3 决策树（Decision Tree）

代启国

大连民族大学
计算机科学与技术系

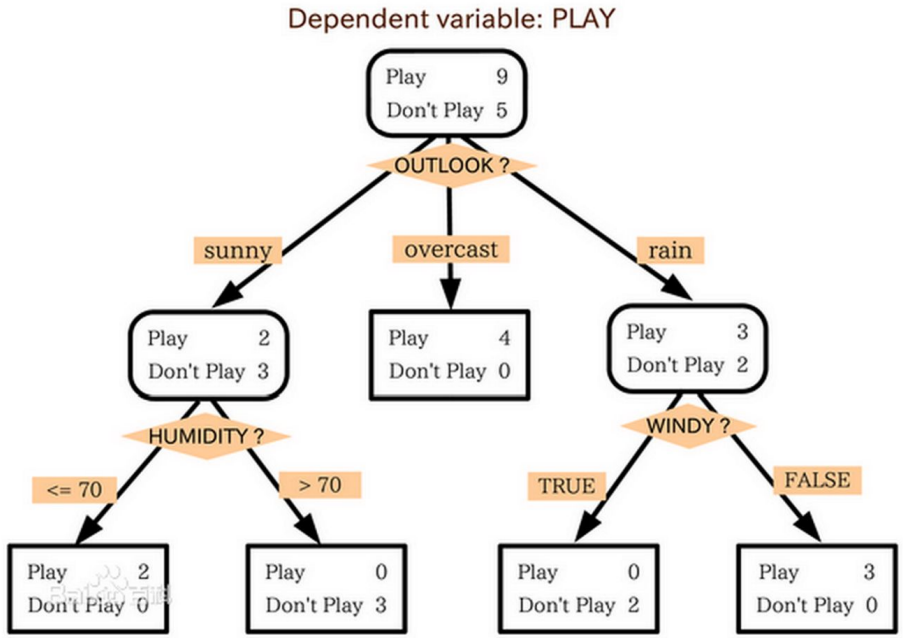
2018 年 11 月 30 日

是否应该去打球?

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

决策树的基本思想

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No



决策树的基本思想

决策树（Decision Tree） 一种经典且十分常用的机器学习方法

基本思想：

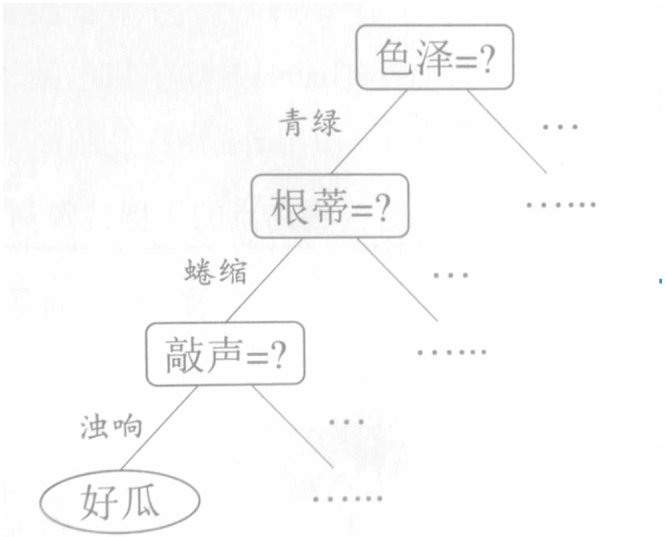
- 决策问题：分类新样本（两类）的任务 == 该样本属于“正类”吗?
- 决策树就是用树结构来进行决策的
  - 对“好瓜”进行决策时，通常会进行一些判断或“子决策”

决策树的基本思想

决策树（Decision Tree） 一种经典且十分常用的机器学习方法

基本思想：

- 决策问题：分类新样本（两类）的任务 == 该样本属于“正类”吗？
- 决策树就是用树结构来进行决策的
  - 对“好瓜”进行决策时，通常会进行一些判断或“子决策”



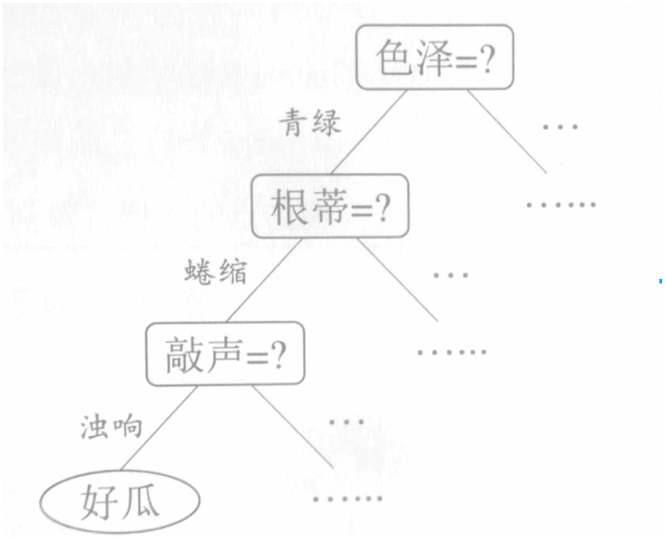
- 如果颜色是青绿色？
- 再看根蒂是否蜷缩？
- 敲声是否浊响？

决策树的基本思想

决策树（Decision Tree） 一种经典且十分常用的机器学习方法

基本思想：

- 决策问题：分类新样本（两类）的任务 == 该样本属于“正类”吗？
- 决策树就是用树结构来进行决策的
  - 对“好瓜”进行决策时，通常会进行一些判断或“子决策”

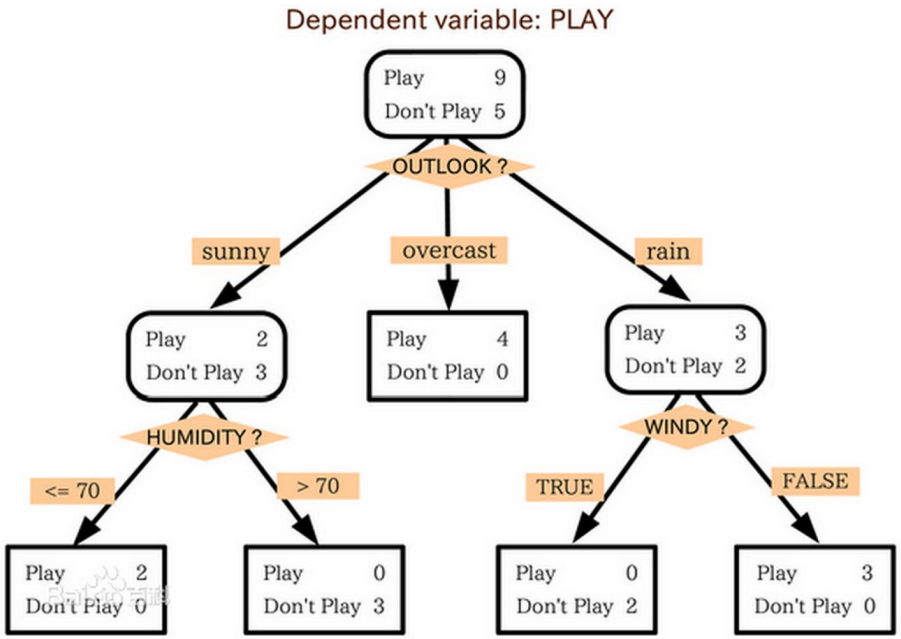


- 如果颜色是青绿色？
- 再看根蒂是否蜷缩？
- 敲声是否浊响？

- 决策树每个判定都是对某个属性的”测试”，最终的结论为“分类”

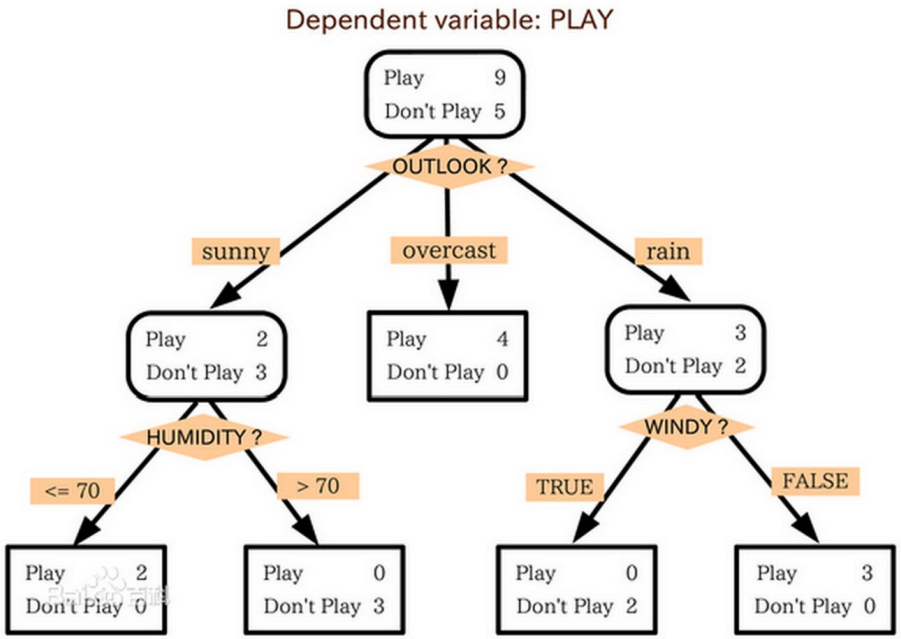
决策树的构造

- 一个根节点、若干个内部节点和若干个叶子节点
- 叶子节点对应于决策结果（类别），其他节点都对于一个属性“测试”



决策树的构造

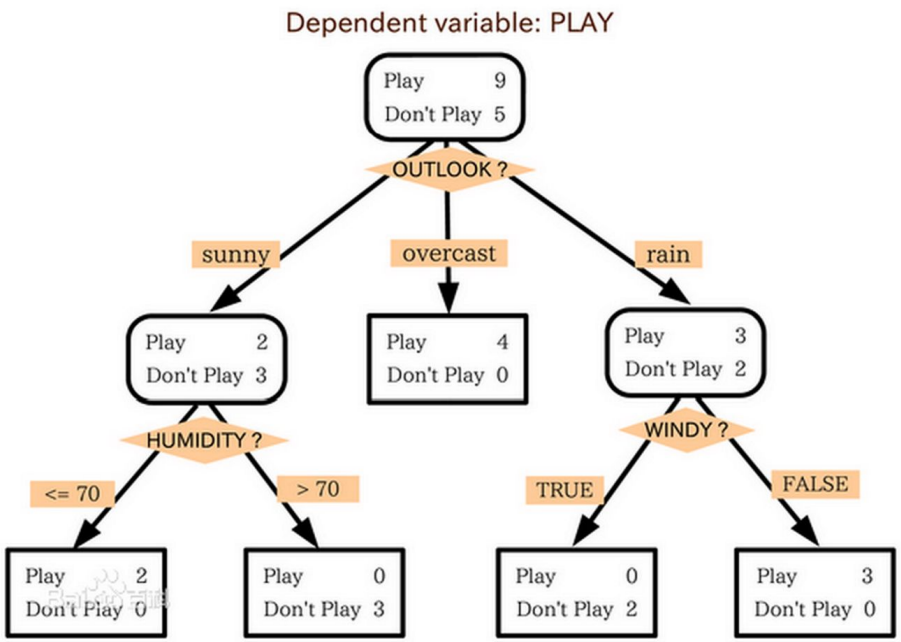
- 一个根节点、若干个内部节点和若干个叶子节点
- 叶子节点对应于决策结果（类别），其他节点都对于一个属性“测试”



- 根据属性测试，样本集被划分至子节点中
- 从根到每个叶节点路径，为判定测试序列

决策树的构造

- 一个根节点、若干个内部节点和若干个叶子节点
- 叶子节点对应于决策结果（类别），其他节点都对于一个属性“测试”



- 根据属性测试，样本集被划分至子节点中
- 从根到每个叶节点路径，为判定测试序列

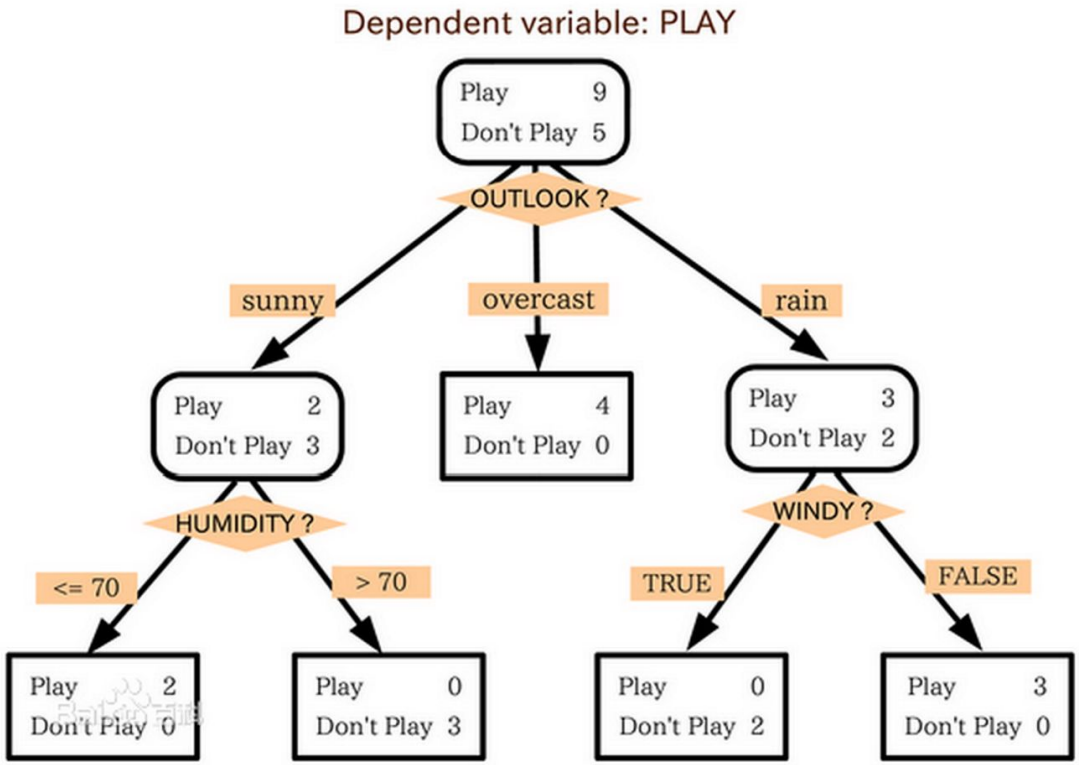
决策树学习

- 目的：产生一颗泛化能力强的树
- 原则：分而治之（divide and conquer）

决策树——划分选择

如何选择最优划分属性（特征）？

- 使得分支结点所包含的样本尽可能属于同一类别
- 即结点的纯度（purity）越来越高



决策树——划分选择

信息熵（Information entropy）

- 度量样本集合纯度最常用的一种指标
- 假定样本集合 D 中第 k 类样本所占比例为 $p_k\,(k=1,2,...,|y|)$
- 则样本集合 D 的信息熵定义为：

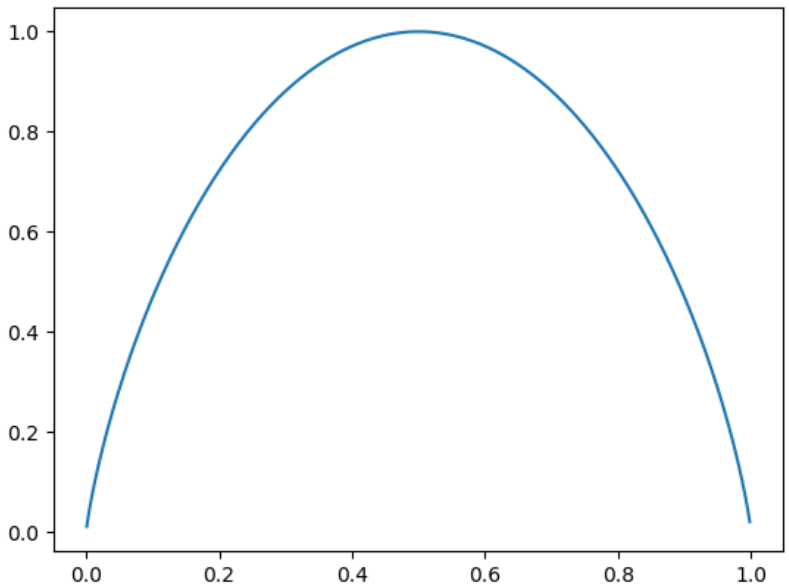
$$Ent\,(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$$

决策树——划分选择

信息熵（Information entropy）

- 样本集合 D 的信息熵定义为：

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$$



意义

$Ent(D)$ 的值越小，则 D 的纯度越高

决策树——划分选择

- 假定离散属性 a 有 V 个可能的取值 $\{a^1, a^1, ..., a^V\}$
- 利用属性 a 对样本集 D 进行划分，产生 V 个分支结点
- 第 v 个分支结点包含了 D 中属性 a 取值为 a^v 的所有样本，记为 D^v
- 计算每个分支权重 $|D^v| / |D|$

信息增益（Information entropy）

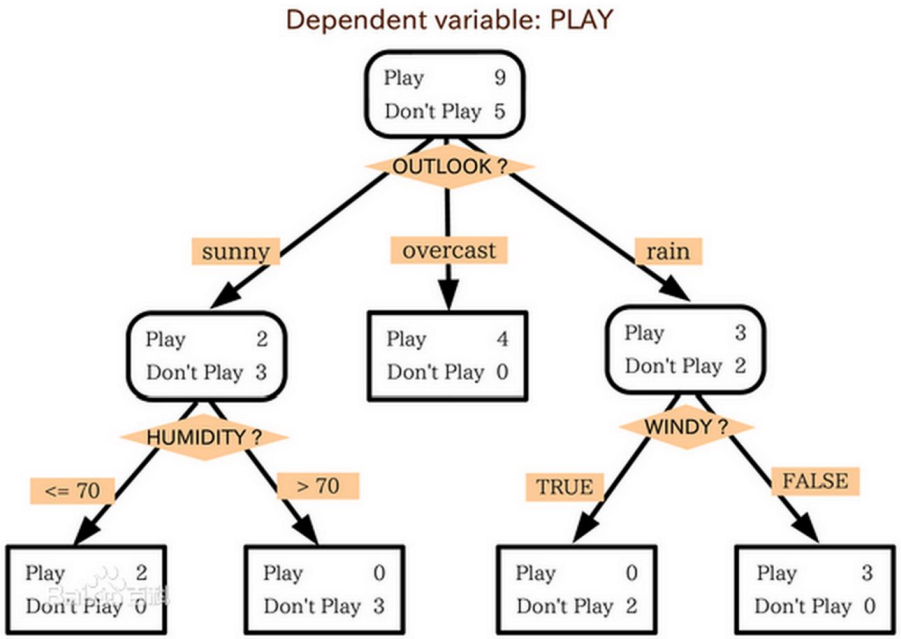
- 对于样本集 D, 属性 a 的信息增益：

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

决策树——划分选择

信息增益最大化

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

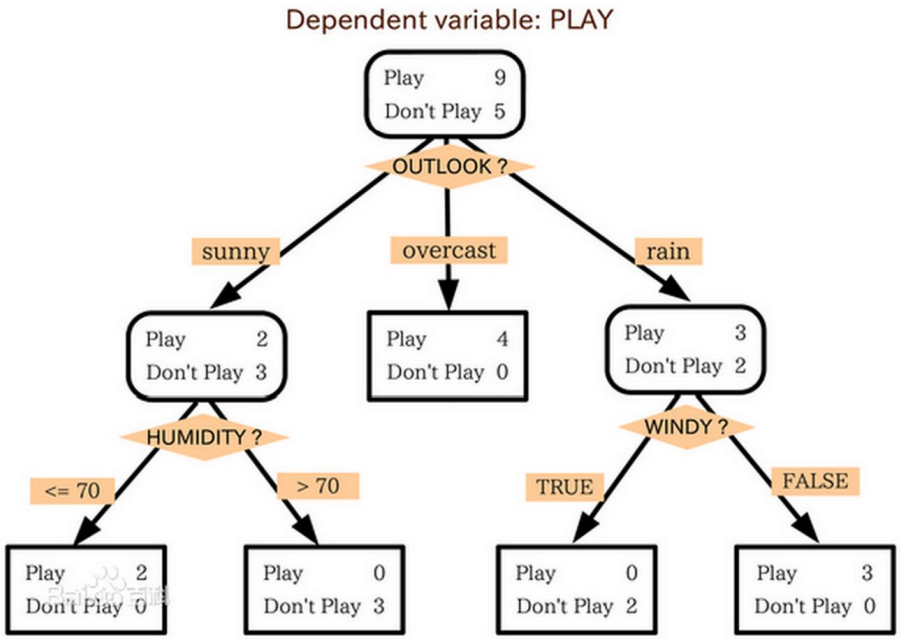


Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

决策树——划分选择

信息增益最大化

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

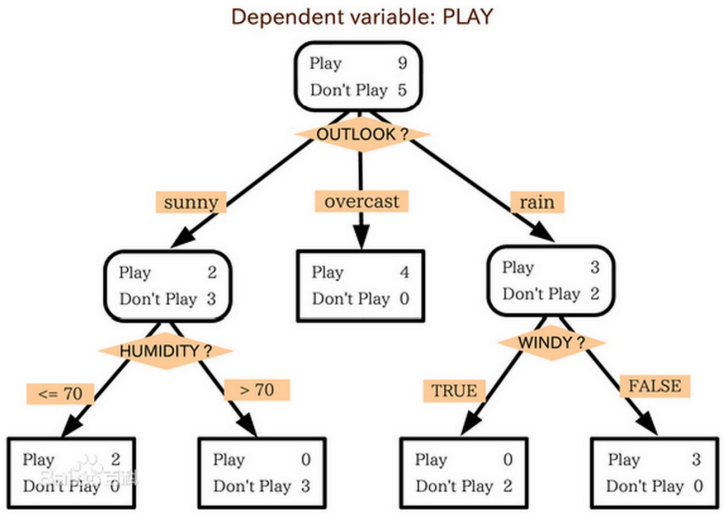


- 信息增益越大，表明使用属性 a 来进行划分所得到的“纯度提升”越大
- 可采用信息增益作为决策树属性划分的依据
- ID3 等决策树学习算法就是以信息增益为准则进行属性划分的

决策树——学习算法

决策树学习基本算法

输入： 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
　　属性集 $A = \{a_1, a_2, \ldots, a_d\}$.
过程： 函数 TreeGenerate(D, A)
1: 生成结点 node;
2: **if** D 中样本全属于同一类别 C **then**
3: 　将 node 标记为 C 类叶结点; **return**
4: **end if**
5: **if** $A = \varnothing$ **OR** D 中样本在 A 上取值相同 **then**
6: 　将 node 标记为叶结点, 其类别标记为 D 中样本数最多的类; **return**
7: **end if**
8: 从 A 中选择最优划分属性 a_*;
9: **for** a_* 的每一个值 a_*^v **do**
10: 　为 node 生成一个分支; 令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集;
11: 　**if** D_v 为空 **then**
12: 　　将分支结点标记为叶结点, 其类别标记为 D 中样本最多的类; **return**
13: 　**else**
14: 　　以 TreeGenerate(D_v, $A \setminus \{a_*\}$)为分支结点
15: 　**end if**
16: **end for**
输出： 以 node 为根结点的一棵决策树



- 寻找当前最优的属性 a_{best}
- 利用递归方法划分子节点
- 如果节点是" 纯" 的，则视为叶子节点

信息增益的局限

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

信息增益的局限

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

- 如果以第一列序号作为划分属性，信息增益是多少？

信息增益的局限

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

- 如果以第一列序号作为划分属性，信息增益是多少？
- “序号”的信息增益会远大于其他属性
- 该属性会产生 14 个分支，每个分支仅包含一个结点（最纯）

信息增益的局限

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

- 如果以第一列序号作为划分属性，信息增益是多少？
- “序号”的信息增益会远大于其他属性
- 该属性会产生 14 个分支，每个分支仅包含一个结点（最纯）
- 所产生的决策树不具有泛化能力，无法对新样本进行有效预测

决策树——划分选择

信息增益率

- 为了避免“信息增益”方法对可取值数目较多属性偏好的不利影响
- C4.5 算法不直接采用“信息增益”，而使用“增益率”

决策树——划分选择

信息增益率

- 为了避免“信息增益”方法对可取值数目较多属性偏好的不利影响
- C4.5 算法不直接采用“信息增益”，而使用“增益率”
- 增益率以信息增益为基础，其定义为

$$Gain\_ratio(D,a) = \frac{Gain(D,a)}{IV(a)}$$

其中，“固有值（intrinsic value）”

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$$

决策树——划分选择

信息增益率

- 为了避免“信息增益”方法对可取值数目较多属性偏好的不利影响
- C4.5 算法不直接采用“信息增益”，而使用“增益率”
- 增益率以信息增益为基础，其定义为

$$Gain\_ratio(D,a) = \frac{Gain(D,a)}{IV(a)}$$

其中，“固有值（intrinsic value）”

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$$

- 属性 a 的可取值数目越多（即 V 越大），通常情况下 $IV(a)$ 也越大
- 增益率则越小

决策树——划分选择

基尼指数（Gini index）

- 除信息熵外，基尼值也常用于选择划分属性

$$Gini(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} p_k^2$$

- $Gini(D)$ 表示样本集中随机抽取两个样本类别不一致的概率
- 该值越小，表示数据 D 的纯度越高

CART 决策树算法采用基尼指数选择划分属性

- 对于数据集 D，属性 a 的基尼指数定义为：

$$Gini\_index(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v)$$

连续值处理

- 连续属性可取值的数目不是有限的
- 无法根据连续属性的可取值对节点进行划分

连续值处理

- 连续属性可取值的数目不是有限的
- 无法根据连续属性的可取值对节点进行划分
- 需要对连续属性进行离散化

连续值处理

- 连续属性可取值的数目不是有限的
- 无法根据连续属性的可取值对节点进行划分
- 需要对连续属性进行离散化
- 常用方法: 寻找某阈值，使得该阈值对应的划分最优

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

连续值处理

- 连续属性可取值的数目不是有限的
- 无法根据连续属性的可取值对节点进行划分
- 需要对连续属性进行离散化
- 常用方法: 寻找某阈值，使得该阈值对应的划分最优

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

小结

- 决策树是一种十分经典的机器学习算法
- 学习决策树的关键在于如何选择划分后样本集合更纯的“属性”
- 划分选择准则：
  - 基于“熵”的增益率
  - 基于“基尼”指数
- 连续值属性的处理