机器学习-尚学堂

1.决策树的基本概念和方法

2.从决策树到随机森林

3.决策树在数值回归中的应用

机器学习-尚学堂

决策树的基本概念和方法

机器学习-尚学堂

决策树是一种非线性有监督分类模型

随机森林是一种非线性有监督分类模型

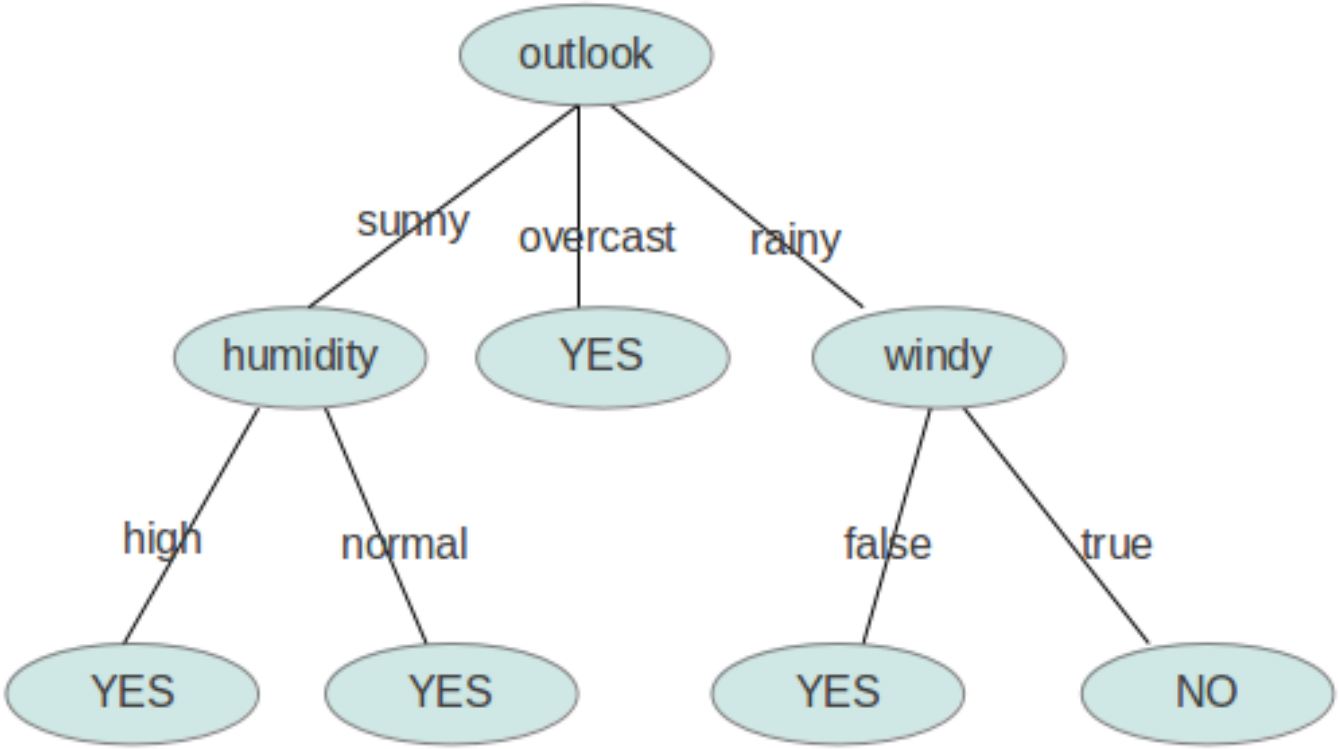
机器学习-尚学堂

案例分析 离散化！！！

天气	温度	湿度	风	车祸
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes

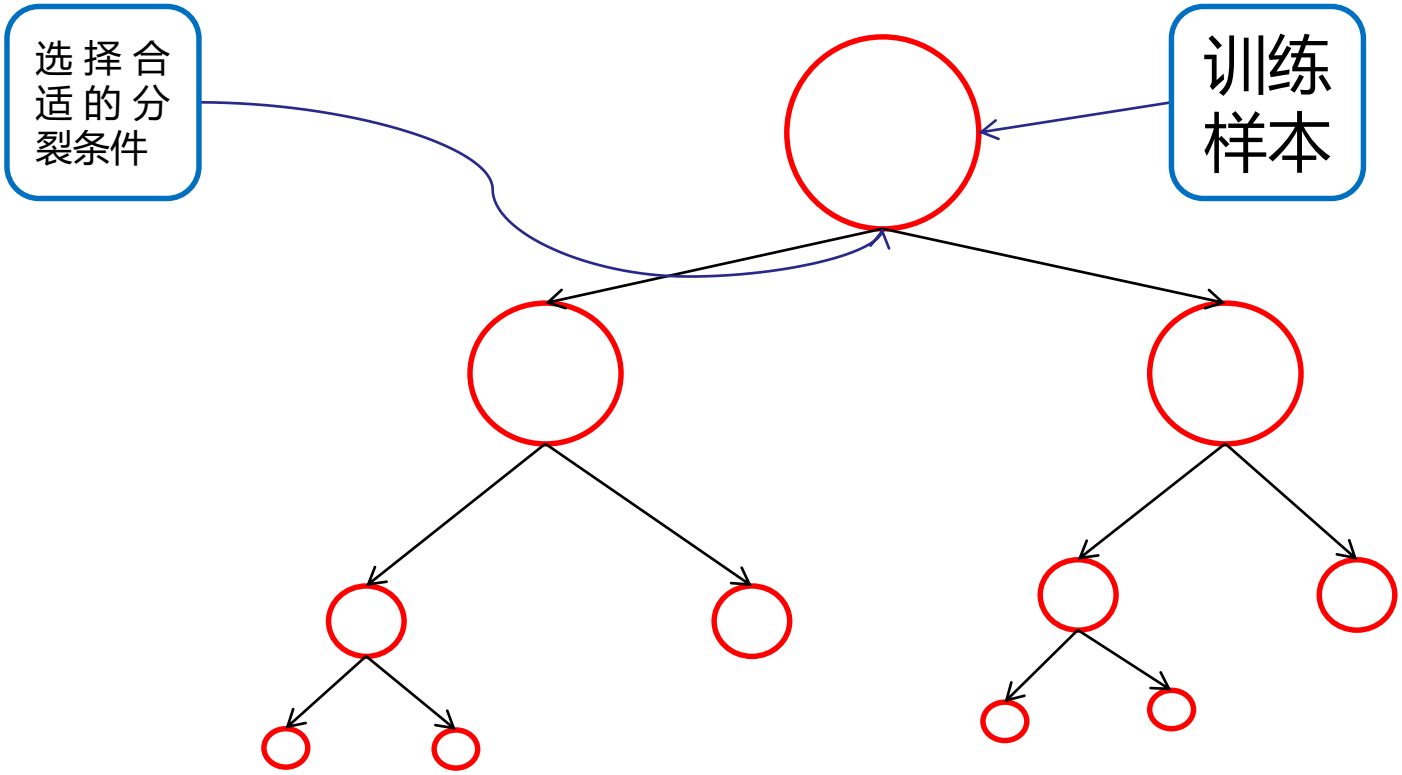
机器学习-尚学堂

决策树是通过固定的条件来对类别进行判断：



机器学习-尚学堂

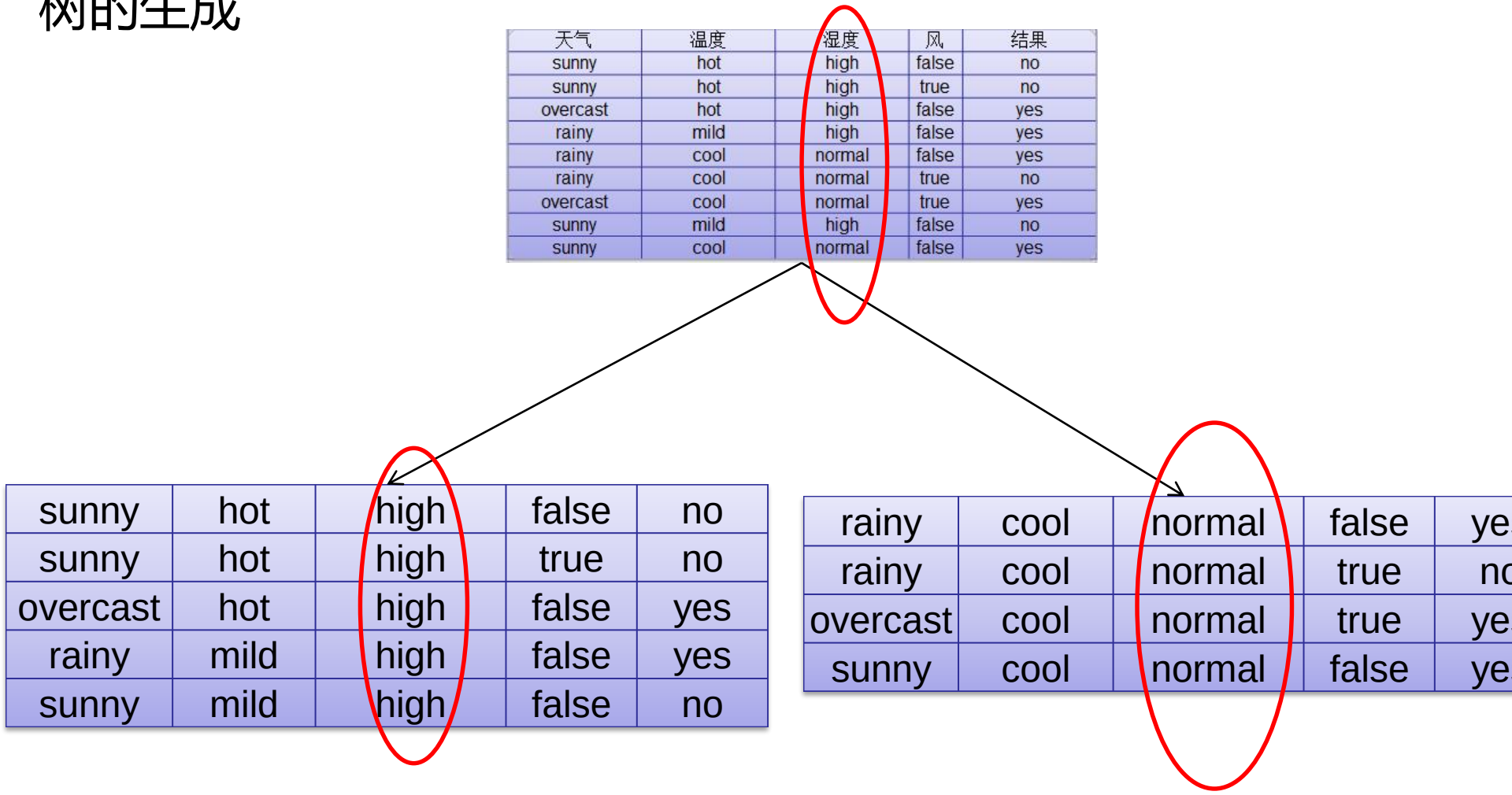
决策树的生成：数据不断分裂的递归过程，每一次分裂，尽可能让类别一样的数据在树的一边，当树的叶子节点的数据都是一类的时候，则停止分类。(if else 语句)



机器学习-尚学堂

树的生成

天气	温度	湿度	风	结果
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes



sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
sunny	mild	high	false	no

rainy	cool	normal	false	ye
rainy	cool	normal	true	no
overcast	cool	normal	true	ye
sunny	cool	normal	false	ye

机器学习-尚学堂

树的生成

天气	温度	湿度	风	结果
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes

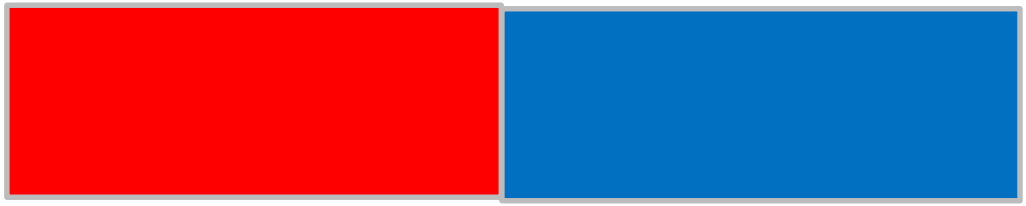
sunny	hot	high	false	no
sunny	hot	high	true	no
sunny	mild	high	false	no
sunny	cool	normal	false	yes

rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no

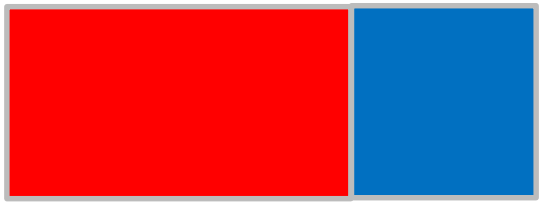
overcast	cool	normal	true	yes
overcast	hot	high	false	yes

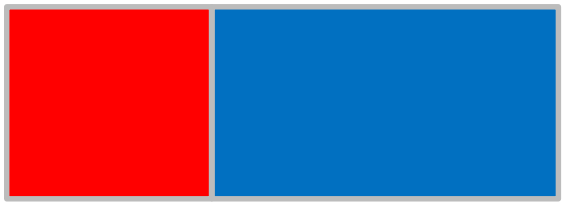
机器学习-尚学堂

数据集



分割方式1





分割方式2

机器学习-尚学堂

树的每一次分类，都有很多种选择标准，每种标准产生不同的分类结果，因此我们需要一个评判指标，看看哪种选择最合适

评判标准是每一个叶子里面的类别尽可能一致。

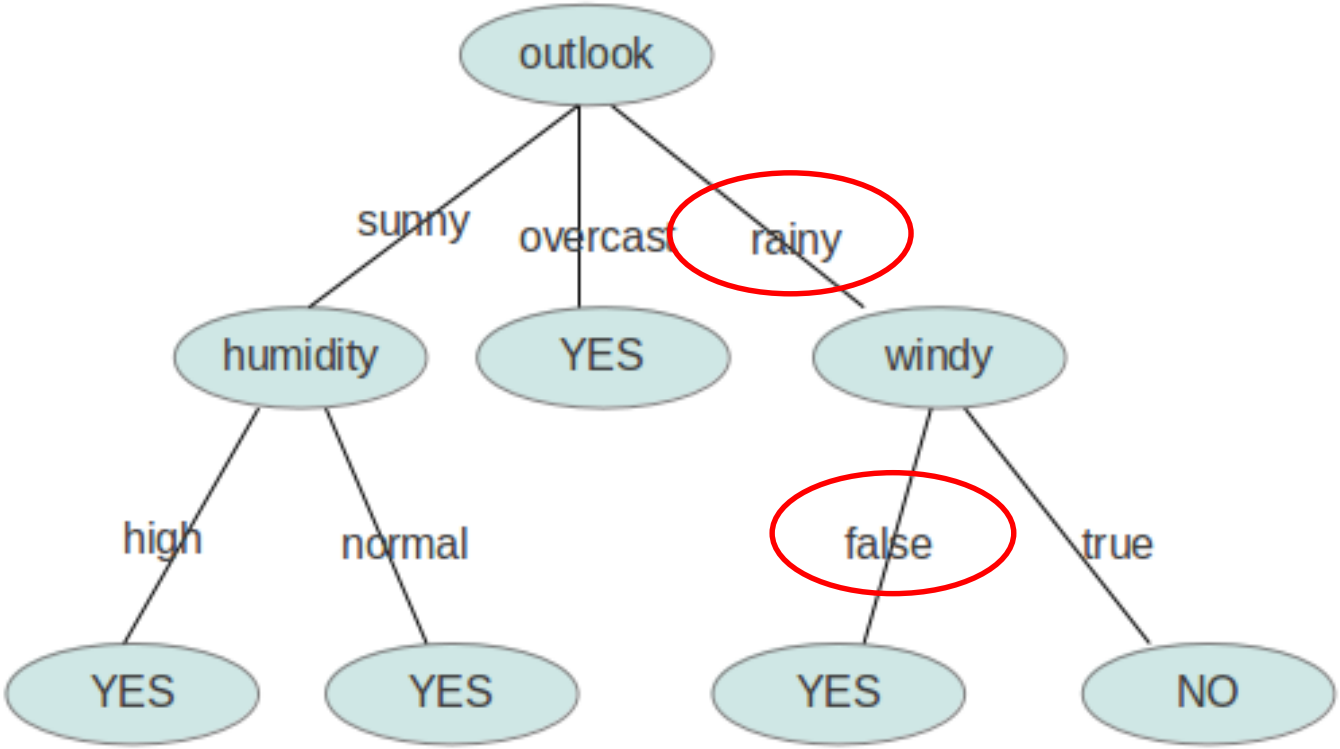
幸运的是，Spark已经将评价标准作了很好的封装，用户只需调用API即可

机器学习-尚学堂

Impurity	Task	Formula	Description
Gini impurity	Classification	$\sum_{i=1}^{C} f_i(1-f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels. 基尼系数：(0,1)，越小表示数据纯度高
Entropy	Classification	$\sum_{i=1}^{C} -f_i log(f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels. 熵：越大越表示不确定性越大
Variance	Regression	$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} x_i$. 方差：越小越纯

机器学习-尚学堂

得到一颗树之后，我们就可以用这棵树来做预测了



机器学习-尚学堂

数据类型（决策是内部有离散化过程）

1.离散数据 需指明取值数量 2^M种分割方式

天气：晴天 雨天 多云

学历：高中 本科 研究生

2.连续数据 需离散化，需指明离散化后的数量

车速:

低速（60 65）中速（70 75）高速（80 90）

M+1种分割方式

机器学习-尚学堂

- 逻辑回归做离散数据
- 性别：男 1　　女 -1
- 宠物：猫 狗 乌龟 1 0 -1

机器学习-尚学堂

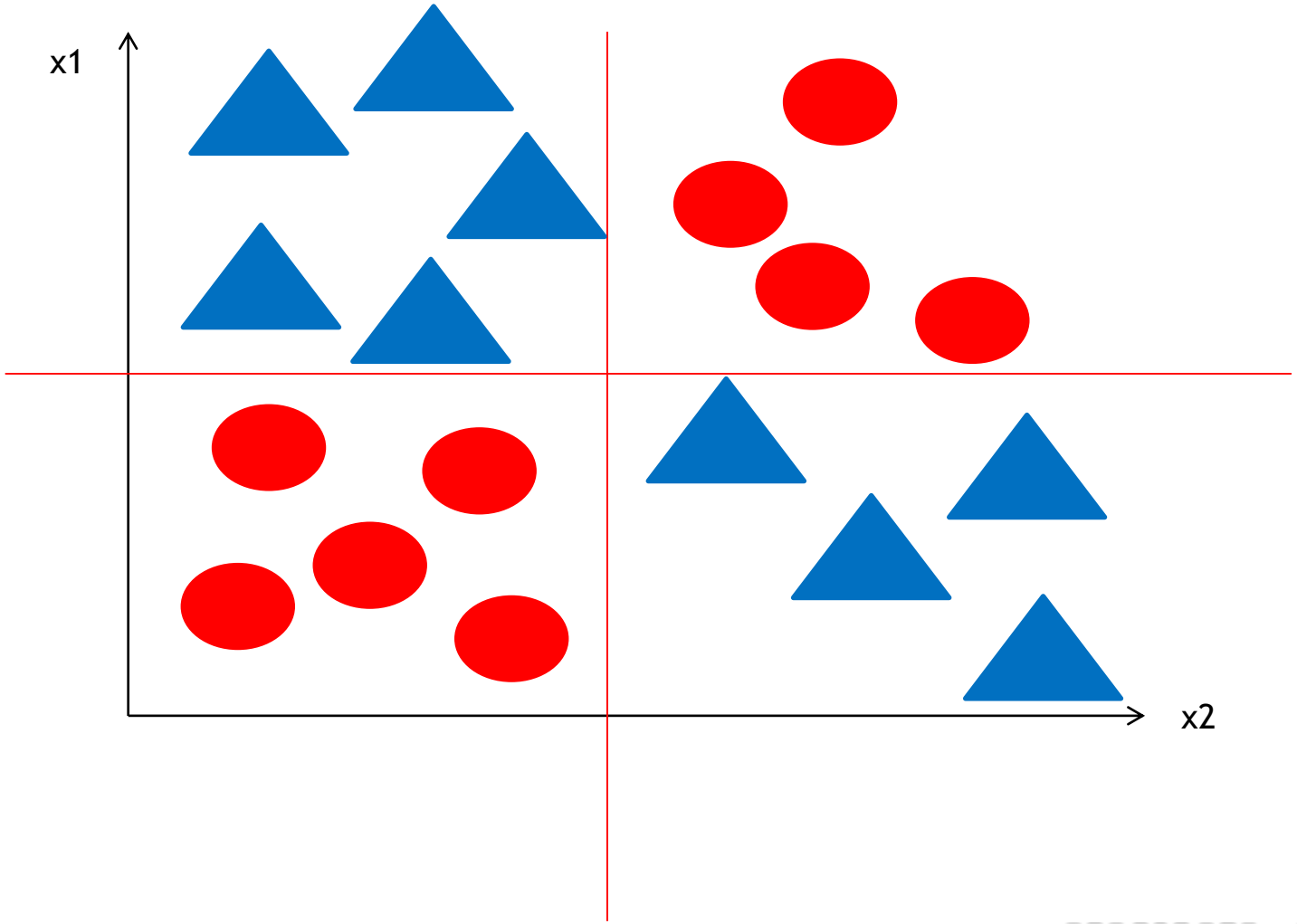
数据展示

代码演示

天气	温度	湿度	风	车速	车祸
sunny	hot	high	false	60	no
sunny	hot	high	true	70	no
overcast	hot	high	false	65	yes
rainy	mild	high	false	90	yes
rainy	cool	normal	false	82	yes
rainy	cool	normal	true	70	no
overcast	cool	normal	true	73	yes
sunny	mild	high	false	50	no
sunny	cool	normal	false	80	yes

机器学习-尚学堂

决策树的分割方式 非线性！



机器学习-尚学堂

单颗决策树的缺点：

1.运算量大，需要一次加载所有数据进内存。并且找寻分割条件是一个极耗资源的工作。

2.训练样本中出现异常数据时，将会对决策树产生很大影响。抗干扰能力差，逻辑回归怎么解决抗干扰能力的?

解决方法：

1.减少决策树所需训练样本

2.随机采样，降低异常数据的影响。

和逻辑回归比，逻辑回归可以告诉我们概率，而决策树只能0，1

机器学习-尚学堂

从决策树到随机森林

机器学习-尚学堂

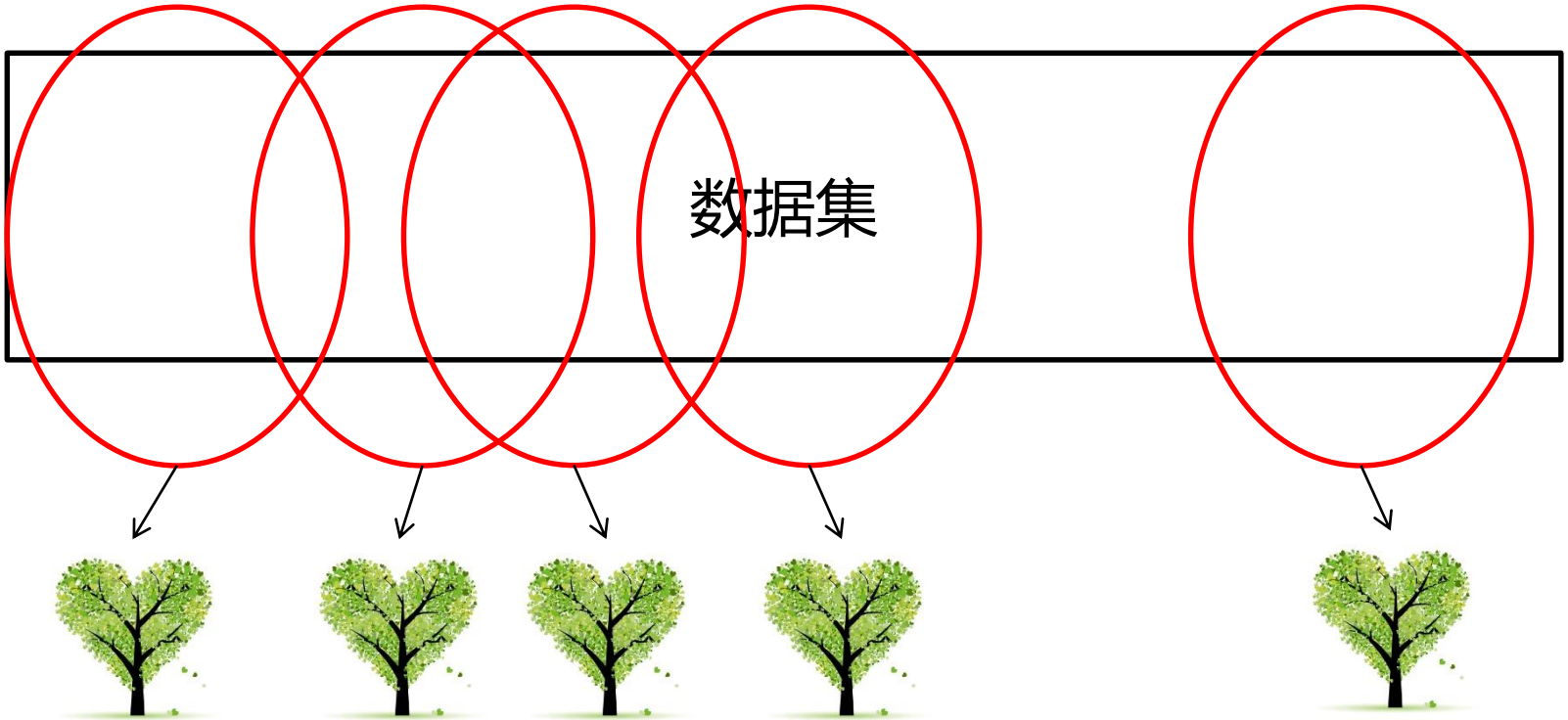
随机森林

森林：由树组成

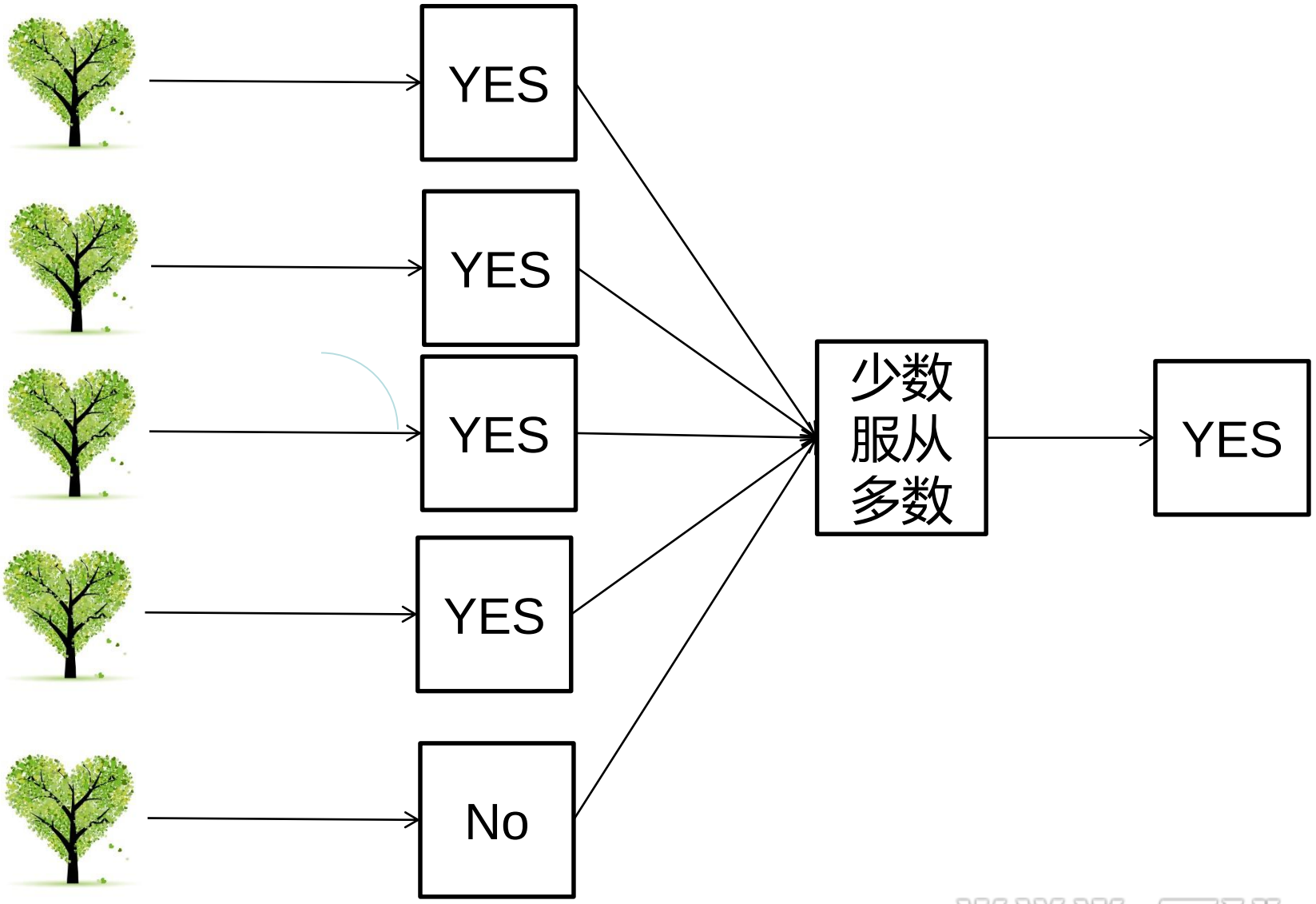
随机：生成树的数据都是从数据集中随机选取的

机器学习-随机森林

当数据集很大的时候，我们随机选取数据集的一部分，生成一棵树，重复上述过程，我们可以生成一堆形态各异的树，这些树放在一起就叫森林



机器学习-尚学堂

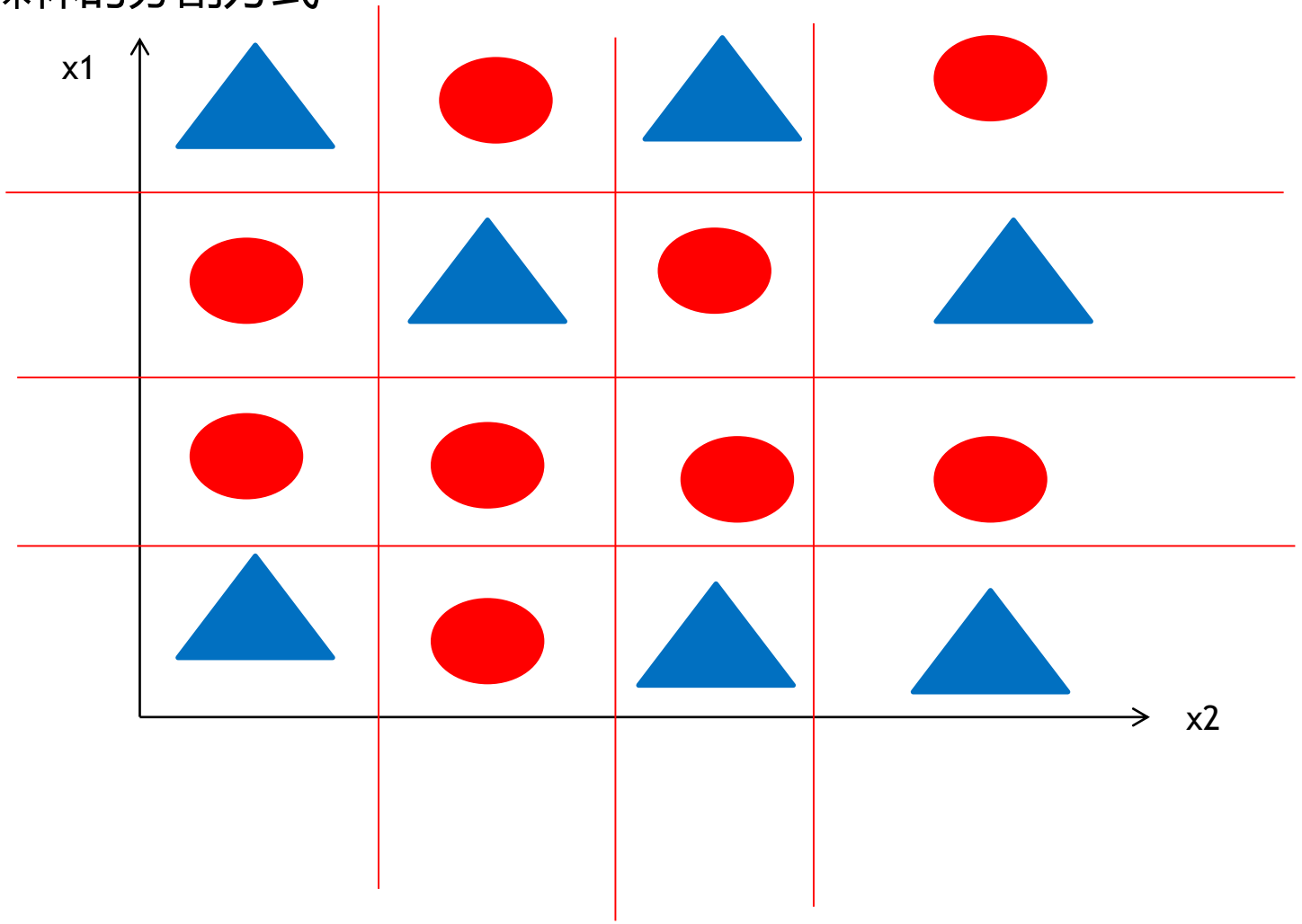


机器学习-尚学堂

随机森林的分类方式

机器学习-尚学堂

随机森林的分割方式



机器学习-尚学堂

随机森林VS逻辑回归

逻辑回归	随机森林
软分类	硬分类
线性模型	非线性模型
输出有概率意义	输出无概率意义
抗干扰能力强	抗干扰能力弱

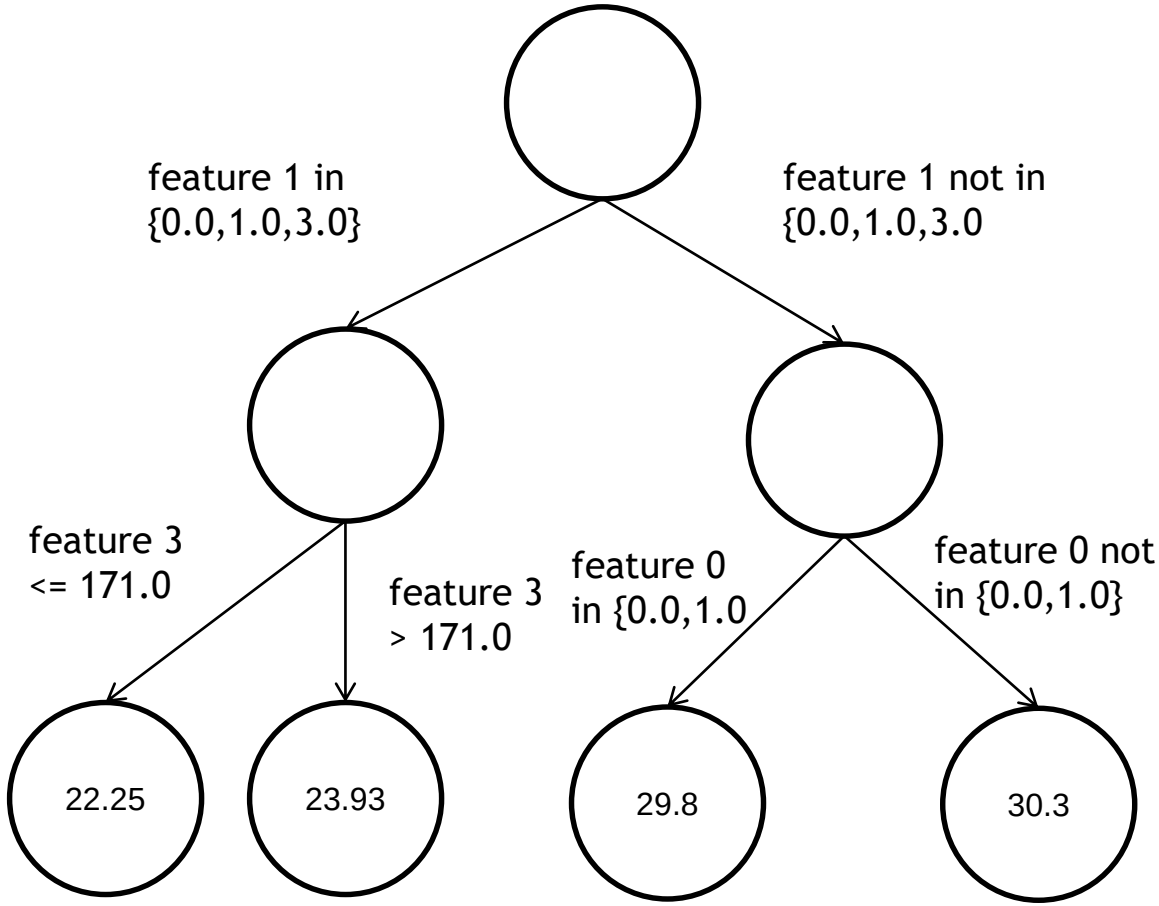
机器学习-尚学堂

决策树的生成

森林是由树组成的，这里的树是决策树，训练的过程就是利用数据集生成决策树的过程

机器学习-尚学堂

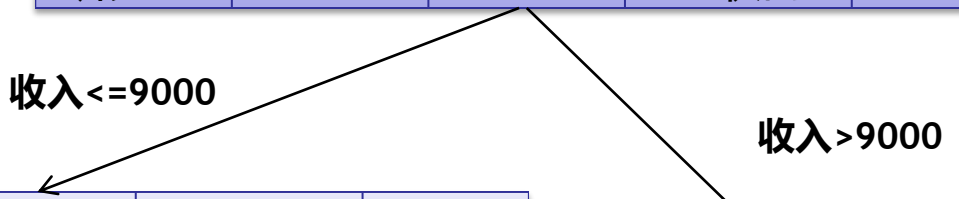
```
If (feature 1 in {0.0,1.0,3.0})
    If (feature 3 <= 171.0)
     Predict: 22.250847457627117
    Else (feature 3 > 171.0)
     Predict: 23.934905660377357
   Else (feature 1 not in {0.0,1.0,3.0})
    If (feature 0 in {0.0,1.0})
     Predict: 29.880434782608695
    Else (feature 0 not in {0.0,1.0})
     Predict: 30.369685767097966
```



机器学习-尚学堂

案例分析

学历	收入	身高	行业	邮件数
本科	16000	177	互联网	23
研究生	21000	182	金融	23
本科	9000	170	公务员	23
研究生	8000	175	传统企业	26
研究生	28000	169	互联网	30



本科	9000	170	公务员	23
研究生	8000	175	传统企业	26

本科	16000	177	互联网	23
研究生	21000	182	金融	23
研究生	28000	169	互联网	30

机器学习-尚学堂

案例分析

学历	收入	身高	行业	结果
本科	16000	177	互联网	23
研究生	21000	182	金融	23
本科	9000	170	公务员	23
研究生	8000	175	传统企业	26
研究生	28000	169	互联网	30

{互联网，金融}

{传统企业，公务员}

本科	16000	177	互联网	23
研究生	21000	182	金融	23
研究生	28000	169	互联网	30

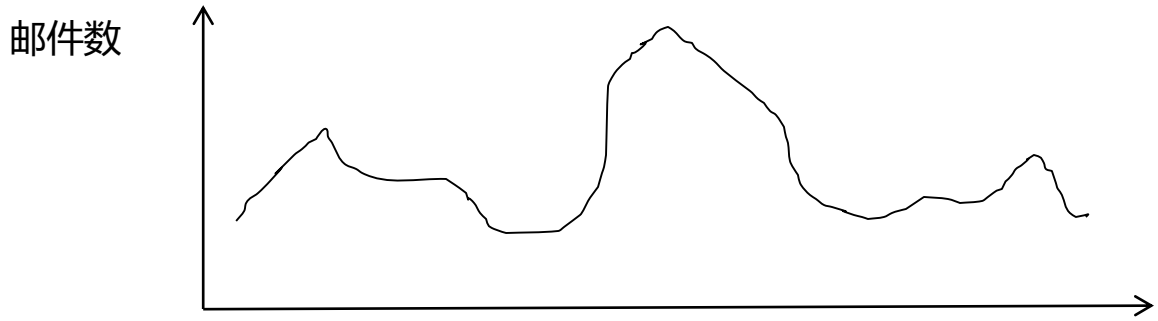
本科	9000	170	公务员	23
研究生	8000	175	传统企业	26

机器学习-尚学堂

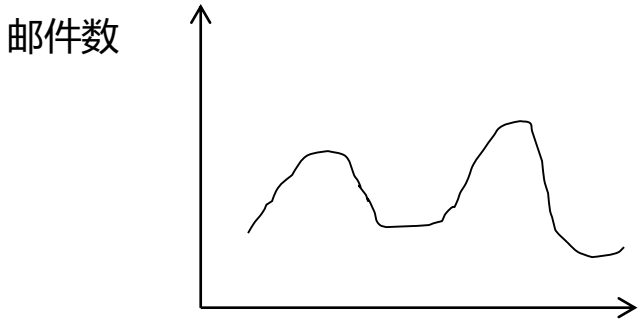
数据的三种类型

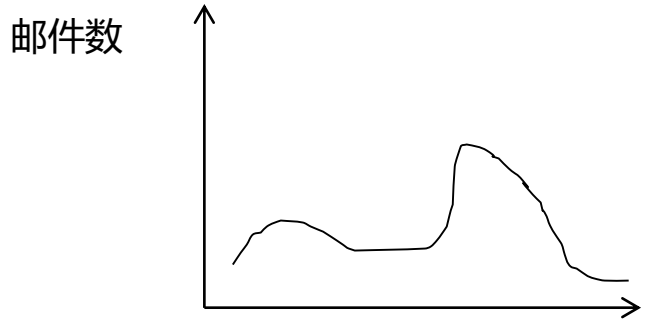
1.连续型可比较：收入，身高
2.离散型可半比较：学历
3.离散型无比较：行业

机器学习-尚学堂

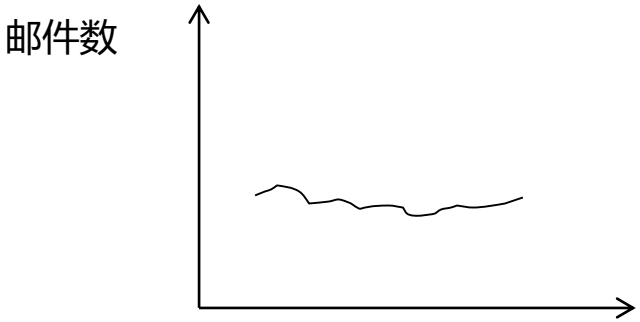


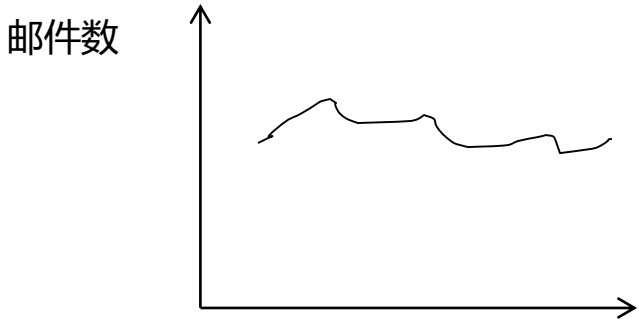
分割方案1





分割方案2





节点选择最优化的方案将自身的数据分成两个子部分，每个子部分按照同样的方式进行分割，直到满足停止条件

停止条件：

1.达到预定深度

2.节点内样本数量足够少

机器学习-尚学堂

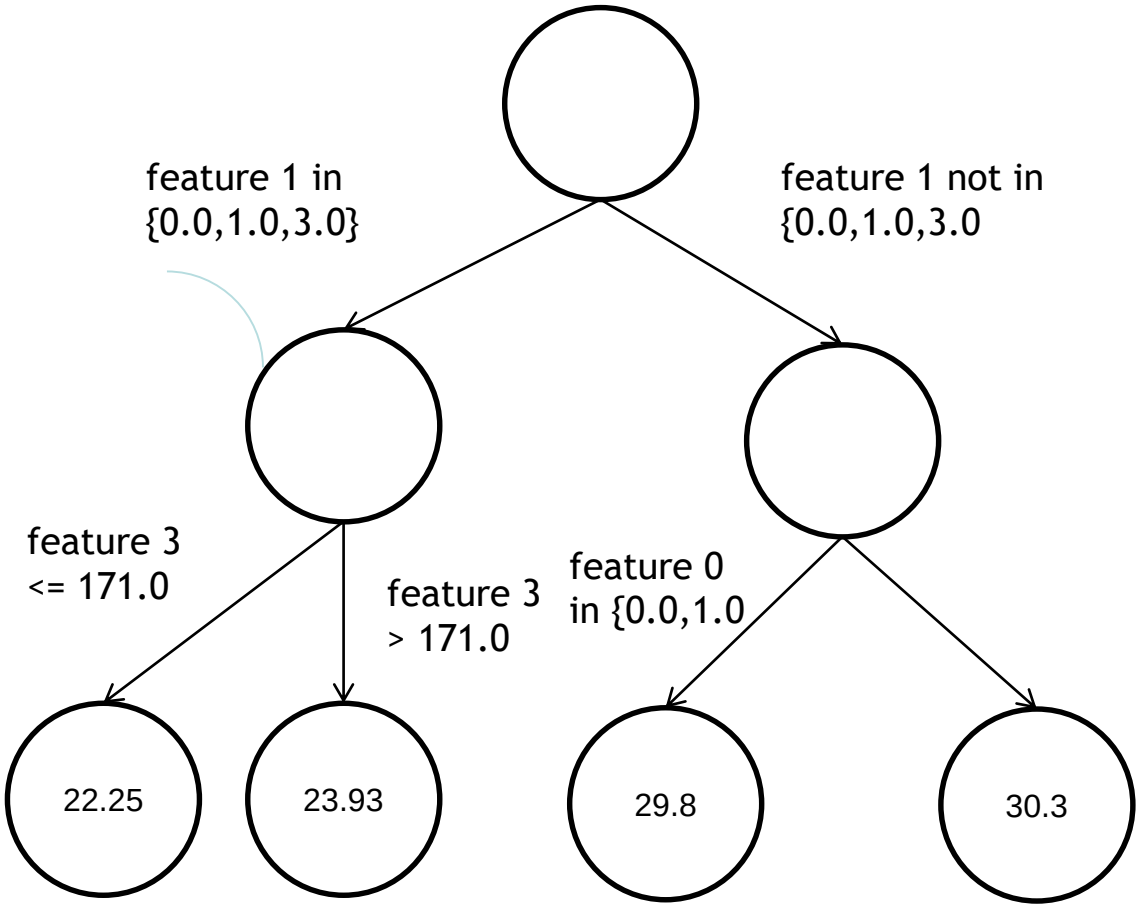
树的每一次分类，都有很多种选择标准，每种标准产生不同的分类结果，因此我们需要一个评判指标，看看哪种选择最合适

评判标准是每一个叶子里面的类别尽可能一致。

幸运的是，Spark已经将评价标准作了很好的封装，用户只需调用API即可

机器学习-尚学堂

得到一颗树之后，我们就可以用这棵树来做预测了



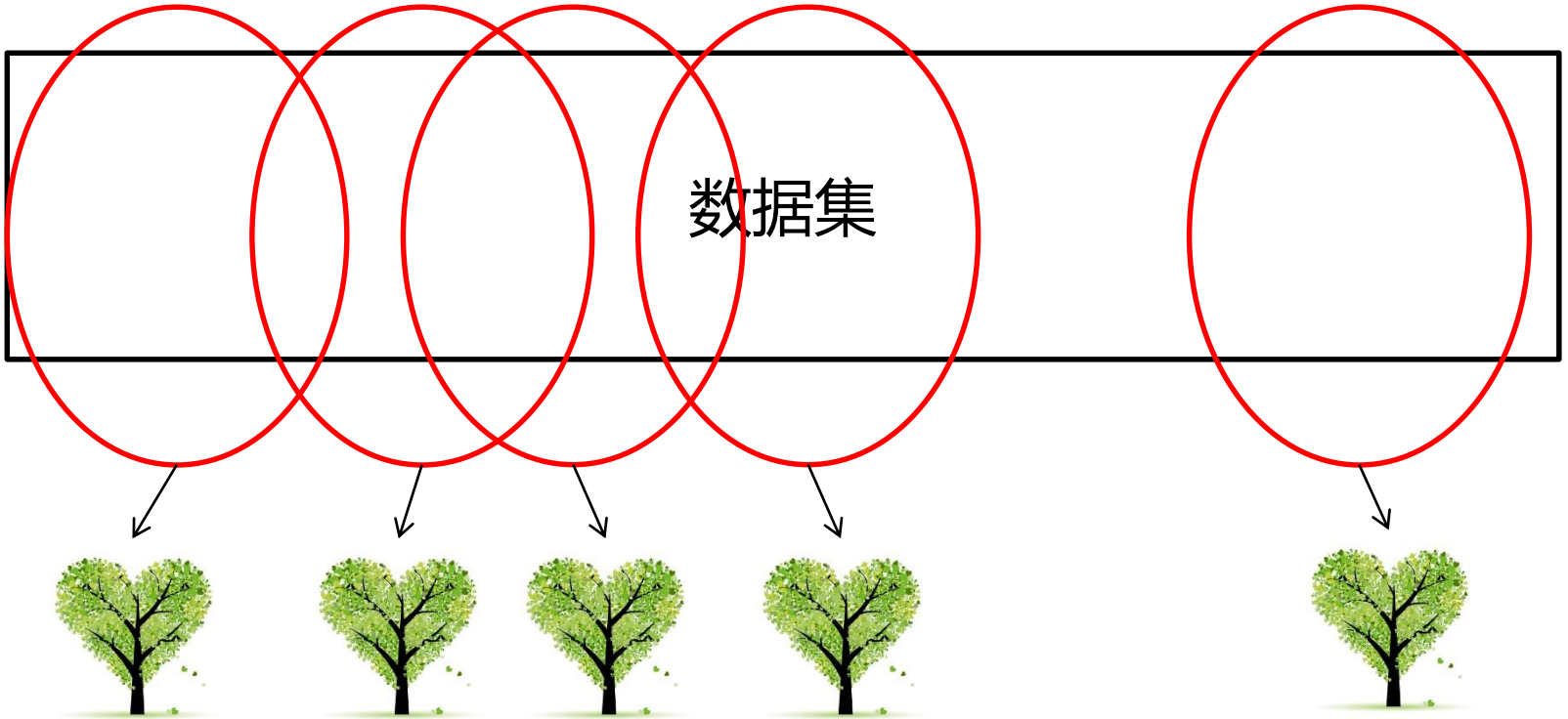
机器学习-随机森林

随机森林

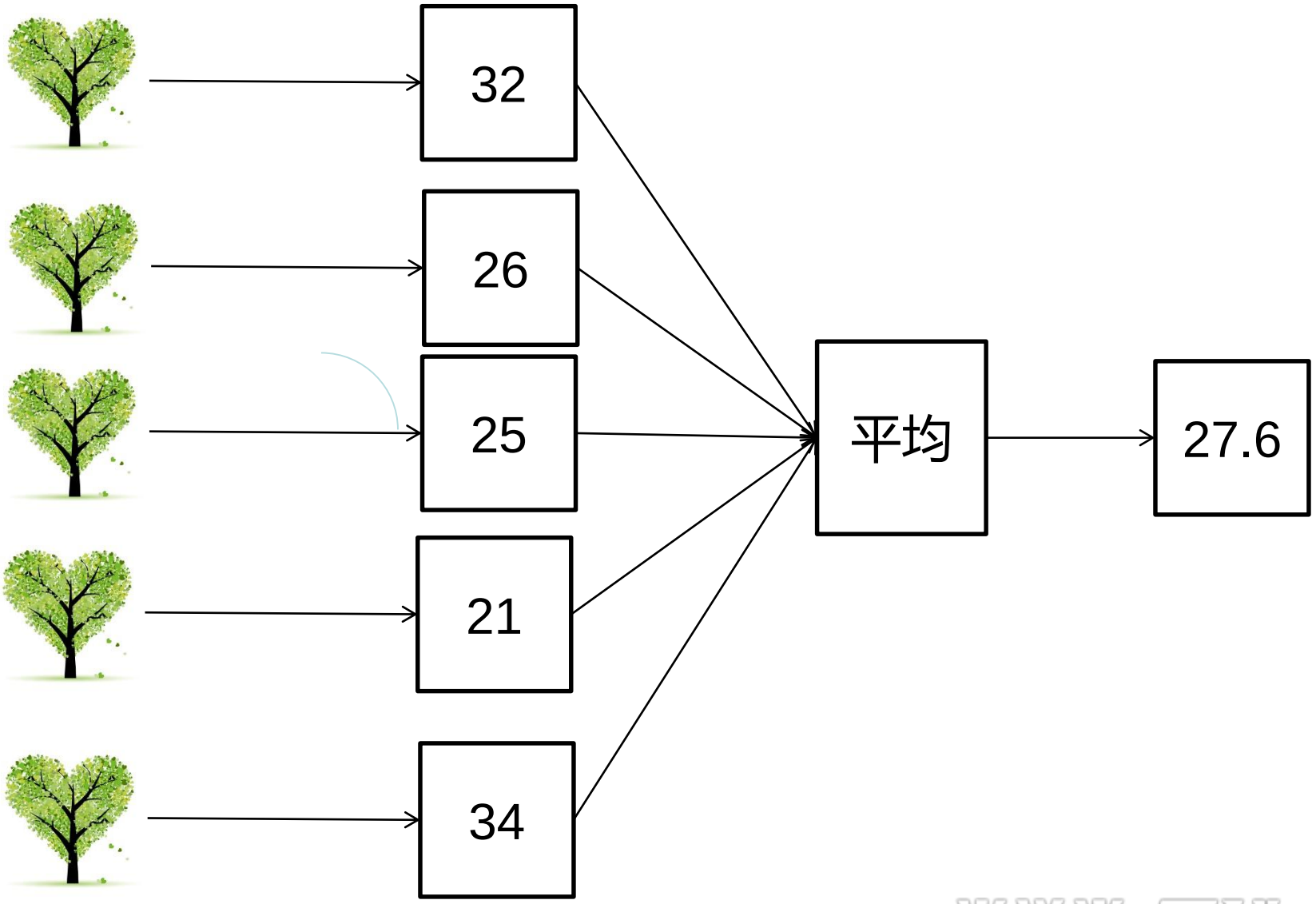
森林：由树组成
随机：生成树的数据都是从数据集中随机选取的

机器学习-尚学堂

当数据集很大的时候，我们随机选取数据集的一部分，生成一棵树，重复上述过程，我们可以生成一堆形态各异的树，这些树放在一起就叫森林



机器学习-尚学堂



机器学习-尚学堂

随机树的生成方式——找寻合适的分割维度和分割点