第6章：决策树学习

蒋良孝 |  中国地质大学（武汉）

机器学习与数据挖掘团队

ljiang@cug.edu.cn

http://www.escience.cn/people/jlx/

本章内容

一、决策树学习基础知识

二、决策树学习基本算法

三、决策树学习常见问题

四、决策树学习理解解释

一、决策树学习基础知识

- 决策树学习(decision tree learning) 是机器学习中一类最常见的方法之一。顾名思义，决策树学习就是学习用来作决策的树。
- 决策树学习是一种逼近离散值目标函数的方法，学习到的函数被表示为一棵决策树。

一、决策树学习基础知识

- 一棵决策树一般包含一个根结点、若干个内部结点和若干个叶子结点。
- ✓ 叶子结点对应于决策结果；
- ✓ 每个内部结点对应于一个属性测试，每个内部结点包含的样本集合根据属性测试的结果被划分到它的儿子结点中；
- ✓ 根结点包含全部训练样本。
- ✓ 从根结点到每个叶子结点的路径对应了一条决策规则。

二、决策树学习基本算法

- 决策树学习的目的就是为了构造一棵泛化能力强，即处理待测样本能力强的决策树，基本算法遵循自顶向下、分而治之的策略，具体步骤如下：

1. 选择最好的属性作为测试属性并创建树的根结点
2. 为测试属性每个可能的取值产生一个分支
3. 训练样本划分到适当的分支形成儿子结点
4. 对每个儿子结点，重复上面的过程，直到所有的结点都是叶子结点

三、决策树学习常见问题

- 可见，决策树的学习是一个递归过程，过程的实现还需要解决以下六个方面的问题：

  1）最佳划分的度量问题

  2）处理缺失属性值问题

  3）处理连续属性值问题

  4）叶子结点的判定问题

  5）怎样解决过拟合问题

  6）待测样本的分类问题

三、决策树学习常见问题

1）最佳划分的度量问题：

- 从决策树学习基本算法的步骤可以看出，决策树学习的关键是如何选择最佳划分属性。一般而言，随着长树过程的不断进行，我们希望决策树的分支结点所包含的样本越来越归属于同一类别，即结点的“不纯度”(impurity) 越来越低。
- 因此，为了确定按某个属性划分的效果，我们需要比较划分前（父亲结点）和划分后（所有儿子结点）不纯度的降低程度，降低越多，划分的效果就越好。

三、决策树学习常见问题

1）最佳划分的度量问题：

- 若记不纯度的降低程度为 Δ，则用来确定划分效果的度量标准可以下面的公式来定义：

$$\Delta_I = I(parent) - \sum_{j=1}^{k} \frac{N(j)}{N} I(j)$$

其中，$I(parent)$是父亲结点的不纯度度量，k是划分属性取值的个数。N是父亲结点上样本的总数，$N(j)$是第j个儿子结点上样本的数目，$I(j)$是第j个儿子结点的不纯度度量。

三、决策树学习常见问题

1）最佳划分的度量问题：

- 接下来的问题是，给定任意结点t，如何来定义它的不纯度度量，令$p(i)$为结点t中第i类样本所占有的比例，则结点t的不纯度度量主要包括：

✓ 熵：
$$Entropy(t) = -\sum_{i=1}^{c} p(i)\log_2 p(i)$$

✓ 基尼指数：
$$Gini(t) = 1 - \sum_{i=1}^{c} p(i)^2$$

✓ 误分类率：
$$Error(t) = 1 - \max_i p(i)$$

其中，c为类别数目，并且在计算熵时，令$0\log_2 0 = 0$

三、决策树学习常见问题

1）最佳划分的度量问题：

- 由此，我们就可以得到以下3种选择最佳划分的度量标准：

✓ 熵减最大：$\Delta_{Entropy\,\mathrm{Re}\,duction} = Entropy(parent) - \sum_{j=1}^{k} \frac{N(j)}{N} Entropy(j)$

✓ 基尼指数减最大：$\Delta_{Gini\,\mathrm{Re}\,duction} = Gini(parent) - \sum_{j=1}^{k} \frac{N(j)}{N} Gini(j)$

✓ 误分类率减最大：$\Delta_{Error\,\mathrm{Re}\,duction} = Error(parent) - \sum_{j=1}^{k} \frac{N(j)}{N} Error(j)$

三、决策树学习常见问题

1）最佳划分的度量问题：

- 需要注意的是，这里的熵减最大度量标准就是信息增益最大度量标准，记为：Δ_{info}
- 信息增益标准存在一个内在的偏置，它偏好选择具有较多属性值的属性，为减少这种偏好可能带来的不利影响，著名的C4. 5决策树算法 [Quinlan, 1993] 不直接使用信息增益，而是使用“增益率” (Gain Ratio) 来选择最佳划分属性。

三、决策树学习常见问题

1）最佳划分的度量问题：

- 增益率度量就是在信息增益度量中引入一个被称为“分裂信息” (Split Information) 的项作分母来惩罚具有较多属性值的属性：

$$GainRatio = \frac{\Delta_{\inf o}}{SplitInfo}$$

其中，*SplitInfo*是划分属性的分裂信息。

三、决策树学习常见问题

1）最佳划分的度量问题：

- 属性的分裂信息*SplitInfo*度量了属性划分数据的广度和均匀性：

$$SplitInfo = -\sum_{j=1}^{k} p(j)\log_2 p(j)$$

其中，$p(j)$是当前结点中划分属性第j个属性值所占有样本的比例。

- 分裂信息实际上就是当前结点关于划分属性各值的熵，它可以阻碍选择属性值均匀分布的属性。

三、决策树学习常见问题

1）最佳划分的度量问题：

- 但同时也产生了一个新的实际问题：当划分属性在当前结点中几乎都取相同的属性值时，会导致增益率无定义或者非常大（分母可能为0或者非常小）。
- 为了避免选择这种属性，C4.5决策树算法[Quinlan, 1993]并不是直接选择增益率最大的划分属性，而是使用了一个启发式方法：先计算每个属性的信息增益及平均值，然后仅对信息增益高于平均值的属性应用增益率度量。

三、决策树学习常见问题

1）最佳划分的度量问题：

- 为了克服信息增益度量和增益率度量的问题，平均增益*AverGain*度量[Wang & Jiang, 2007]被提出。平均增益度量用划分属性取值的个数来替换属性的分裂信息，不仅惩罚了属性值较多的属性，还避免了增益率度量的实际问题。具体的度量公式如下：

$$AverGain = \frac{\Delta_{Info}}{k}$$

- 其中，k是划分属性取值的个数。

三、决策树学习常见问题

1）最佳划分的度量问题：

- 增益率和平均增益度量改进信息增益度量的方法同样适合于基尼指数和误分类率，由此，我们又可以得到以下6种选择最佳划分的度量标准：

✓ 熵减率最大：$$\Delta_{EntropyReductionRatio} = \frac{\Delta_{EntropyReduction}}{SplitInfo}$$

✓ 基尼指数减率最大：$$\Delta_{GiniReductionRatio} = \frac{\Delta_{GiniReduction}}{SplitInfo}$$

✓ 误分类率减率最大：$$\Delta_{ErrorReductionRatio} = \frac{\Delta_{ErrorReduction}}{SplitInfo}$$

三、决策树学习常见问题

1）最佳划分的度量问题：

- 增益率和平均增益度量改进信息增益度量的方法同样适合于基尼指数和误分类率，由此，我们又可以得到以下6种选择最佳划分的度量标准：

✓ 平均熵减最大：$$\Delta_{AverEntropyReduction} = \frac{\Delta_{EntropyReduction}}{k}$$

✓ 平均基尼指数减最大：$$\Delta_{AverGiniReduction} = \frac{\Delta_{GiniReduction}}{k}$$

✓ 平均误分类率减最大：$$\Delta_{AverErrorReduction} = \frac{\Delta_{ErrorReduction}}{k}$$

三、决策树学习常见问题

1）最佳划分的度量问题：

- 给定训练集S，下面以信息增益度量作为最佳划分的标准，演示信息增益的计算和决策树生长的过程：

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

三、决策树学习常见问题

假如“*Outlook*”被选作划分属性，信息增益的计算：

{D1,D2,D3,D4,D5,D6,D7,D8,
D9,D10,D11,D12,D13,D14}
[9+,5-]

Outlook

Sunny *Overcast* *Rain*

{D1,D2,D8,
D9,D11}
[2+,3-]

{D3,D7,
D12,D13}
[4+,0-]

{D4,D5,D6,
D10,D14}
[3+,2-]

$E(S)=-9/14\log_2 9/14-5/14\log_2 5/14$

$E(S_{Sunny})=-2/5\log_2 2/5-3/5\log_2 3/5$

$E(S_{Overcast})=-4/4\log_2 4/4-0/4\log_2 0/4$

$E(S_{Rain})=-3/5\log_2 3/5-2/5\log_2 2/5$

$Gain(S,Outlook)=E(S)-[5/14E(S_{Sunny})+4/14E(S_{Overcast})+5/14E(S_{Rain})]=0.246$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

三、决策树学习常见问题

可见，用属性 “*Outlook*”划分样本集S的信息增益为：

- Gain(S,Outlook) = **0.246**

以同样的方法，我们可以得到分别以“*Temperature*”、“*Humidity*”、“*Wind*”作为划分属性的信息增益：

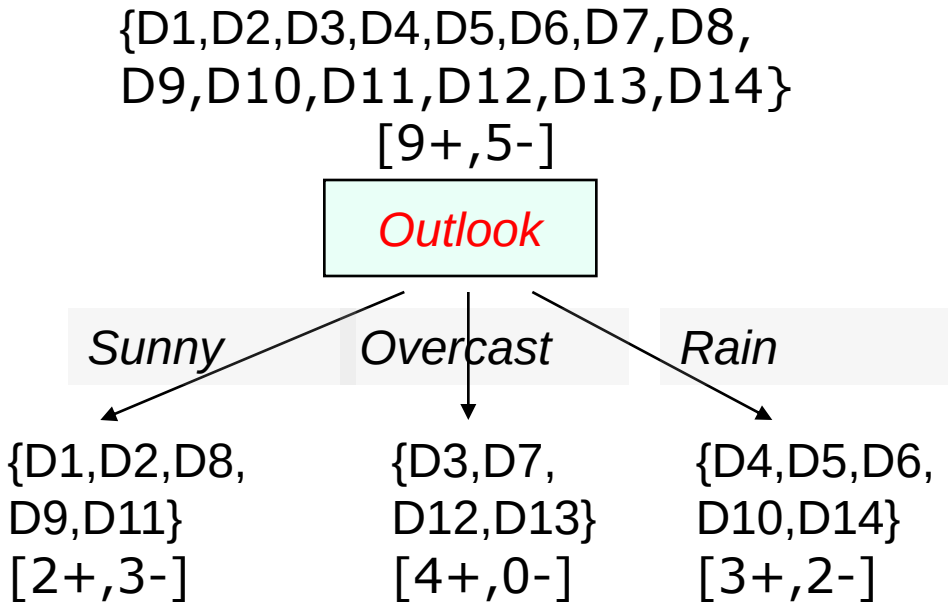
- Gain(S,Temperature) = 0.029
- Gain(S,Humidity) = 0.151
- Gain(S,Wind) = 0.048

因此，对于当前结点，用 “*Outlook*”划分样本集S的信息增益最大，被选为划分属性。

对于生成的每一个儿子结点，重复上面的过程，直到所有的结点为叶子结点。长树过程演示如下：

三、决策树学习常见问题

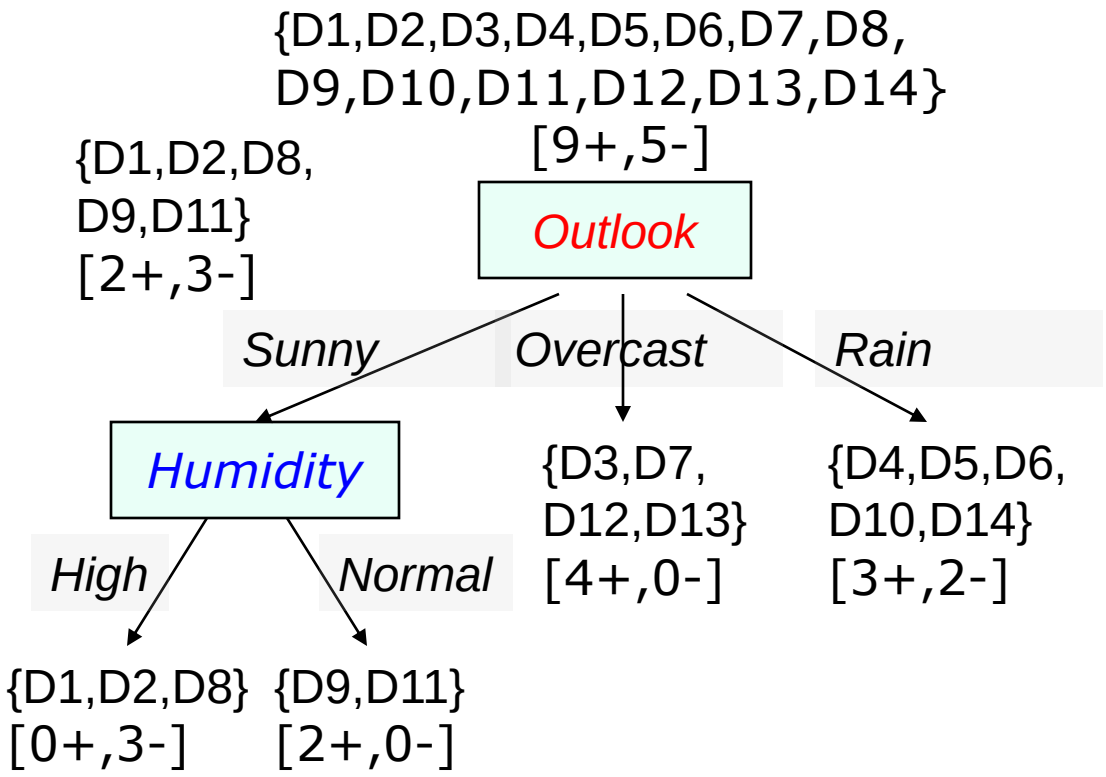
对于根结点，“*Outlook*”被选作最佳划分：



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

三、决策树学习常见问题

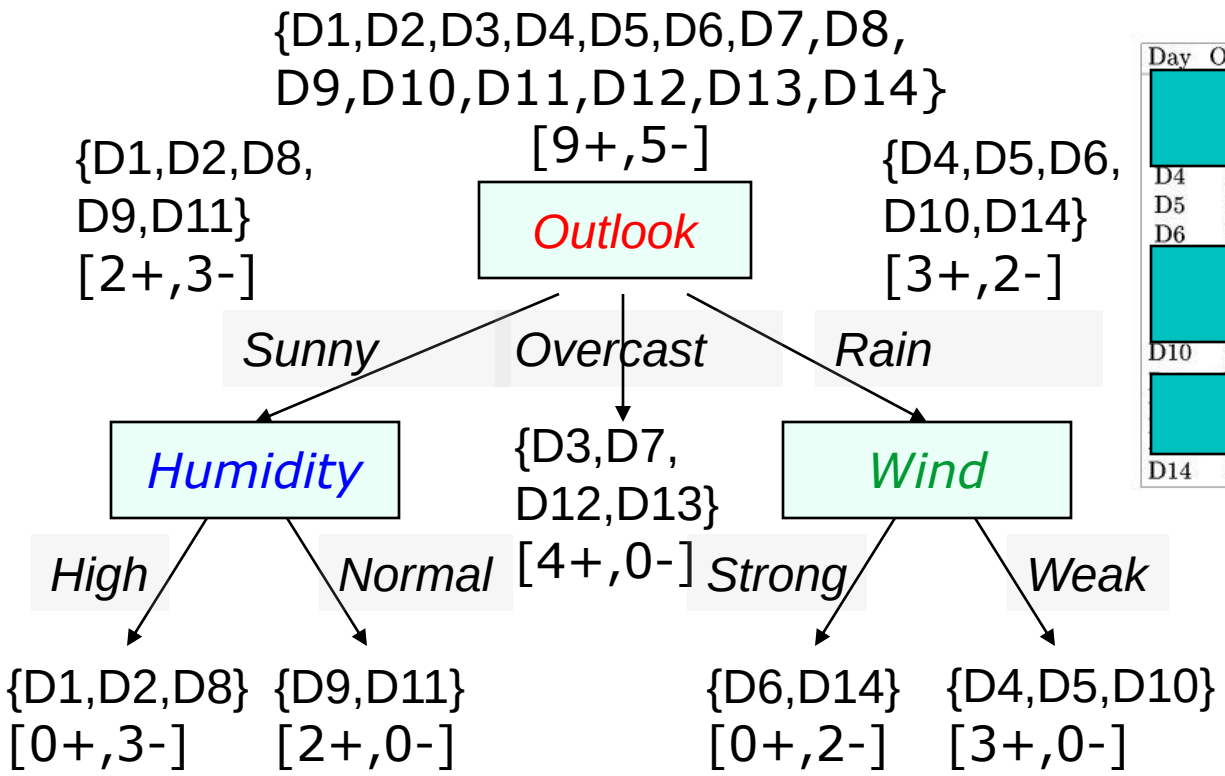
对于左儿子{D1,D2,D8,D9,D11}, “*Humidity*”被选：



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

三、决策树学习常见问题

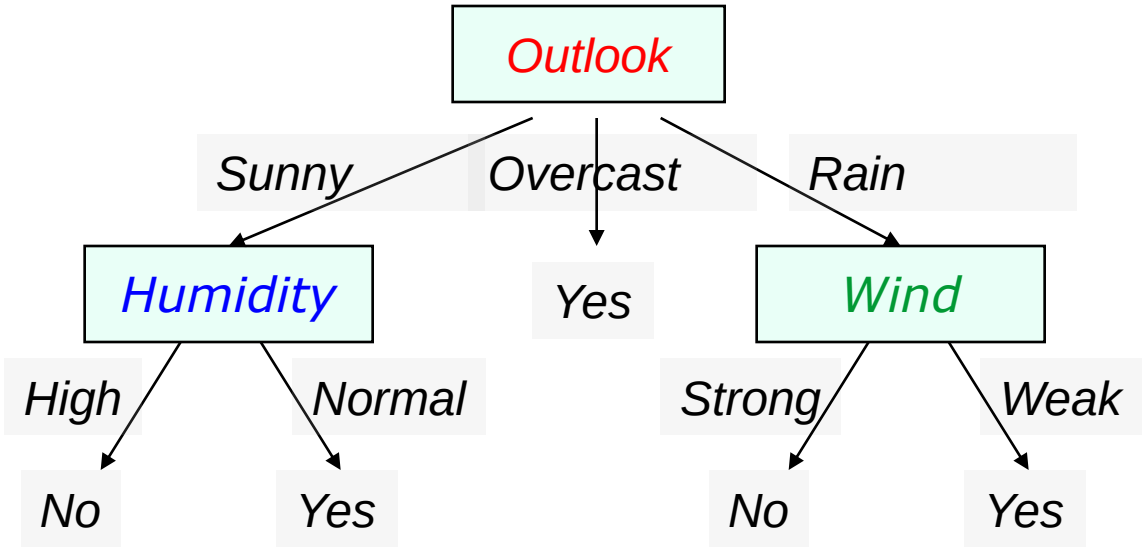
对于右儿子{D4,D5,D6,D10,D14}, “*Wind*”被选：



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D10	Rain	Mild	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

三、决策树学习常见问题

对于整个样本集S，按信息增益度量生成的决策树：



三、决策树学习常见问题

2）处理缺失属性值问题：

- 现实任务中常会遇到不完整样本，即样本的某些属性值缺失，尤其是在属性数目较多的情况下，往往会有大量样本出现缺失值。面对缺失属性值，决策树学习会面临两个方面的问题：

✓ 如何计算含缺失值属性的划分度量、并进行最佳划分的选择？

✓ 选择好最佳划分后，若样本在该属性上的值缺失，如何对样本进行划分？

三、决策树学习常见问题

2）处理缺失属性值问题：

- 处理缺失属性值的问题，通常有两个办法：
- ✓ 放弃函缺失值的样本，仅使用无缺失值的样本来进行学习，这种方法造成了数据信息的浪费。
- ✓ 根据此属性值已知的其他样本，来估计这个缺失的属性值。
  - ➢ 赋给它当前结点所有样本中该属性最常见的值
  - ➢ 赋给它当前结点同类样本中该属性最常见的值
  - ➢ 为含缺失值属性的每个可能值赋予一个概率，而不是简单地将最常见的值赋给它。

三、决策树学习常见问题

3）处理连续属性值问题：

- 到目前为止我们仅讨论了基于离散属性来生成决策树。现实学习任务中常会遇到连续属性，有必要讨论如何在决策树学习中使用连续属性。
- 由于连续属性的可取值数目不再有限，因此，不能直接根据连续属性的可取值来对结点进行划分。此时，连续属性的离散化技术可派上用场。
- 数据离散化是一个很大的研究主题，学者们提出的离散化技术也很多，可以分为：无监督离散化和有监督离散化。

三、决策树学习常见问题

3）处理连续属性值问题：

- 无监督离散化常用的有等深分箱法和等宽分箱法：等深分箱法让每个分箱中的样本数目一致；等宽分箱法让每个分箱中的取值范围一致。
- 等宽分箱法也叫均分法，就是把一个连续取值的区间等分成若干段，每一段赋一个离散值，常用的有ten-binning。

三、决策树学习常见问题

3）处理连续属性值问题：

- 有监督离散化常用的有二分法(bi-partition)，即将连续取值的属性按选定的阈值分割成布尔属性（二值属性）：

✓ 按照某个连续属性A排列训练样本，找出类标记不同的相邻样本

✓ 计算类标记不同的相邻样本的属性A的取值的中间值，产生一组候选阈值，可以证明产生最大信息增益的阈值一定在这样的边界中（Fayyad, 1991）

✓ 计算与每个候选阈值关联的信息增益，选择具有最大信息增益的阈值来离散化连续属性A

- 二分法的扩展是最小描述长度法 (Minimum Description Length, MDL) (Fayyad & Irani, 1993)。MDL法将连续取值的属性分割成多个区间，而不是单一阈值的两个区间。

三、决策树学习常见问题

4）叶子结点的判定问题：

- 上述三个问题的解决都是围绕树要长大的问题来展开的，那到底什么时候才停止树的生长，也就是递归过程什么时候返回？当前结点会被判定为叶子结点？
- 如果我们暂且不考虑树的规模过大而导致的过拟合问题，在决策树学习基本算法中，有三种情形会判定为叶子结点：

✓ 当前结点中的样本集合为空，即空叶子；

✓ 当前结点中的所有样本全部归属于同一类别，即纯叶子；

✓ 当前结点中的所有样本在所有属性上取值相同，即属性被测试完的叶子。

对于后面两种情形，可合并等价为最佳划分的度量值为0。

三、决策树学习常见问题

5）怎样解决过拟合问题：

- 上述对叶子结点判定的情形，都太过苛刻和完美，从而造成决策树的规模过大，以致于把训练集自身的一些特点当作所有数据都具有的一般性质而导致过拟合问题。
- 剪枝(pruning)是解决过拟合问题的主要手段，基本策略有“预剪枝”(prepruning)和“后剪枝” (post pruning) 。

✓ 预剪枝：在算法完美划分训练数据之前就停止树生长；

✓ 后剪枝：允许树过度拟合训练数据，然后对树进行后修剪。

三、决策树学习常见问题

5）怎样解决过拟合问题：

- 尽管预剪枝可能看起来更直接，但是对过拟合的树进行后剪枝被证明在实践中更成功。这是因为在预剪枝中精确地估计何时停止增长树是非常困难的。
- 无论是通过预剪枝还是后剪枝来得到正确规模的树，一个关键的问题是使用什么样的准则来确定最终正确树的规模。

三、决策树学习常见问题

5）怎样解决过拟合问题：

- 解决这个问题的方法包括：

✓ 使用与训练样例截然不同的一套分离的样例，来评估通过后剪枝从树上修剪结点的效果。

✓ 使用所有可用数据进行训练，但进行统计测试来估计生长或修剪一个特定的结点是否有可能改善在训练集以外的样例上的性能。

✓ 使用一个明确的标准来衡量训练样例和决策树的复杂度，当这个编码的长度最小时停止树增长。

三、决策树学习常见问题

5）怎样解决过拟合问题：

- 上面的第一种方法是最普通的，常被称为训练和验证集法。它将可用数据分成两个样例集合：训练集用于形成学习到的假设；验证集用于评估这个假设在后续数据上的精度。
- 训练和验证集法的动机：即使学习器可能会被训练集误导，但验证集不大可能表现出同样的随机波动。
- 通常的做法是，所有样例的三分之二作训练集，三分之一作验证集。
- **训练和验证集法主要包括：错误率降低修剪和规则后修剪。具体算法可参阅 [Mitchell, 1997]。**

三、决策树学习常见问题

6）待测样本的分类问题：

- 到此，我们解决了决策树生长的相关问题，那么，决策树学习学到后，怎样应用决策树进行待测样本的分类？
- 分类待测样本的方法：从决策树的根结点开始，测试这个结点指定的划分属性，然后按照待测样本的该属性值对应的树枝向下移动。这个过程再在以新结点为根的子树上重复，直到将待测样本划分到某个叶子结点为止。然后根据该叶子结点上的训练样本集计算其后验概率，最后把具有最大后验概率的类赋给待测样本。

三、决策树学习常见问题

6）待测样本的分类问题：

- 给定一个叶子结点（其本质就是一个训练样本的集合），计算其后验概率的常用方法包括：投票法、加权投票法、局部概率模型法。当计算得到的后验概率出现相同的情况下，可以采用随机分类或者拒判的方法进行处理。
- **在计算后验概率的过程经常会采用一些常用的概率估计方法：基于频率的极大似然估计、拉普拉斯估计、基于相似度（距离）加权的拉普拉斯估计、m-估计，朴素贝叶斯估计等等。**

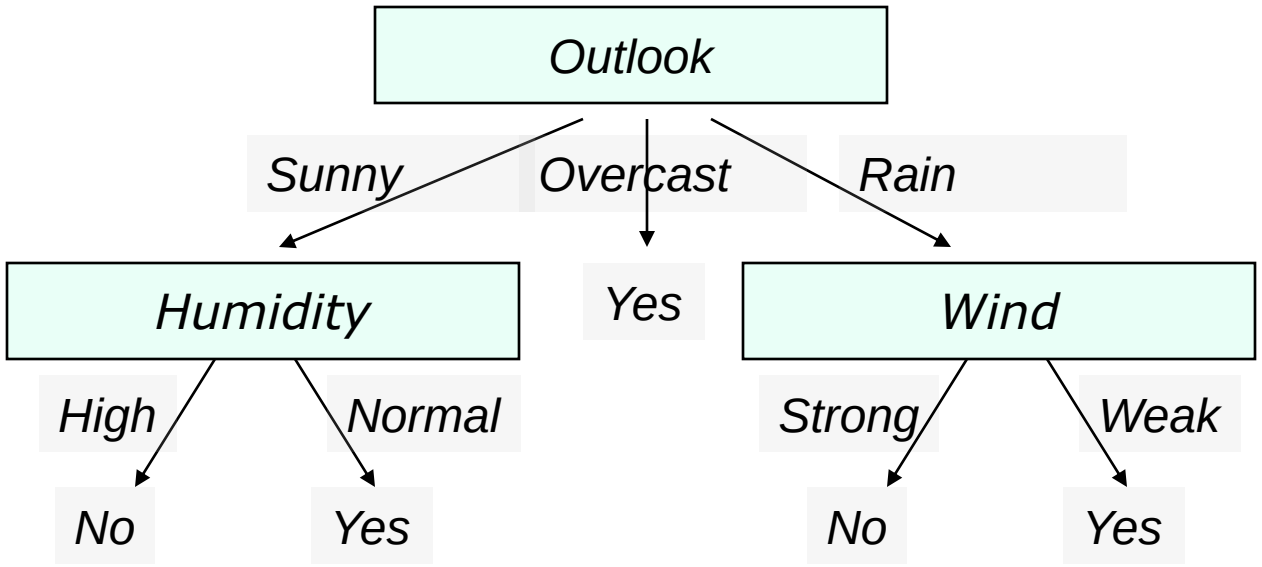
三、决策树学习常见问题

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Classify an unseen example: <*sunny*,*hot*,*normal*,*weak*>=?

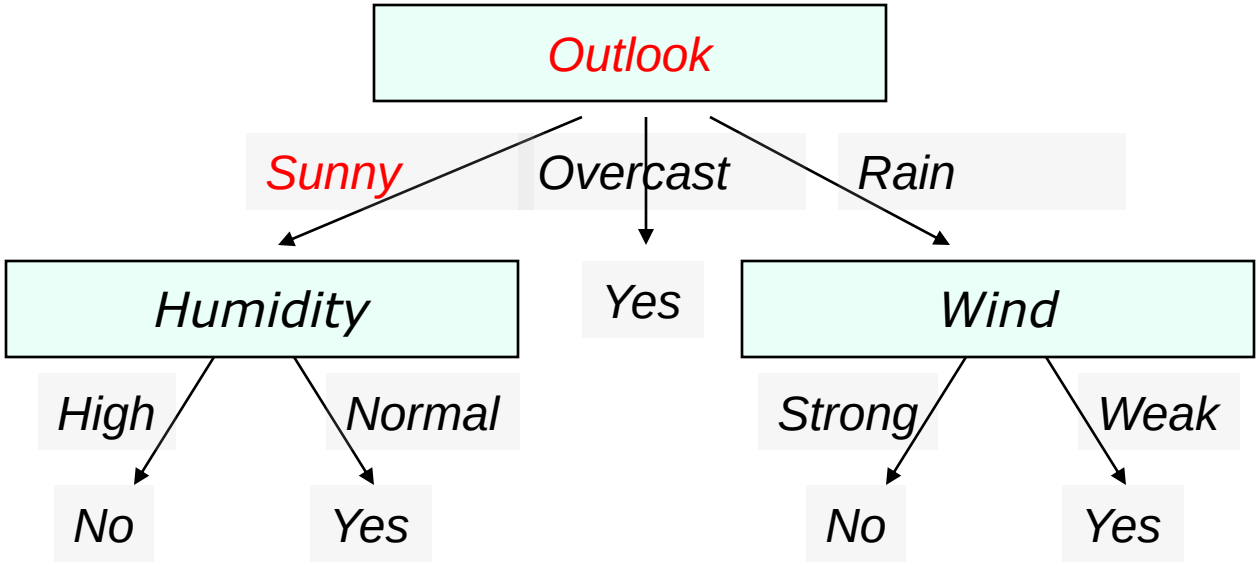
三、决策树学习常见问题

Classify an unseen example: <*sunny*,*hot*,*normal*,*weak*>=**?**



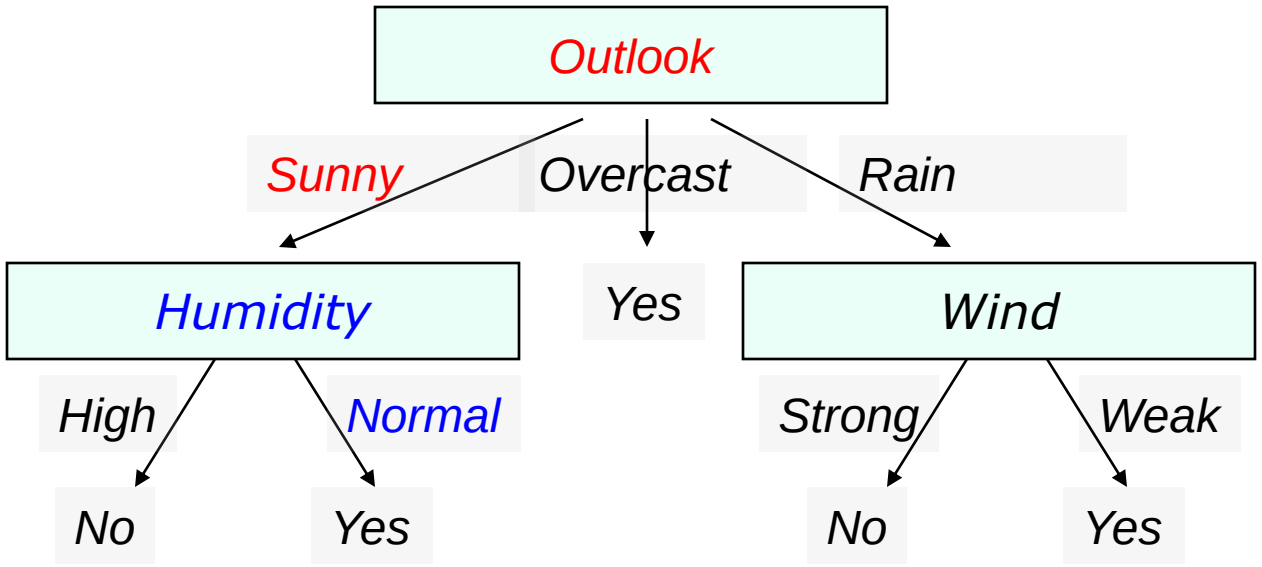
三、决策树学习常见问题

Classify an unseen example: <*sunny*,*hot*,*normal*,*weak*>=**?**



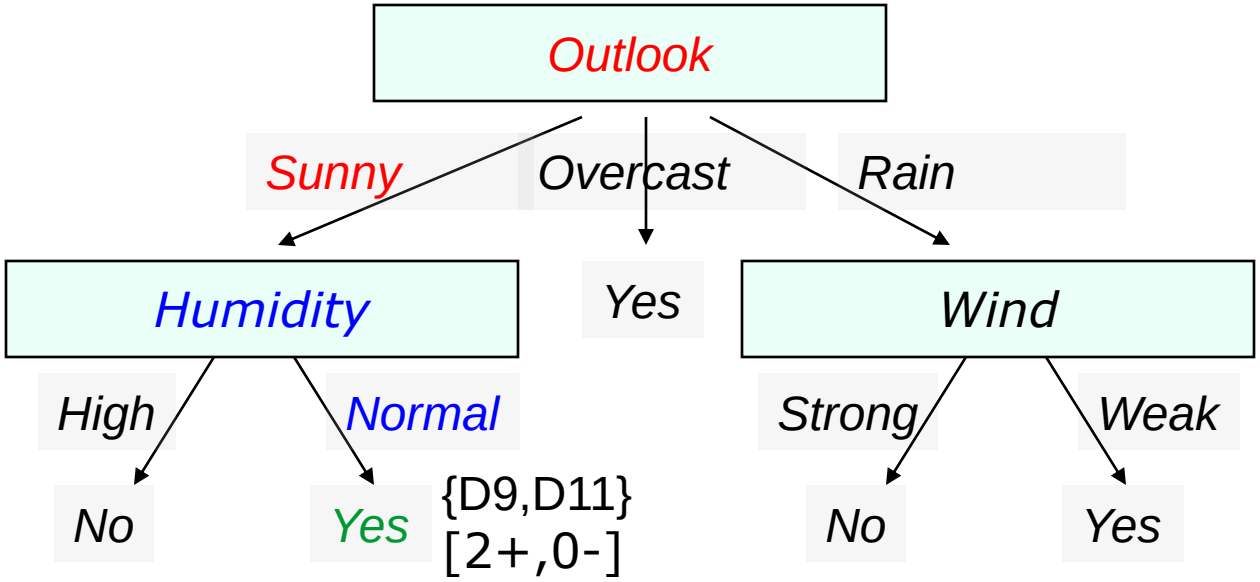
三、决策树学习常见问题

Classify an unseen example: <*sunny*,*hot*,*normal*,*weak*>=**?**



三、决策树学习常见问题

Classify an unseen example: <*sunny*,*hot*,*normal*,*weak*>=*Yes*



应用拉普拉斯估计得到待测样本x属于Yes和No的概率分别为：

$$P(Yes|x)=\frac{2+1}{2+2}=\frac{3}{4}\ ;\ P(No|x)=\frac{0+1}{2+2}=\frac{1}{4}$$

四、决策树学习理解解释

- 决策树学习是以样本为基础的归纳学习方法，它采用自顶向下的递归方式来生长决策树。随着树的生长，完成对训练样本集的不断细分，最终都被细分到了每个叶子结点上。
- 决策树的每个结点都是样本的集合，熵等度量刻画了样本集的不纯度，决策树的生长过程是一个熵降低、信息增益、从混沌到有序的过程。
- 决策树学习对噪声数据具有很好的鲁棒性，而且学习得到的决策树还能被表示为多条if-then形式的决策规则，因此具有很强的可读性和可解释性。