二元分类树

《递归划分》第四章报告

任宣霏

数学学院

2022 年 10 月 16 日

目录

结点不纯度

我们之前用熵值来刻画结点的不纯度，并作为划分依据。现在我们来讨论更抽象的情况：
我们用 $\mathbb{P}\{Y=1|\tau\}$ 表示结点中正例的概率。那么不纯度可以表示为它的一个函数（一般是凹函数），满足以下性质：

- $\phi \geq 0$;
- 对任意 $p \in (0,1)$,$\phi(p) = \phi(1-p)$ 且 $\phi(0) = \phi(1) < \phi(p)$.

结点不纯度

几个可能的选择。

- 贝叶斯误差，即最小误差：

$$\phi(p) = min(p, 1-p). \tag{1}$$

- 熵函数：

$$\phi(p) = -p\log(p) - (1-p)\log(1-p). \tag{2}$$

- 基尼指数：

$$\phi(p) = p(1-p). \tag{3}$$

其中，贝叶斯误差和基尼指数都存在一些问题，所以后面所述的不纯度都指的是熵准则。

终端结点的确定

在上一节中，我们已经可以通过熵值得到一棵较充分划分的树。
但有些划分可能导致过拟合，我们通过剪枝步骤选取这棵树的最优子树作为最终结果。
剪枝则需要在这棵充分划分的树中确定终端结点。

终端结点的确定

递归划分的目的是提取样本中的同质子群，我们是否达到这个目标取决于终端结点是否确实是同质的。因此我们对一棵树进行如下定义：

定义 (刻画树好坏程度的指标：)

$$R(\mathcal{T}) = \sum_{\tau \in \tilde{\mathcal{T}}} \mathbb{P}\{\tau_L\} r(\tau).$$

这里 $\tilde{\mathcal{T}}$ 是 $\mathcal{T}$ 终端结点的集合， $r(\tau)$ 用来衡量结点 τ 的分类好坏，即结点 τ 的分类质量。
本节主要研究，得到一颗充分划分的树以后，如何选取终端结点，即在何处剪枝的到最终树。而剪枝的目的是选择使得 $R(\mathcal{T})$ 最小的最优子树 $\mathcal{T}^*$.

误分类成本

首先我们需要定义以上的 $r(\tau)$. 显然我们可以使用结点不纯度来定义 $r(\tau)$, 但通常会选为误分类成本。

问题

如果想要一棵误分类成本最小，而不是结点不纯度最小的树，那么在划分时为何以结点不纯度（即熵值）为指标，而不选用误分类成本?

- 在任何树生长之前，难以分配成本函数;
- 经验证据表明，使用熵不纯度来划分，通常只需要合理的样本量，就能得到有用的树。

我们用熵值来对于树进行较为充分的划分，并通过误分类成本来选取终端结点，即剪枝。

误分类成本

把正常婴儿分类为早产儿会使这个孩子受到不必要的照顾，浪费资源；把早产儿分类为正常婴儿可能使他得不到必要的特殊照顾。
因此，假阳性和假阴性错误的代价是不同的，在大多数应用中，假阴性错误比假阳性更严重，因此我们有必要考虑误分类成本以权衡错误的严重程度，并以此来衡量一棵树的划分效果。
记 $c(i|j)$ 为第 j 类个体被归为 i 类的一个单位误分类成本。显然 $c(i|i)$ 应为 0.
不失一般性，假阳性错误成本 $c(1|0)$ 取 1，而 $c(0|1)$ 的相对成本需要临床医生和统计人员一起评估。

误分类成本

一个结点被分类为第 1 类还是第 0 类，取决于假阳性错误的成本是否低于假阴性错误的成本。
形式上，若

$$\sum_i [c(j|i)\mathbb{P}\{Y = i|\tau\}] \leq \sum_i [c(1-j|i)\mathbb{P}\{Y = i|\tau\}], \tag{4}$$

则结点被分配到第 j 类。
例如，一个终端结点里有 3656 个阴性和 205 个阳性，假阴性成本 $c(0|1)$ 取 10. 则比较成本：

$$3656 \times 1 = 3656 > 2050 = 205 \times 10 \tag{5}$$

所以把这个结点分为类 0 的成本小于分类为第 1 类的成本，把他分为第 0 类，即作出阴性的判断。

误分类成本

下面是一些定义和记号：

定义

- 结点内（误分类）成本，或称条件误分类成本：
  $r(\tau) = \sum_i [c(j|i)\mathbb{P}\{Y = i|\tau\}]$,
- 权重 $\mathbb{P}\{\tau\}$
- 非条件误分类成本 $R(\tau) = \mathbb{P}\{\tau\}r(\tau)$,
- $R^s(\tau)$ 表示结点 τ 的误分类成本的回代估计。

注记

误分类成本的回代估计往往过于乐观，因此需要测试集，交叉验证。

成本复杂度

在误分类成本的基础上，定义树的复杂度，事实上是对误分类成本的修正，对于大树给予惩罚。

定义 (成本复杂度)

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

我们可以据此计算一棵树和它的各个子树的成本复杂度。下面定理保证了不会有两个成本复杂度相同的最小子树。

定理

对于任意复杂度参数 α，存在 T_0 唯一最小子树，使其成本复杂度最小。

成本复杂度

关于成本复杂度，我们注意到以下几点：

- 一般情况下，对应于 $\alpha = 0$ 的最优子树不太可能是初始树。
- 复杂度相同时，最小的树更优。
- 并非所有子树都有对应的复杂度参数，使其最优。
- 一个最优子树对应于一个复杂度参数的区间，在其范围内都是最优的。

例

这是一棵未充分剪枝的树。
结点内的数字表示标号和误分类成本的单位。
结点旁边的数字分别表示结点内的非正常个体数（阳性）和正常个体（阴性）数。

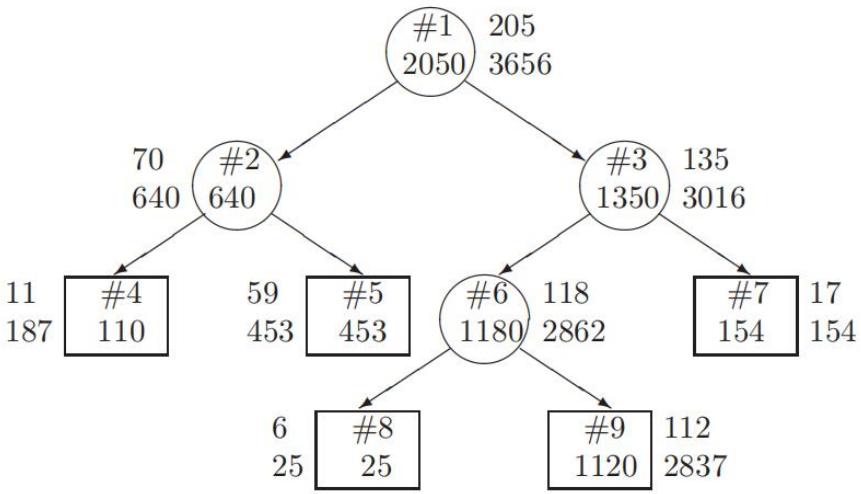

图: 原始树

嵌套最优子树

对于每一个结点，我们可以计算出其回代误分类成本。
对于某个特定的中间结点（包括根结点），可以比较其回代误分类成本，与后代终端结点的回代误分类成本总和。
通过选取 α，使得剪枝（减掉其子节点，让它变成根节点）与不剪枝（令其仍为中间结点）的成本复杂度相同。容易看出，此时 α 更大时，大树复杂度高，剪枝更好； α 更小时，小树复杂度高，应该保留。

例

例如在上述树中，考虑结点 3. 结点 3 的误分类成本为 $R^s(3) = 1350/3861 = 0.350$，它的后代误分类成本之和为 $R^s(\tilde{T}_3) = (154 + 25 + 1120)/3861 = 0.336$.
考虑成本复杂度，结点 3 和后代之间的复杂度之差为 $3 - 1 = 2$. 引入 α 使得 $R_\alpha(3) = R_\alpha(\tilde{T}_3)$ 得到的 $\alpha = 0.007$

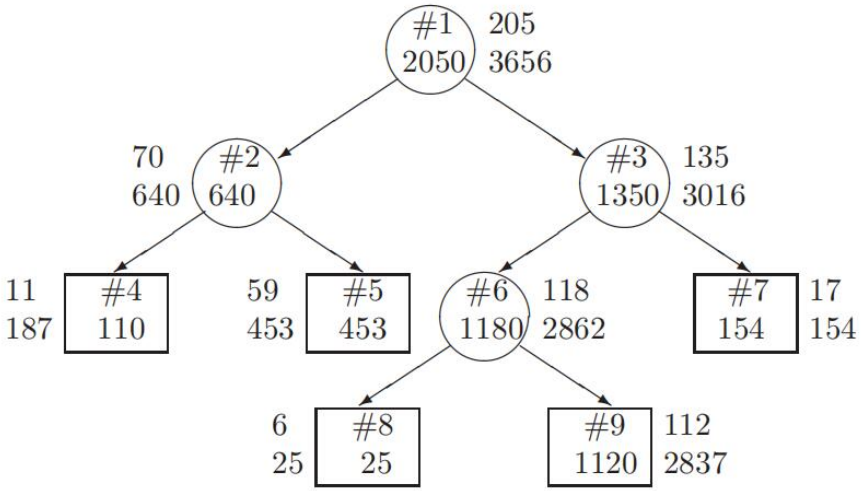

图: 原始树

例

每个中间结点都面临剪枝选择，对于每个中间结点都进行这样的计算。得到成本复杂度如下表：

Node	$R^s(\tau)$	$R^s(\tilde{T}_\tau)$	$\lvert\tilde{T}_\tau\rvert$	α
9	0.290	0.290	1	
8	0.006	0.006	1	
7	0.040	0.040	1	
6	0.306	0.296	2	0.010
5	0.117	0.117	1	
4	0.028	0.028	1	
3	0.350	0.336	3	0.007
2	0.166	0.145	2	0.021
1	0.531	0.481	5	0.013
		Minimum		0.007

图: 成本复杂度

例

α 的值使新旧子树具有相同的成本复杂度。实际上，对于任何内部节点， $\mathcal{T}$ ，α 的值正好为

$$\frac{R^s(\mathcal{T}_\tau) - R^s(\tilde{\mathcal{T}}_\tau)}{|\tilde{\mathcal{T}}_\tau| - 1}. \tag{6}$$

选取最小的正阈值参数 $\alpha_1 = 0.007$ ，相对应的结点 3 优于其终端结点之和，所以我们减掉 3 的子节点，得到优化子树。
除非一次修剪之后，我们的树变成了只有根节点的树，否则，与寻找第一个正阈值参数 α_1 相同，我们继续寻找第二个复杂度参数的值 $\alpha_2 = 0.018$ ，此时修剪后的子树为根节点自己。

嵌套最优子树

一般来说，假设我们找到了 m 个阈值之后，结束了寻找：

$$0 < \alpha_1 < \alpha_2 < \cdots < \alpha_m \tag{7}$$

并且令 $\alpha_0 = 0$，对应的最优子树为：

$$\mathcal{T}_{\alpha_0} \supset \mathcal{T}_{\alpha_1} \supset \mathcal{T}_{\alpha_2} \supset \cdots \mathcal{T}_{\alpha_m}, \tag{8}$$

这就是所谓的嵌套最优子树。同时也证明了以下结论：

定理 (嵌套最优子树)

若 $\alpha_1 > \alpha_2$,α_1 对应的最优子树是 α_2 对应的最优子树的子树。

嵌套最优子树

当有测试样本时，我们把子树应用到测试样本中，去估计误分类成本 $R(T)$，最终选择误分类成本最小的一个。
如果没有测试样本，用交叉验证法。
对于已经确定的 $\{\alpha_k\}_0^m$ 和对应的最优子树 $\{T_{\alpha_k}\}$，我们把样本划成训练集和测试集。用训练集训练出的树，对于每个 α_k，都有一个唯一的最优子树。
再把测试集的样本代入这个最优子树中，得到其误分类成本 $R^{ts}(T_k)$ 作为 $R(T_{\alpha_k})$ 的无偏估计。

R^{cv^*} 的标准差

我们当然可以选取对应于最小 R^{cv} 的子树，但考虑到交叉验证法中估计过程的不确定性和构造简单树结构的目标，可提出一个修改后的策略。

$R^{cv}(T_{\alpha_k})$ 的启发式标准误差（heuristic standard error）为：

$$SE_k = \{\sum_{j=0,1}(\frac{\mathbb{P}\{Y=j\}}{n_j})^2(\sum_{i\in S_j} C_{i,k}^2 - n_j\bar{C}_{k|j}^2)\}^{1/2} \tag{9}$$

其中， $C_{i,k}$ 是第 i 个个体在第 k 个子树上，作为测试个体时的误分类成本。

$$\bar{C}_{k|j}^2 = \frac{1}{n_j^2}\sum_{i\in S_j} C_{i,k} \tag{10}$$

修改后策略选择交叉验证估计值在 $R^{cv}(T_{\alpha_k})$ 的最小值 $R^{cv}(T_{\alpha_{k^*}})$ 的预定范围内的最小子树，通常是一个标准差单位 SE_{k^*} .

替代剪枝的方法

程序自动产生的树结构不一定是我们想要的，真实情况下要更多考虑变量的可解释性。所以还有一些剪枝的替代方法和局部交叉验证调整偏差的方法。
我们可以为每个内部结点 τ 分配一个统计量 S_τ，然后按照递增的顺序排列这些统计量：

$$S_{\tau_1} \leq S_{\tau_2} \leq \cdots S_{\tau_n} \tag{11}$$

之后，选择一个阈值，如果内部结点所对应的统计量小于阈值所对应的，则将内部结点更改为终端结点。

比较树和逻辑斯蒂回归

应用树方法和逻辑斯蒂回归方法后，都可以画出 ROC 曲线图。从曲线可以看出，还需要做很多改进来提高我们的预测效果。

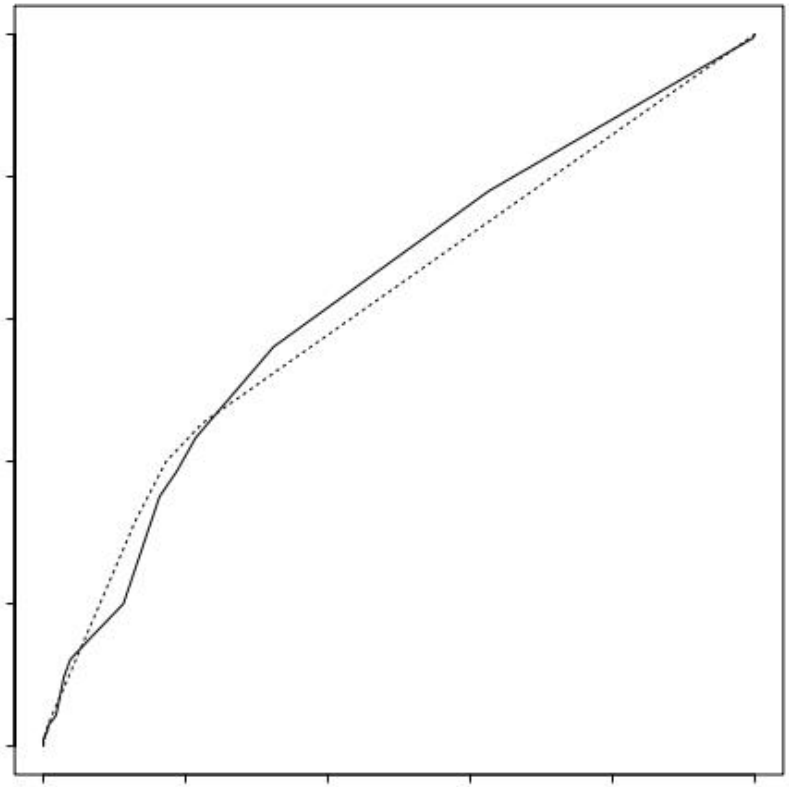

图: 成本复杂度

综合树和逻辑斯蒂回归的方法

下面是综合使用树和逻辑斯蒂回归的两种方式：

- 先逻辑斯蒂回归，得到的线性方程作为一个新的预测变量，加入到原有的变量中，来构建分类树；
- 先使用树方法，对于每一个得到的终端结点，可以诱导产生一个新的变量，纳入到逻辑斯蒂回归方程中。这个新的变量具体为一个示性函数，为 1 表示样本满足落入此终端结点的条件。

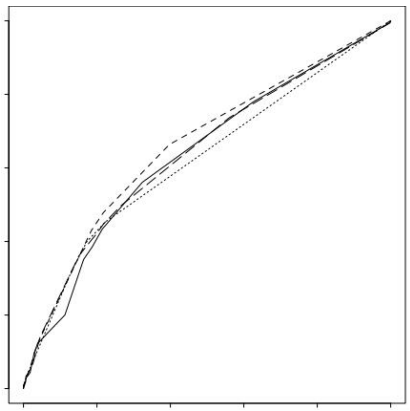

图: 各种方法比较

缺失数据

本书介绍处理缺失数据的三种方法，会在后面详述：

- 联合缺失法；
- 代理分裂；
- 填补缺失值。

缺失值处理

还有一种较为简单的办法是对于某一特定的特征 a，在计算信息增益时可以考虑在全部数据 D 中，该特征上未缺失的所有数据 $\tilde{D}$，得到信息增益 $Gain(\tilde{D}, a)$，然后定义信息增益为：

$$Gain(D, a) = \rho \times Gain(\tilde{D}, a) \tag{12}$$

其中，$\rho = \frac{\sum_{x \in \tilde{D}} w_x}{\sum_{x \in D} w_x}$，$w_x$ 为 x 的权重，在根结点中所有样本权重定义为 1.

我们用新定义的允许缺失值的信息增益来作划分的准则，并将有缺失值的结点按比例分配到两个子节点中。

其他处理方法和无缺失完全相同。

联合缺失法

联合缺失法的核心是通过简单的处理把数据化成没有缺失值的样子，以便仍然可以使用针对无缺失数据的递归划分方法。

- 对于离散（定性）的变量 x_j，若样本中有缺失值，则将缺失值 NA 视为 x_j 的一个水平，进行划分。
- 对于连续（有序）的变量 x_j，我们把它复制两个不含缺失值变量 $x_j^{(1)}$ 和 $x_j^{(2)}$，二者缺失值分别用 ∞ 和 $-\infty$ 代替。将这两个变量纳入递归划分的算法中，可以看出最终若按 x_j 这一指标划分，所有含缺失值个体都被划分到同一个子节点中。

注记

上述的 ∞ 在实际操作中可以任意选择一个比所有样本都大（小）的数来代替。

代理分裂

问题依旧是含缺失值的数据如何划分。这次我们讨论代理分裂的方法。
代理分裂利用其他预测因子中的信息，来帮助我们做出如何分裂的决定。
假如我们要对“种族”变量进行划分，那么要寻找与其最相似的预测因子。
两个因子的相似程度定义为：将同一个主体分配给同一个结点的概率。

代理分裂

比如我们来看“种族”和“年龄”的相似性。

	Black	Others
Age ≤ 35	702	8
Age > 35	3017	134

图: 相似性

3861 名受试者中的 $702 + 134 = 836$ 名被分配到相同的结点，因此 $836/3861 = 0.217$ 为分裂相似性，也叫做这两个分裂置信概率的估计。

代理分裂

一般情况下，如果受试者不是从一般人群中随机抽取，例如对照研究，则应将先验信息纳入置信概率的估计中。
我们有先验信息患病率 $\mathbb{P}\{Y=1\}$ 和相对应的 $\mathbb{P}\{Y=0\}$，则纳入先验信息的置信概率为样本中正例分配给同一结点的概率和负例分配给同一结点的概率关于先验信息患病率的加权平均，即：

$$\mathbb{P}\{Y=0\}M_0(\tau)/N_0(\tau)+\mathbb{P}\{Y=1\}M_1(\tau)/N_1(\tau) \qquad (13)$$

定义 (最优代理分裂)

对于任意分裂 s^*，若基于不同的预测因子时，在所有可行的分裂中 s' 与 s^* 产生的置信概率最大，则称分裂 s' 是 s^* 的最优代理分裂。

代理分裂

评价：

- 因最佳分裂点数据缺失找到的预测因子也可能面临缺失值问题，因此需要继续找这个预测因子的预测因子，即次代理分裂；
- 用计算机不难解决，但对于分析人员而言，需要对数据进行全面了解，分析代理分裂可能包含的有用信息，工作量大；
- 由于计算机空间限制，代理分裂很少在文献中发表，实用性也受到实际情况的限制；
- 如果对特定分裂，代理分裂无法提高预测能力，应该被舍弃；
- 对最佳代理分裂的全面研究可能发现其他重要的预测因素，虽然这些因素可能未在最终树结构中出现。还可能提供替代的树结构，原则上误分类成本更低。