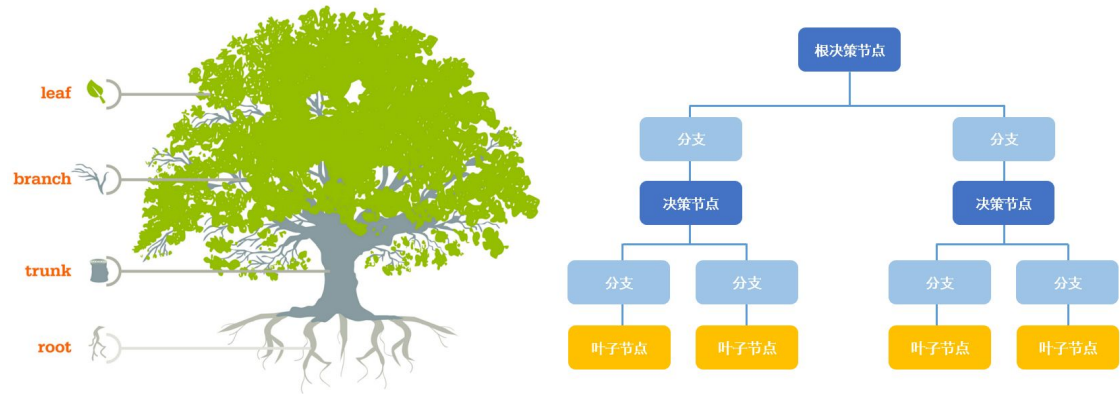



# 决策树



- 决策树(Decision Tree ) 是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系
- 作者：andyham链接：<https://www.jianshu.com/p/9688b2741d43>来源：简书著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。

- 构建决策树的算法主要有以下三种，且根据决策树的输出结果，决策树可以分为**分类树**和**回归树**，分类树输出的结果为具体的类别，而回归树输出的结果为一个确定的数值。其中 **ID3** 和 **C4.5** 是分类树，**CART** 是分类回归树，且在 **ID3** 和 **C4.5** 中，特征（属性）只能选一次，而 **CART** 没有这样的要求：
  - a. ID3 在决策树生成过程中，以**信息增益**为特征选择的准则。
  - b. C4.5 在决策树生成过程中，以**信息增益比**为特征选择的准则。
  - c. CART 对回归树用**平方误差最小化准则**，对分类树用**基尼指数**（Gini index）**最小化准则**，进行特征选择，生成二叉树。



## 随机森林

- 如果大家理解了决策树的话，那么会很容易理解什么是随机森林。随机森林就是通过集成学习的思想将多棵树集成的一种算法，它的基本单元是决策树，而它的本质属于机器学习的一大分支——集成学习（Ensemble Learning）方法。随机森林的名称中有两个关键词，一个是“随机”，一个就是“森林”。“森林”我们很好理解，一棵叫做树，那么成百上千棵就可以叫做森林了，这样的比喻还是很贴切的，其实这也是随机森林的主要思想--集成思想的体现。“随机”的含义我们会在下边部分讲到。
- 其实从直观角度来解释，每棵决策树都是一个分类器（假设现在针对的是分类问题），那么对于一个输入样本， $N$ 棵树会有 $N$ 个分类结果。而随机森林集成了所有的分类投票结果，将投票次数最多的类别指定为最终的输出，这就是一种最简单的 Bagging 思想。

## 随机森林



- 入分底的鼠，分，消。一  
 输入到题松的做抵出  
 个进物问是林树此做  
 一中动个还森的关系，  
 将树个这鼠是相会”  
 要棵某对老就不将音  
 们每论己是别不果噪  
 我到讨自底类.9%结“  
 。入，表到的99测芸  
 树输入发物多，预芸  
 类本会地动最的此于  
 分样并立该数立这脱  
 的入召独。票独，超  
 多输中要票得是况会  
 许将林都投获都情将  
 有要森树要，树的果  
 中需：棵都定棵有结  
 林们喻每树确每所测  
 森我比，棵来的盖预  
 机的鼠每况中涵的  
 随类象松是情林果树的  
 ，分形是就票森结的测  
 到行个还也投。测秀预  
 提进打鼠，据果预优的  
 面本。老法依结的数好  
 前样类是看要类出少个

- 有了树我们就可以分类了，但是森林中的每棵树是怎么生成的呢？

每棵树的按照如下规则生成：

1) 如果训练集大小为 $N$ ，对于每棵树而言，随机且有放回地从训练集中的抽取 $N$ 个训练样本作为该树的训练集；

从这里我们可以知道：每棵树的训练集都是不同的，而且里面包含重复的训练样本。

为什么要随机抽样训练集？

如果不进行随机抽样，每棵树的训练集都一样，那么最终训练出的树分类结果也是完全一样的，这样的话完全没有bagging的必要；

为什么要有放回地抽样？

我理解的是这样的：如果不是有放回的抽样，那么每棵树的训练样本都是不同的，都是没有交集的，这样每棵树都是“有偏的”，都是绝对“片面的”，也就是说每棵树的训练出来都是有很大的差异的；而随机森林最后分类取决于多棵树的投票表决，这种表决应该是“求同”，因此使用完全不同的训练集来训练每棵树这样对最终分类结果是没有帮助的，这样无异于是“盲人摸象”。

2) 如果每个样本的特征维度为 $M$ ，指定一个常数 $m \ll M$ ，随机地从 $M$ 个特征中选取 $m$ 个特征子集，每次树进行分裂时，从这 $m$ 个特征中选择最优的；

3) 每棵树都尽最大程度的生长，并且没有剪枝过程。

一开始我们提到的随机森林中的“随机”就是指的这里的两个随机性。两个随机性的引入对随机森林的分类性能至关重要。由于它们的引入，使得随机森林不容易陷入过拟合，并且具有很好得抗噪能力

- 随机森林相对于决策树的优点主要是：1) 降低异常值所带来的影响：因为随机森林选取了部分数据建立了多个决策树，即使有个别决策树会因为异常值的影响导致预测不准确，但预测结果是参考多个决策树得到的结果，降低了异常值带来的影响。2) 降低了过拟合的可能性，因为决策树是采用了所有的特征及样本，容易出现过拟合（即对训练样本有很好的效果，对测试集的效果很差），随机森林是采用了部分样本的部分特征而构造的很多个决策树（采取的有放回抽样），特征和数据在单个决策树上变少了，降低了过拟合的可能性。随机森林相对于决策树的缺点主要是：1) 计算量相对于决策树很大，性能开销很大。2) 可能会导致有些数据集没有训练到，但这种几率很小。
- 版权声明：本文为CSDN博主「qq\_24497419」的原创文章，遵循CC 4.0 BY-SA 版权协议，转载请附上原文出处链接及本声明。
- 原文链接：  
[https://blog.csdn.net/qq\\_24497419/article/details/89092682](https://blog.csdn.net/qq_24497419/article/details/89092682)