树的构建与逻辑斯谛回归

《递归划分》前三章读书报告

任宣霏

数学学院

2022 年 10 月 10 日

目录

导论

Question

为什么要研究递归划分?

参数回归，比如线性回归，逻辑斯蒂回归等对于模型假设的要求比较高，当基本的模型假设不成立时，不能很好地刻画变量之间的关系。
常见的诊断方法，例如残差图，随着模型复杂度上升，方法会变得高度复杂。所以我们要研究非参数回归。
递归划分包括两类非参数回归方法：基础分类/回归树（CART）和多元自适应性样条回归（MARS），这次我会重点介绍前者。

应用实例

CART 方法

递归划分广泛应用在生物、物理和社会科学等领域。

- 胸痛
  根据临床指标将患者分成相对相似的小组，帮助医生准备药物和治疗方案。
- 昏迷
  根据年龄、性别、言语反应等指标，预测患者的结果。
- 哺乳动物的精子、婴儿高烧、妊娠结果、头部损伤、基因表达、市场营销与管理、化学成分、音乐音频……

统计问题

以上所有例子都可以被总结为一个统计问题：
解释变量 $x_1, \cdots, x_p$ 对于因变量 Y 的解释。
在数学上，我们希望用 x 的值预测 Y 的值。希望估计条件概率

$$\mathbb{P}\{Y = y | x_1, \cdots, x_p\} \tag{1}$$

或者这个概率的函数，例如条件期望：

$$\mathbb{E}\{Y | x_1, \cdots, x_p\} \tag{2}$$

这个问题的参数模型研究比较多，如果 Y 是连续值可以采用回归分析中线性回归等方法，如果 Y 是离散的分类变量，可以采用逻辑斯蒂回归法。
后面讨论的大多是分类与决策问题，所以我们希望比较逻辑斯蒂回归和非参数方法（例如 CART）的效果。

逻辑斯蒂回归

这是一个我之前比较陌生的模型，所以我想重点介绍。把关于构建树的内容放到后面。
逻辑斯蒂回归（Logistic Regression），又叫对数几率回归，是分析二元型数据的一个标准方法。
对于样本 i，我们假设其因变量 Y_i 服从伯努利分布，即

$$\mathbb{P}\{Y_i = y_i\} = \theta_i^{y_i}(1-\theta_i)^{1-y_i} \tag{3}$$

便可以用分对数 logit 连接函数估计概率 θ_i，即

$$\theta_i = \frac{exp(\beta_0 + \Sigma_{j=1}^{p}\beta_j x_{ij})}{1 + exp(\beta_0 + \Sigma_{j=1}^{p}\beta_j x_{ij})} \tag{4}$$

问题引出

在进一步叙述逻辑斯蒂回归的原理之前，我想先引出本次重点讨论的一个问题。后续以这个实例来讲解各种研究方法。
这个问题我们感兴趣的因变量是孕妇“是否早产”，数据来自耶鲁大学一研究，共有 3861 名孕妇。目前有 15 个候选自变量：

Variable name	Label	Type	Range/levels
Maternal age	x_1	Continuous	13–46
Marital status	x_2	Nominal	Currently married, divorced, separated, widowed, never married
Race	x_3	Nominal	White, Black, Hispanic, Asian, others
Marijuana use	x_4	Nominal	Yes, no
Times of using marijuana	x_5	Ordinal	>= 5, 3–4, 2, 1 (daily) 4–6, 1–3 (weekly) 2–3, 1, < 1 (monthly)
Years of education	x_6	Continuous	4–27
Employment	x_7	Nominal	Yes, no
Smoker	x_8	Nominal	Yes, no
Cigarettes smoked	x_9	Continuous	0–66
Passive smoking	x_{10}	Nominal	Yes, no
Gravidity	x_{11}	Ordinal	1–10
Hormones/DES used by mother	x_{12}	Nominal	None, hormones, DES, both, uncertain
Alcohol (oz/day)	x_{13}	Ordinal	0–3
Caffeine (mg)	x_{14}	Continuous	12.6–1273
Parity	x_{15}	Ordinal	0–7

图: A List of Candidate Predictor Variables

原理叙述

其中 y_i 的取值为 0 或 1， x_i 有离散或连续值。对于离散的自变量值，我们将其拆成多个 $0/1$ 自变量。
比如 x_2 “婚姻状况”可以取“已婚”“离婚”“分居”“丧偶”“未婚”，我们取示性函数

$$x_{21} = \mathbb{1}_{married},$$

即已婚是 1，未婚是 0. 其他也是同样操作。

原理叙述

这时，我们希望用线性模型

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \tag{5}$$

来预测 y 值，但由于是分类问题，我们更希望得到一个 0/1 值来预测结果。
一个尝试是令 $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.
单位阶跃函数

$$y = \begin{cases} 0 & z<0, \\ 0.5 & z=0, \\ 1 & z>0. \end{cases}$$

我们希望找到一个类似于单位阶跃函数的连续函数，对数几率函数

$$y = \frac{1}{1+e^{-z}} \tag{6}$$

正是这样一个好的替代。

原理叙述

反解出

$$z = ln\frac{y}{1-y} \tag{7}$$

于是模型可化为

$$ln\frac{y}{1-y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{8}$$

但这个式子不能当作一般的线性模型来求解，因为 $y = 0/1$ 时左边没有意义。
把预测的结果值 y 看作样本被判定为正例的概率，所以我们有：

$$p_1 = \mathbb{P}(y=1|x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}, \tag{9}$$

$$p_0 = \mathbb{P}(y=0|x) = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}, \tag{10}$$

原理叙述

两式综合得到

$$\mathbb{P}(Y=y|x)=yp_1+(1-y)p_0. \tag{11}$$

最大化

$$l(x,y|b)=\Sigma_{i=1}^{p} lnp_i \tag{12}$$

等价于最小化

$$l=\Sigma_{i=1}^{p}(-y_i(\beta_0+\beta_1x_1+\cdots+\beta_px_p)+ln(1+e^{\beta_0+\beta_1x_1+\cdots+\beta_px_p})). \tag{13}$$

这个是关于参数的高阶可导连续凸函数，可以用经典数值优化算法，比如梯度下降法、牛顿法求解。

参数解释

定义

上述因变量 y，即最初伯努利模型中的概率 θ 解释为判定为正例的概率，那么第 i 个样本是异常情况的优势 (Odds) 定义为：

$$\frac{\theta_i}{1-\theta_i} = exp(\beta_0 + \Sigma_{j=1}^{p} \beta_j x_{ij}) \tag{14}$$

考虑两个独立的个体 i 和 k，$x_{i1} = 1$，$x_{k1} = 0$，其他协变量都相同，则个体 i 和 k 的优势比（Odds ratio）为

$$\frac{\theta_i/(1-\theta_i)}{\theta_k/(1-\theta_k)} = exp(\beta_1) \tag{15}$$

取对数，可以看出 β_1 就是两个个体只有第一个协变量有一个单位差距，其他协变量均相同时的优势比的 log 值。

实例

我们用逻辑斯蒂回归分析耶鲁妊娠结果数据，运用逐步向后回归选择显著的变量。
这里介绍一些具体的细节性问题。
初选模型如下：

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	−2.172	0.6912	0.0017
x_1(age)	1	0.046	0.0218	0.0356
z_6(Black)	1	0.771	0.2296	0.0008
x_6(educ.)	1	−0.159	0.0501	0.0015
z_{10}(horm.)	1	1.794	0.5744	0.0018

图: MLE for an Initially Selected Model

实例

由于含缺失值的数据不参与回归，构建这个模型实际时在 3861 个样本中有 1797 个未被使用。
但我们发现，如果不考虑 x_7 (就业) 和 x_8 (吸烟) 这两个变量，只有 24 个带缺失值的样本被移除，样本信息被更充分利用。同时，从初步回归结果中可以看出这两个变量都不够显著，所以我们不考虑它们。不考虑这两个变量的回归结果为：

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	−2.334	0.4583	0.0001
x_6(educ.)	1	−0.076	0.0313	0.0151
z_6(Black)	1	0.705	0.1688	0.0001
x_{11}(grav.)	1	0.114	0.0466	0.0142
z_{10}(horm.)	1	1.535	0.4999	0.0021

图: MLE for a Revised Model

实例

观察上述结果，我们发现有一些二级变量，例如 z_6 (黑人)，我们更希望纳入相对应的 x_3 人种这个变量。
所以我们选择在逐步向后回归的过程中，将其原本变量考虑进来，但回归发现加进来的变量显著水平都达不到 0.05，因此筛选得到的变量依然如上。
以上是去掉了 24 个有缺失值的样本得到的结果。如果我们只考虑这四个最终筛选出来的变量，则只需去掉 3 个关于这四个变量有缺失的样本。结果如下：

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	−2.344	0.4584	0.0001
x_6(educ.)	1	−0.076	0.0313	0.0156
z_6(Black)	1	0.699	0.1688	0.0001
x_{11}(grav.)	1	0.115	0.0466	0.0137
z_{10} (horm.)	1	1.539	0.4999	0.0021

图: MLE for the Final Model

实例

观察系数可以做出一些推断，比如黑种人 (z_6) 的早产优势是其他的两倍，因为优势比 $exp(0.699) \approx 2.013$
同样可以说明其他几个变量的影响。

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	−2.344	0.4584	0.0001
x_6(educ.)	1	−0.076	0.0313	0.0156
z_6(Black)	1	0.699	0.1688	0.0001
x_{11}(grav.)	1	0.115	0.0466	0.0137
z_{10} (horm.)	1	1.539	0.4999	0.0021

图: MLE for the Final Model

基于这个结果，我们可以估计样本 i 的早产风险：

$$\hat{\theta}_i = \frac{exp(-2.344 - 0.076x_{i6} + 0.699z_{i6} + 0.115x_{i,11} + 1.539z_{i,10})}{1 + exp(-2.344 - 0.076x_{i6} + 0.699z_{i6} + 0.115x_{i,11} + 1.539z_{i,10})}. \tag{16}$$

模型评估

- 数据的缺失值会导致大量信息损失，可能导致不精确甚至错误的决定。上述讨论中最终模型和原本模型得到的影响变量就不同！后面可以看到基于树模型可以高效处理缺失值，专门创建一个分类或使用替代变量，避免产生不好的结果。
- ROC 曲线评估预测能力，可以看出还有大量差异未被解释，模型需要进一步改进。
- 基于 ROC 曲线的结果可能过于乐观，对于训练集拟合合较好，但没有验证集无法评价泛化能力！后面将大量使用交叉验证法。

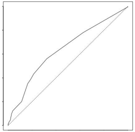

图: ROC curve for the final logistic regression model

树初步

树包含根结点，中间结点和终端结点。每一个分叉代表一次判断。这里研究树是为了分类/决策，我们希望得到一个模型，使得当我们获得一个新的样本时，可以根据其协变量的信息依次在树中检索，最后作出其是正例/反例的判断。

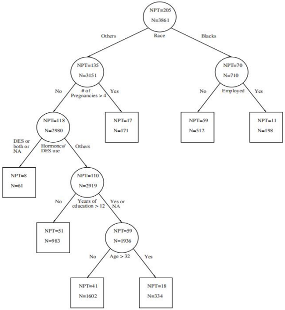

图: 决策树

生成树

树模型由训练样本生成，是一个学习过程。树模型中的每一个结点都是训练样本的一个子集。
考虑假想模型，我们将样本划分成不同区域，便可以得到一个决策树，在训练集上没有误差！

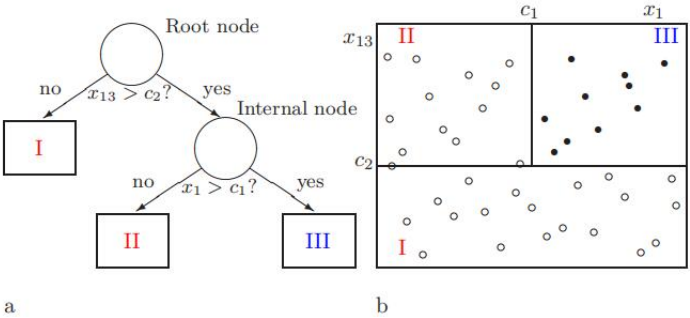

图: 假想模型生成的树

生成树

从这个假想模型中我们可以看到，生成树的过程大致就是把样本依据各个参数，划分成几个同质区间的过程。
但实际分析中，难以实现完全同质，我们希望结点不纯度，即正例/结点容量的比值接近 0 或 1. 还可以引入熵的概念。

定义

假定一个结点中正例比例为 p_1，反例比例为 p_2，那么结点的熵 (entropy) 可以定义为：

$$entropy = -p_1 log(p_1) - p_2 log p_2 \tag{17}$$

熵值越大，结点越不纯。

结点分裂

Question

我们将所有样本放入根节点中（当然，后续过程也是类似如此），如何训练出第一个分裂?

所有分裂方式：

- 离散型变量：每两个值之间都可以分裂；
- 有序或者连续型变量：看样本中有多少个不同的取值，分裂种类数为取值数-1；
- 分类变量：任意一个 k 分类变量有 $2^{k-1}-1$ 种可能的分裂方式。

需要一种最好的分裂。

结点分裂

定义

分裂的好坏程度可以定义为，熵减最大的分裂，具体是：

$$\Delta I = i_0 - p_1 i_1 - p_2 i_2 \tag{18}$$

这个量越大越好。其中 i_0 ，i_1 ，i_2 分别为父节点和两个子节点的熵值， p_i 是分到第 i 个子节点的概率，是针对 i_0 中不同样本的占比而言。

这样，在计算所有可能分裂的熵减以后，便可以选出一个最佳解，生成两个子节点，对于子节点可以继续重复上述划分。这就是是递归划分方法。

结点分裂注

- 子节点可以重复使用先前结点使用过的分类变量。
- 多种分裂好坏接近时，考虑可解释性问题。倾向于选择可解释的变量继续分裂过程。
- 如果所有都不够好，可以强制加入一个和父节点相关的变量，再下一次分裂时一般可以找到一个很好的分裂。
- 多叉分裂，加入惩罚因子，防止一直多叉而不是二叉分裂。并没有足够文献评估多叉和二叉树的性能。

终端节点

任何递归方法都要有终止条件。

- 可以选择分裂到只有一个样本或者允许分裂数目降为 0，但这样可能无实际意义，因为终端节点样本量太小了，统计推断不合理。
- 可以选择设立阈值，例如不少于样本量的 1% 或五个样本.
- 完全分裂再剪枝, 后面详述。

划分和剪枝可以视为线性模型中向前和向后逐步回归的变体。