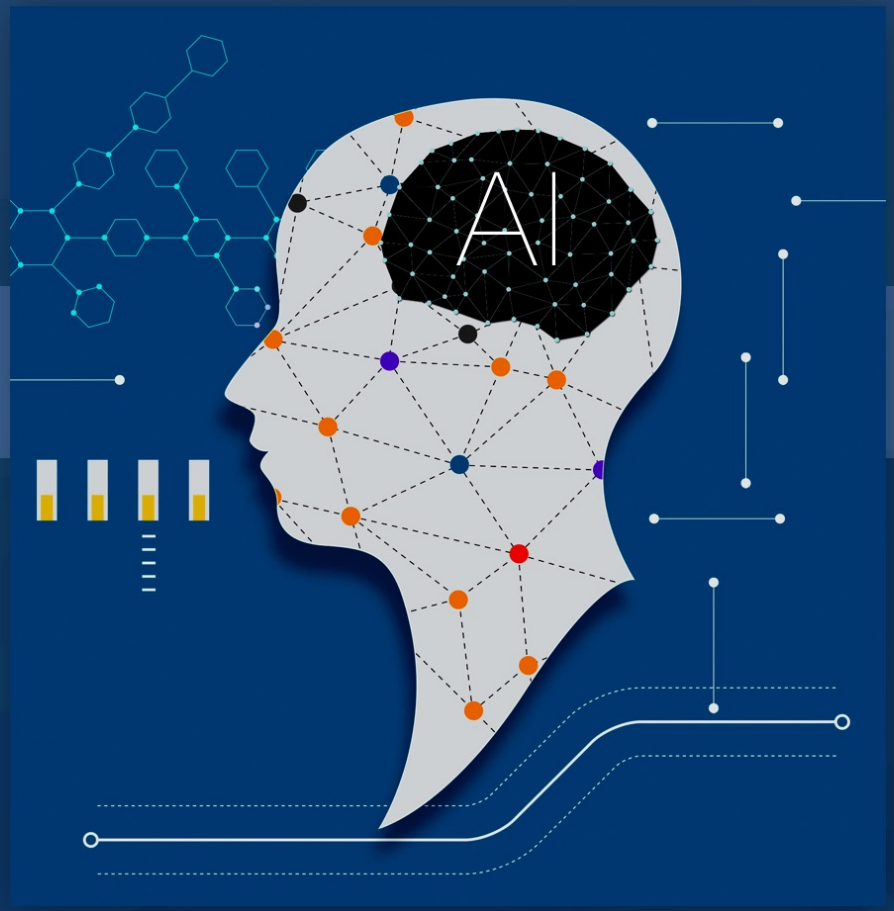
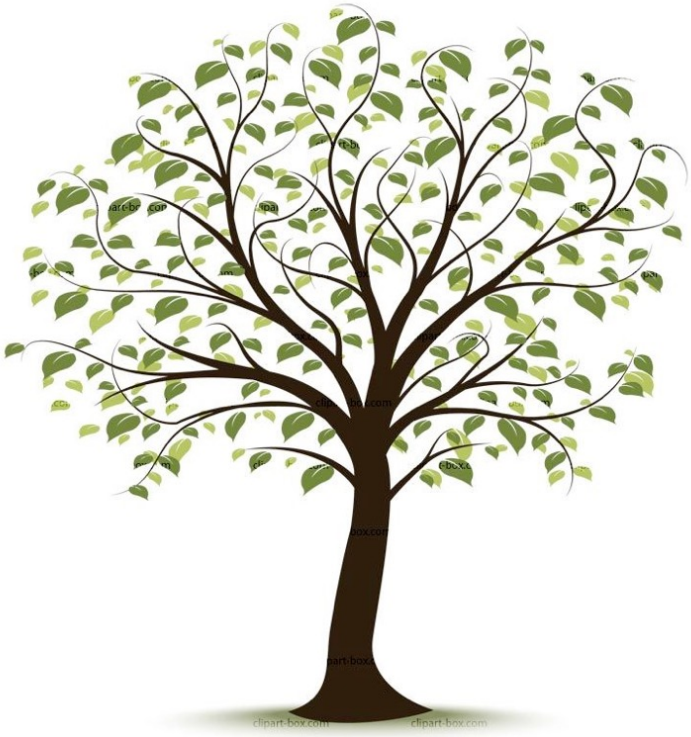
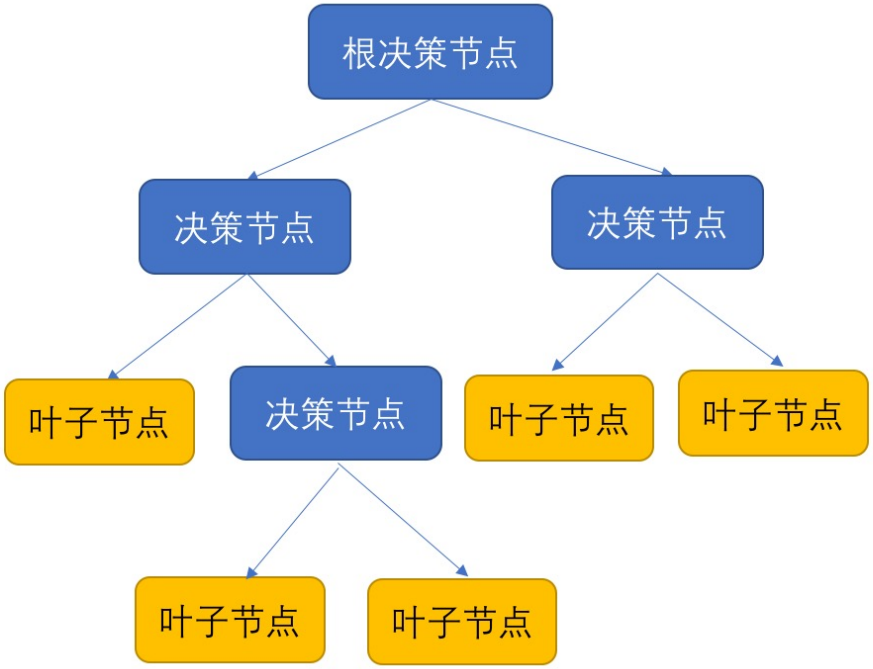
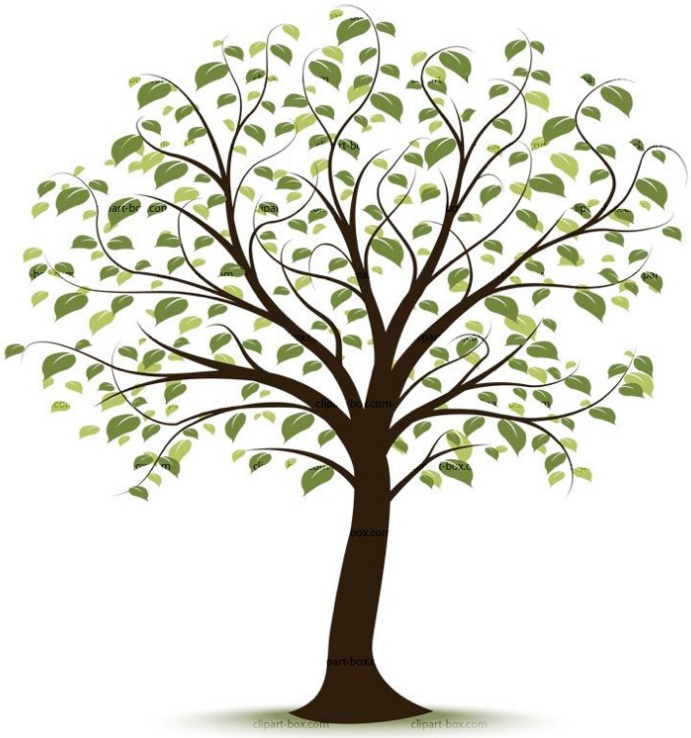


贪心科技
让每个人享受个性化教育服务

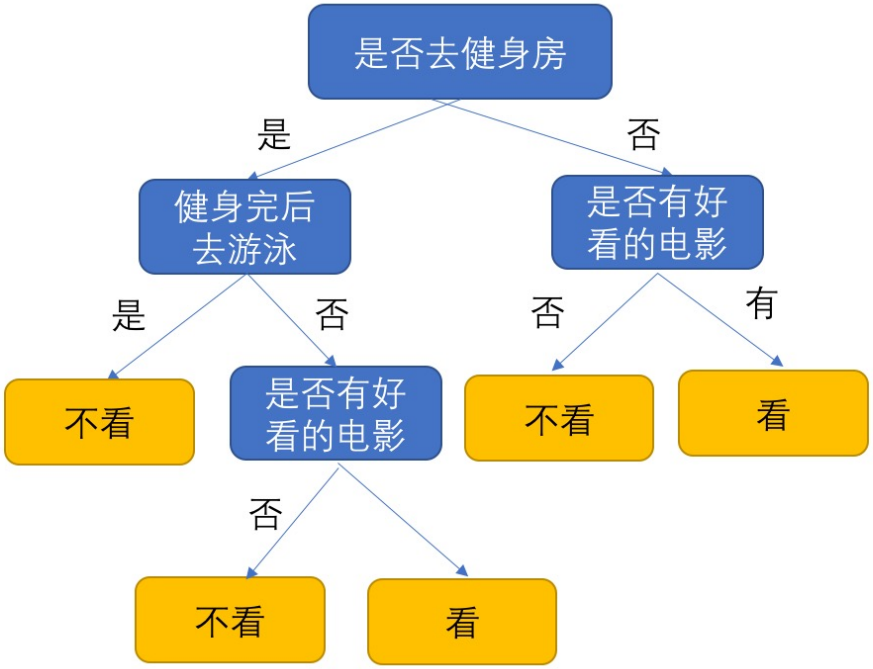
决策树




根决策节点
决策节点
决策节点
叶子节点
决策节点
叶子节点
叶子节点
叶子节点
叶子节点



晚上是否看电影



信息熵

- 衡量信息的不确定性或混乱程度的指标
- 信息的不确定性越大，熵越大。

- 三个概念:
  - 信息熵
  - 条件熵
  - 信息增益

信息熵

- 信息熵是用来衡量信息不确定性的指标
- 而不确定性是一个事件出现不同结果的可能性

- 例如抛硬币，可能出现的结果有两个，分别是正面和反面。而每次抛硬币的结果是一个非常不确定的信息。因为根据我们的经验或者历史数据来看，一个均匀的硬币出现正面和反面的概率相等，都是50%。因此很难判断下一次出现的是正面还是反面。这时抛硬币这个事件的熵值也很高。
- 而如果历史数据告诉我们这枚硬币在过去的100次试验中99次都是正面，也就是说这枚硬币的质量不均匀，出现正面结果的概率很高。那么我们就很容易判断下一次的结果了。这时的熵值很低。

信息熵

- 熵的计算公式：

$$H(X) = -\sum_{i=1}^{n} P(X=i)\log_2 P(X=i)$$

- 其中: P (X =i) 为随机变量 X 取值为 i 的概率.

信息熵（举例）

- 我们把抛硬币这个事件看做一个随机变量X，它可能的取值有2种，分别是正面 1 和反面 0 。每一种取值的概率分别为P(X=1) 和P(X=0)。

硬币	P
正面	0.5
反面	0.5

$$\text{Entropy} = -0.5 * log0.5 - 0.5 * log0.5 = 1$$

硬币	P
正面	0.99
反面	0.01

$$\text{Entropy} = -0.99 * log0.99 - 0.01 * log0.01 = 0.08$$

条件熵

- 条件熵是通过获得更多的信息来减小不确定性。我们知道的信息越多，信息的不确定性越小。
- 例如，我们无法只是根据用户历史数据中的购买某种商品的频率来判断这个用户在近期是否会再次购买某种商品，因为不确定性太大。在加入了促销活动，商品价格等信息后，我们可以更容易发现用户购买某种商品与促销活动，或者商品价格变化之间的联系。并通过在不同促销活动时购买行为出现的概率来降低不确定性。

条件熵

$$H(X|Y) = \sum_{v \in values(Y)} P(Y=v) H(X|Y=v)$$

$$H(X|Y=v) = -\sum_{i=1}^{n} P(X=i|Y=v) \log_2 P(X=i|Y=v)$$

信息增益

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

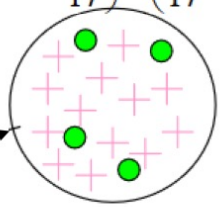
$$H(X) = -\sum_{i=1}^{n} P(X=i)\log_2 P(X=i)$$

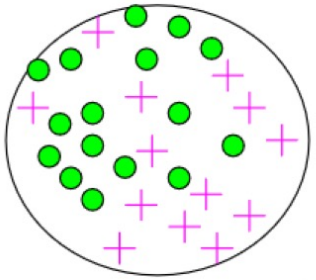
$$H(X|Y) = \sum_{v \in values(Y)} P(Y=v)H(X|Y=v)$$

$$H(X|Y=v) = -\sum_{i=1}^{n} P(X=i|Y=v)\log_2 P(X=i|Y=v)$$

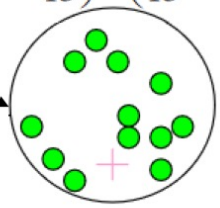
计算信息增益(举例)

子节点熵 $-\left(\frac{13}{17}\cdot\log_2\frac{13}{17}\right)-\left(\frac{4}{17}\cdot\log_2\frac{4}{17}\right)=0.787$





子节点熵 $-\left(\frac{1}{13}\cdot\log_2\frac{1}{13}\right)-\left(\frac{12}{13}\cdot\log_2\frac{12}{13}\right)=0.391$



父节点熵 $-\left(\frac{14}{30}\cdot\log_2\frac{14}{30}\right)-\left(\frac{16}{30}\cdot\log_2\frac{16}{30}\right)=0.996$

条件熵: 子节点加权熵 $\left(\frac{17}{30}\cdot 0.787\right)+\left(\frac{13}{30}\cdot 0.391\right)=0.615$

信息增益 = 父节点熵 − 子节点加权熵 = 0.996 − 0.615

例子

- 拥有不同天气状况某一位用户来打高尔夫球的历史记录。我们要做的是通过构建决策树来预测此用户在某种天气情况下是否会来打高尔夫球
- 我们无法仅通过Yes和No的历史频率来判断用户明天是否会来。因此，需要借助天气的信息来减少不确定性。下面分别记录到了4种天气情况，我们通过计算条件熵来开始构建决策树的第一步：构建根决策点。

Outlook	Temp	Humidity	Wind	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TURE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TURE	No
Overcast	Cool	Normal	TURE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TURE	Yes
Overcast	Mild	High	TURE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TURE	No

父节点的熵

- 构建根决策点的方法就是寻找4种天气情况中与打高尔夫球相关性最高的一个。首先我们来看Play Golf的熵，来看看这件事的不确定性有多高.
- 仅通过历史数据的频率来看预测Play Golf是一件非常不确定的事情，在14条历史数据中，打球的概率为64%，不打球的概率为36%。熵值达到了0.940。

Play Golf	Frequency	P	Entropy
No	5	0.36	-0.531
Yes	9	0.64	-0.410
	14		0.940

$$H(X) = -\sum_{i=1}^{n} P(X=i)\log_2 P(X=i)$$

4种条件下的条件熵（一. 晴朗程度）

Outlook	Yes	No	Frequency	P
Rainy	2	3	5	0.36
Overcast	4	0	4	0.29
Sunny	3	2	5	0.36
			14	

Outlook	Yes	No	Entropy
E(Rainy)	0.40	0.60	0.971
E(Overcast)	1.00	0.00	0.000
E(Sunny)	0.60	0.40	0.971

使用Outlook 的条件熵　0.36 *0.971 + 0.29 * 0 + 0.36 *0.971 =0.69

信息增益: 0.940 – 0.69 = 0.25 (最佳分隔特征)

4种条件下的条件熵（二. 温度）

Temp	Yes	No	Frequency	P
Hot	2	2	4	0.29
Mild	4	2	6	0.43
Cool	3	1	4	0.29
			14	

Temp	Yes	No	Entropy
E(Hot)	0.50	0.50	1.000
E(Mild)	0.67	0.33	0.918
E(Cool)	0.75	0.25	0.811

使用Temp分隔的熵 0.29 * 1 + 0.43 * 0.918 + 0.29 * 0.811 = 0.92

信息增益: 0.940 – 0.92 = 0.02

4种条件下的条件熵（三. 湿度）

Humidity	Yes	No	Frequency	P
High	3	4	7	0.50
Normal	6	1	7	0.50
			14	

Humidity	Yes	No	Entropy
E(High)	0.43	0.57	0.985
E(Normal)	0.86	0.14	0.592

使用Humidity分隔的熵 0.5 * 0.985 + 0.5 * 0.592 = 0.79
信息增益: 0.940 – 0.79 = 0.15

4种条件下的条件熵（四. 风力）

Wind	Yes	No	Frequency	P
FALSE	6	2	8	0.57
TURE	3	3	6	0.43
			14	

Wind	Yes	No	Entropy
E(FALSE)	0.75	0.25	0.811
E(TRUE)	0.50	0.50	1.000

使用Wind分隔的熵 0.57 * 0.811 + 0.43 * 1 = 0.89,
信息增益: 0.940 – 0.89 = 0.05

4种条件下的条件熵

Outlook	Yes	No	Frequency	P
Rainy	2	3	5	0.36
Overcast	4	0	4	0.29
Sunny	3	2	5	0.36
			14	

Temp	Yes	No	Frequency	P
Hot	2	2	4	0.29
Mild	4	2	6	0.43
Cool	3	1	4	0.29
			14	

Humidity	Yes	No	Frequency	P
High	3	4	7	0.50
Normal	6	1	7	0.50
			14	

Wind	Yes	No	Frequency	P
FALSE	6	2	8	0.57
TURE	3	3	6	0.43
			14	

Outlook	Yes	No	Entropy
E(Rainy)	0.40	0.60	0.971
E(Overcast)	1.00	0.00	0.000
E(Sunny)	0.60	0.40	0.971

Temp	Yes	No	Entropy
E(Hot)	0.50	0.50	1.000
E(Mild)	0.67	0.33	0.918
E(Cool)	0.75	0.25	0.811

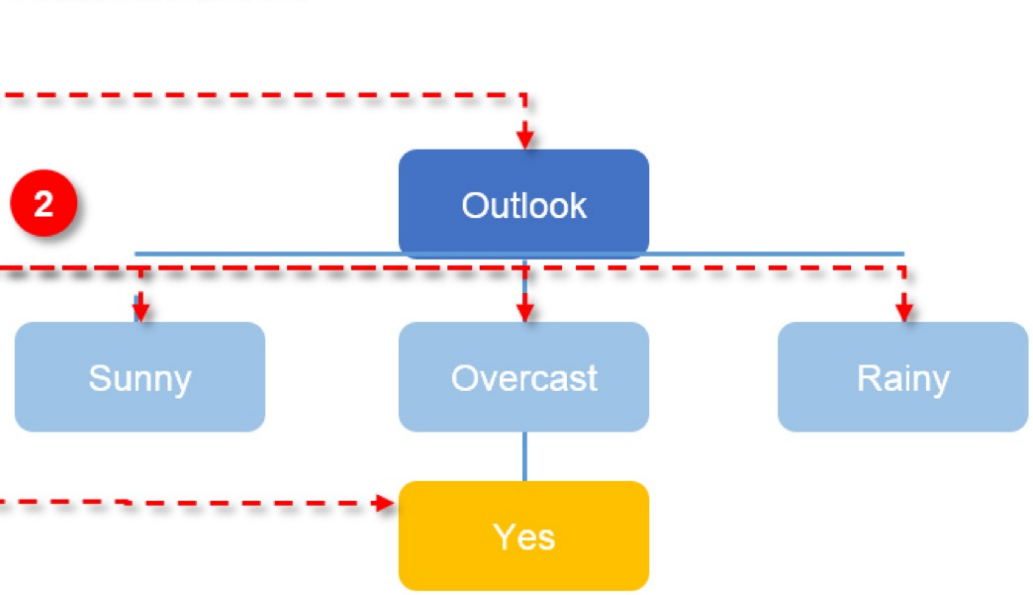
Humidity	Yes	No	Entropy
E(High)	0.43	0.57	0.985
E(Normal)	0.86	0.14	0.592

Wind	Yes	No	Entropy
E(FALSE)	0.75	0.25	0.811
E(TRUE)	0.50	0.50	1.000

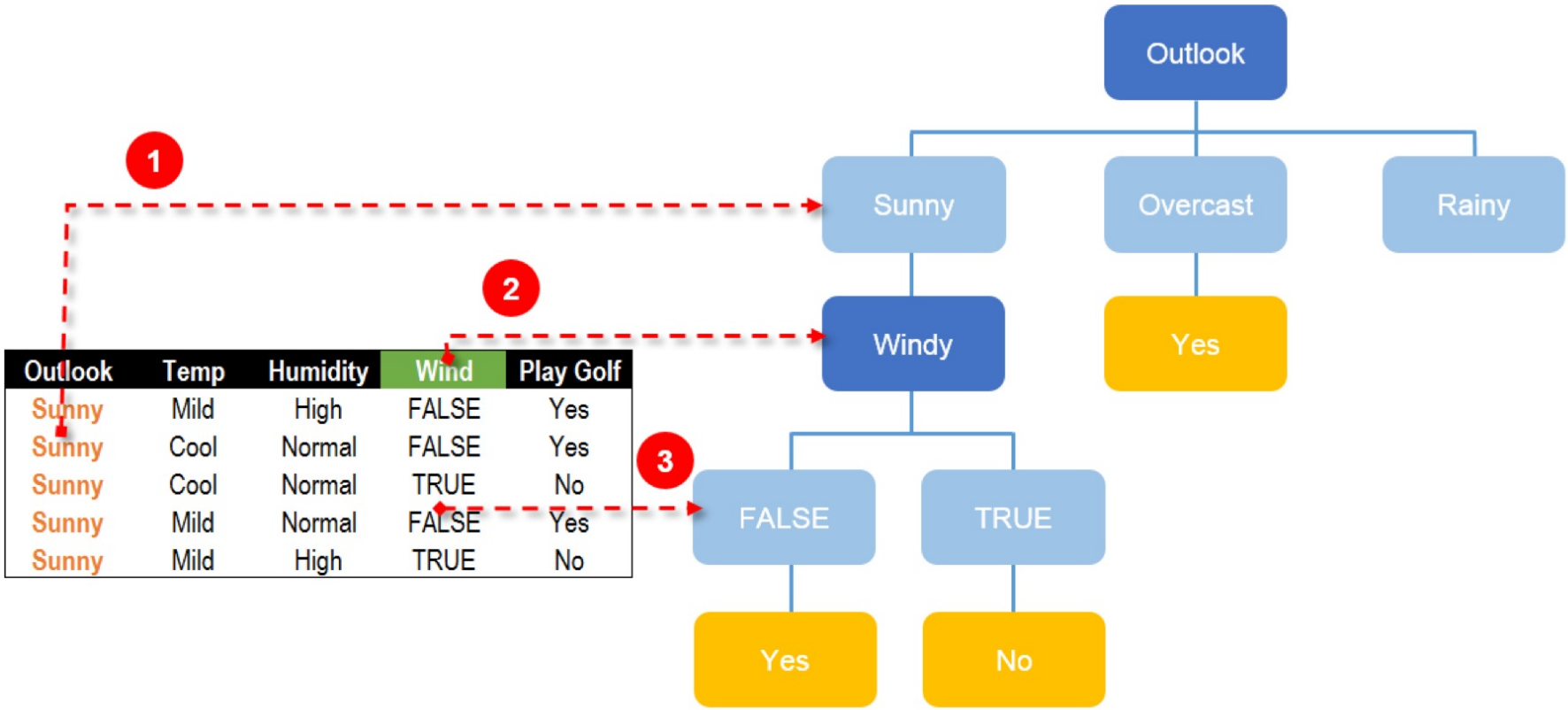
使用Outlook分隔的熵　0.36 *0.971 + 0.29 * 0 + 0.36 *0.971 =0.69, 信息增益: 0.940 – 0.69 = 0.25 (最佳分隔特征)
使用Temp分隔的熵 0.29 * 1 + 0.43 * 0.918 + 0.29 * 0.811 = 0.92,　信息增益: 0.940 – 0.92 = 0.02
使用Humidity分隔的熵 0.5 * 0.985 + 0.5 * 0.592 = 0.79,　信息增益: 0.940 – 0.79 = 0.15
使用Wind分隔的熵 0.57 * 0.811 + 0.43 * 1 = 0.89,　信息增益: 0.940 – 0.89 = 0.05

根节点的分隔

Outlook	Temp	Humidity	Wind	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No



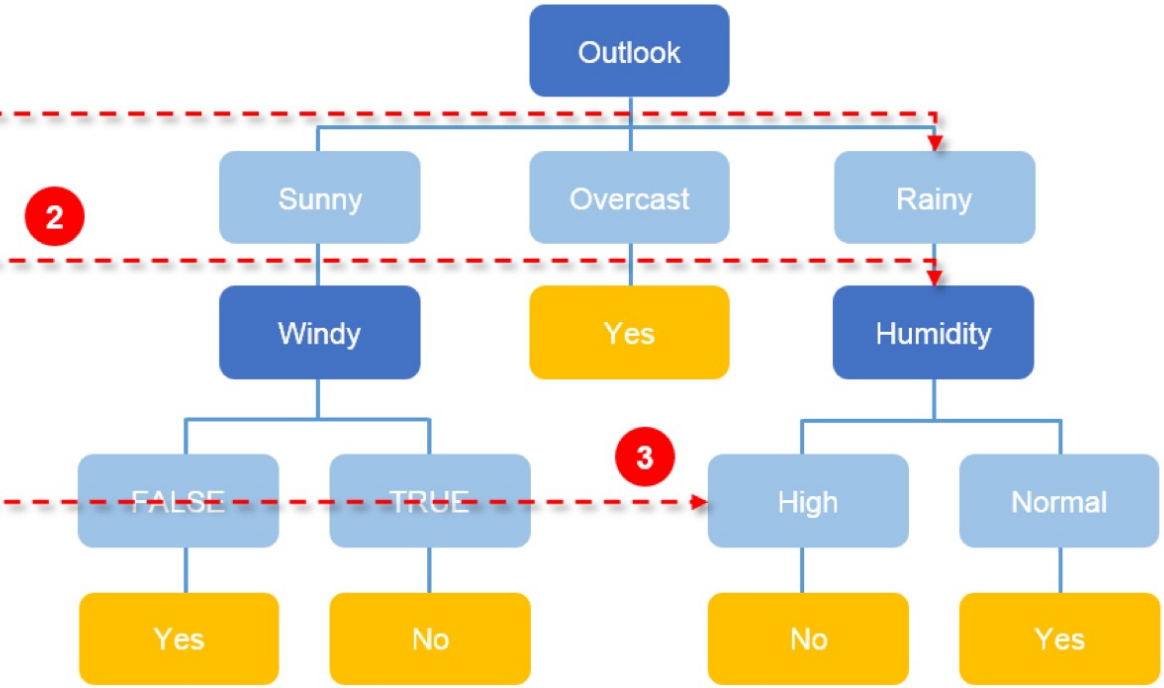
第二层，都属于Sunny情况下的再次分隔



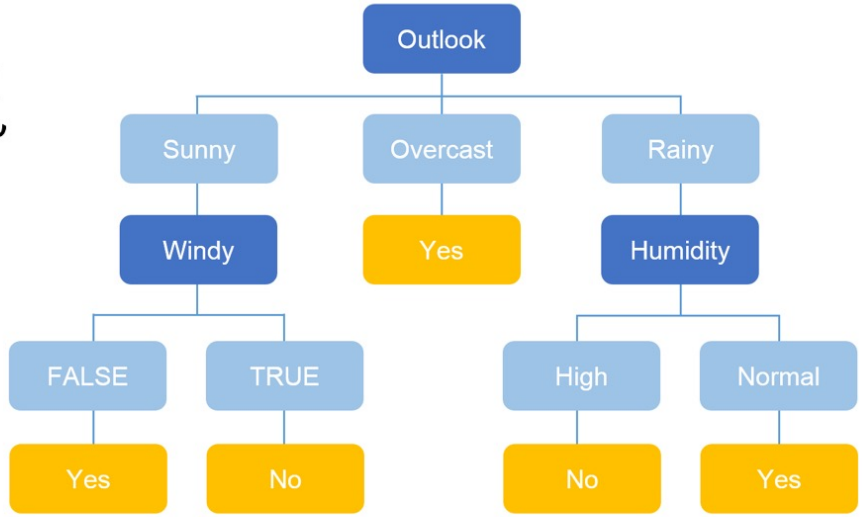
Outlook	Temp	Humidity	Wind	Play Golf
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

第二层，都属于Rainy情况下的再次分隔

Outlook	Temp	Humidity	Wind	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes



第二层, 都属于Overcast情况



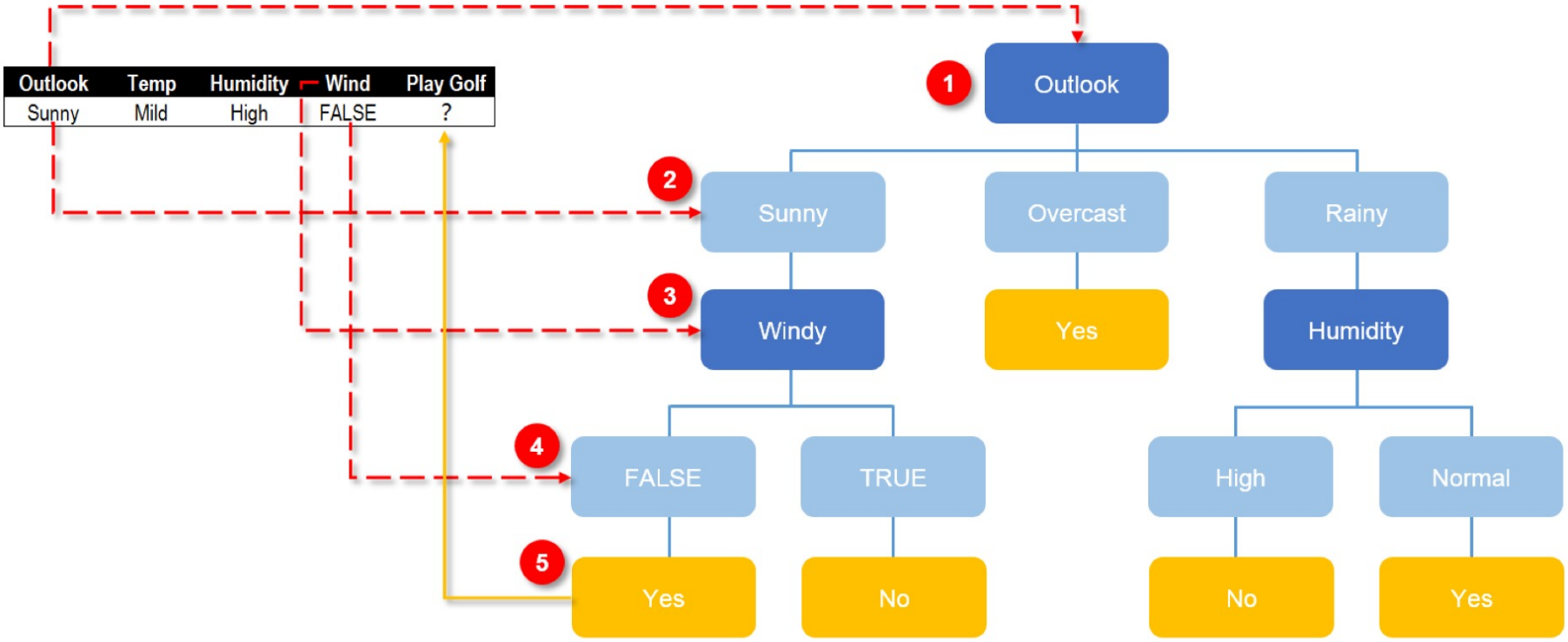
Outlook	Temp	Humidity	Wind	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

Outlook	Temp	Humidity	Wind	Play Golf
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

Outlook	Temp	Humidity	Wind	Play Golf
Overcast	Hot	High	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes

Outlook	Temp	Humidity	Wind	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

使用决策树做出预测


Outlook
Temp
Humidity
Wind
Play Golf
Sunny
Mild
High
FALSE
?
1
Outlook
2
Sunny
Overcast
Rainy
3
Windy
Yes
Humidity
4
FALSE
TRUE
High
Normal
5
Yes
No
No
Yes

对于属性是连续值而非类别的情况

如果温度Temp不是使用Hot, Mild, Cool来表达, 而是实际温度, 例如 31度等实数, 怎么办?

常用的解决方案是将实数的值等分为n个区间, 然后把一个区间内的值看作是一个值, 这样的话可以当作离散的特征值来处理了.

例如:

Temp <= 15 -> Cool

15 < Temp < 28 -> Mild

>= 28 -> Hot

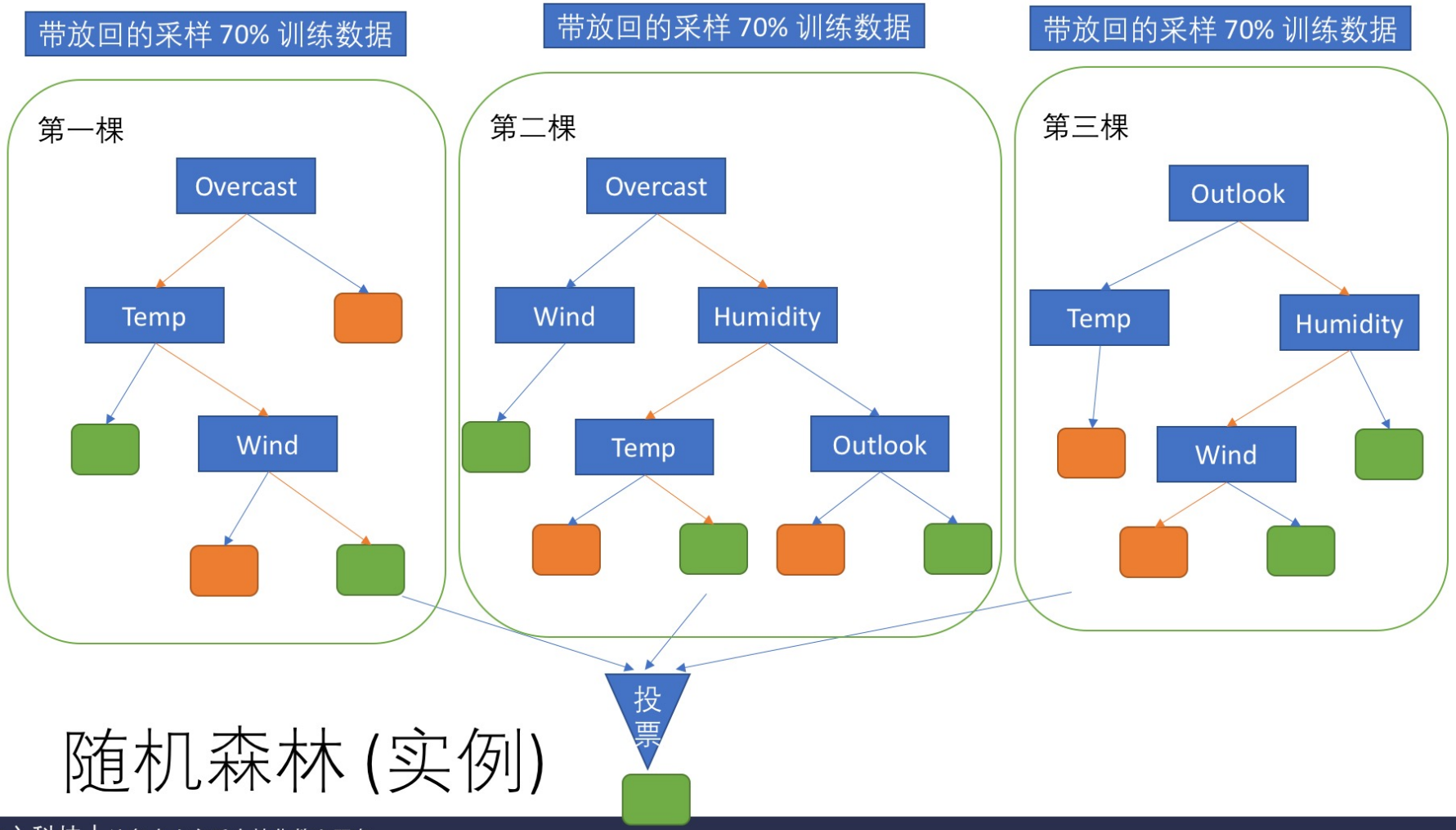
Outlook	Temp	Humidity	Wind	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TURE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TURE	No
Overcast	Cool	Normal	TURE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TURE	Yes
Overcast	Mild	High	TURE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TURE	No

决策树总结

- 优势:
  - 可解释性高
  - 能处理非线性的数据
  - 不需要数据归一化
  - 可以用于特征工程, 特征选择
  - 对数据分布没有偏好
  - 广泛使用
  - 容易软件实现
  - 可以转化为规则
- 弱点:
  - 启发式生成, 不是最优解
  - 容易过拟合
  - 微小的数据改变会改变整个树的形状
  - 对类别不平衡的数据不友好

随机森林

- 集成学习 (Ensemble Learning)
- 同时训练多个决策树, 预测的时候, 综合考虑多个结果做预测. 例如取多个结果的均值, 众数
- 随机性体现在两点:
  - 1. 从原来的训练数据集随机(带放回)取一个子集作为森林中某一个决策树的训练数据集
  - 2. 每一次选择分叉的特征时, 限定为在随机选择的特征的子集中寻找一个特征.
- 两个优势:
  - 消除了决策树容易过拟合的缺点
  - 减小了预测的variance:预测值不会因为训练数据的小变化而剧烈变化


带放回的采样 70% 训练数据
带放回的采样 70% 训练数据
带放回的采样 70% 训练数据
第一棵
Overcast
Temp
Wind
第二棵
Overcast
Wind
Humidity
Temp
Outlook
第三棵
Outlook
Temp
Humidity
Wind
投票

随机森林 (实例)

实例

贪心科技 GREEDY TECHNOLOGY

贪心科技　让每个人享受个性化教育服务

THANKS

贪心学院讲师：袁源

Appendix 课后阅读

决策树特征的重要性

- 计算决策树特征重要性的步骤, 假设数据有 M 个特征, 使用信息熵来决定每一个分叉情况:
  - 1. 初始化一个数组 A[], 长度为 M, 所有的值都为 0 .
  - 2. 遍历树的每一个节点:
    - 假设某一个分叉点是基于第 m (0 <= m <= M) 个特征来分叉的;
    - 假设这一个节点分叉造成的信息熵减小值为 e;
    - 假设训练数据中通过这个分叉的数据个数为 d;
    - 令 A[m] += d * e;
  - 3. 假设数组A[] 中所有的值求和的结果为 S, 将数值 A[] 中的每个元素除以 S
  - 4. 最终数组A[] 中第 m 个元素的值就是数据中第 m 个特征的重要性.

- 注意: 决策树的特征重要性是取决于特定数据集的, 即使同一棵树, 换了一个数据集, 特征重要性也会改变.