决策树的连续与缺失值

前言

我们在决策树划分选择和预剪枝与四种后剪枝算法中已经把“单变量决策树”（多变量决策树需要你学完这篇文章后去看多变量决策树）的内容做到了详尽，但是我们有时候可能面对突发状况，比如***假如我们的决策事物的属性不是离散的，而是连续的（比如“下雨的可能 > 10%”），那我们该怎么办？或者当我们遇到属性值缺失该怎么办？*** 实际上这些情况***在实际样本集中都是很常见的***。我们这篇文章的目的就是为了解决这两个问题。

这部分我将照搬西瓜书大部分内容，并作注解（没办法，西瓜书这部分写得不错）

连续属性的处理

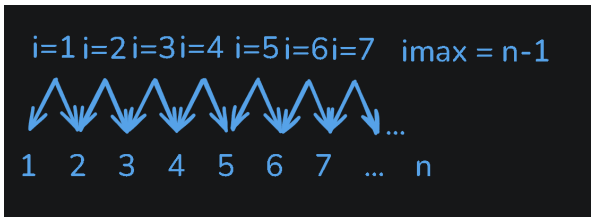
公式推导

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，*连续属性离散化技术可派上用场。最简单的策略是采用二分法（bi-partition）对连续属性进行处理，这正是 C4.5 决策树算法中采用的机制*。

给定样本集 D 和连续属性 a，假定 a 在 D 上出现了 n 个不同的取值，将这些值从小到大进行排序，记为 $a^1, a^2, \ldots, a^n$。基于划分点 t 可将 D 分为子集 D_t^- 和 D_t^+，其中 D_t^- 包含那些在属性 a 上取值不大于 t 的样本。而 D_t^+ 则包含那些在属性 a 上取值大于 t 的样本。显然，对相邻的属性取值 a^i 与 a^{i+1} 来说，t 在区间 $[a^i, a^{i+1})$ 中取任意值所产生的划分结果相同。因此，对连续属性 a，我们可考察包含 $n-1$ 个元素的候选划分点集合

$$T_a = \left\{ \frac{a^i + a^{i+1}}{2} \mid 1 \leq i \leq n-1 \right\}, (4.7)$$

关于为什么是 $n-1$ 很简单，直接看图：



这部分：***可将划分点设为该属性在训练集中出现的不大于中位点的最大值，从而使得最终决策树使用的划分点都在训练集中出现过。***

即把区间 $[a^i, a^{i+1})$ 的***中位点*** $\frac{a^i + a^{i+1}}{2}$ ***作为候选划分点（核心思想）***，然后，我们就可像离散属性值一样来考察这些划分点，选取最优的划分点进行样本集合的划分。例如，可对信息增益公式稍加改造：

$$\begin{aligned}\mathrm{Gain}(D,a) &= \max_{t\in T_a} \mathrm{Gain}(D,a,t) \\ &= \max_{t\in T_a} \mathrm{Ent}(D) - \sum_{\lambda\in\{-,+\}} \frac{|D_t^\lambda|}{|D|}\mathrm{Ent}(D_t^\lambda) \ , (4.8)\end{aligned}$$

其中 $\mathrm{Gain}(D,a,t)$ ***是样本集*** D ***基于划分点*** t ***二分后的信息增益***。于是我们就可选择使 $\mathrm{Gain}(D,a,t)$ 最大化的划分点。

相关的例子可以查看西瓜书P84-85，当然我这里也有一个：

案例（神一样的精度，别骂了，用python算的）

数据（6 个样本，属性 X 为连续值，类标签为 + 或 -）：

样本编号	1	2	3	4	5	6
X	2.5	3.0	4.0	7.0	9.0	11.0
label	+	+	−	−	−	+

先计算总体熵 $H(S)$：

总体正类 3 个，负类 3 个，因此 $p_+ = p_- = 0.5$：

$$H(S) = -0.5\log_2 0.5 - 0.5\log_2 0.5 = 1.0000000000 \text{ (bits)}.$$

排序后看相邻样本标签变化的位置：序列为 $++,-,-,-,+$。相邻不同的分界有：

- 在 3.0 与 4.0 之间（从 + 变为 -）：候选阈值 $t_1 = \dfrac{3.0+4.0}{2} = 3.5$；
- 在 9.0 与 11.0 之间（从 - 变为 +）：候选阈值 $t_2 = \dfrac{9.0+11.0}{2} = 10.0$。

对每个候选阈值计算信息增益。

1. 候选 $t_1 = 3.5$

左子集 S_L $(X \le 3.5)$：样本 $\{2.5(+),\ 3.0(+)\}$ —— 2 个样本，都是正类

$$H(S_L) = -1\cdot\log_2 1 = 0.$$

右子集 S_R $(X > 3.5)$：样本 $\{4.0(-), 7.0(-), 9.0(-), 11.0(+)\}$ —— 4 个样本，正类 1，负类 3

$$p_+^{(R)} = \frac{1}{4} = 0.25, \quad p_-^{(R)} = \frac{3}{4} = 0.75,$$
$$H(S_R) = -0.25\log_2 0.25 - 0.75\log_2 0.75 = 0.8112781244591328 \text{ (approx)}.$$

加权平均子熵：

$$\frac{|S_L|}{|S|}H(S_L) + \frac{|S_R|}{|S|}H(S_R) = \frac{2}{6}\cdot 0 + \frac{4}{6}\cdot 0.8112781244591328 = 0.5418520829694219 \text{ (approx)}.$$

信息增益：

$$\mathrm{Gain}(t_1) = H(S) - \text{加权子熵} = 1.0 - 0.5418520829694219 = 0.4581479170305781.$$

（取到若干位数后，四舍五入可写作 \approx 0.45915。）

> 注：我们给出高精度计算，实际实现中保留 4–6 位小数已足够。

2. 候选 $t_2 = 10.0$

左子集 S_L（$X \le 10.0$）：样本 $\{2.5(+), 3.0(+), 4.0(-), 7.0(-), 9.0(-)\}$ —— 5 个样本，正 2，负 3

$$p_+ = \frac{2}{5} = 0.4, \quad p_- = \frac{3}{5} = 0.6,$$
$$H(S_L) = -0.4 \log_2 0.4 - 0.6 \log_2 0.6 = 0.9709505944546686.$$

右子集 S_R（$X > 10.0$）：样本 $\{11.0(+)\}$ —— 单个正类，熵 0。

加权子熵：

$$\frac{5}{6} \cdot 0.9709505944546686 + \frac{1}{6} \cdot 0 = 0.8091254953712238.$$

信息增益：

$$\text{Gain}(t_2) = 1.0 - 0.8091254953712238 = 0.1908745046287762.$$

比较两候选增益：$\text{Gain}(t_1) \approx 0.4591$ 大于 $\text{Gain}(t_2) \approx 0.1909$，因此选择 $t_1 = 3.5$ 作为最佳二分阈值。这就是对于连续属性的处理。

缺失值的处理

这部分按照西瓜书的讲法是*直接上权重*，但是事实上我们需要考虑的有很多（~~*西瓜书只需要讲最经典的算法就好了，而作者我可要考虑的就多了*~~（

比如在统计学当中，对于缺失值的分类，或者叫缺失的方式（我是这么想的），有这么几种：***MCAR（Missing Completely At Random）：缺失与任何观测或潜在变量都无关（最“安全”）；MAR（Missing At Random）：缺失与观测到的变量有关，但与潜在的未观测变量无关；MNAR（Missing Not At Random）：缺失与未观测信息（如变量自身真实值）有关（处理最困难）***。合理区分这三类很重要，因为不同假设下处理方式效果差异大。

那我们来看看究竟怎么处理这些“不老实的样本”：

第一种，直接扔掉，忽略这一整个样本，很显然，直接删掉会直接导致训练集数量减少（尤其当缺失不是 MCAR 时）

第二种，见缝插针版。你说少样本？简单，那就加啊。你问我怎么加？这么加：

a. *简单插补：连续用均值/中位数，类别用众数。*

b. *回归/分类插补：用其它属性建立模型预测缺失值。*

c. *kNN 插补、MICE（多重插补）等更复杂方法。*

值得注意的是：如果你插补用了标签信息或未来信息，会导致数据泄露；同时插补忽略了插补不确定性（MICE 多重插补可缓解）。当然这些方法不在本篇文章的讨论范围，我们讲解的目标是下面的内容

第三种，直接在决策树内建处理。怎么办？就是我上边提到的加权重，我会在下文详细介绍。

公式推导

首先我们需要搞清楚三件事，当我们缺失值时，会影响什么？最直观的就是会***影响存在样本和总样本的比值（1）***。除了这些呢？是否还记得存在类别这一说法？也就是说还会***影响存在样本中某一类数量的占比（2） 以及 影响存在样本中某一类的某个值数量的占比（3）。***很抽象的描述，没关系我们举个例子（作为参考，不是特别准确，但是帮助你们理解足够了）：

现在存在一个总样本 Bob 这个人的数据集 D，现在 Bob 脱下了衣服和鞋子，剩下了存在样本子集 $\tilde{D}$，这个时候（1）就是 $\frac{\tilde{D}}{D}$；而存在样本子集 $\tilde{D}$ 中又进行了分类，有“项链”（a^1）、“戒指”（a^2）、“手镯”（$a^3 ... a^n, i \subset n$）等，每一个类别的个数与存在样本子集 $\tilde{D}$ 的比值就是 $\frac{a^i}{\tilde{D}}$ (2)；而对于“手镯”（a^3），我们又可细分为“金手镯”（b^1）、“银手镯”（$b^2, j \subset \{1,2\}$），所以单个值所存在的个数与存在样本子集 $\tilde{D}$ 的占比就是 $\frac{b^j}{\tilde{D}}$

当我们搞清楚这个之后，我们就可以上西瓜书的公式了：

给定训练集 D 和属性 a，令 $\tilde{D}$ 表示 D 中在属性 a 上没有缺失值的样本子集。显然我们仅可根据 $\tilde{D}$ 来判断属性 a 的优劣. 假定属性 a 有 V 个可取值 $\{a^1, a^2, \ldots, a^V\}$，令 $\tilde{D}^v$ 表示 $\tilde{D}$ 中在属性 a 上取值为 a^v 的样本子集，$\tilde{D}_k$ 表示 $\tilde{D}$ 中属于第 k 类 $(k = 1, 2, \ldots, |\mathcal{Y}|)$ 的样本子集，则显然有 $\tilde{D} = \bigcup_{k=1}^{|\mathcal{Y}|} \tilde{D}_k$，$\tilde{D} = \bigcup_{v=1}^{V} \tilde{D}^v$。假定我们为每个样本 x 赋予一个权重 w_x，并定义

$$\rho = \frac{\sum_{x \in \tilde{D}} w_x}{\sum_{x \in D} w_x}, \quad (4.9)$$

$$\tilde{p}_k = \frac{\sum_{x \in \tilde{D}_k} w_x}{\sum_{x \in \tilde{D}} w_x} \quad (1 \leq k \leq |\mathcal{Y}|), \quad (4.10)$$

$$\tilde{r}_v = \frac{\sum_{x \in \tilde{D}^v} w_x}{\sum_{x \in \tilde{D}} w_x} \quad (1 \leq v \leq V). \quad (4.11)$$

这里，(4.9)对应(1)，(4.10)对应(2)，(4.11)对应(3)，并且 w_x 的初始值为 1。~~*周志华老师以为我们都会呢，仅给出了一点解释*~~，下图：

> 直观地看, 对属性 a, ρ 表示无缺失值样本所占的比例, $\tilde{p}_k$ 表示无缺失值样本中第 k 类所占的比例, $\tilde{r}_v$ 则表示无缺失值样本中在属性 a 上取值 a^v 的样本所占的比例. 显然, $\sum_{k=1}^{|\mathcal{Y}|} \tilde{p}_k = 1$, $\sum_{v=1}^{V} \tilde{r}_v = 1$.

当我们搞清楚了我们的这三个公式，我们就可以无脑带入了，因为我们计算的是权重，所以直接乘原始公式。将信息增益的计算式推广为：

$$\begin{aligned}\operatorname{Gain}(D, a) &= \rho \times \operatorname{Gain}(\tilde{D}, a) \\ &= \rho \times \left(\operatorname{Ent}\left(\tilde{D}\right) - \sum_{v=1}^{V} \tilde{r}_v \operatorname{Ent}\left(\tilde{D}^v\right) \right) \end{aligned}, (4.12)$$

其中由信息熵公式得到：

$$\operatorname{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k \rightarrow \operatorname{Ent}(\tilde{D}) = -\sum_{k=1}^{|\mathcal{Y}|} \tilde{p_k} \log_2 \tilde{p_k}$$

现在我们使用权重已经解决了缺失值问题，但是话又说回来，我们怎么划分呢？书上是这么讲的：***对于缺失的值，我们需要将其在划分后引入每一个子叶当中，并将权重修改为当前子节点的个数与本类的总存在样本个数比值***

大概在这个位置：

“纹理”在所有属性中取得了最大的信息增益, 被用于对根结点进行划分. 划分结果是使编号为 $\{1,2,3,4,5,6,15\}$ 的样本进入“纹理= 清晰”分支, 编号为 $\{7,9,13,14,17\}$ 的样本进入“纹理=稍糊”分支, 而编号为 $\{11,12,16\}$ 的样本进入“纹理=模糊”分支, 且样本在各子结点中的权重保持为 1. 需注意的是, 编号为 $\{8\}$ 的样本在属性“纹理”上出现了缺失值, 因此它将同时进入三个分支中, 但权重在三个子结点中分别调整为 $\frac{7}{15}$、$\frac{5}{15}$ 和 $\frac{3}{15}$. 编号为 $\{10\}$ 的样本有类似划分结果.

上述结点划分过程递归执行, 最终生成的决策树如图 4.9 所示.

案例

这部分的案例详见西瓜书P87-88。太懒了，我就不编例子了（