决策树

1. **信息论基础**

概念的引出：一个离散的随机变量 x，观察到这个变量的一个具体值的时候，我们接受到了多少信息？信息量可以被看成在学习 x 的值的时候的“惊讶程度”。一个很可能发生的事情 A 和一个相当不可能发生的事情 B，都发生了，那么收到的 B 的信息要多于 A 的信息，如果一个事情一定会发生，那么我们就不会接收到信息。因此，对于信息内容的度量依赖于概率分布 p（x），找一个函 h(x),是 p(x)的单调递减函数。h(x)还要满足：两个不相关的事件 x 和 y，观察到两个事件同时发生时获得的信息应该等于观察到事件各自发生时获得的信息之和$\mathrm{h(x,y)=h(x)+h(y)}$,而$\mathrm{p(x,y)=p(x)p(y)}$。所以$\mathrm{h(x)=-\log_2 p(x)}$，负号确保信息一定是正数或者 0，且是 p（x）的单调递减函数，2 为底是遵循信息论的普遍传统。

熵：随机变量的熵，表示一个随机变量 x 的平均信息量

$$\mathrm{H[x]=-\sum_x p(x)\log_2 p(x)}$$

例子：具体化随机变量的状态所需要的平均信息
例子：统计力学的无序程度的度量

相对熵（或 Kullback-Leibler 散度或 KL 散度）

$$\begin{aligned}\mathrm{KL}(p\parallel q) &= -\int p(x)\ln q(x)\,\mathrm{d}x-\left(-\int p(x)\ln p(x)\,\mathrm{d}x\right)\\ &= -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}\mathrm{d}x\end{aligned}$$

KL 散度大于等于 0，当且仅当 p(x)=q(x)时，等号成立
KL 散度：两个分布不相似程度的度量

p(x,y)给出两个变量 x 和 y 的数据集，考虑联合分布于边缘分布乘积之间的 KL 散度来判断它们是否相互独立
互信息（ mutual information ）：由于知道 y 值而造成 x 的不确定性的减少的量

$$\begin{aligned}I[x,y] &\equiv \mathrm{KL}(p(x,y)\parallel p(x)p(y))\\ &= -\iint p(x,y)\ln\left(\frac{p(x)p(y)}{p(x,y)}\right)\mathrm{d}x\,\mathrm{d}y\end{aligned}$$

联合熵：联合熵就是度量一个联合分布的不确定度，两个随机变量的联合熵定义为：分布为

> 分布为 $p(x,y)$ 的一对随机变量 (X,Y) ,其联合熵定义为：
> $$H(X,Y)=-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p(x,y)\log p(x,y)=E\left[\log\frac{1}{p(x,y)}\right]$$

联合熵的物理意义为，观察一个多维随机变量的随机系统获得的信息量，可以对其分解如下：

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x)p(y|x)$$

$$= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)$$

$$= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)$$

$$= H(X) + H(Y|X)$$

其中，**条件熵** $H(Y|X)$ 由 $-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)$ 所定义，其物理意义就是，在得知某一确定信息的基础上获取另外一个信息时所获得的信息量。

信息增益

信息增益就是熵-条件熵，Y 对 X 的信息增益为：H(X) − H(X|Y)，表示在已知 Y 的情况下，X 的不确定性的减少。例如，X 代表明天下雨，Y 代表明天阴天，则 Y 对 X 的信息增益就表示，已知明天阴天的情况下，明天下雨的不确定性的减少。如果 Y 对 X 的信息增益越小，说明 Y 对提供 X 的信息的贡献越小，也说明 Y 与 X 的关系越小。

基尼不纯度

从一个数据集中随机选取子项，度量其被错误的划分到其他组里的概率

在一个集合中，概率为 p(i)，标签为 i 的事件被划分到其他标签组的概率为

$$\sum\nolimits_{j\neq i} p(j) = 1 - p(i)$$

那么整个集合中，平均分错的概率为每个 i 事件的加权平均:

$$\sum_{i=1}^{m} p(i)\sum_{j\neq i} p(j) = \sum_{i=1}^{m} p(i)(1-p(i)) = 1 - \sum_{i=1}^{m} p(i)^2$$

这就是基尼不纯度的定义，它描述了一个集合的混乱程度，基尼不纯度越小，集合的纯度越高，即集合的有序程度越高，当基尼不纯度为 0 使，表示集合类别一致。在决策树中，比价基尼不纯度的大小，可以选择更好的决策条件。

2. 决策树的不同分类算法（ID3 算法、C4.5、CART 分类树）的原理及应用场景

ID3 算法

使用信息增益作为属性的选择度量。

在分类问题中，给定训练集 D，D 的熵（对 D 中元组分类所需要的期望信息）为

$$\mathrm{H(D)} = -\sum\nolimits_{i=1}^{m} p_i \log_2(p_i)$$

现在假设要按某属性 A 划分 D 中的元组，A 有 v 个不同的取值$\{\mathrm{a}_1, a_2, \dots, a_v\}$。

如果 A 的离散的，那么根据 A 的取值，可以将 D 划分为 v 个子集$\{\mathrm{D}_1, D_2, \dots, D_v\}$。则条件熵可以如下计算:

$$\mathrm{H(D|A)} = \sum\nolimits_{j=1}^{v} \frac{|D_j|}{|D|} \times H(D_j)$$

则信息增益为

$$\mathrm{Gain(A)} = \mathrm{H(D)} - \mathrm{H(D|A)}$$

选择信息增益高的属性作为分裂属性。
如果 A 为连续变量，则需要确定 A 的最佳分裂点，将 A 的取值按从小到大的顺序排列，相邻连个取值的中点$\frac{a_i+a_{i+1}}{2}$,作为候选分裂点。对于每个这样的候选分裂点，计算 $\mathrm{H(D|A)}$，每个分裂点将 D 划分为两个子集。取得最小$\mathrm{H(D|A)}$的候选分裂点作为 A 的最佳分裂点。根据最佳分裂点 split_point 将 D 分为两个子集D_1, D_2，其中D_1是满足 A≤split_point 的元组组合，其中D_2是满足 A>split_point 的元组组合。

信息增益度量偏向具有许多输出的测试，换句话说，它倾向于选择具有大量的属性。例如，如果 D 中的唯一标识符的属性 ID，ID 的划分将导致大量分区（与值一样多），每个分区只包含一个元组，由于每个分区都是纯的，所以基于该划分对数据集 D 分类所需要的信息$\mathrm{H(D|ID)} = 0$，因此得到的信增益最大。但是显然，这种划分对分类没有用。

C4.5
继 ID3 之后的 C4.5 使用一种称为增益率的信息扩充，试图克服 ID3 的偏倚。它用“分裂信息”值将信息增益规范化，分裂信息的定义如下：

$$\mathrm{SplitInfo(D|A)} = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|})$$

该值代表由训练数据集 D 划分成对应于属性 A 的 v 个输出值得 v 个分区产生的信息。
增益率定义为：

$$\mathrm{GainRate(A)} = \frac{Gain(A)}{SplitInfo(D|A)}$$

选择具有最大增益率的属性作为分裂属性。
注意：随着划分信息趋向于 0，该比例变得不稳定。为了避免这种情况，增加一个约束：选取的信息增益$\mathrm{Gain(A)}$必须较大，至少与考察的所有测试的平均增益一样大。

GART 分类树
使用基尼不纯度。集合 D 的基尼不纯度为

$$\mathrm{Gini(D)} = 1 - \sum_{i=1}^{m} p_i^2$$

对于属性 A，基尼不纯度考虑每个属性的二元划分。A 有 v 个取值$\{\mathrm{a}_1, a_2, \dots, a_v\}$，根据这 v 个取值，将 D 划分为两个分区，可以这样来看，将 A 的取值分类两个子集 $\mathrm{S_A}$和其余集。

根据每个元组中 A 的属性的取值在哪个 A 的子集，就将 D 分成两个分区。这样的$\mathrm{S_A}$有$2^{\mathrm{v}} - 2$个。
当考虑二元划分时，A 将 D 划分为D_1, D_2，给定该划分，D 的基尼不纯度为

$$\mathrm{Gini(D|A)} = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

对于每个属性，考虑每种可能的二元划分。对于离散属性，选择该属性产生最小基尼不纯度的子集作为它的分裂子集。对于连续型属性，处理类似上面。
属性 A 的二元划分导致的基尼不纯度的降低为

$$\Delta\mathrm{Gini(A)} = \mathrm{Gini(D)} - \mathrm{Gini(D|A)}$$

选择最大化不纯度降低（等价于具有最小条件基尼不纯度Gini(D|A)）的属性选为分裂属性。

对比：信息增益偏向于多值属性，尽管增益率调整了这种偏倚，但是她倾向于产生不平衡的划分，其中一个分区比其他分区小得多；基尼不纯度偏向于多值属性，并且当类的数量很大时，会有困难，它还倾向于到导致相等大小的分区和纯度。尽管如此，这些度量在实践中产生相当好的结果。

3 决策树防止过拟合的手段

树剪枝，分为先剪枝和后剪枝

先剪枝：通过提前停止树的构建（例如，通过决定在给定的结点不再）。

后剪枝：由“完全生长”的树剪去子树，通过删除节点并用树叶替换它。