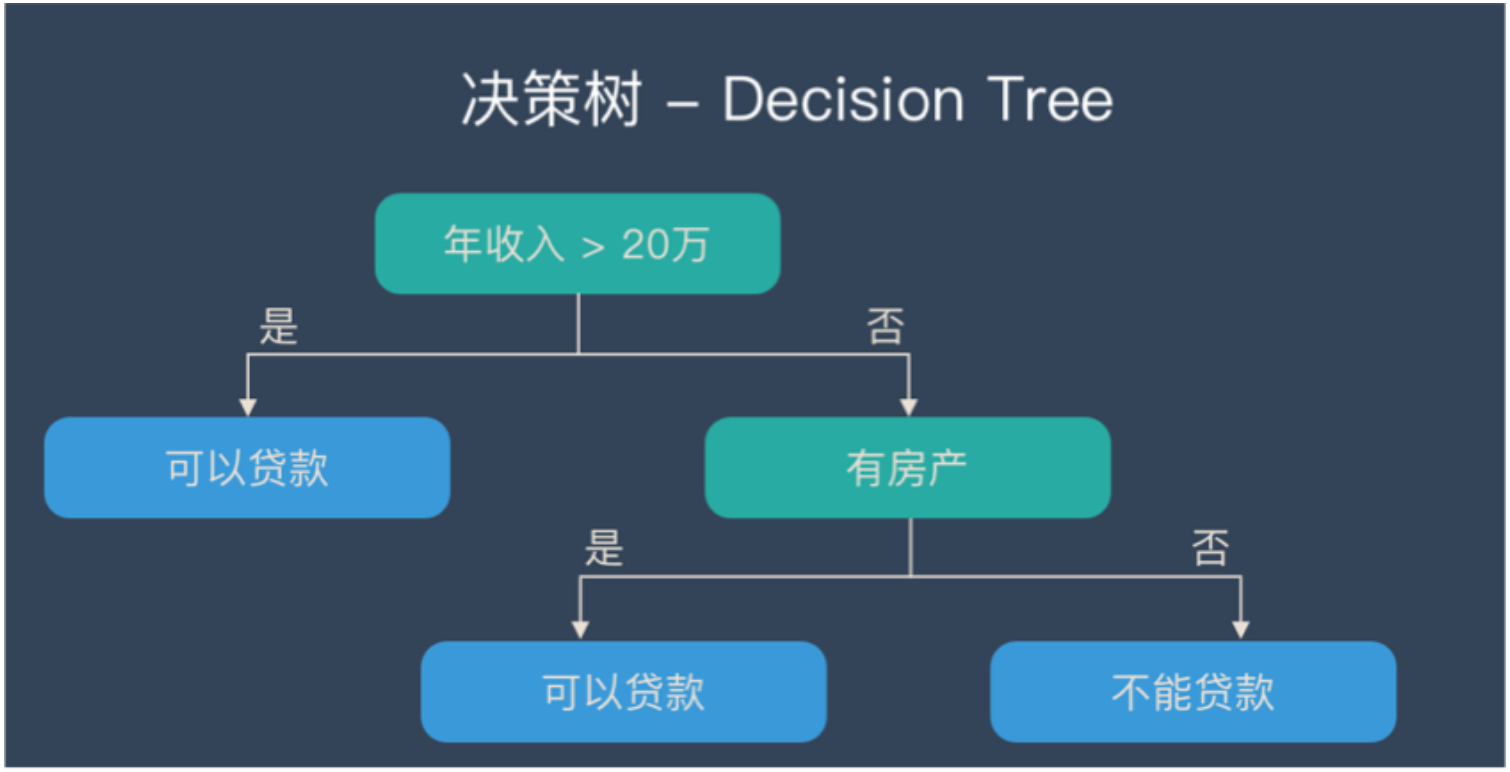
集成分类器之随机森林

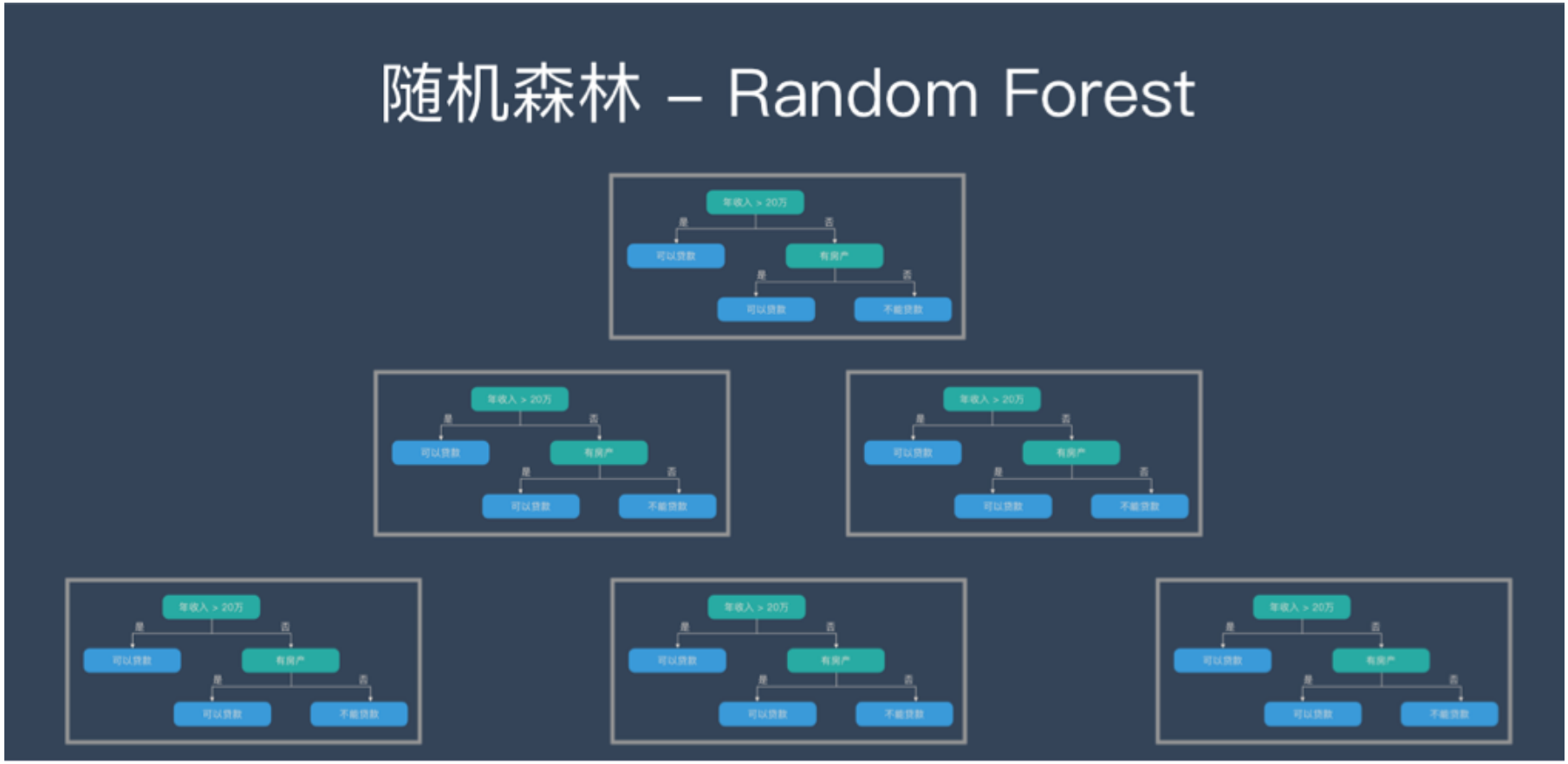
什么是随机森林？



决策树回顾



随机森林 – Random Forest



构造随机森林的 4 个步骤

Step 1

随机抽样

训练决策树

Step 2

随机选取属性

做节点分裂属性

Step 3

重复步骤 2

直到不能再分裂

Step 4

建立大量决策树

形成森林

随机森林的优点

- 它可以出来很高维度（特征很多）的数据，并且不用降维，无需做特征选择
- 它可以判断特征的重要程度
- 可以判断出不同特征之间的相互影响
- 不容易过拟合
- 训练速度比较快，容易做成并行方法
- 实现起来比较简单
- 对于不平衡的数据集来说，它可以平衡误差。
- 如果有很大一部分的特征遗失，仍可以维持准确度。

随机森林的缺点

- 随机森林已经被证明在某些噪音较大的分类或回归问题上会过拟合。
- 对于有不同取值的属性的数据，取值划分较多的属性会对随机森林产生更大的影响，所以随机森林在这种数据上产出的属性权值是不可信的