集成学习

什么是集成学习

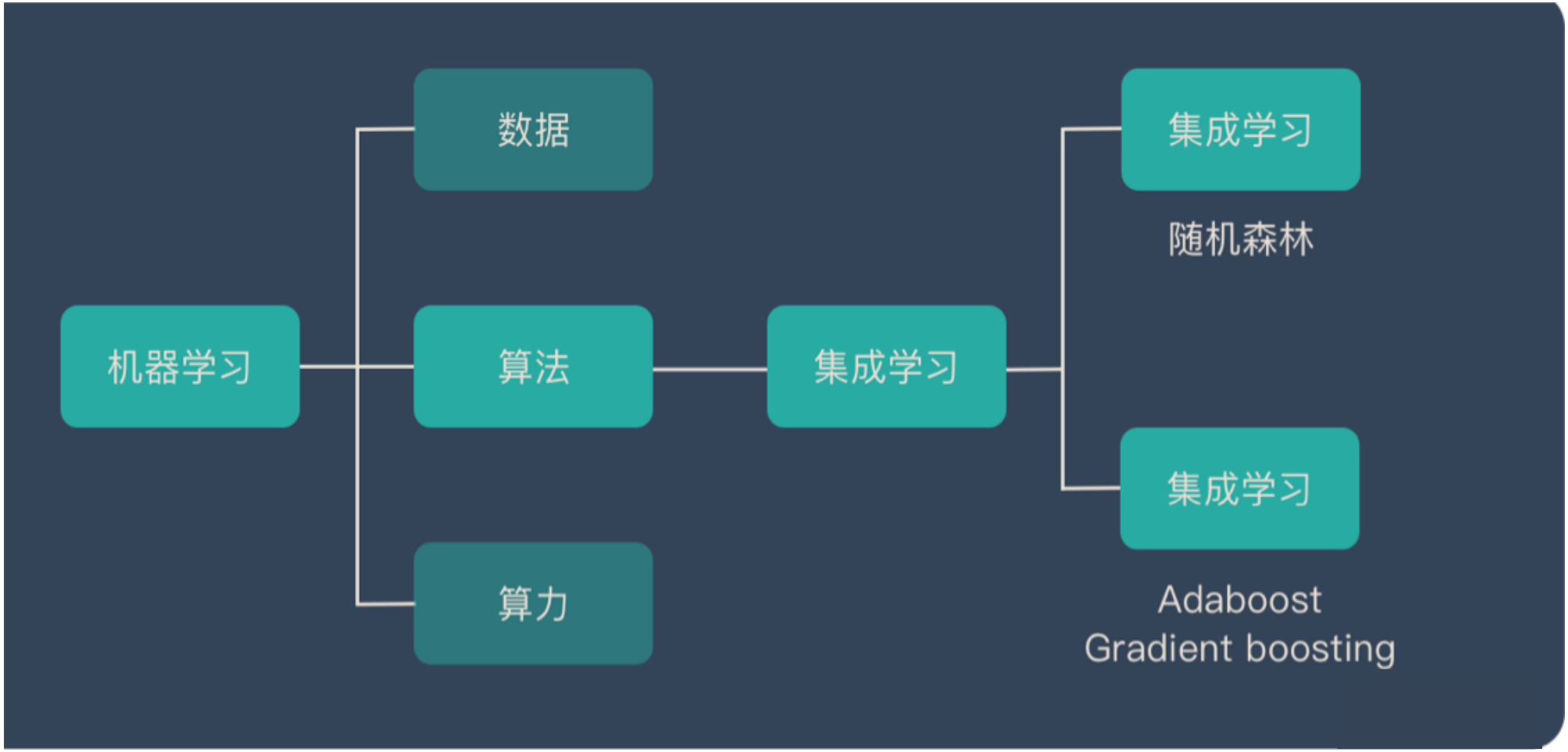
- **集成学习归属于机器学习，他是一种「训练思路」，并不是某种具体的方法或者算法。**

什么是集成学习

- 集成学习会挑选一些简单的基础模型进行组装，组装这些基础模型的思路主要有 2 种方法：

1. bagging（称作“套袋法”）
2. boosting

什么是集成学习


机器学习
数据
算法
算力
集成学习
集成学习
随机森林
集成学习
Adaboost
Gradient boosting

Bagging

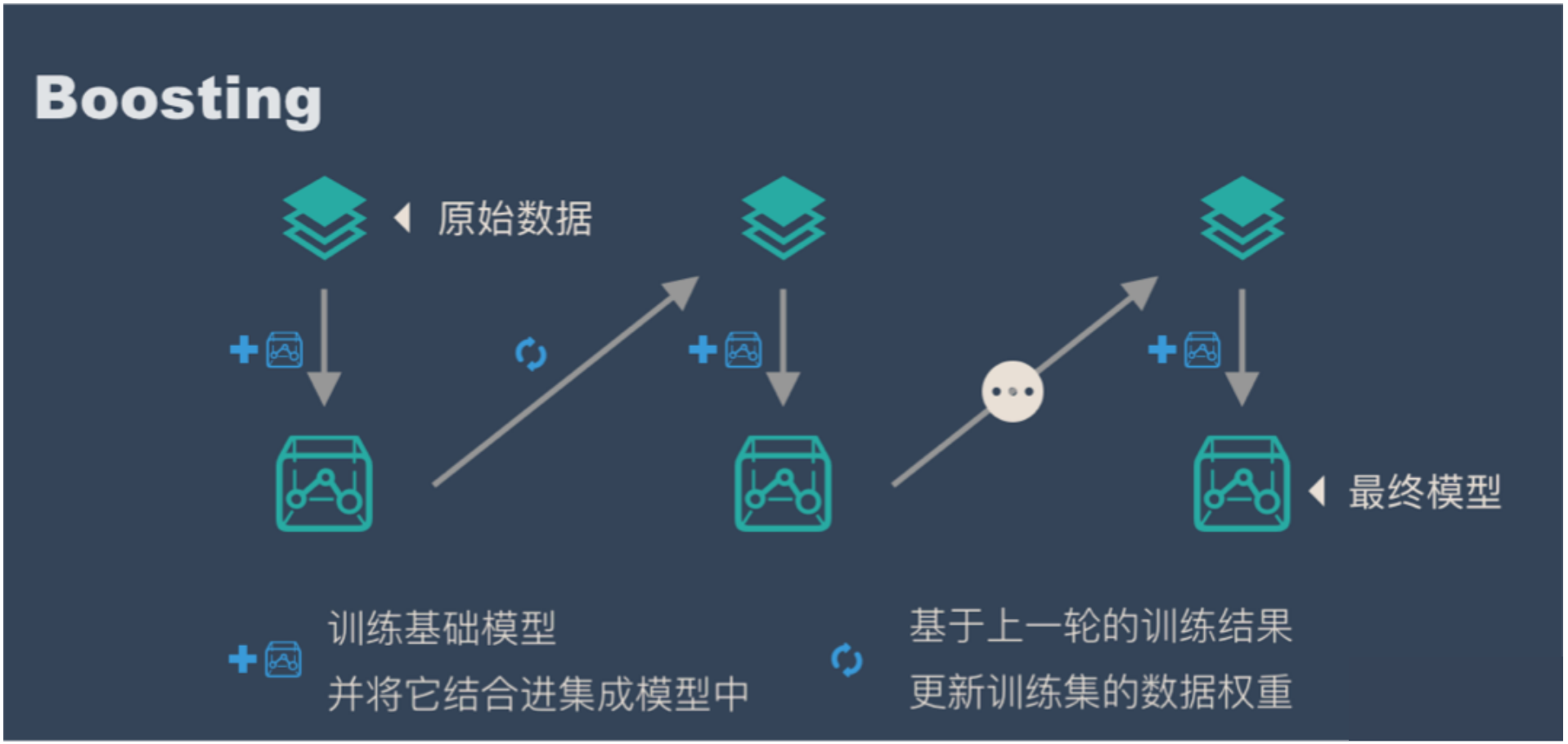
Bagging

Bagging

- 从原始样本集中抽取训练集。每轮从原始样本集中使用Bootstraping的方法抽取n个训练样本（在训练集中，有些样本可能被多次抽取到，而有些样本可能一次都没有被抽中）。共进行k轮抽取，得到k个训练集。（k个训练集之间是相互独立的）
- 每次使用一个训练集得到一个模型，k个训练集共得到k个模型。（注：这里并没有具体的分类算法或回归方法，我们可以根据具体问题采用不同的分类或回归方法，如决策树、感知器等）
- 对分类问题：将上步得到的k个模型采用投票的方式得到分类结果；对回归问题，计算上述模型的均值作为最后的结果。（所有模型的重要性相同）

Boosting

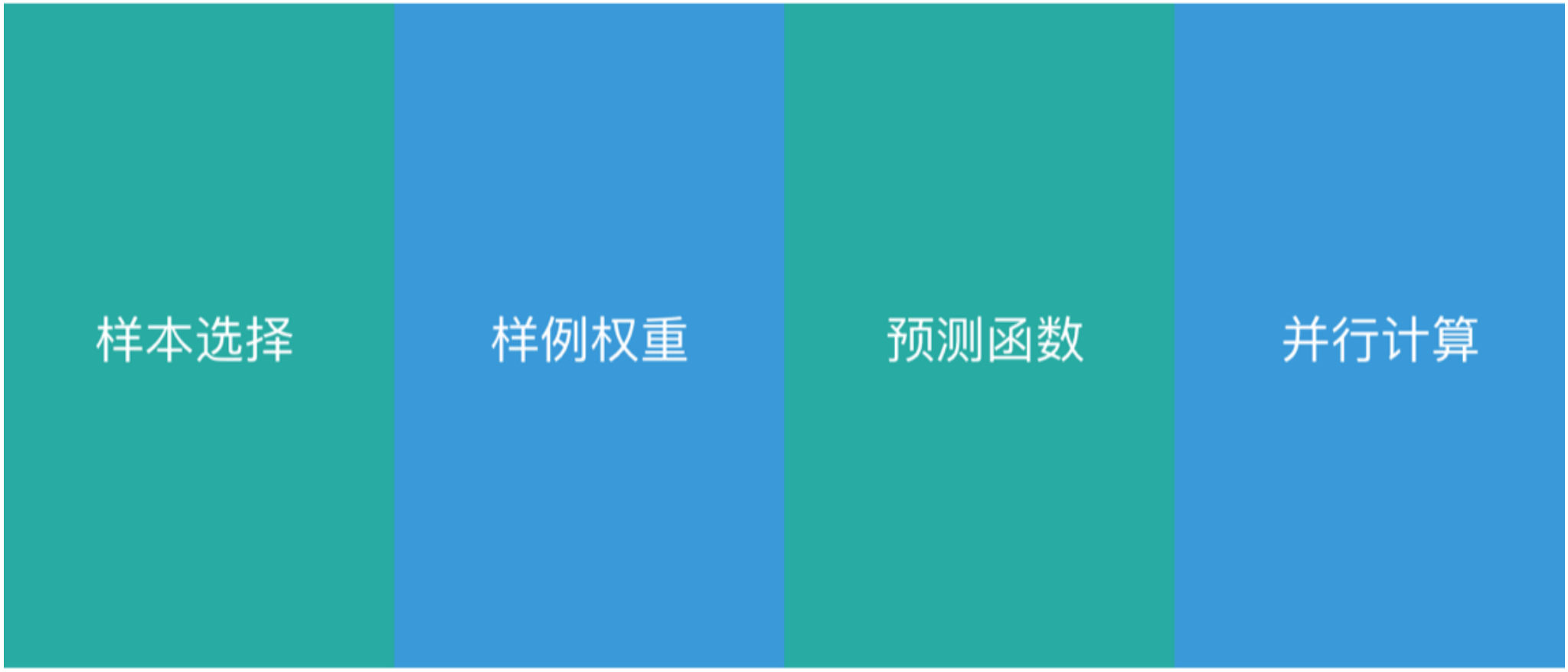
Boosting


Boosting
原始数据
最终模型
训练基础模型
并将它结合进集成模型中
基于上一轮的训练结果
更新训练集的数据权重

Boosting

- 通过加法模型将基础模型进行线性的组合。
- 每一轮训练都提升那些错误率小的基础模型权重，同时减小错误率高的模型权重。
- 在每一轮改变训练数据的权值或概率分布，通过提高那些在前一轮被弱分类器分错样例的权值，减小前一轮分对样例的权值，来使得分类器对误分的数据有较好的效果。

Bagging 和 Boosting 的4 点差别



Bagging 和 Boosting 的4 点差别

- **样本选择上：**
- Bagging：训练集是在原始集中有放回选取的，从原始集中选出的各轮训练集之间是独立的。
- Boosting：每一轮的训练集不变，只是训练集中每个样例在分类器中的权重发生变化。而权值是根据上一轮的分类结果进行调整。

Bagging 和 Boosting 的4 点差别

- **样例权重：**
- Bagging：使用均匀取样，每个样例的权重相等
- Boosting：根据错误率不断调整样例的权值，错误率越大则权重越大。

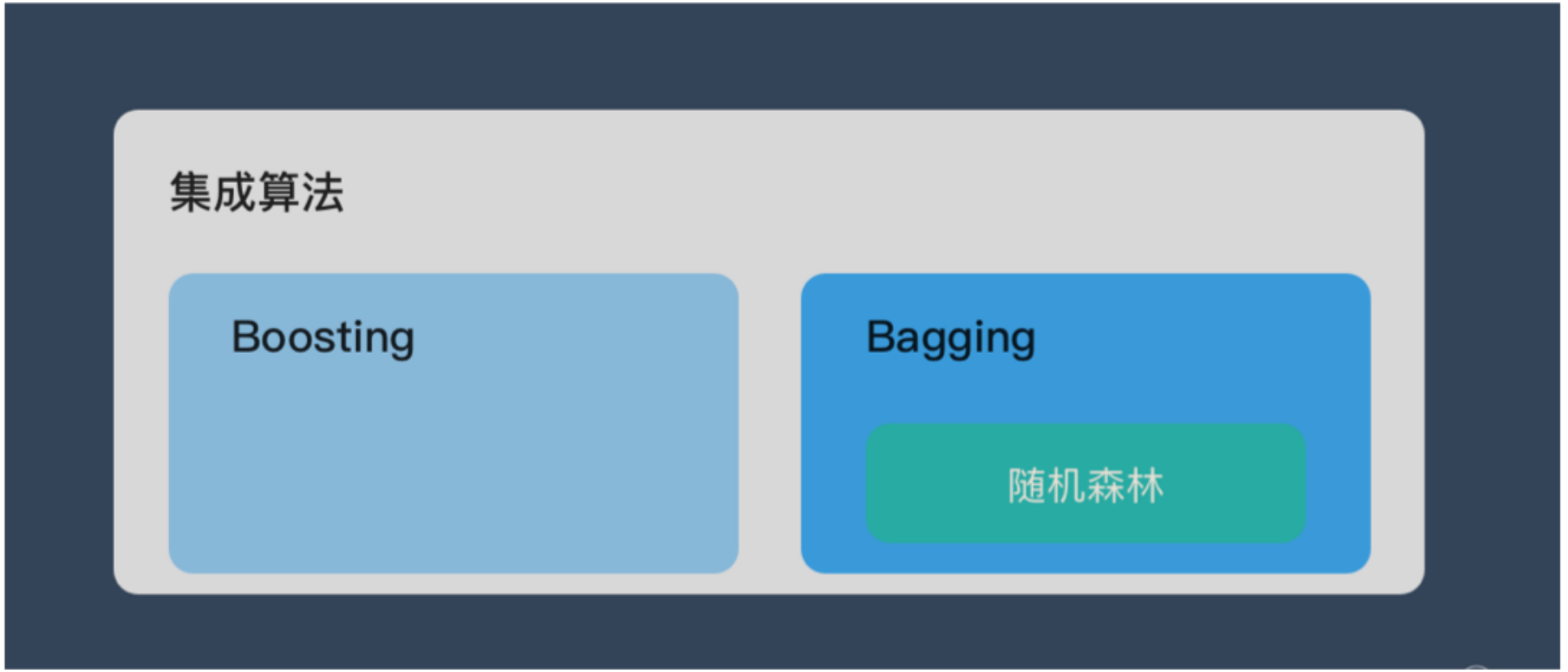
Bagging 和 Boosting 的4 点差别

- **预测函数：**
- Bagging：所有预测函数的权重相等。
- Boosting：每个弱分类器都有相应的权重，对于分类误差小的分类器会有更大的权重。

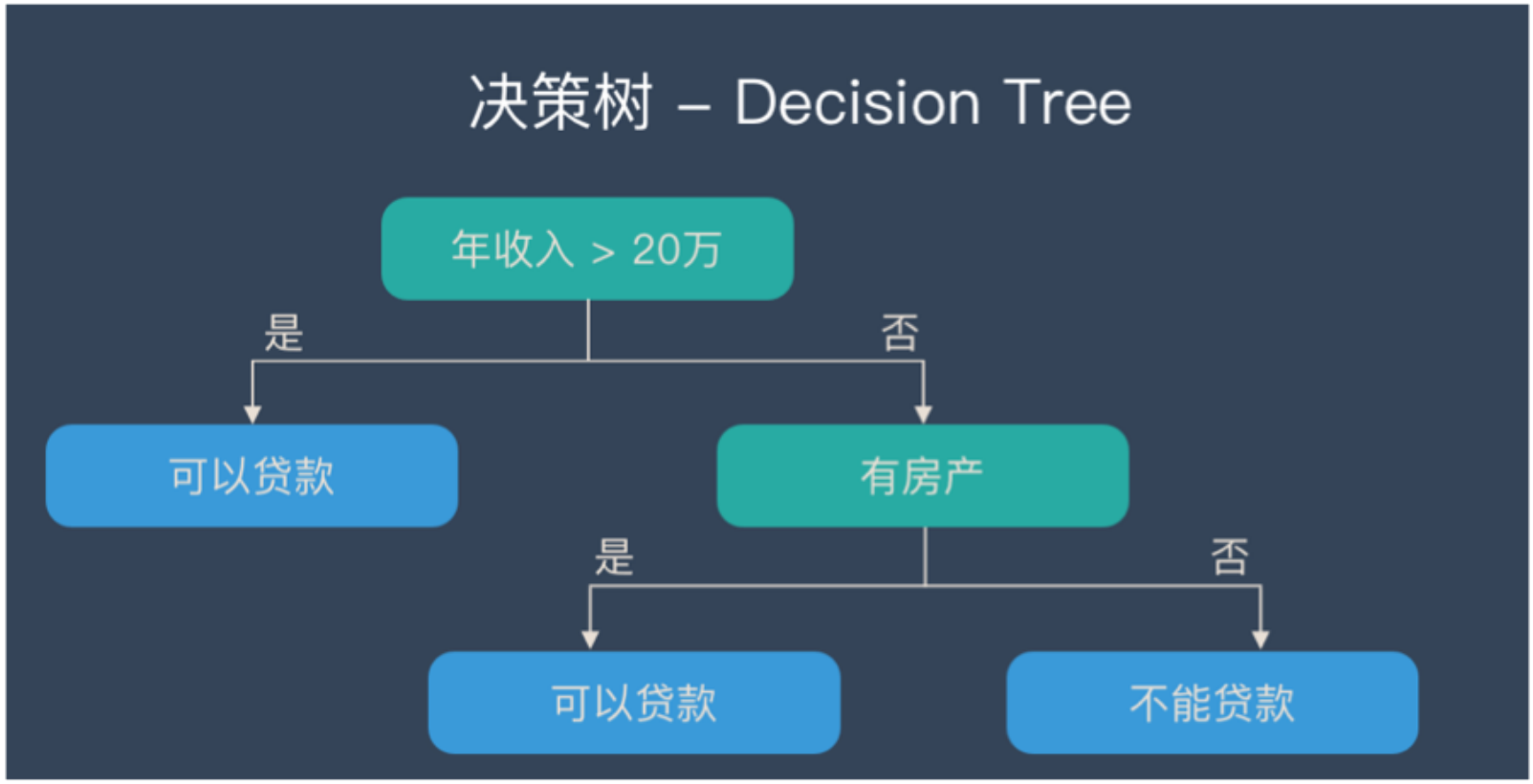
Bagging 和 Boosting 的4 点差别

- **并行计算：**
- Bagging：各个预测函数可以并行生成
- Boosting：各个预测函数只能顺序生成，因为后一个模型参数需要前一轮模型的结果。

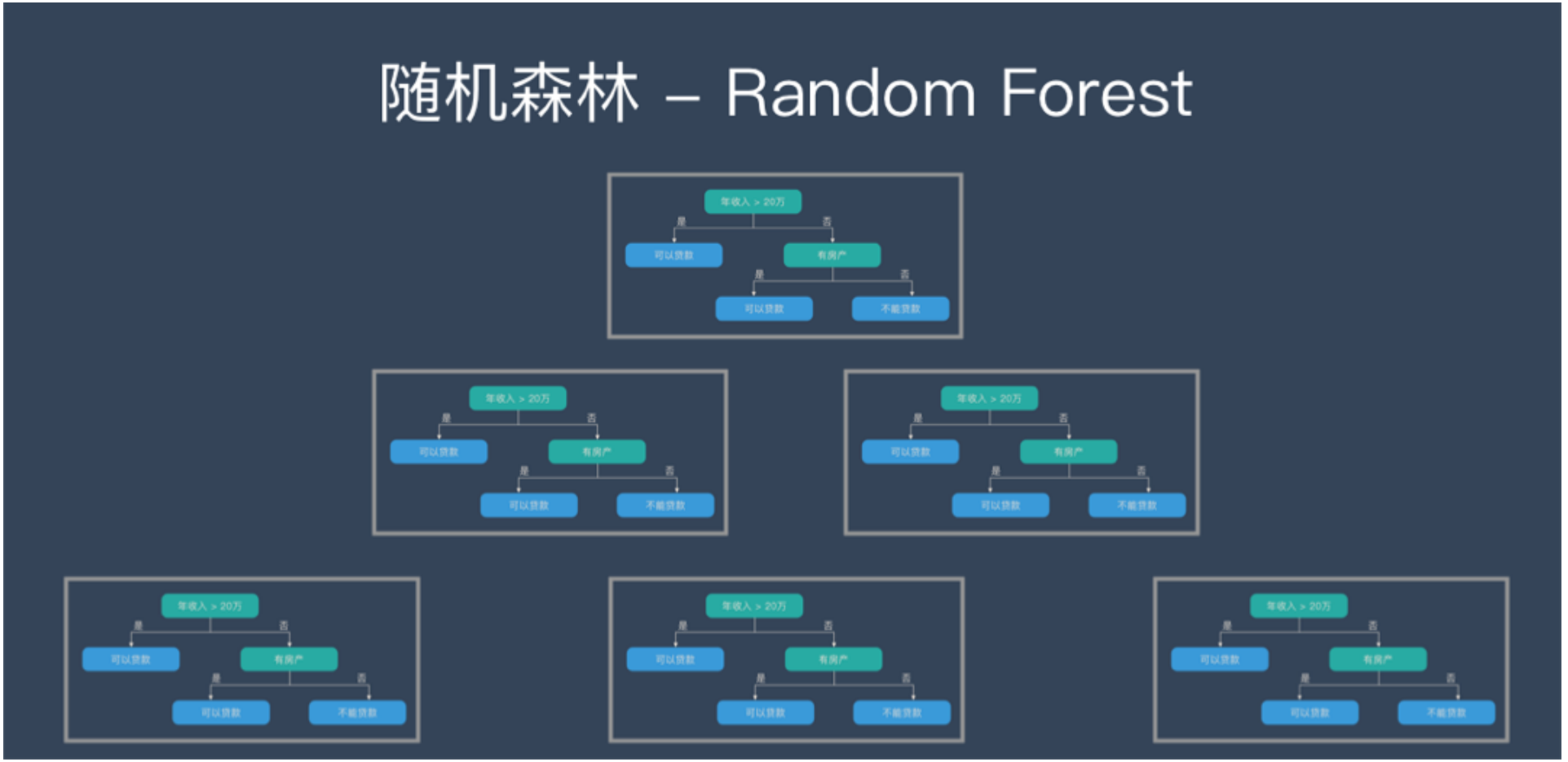
什么是随机森林？



决策树回顾



随机森林 – Random Forest



构造随机森林的 4 个步骤

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。

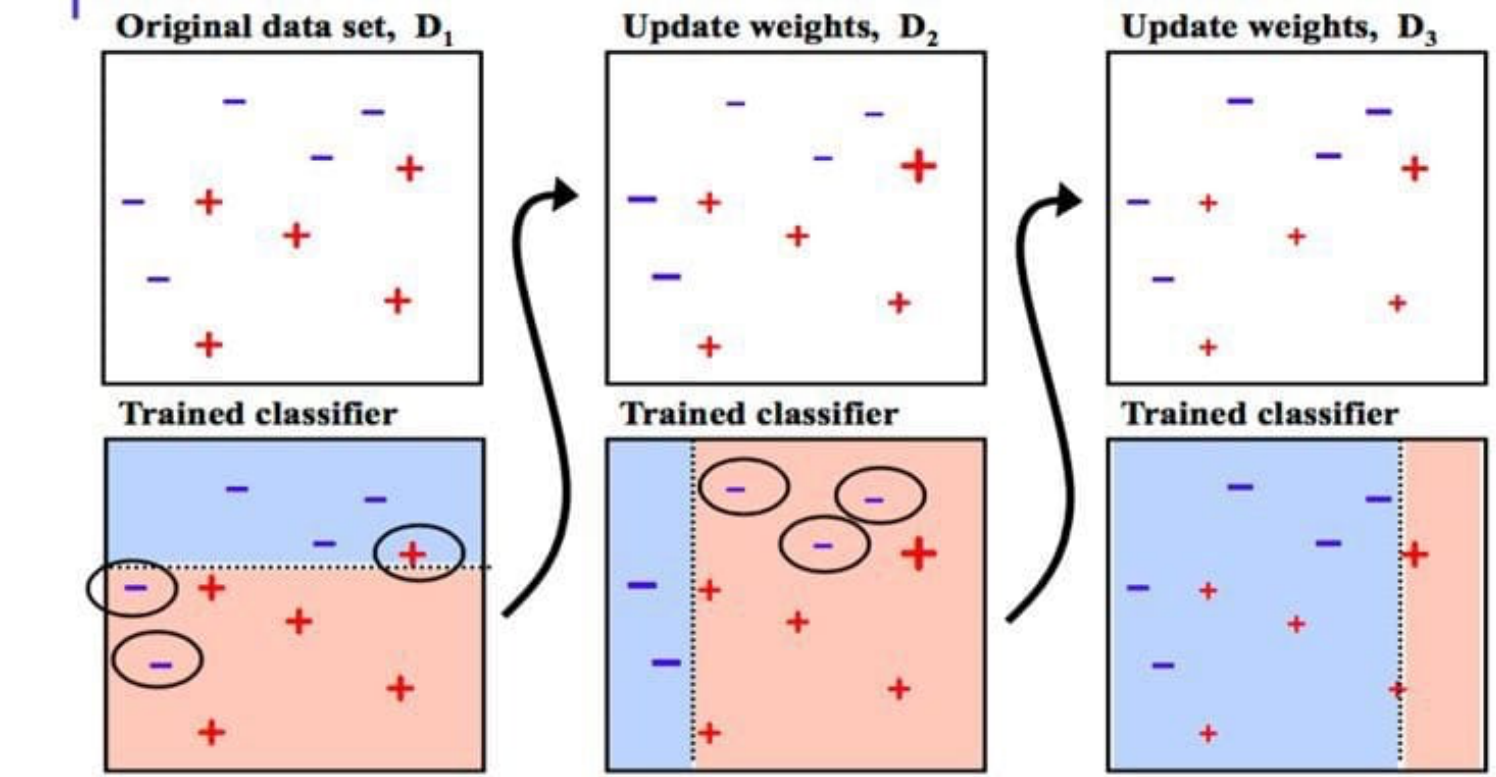
随机森林的缺点

- 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

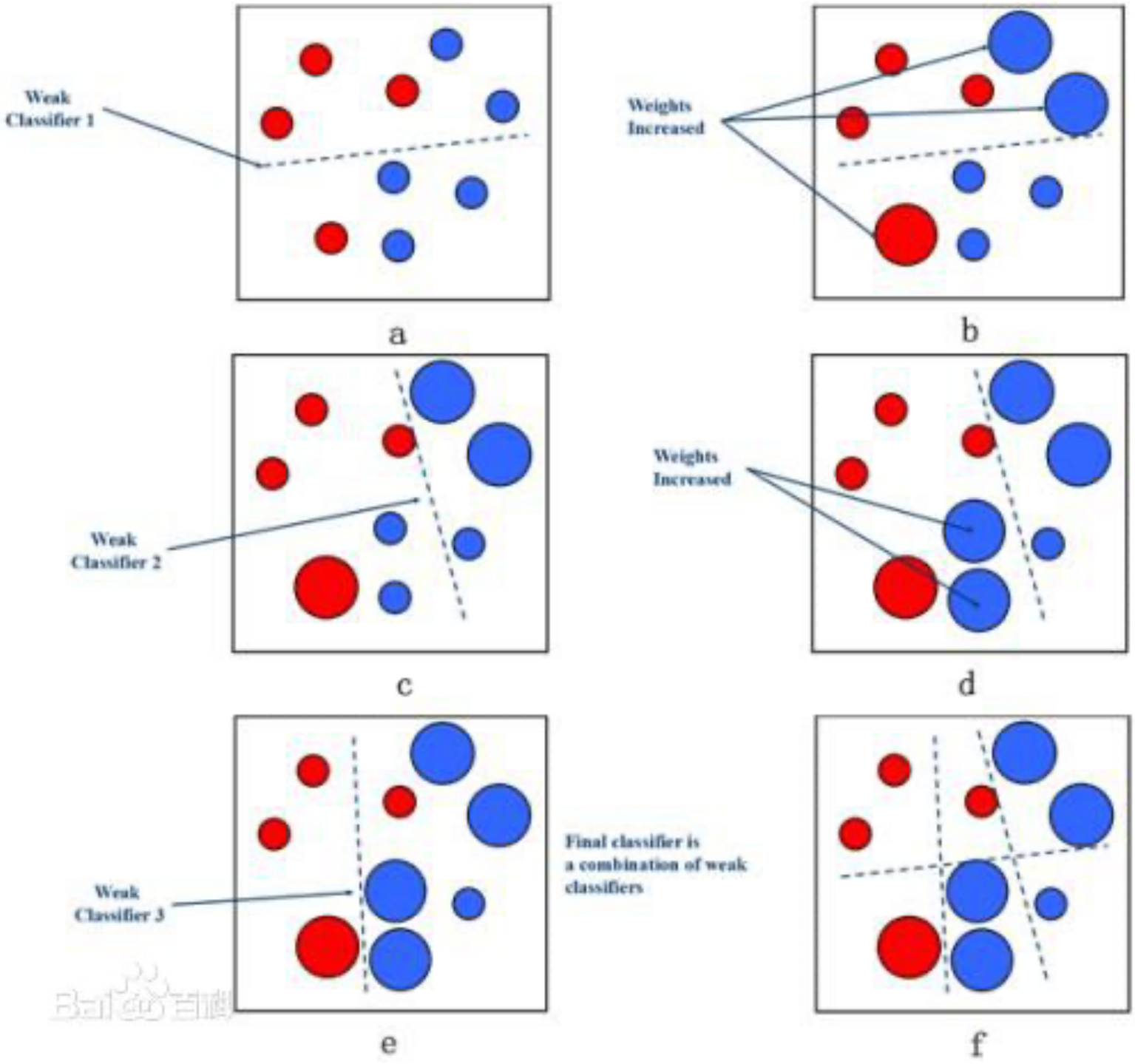
什么是AdaBoost？

- Boosting是一种集合技术，试图从许多弱分类器中创建一个强分类器。这是通过从训练数据构建模型，然后创建第二个模型来尝试从第一个模型中纠正错误来完成的。添加模型直到完美预测训练集或添加最大数量的模型。
- AdaBoost是第一个为二进制分类开发的真正成功的增强算法。这是理解助力的最佳起点。现代助推方法建立在AdaBoost上，最着名的是随机梯度增强机。

Algorithm Adaboost - Example



courtesy to Alexander Ihler http://sli.ics.uci.edu/Classes/2012F-273a?action=download&upname=10-ensembles.pdf


Weak
Classifier 1
a
Weights
Increased
b
Weak
Classifier 2
c
Weights
Increased
d
Weak
Classifier 3
e
Final classifier is
a combination of weak
classifiers
f

Adaboost的优缺点

- **AdaBoost算法优点：**
- 很好的利用了弱分类器进行级联；
- 可以将不同的分类算法作为弱分类器；
- AdaBoost具有很高的精度；
- 相对于bagging算法和Random Forest算法，AdaBoost充分考虑的每个分类器的权重；

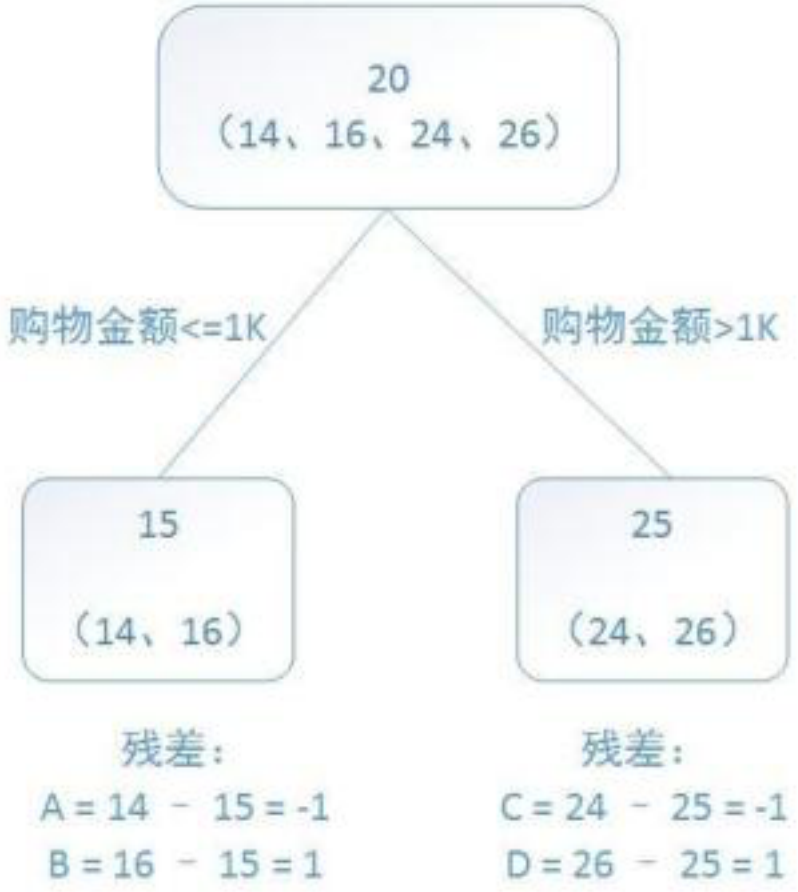
Adaboost的优缺点

- **Adaboost算法缺点：**
- AdaBoost迭代次数也就是弱分类器数目不太好设定，可以使用交叉验证来进行确定；
- 数据不平衡导致分类精度下降；
- 训练比较耗时，每次重新选择当前分类器最好切分点；

梯度提升决策树(GBDT)

- 4个人，A、B、C、D年龄分别是14、16、24、26。样本中有购物金额、上网时长、经常到百度知道提问等特征。预测4个人的年龄

梯度提升决策树(GBDT)



梯度提升决策树(GBDT)