离散选择模型——逻辑回归分类学习笔记

Supplement of Classification and Discrete Choice Model

姚炜彤

版本：*1.0.0*

最后更新：*April 14, 2019*

Classification and Discrete Choice Model(DCM) 介绍了如何应用逻辑回归模型对离散选择模型进行分类，本文主要对 Lecture 的内容做学习回顾和框架总结（参考 TA Session）并对已有框架的内容进行包括有序因变量、Logit 两类模型—嵌套 logit 模型 (the Nested Logit Regression) 和混合 Logit 模型的衍生潜类模型进行补充。

1 离散选择模型的原理回顾

离散选择模型的原理是随机效用理论（random utility theory）（属于 Probabilistic Choice Theory 的范畴，用于衡量未关测到的或已观测到但无法测度的选择）。对于个体 i 每个选择 j 的效用公式为 $U_{ij} = V_{ij} + e_{ij}$, 其中 V_{ij} 是驱动因素 X 的线性函数，因为可以被观测到或测度被称为系统性或代表性效用；e_{ij} 是不可观测的，假定为随机干扰项；选择 j 的概率 $\mathrm{P}(U_{ij}|X)$ 的分布取决于 e_{ij}，利用效用最大化理论选择一般是 $j = \mathrm{argmax}_j \mathrm{P}(U_{ij}|X)$。

2 变量的学习补充：定序型因变量

在一些经济领域的调查问卷我们通常会设置五分评级法，比如某一个地区的财政收入和该地区环境污染的研究调查，对于环境的评价会设置“非常严重”“严重”“中度”“轻度“”微弱“等选项，这些变量的设置存在单调趋势的关联，而且内在排序是统计推断的重要信息来源。尤其在列联表里，如果随着 X 的上升，Y 也上升，称为相协（concordant）；反之 X 上升而 Y 下降，则是相异（discordant）；如果在 X 和 Y 的取值上相同则称相平（tied）。本小节主要介绍定序变量的相关性检验，分类赋值和相关的 logit 回归模型。

2.1 定序趋势检验：γ 系数和 Spearman 系数

设 N_c 为相协对总数，N_d 为相异对总数（N_c>N_d 可以判断 X 上升伴随 Y 上升），用概率判断 $\gamma = \frac{N_c - N_d}{N_c + N_d}$，被称为 γ 系数（Goodman and Kruskal, 1954）。γ 的取值范围为 $[-1, 1]$，当绝对值为 1

时意味着 X 和 Y 完全线性相关，绝对值越大相关性越强，如果完全独立则 $|y| = 0$，但反之不成立。*Gamma* 系数检验是定序变量相关性的最宽松检验，其他的常用检验系数还有 Spearman 相关系数。Spearman 相关系数是用单调方程来评价两个统计变量的相关性，这里根据原始数据的排序位置（称为秩）代替原始数据进行计算。首先对变量（X,Y）进行排序，重新排序后的位置为秩，秩差 d_i, 数据个数为 n, Spearman 相关系数为：$\rho_s = 1\text{-}\frac{6\sum d_i^2}{n(n^2-1)}$

2.2 定序变量的赋值

赋值后的变量可以反应类别之间的距离。一种方法是等间距赋值，这种赋值无法反映类别之间的距离，但是适用于没有明显的类别距离或赋值选择的情况，比如说态度分类；反映距离的一种赋值方法是根据类别自动生成赋值，再使用每一类别所包括的对象的平均排序，又称中位秩作为赋值，比如类别 1 的对象排序为 1 到 N，则排序的中位秩为 $(1 + N)/2$。缺点时当一个类别的数值较少时，相邻的类别有相近的中位秩。

2.3 定序变量的 Logit 回归

这里介绍三类简单的 Logit 模型，此外还有比例发生模型，结合累计概率和线性预测的累积连结模型等。

1. 累积 logit 模型 (Cummulative Logit Model)

   累积 logit 模型的特点是可以反映 Y 类别整体排序特征，估计不会随变量的类别数量和切点位置变化。假设一共有 J 个类别，定序因变量为 Y(Y=1,2,...,J), 累积概率的 CDF 为：

$$P(Y \le j|x) = \pi_1(x) + \pi_2(x) + ... + \pi_j(x), j = 1, 2, ...J \tag{1}$$

   则累积 Logit 定义为:

$$\begin{aligned} L(Y \le j|x) &= log\frac{P(Y \le j|x)}{1 - P(Y \le j|x)} \\ &= log\frac{\pi_1(x) + \pi_2(x) + ... + \pi_j(x)}{\pi_{j+1}(x) + \pi_{j+2}(x) + ... + \pi_J(x)}, j = 1, 2...J - 1 \end{aligned} \tag{2}$$

   其分析整体排序特征的 Logit 表示如下，当 Y 和 X 相互独立时该式只剩下常数项，当 X,Y 相关时 β 不为零矩阵。特别的对于不同的 j，该模型的截距项不同但是所有变量的系数 β 时相同的，不同类别的概率曲线在水平方向上平移。

$$L(Y \le j|x) = \alpha_j + \beta X, j = 1, ..., J - 1 \tag{3}$$

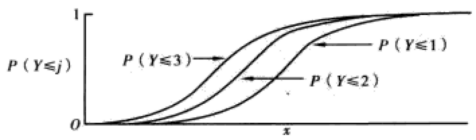


图 7.2 效应与切点无关的累积 Logit 模型

对于同一类别 j，不同自变量的相对发生比为不同 x_1，x_2 两个 CDF 比率的对数，在 Classification and Discrete Choice Model Lecture 的 Blue Bus,Red Bus 即是相对比例模型的应用（Blue Bus,Red Bus 可以视为一个最简单的定序变量问题！）：

$$L(Y \leq j|x_1) - L(Y \leq j|x_2) = \beta(x_1 - x_2) \tag{4}$$

2. 连续 logit 模型 (Continuation-ratio Logit Model)
   连续比 logit 模型适用于变量的连续过程存在序列特征的情况，主要应用在事件史分析上（e.g. 金融危机脱欧、中美贸易战,.etc？），定义为：

$$log\frac{P(Y = j|Y \geq j, X)}{P(Y \geq j + 1|Y \geq j, X)} \tag{5}$$

   其中 β 会随分类 j 变化

3. 相邻类别 logit 模型 (Adjecent-categories Logit Model)
   相邻类别 logit 反映 Y 类别之间的排序特征，反过来也意味着估计会随变量的类别数量和切点位置变化：

$$\begin{aligned} & logP(Y = j|Y = jorj + 1, X) \\ & = log\frac{\pi_j(x)}{\pi_{j+1}(x)} \\ & = \alpha_j + \beta X \end{aligned} \tag{6}$$

   反映了每改变一个类别发生的概率和相应的 X 的效应是多少。

3 Logit 模型的学习补充

3.1 混合 Logit 模型的衍生

在 Classification and Discrete Choice Model 里已经介绍了几种解释变量：X_i(case-specific or individual specific)，X_j(individual-Alternative specific) 另外还有具有个体和选择效应的变量 X_{ij}，并介绍了几种 variable 的模型构造情况。对于反映随机偏好差异的解释变量为 X_{ij} 变量的 logit 模型称为混合 Logit 模型 (the Mixed logit)。另一类同样反映随机偏好差异，但相较混合 Logit 模型具有避免切点人为划分造成无意义分类优势的模型称为潜在类别模型（Latent class model, LC）。

3.1.1 潜类模型介绍

潜在类别模型反映了潜在类本身的发生概率 $P(Z = z)$ 和变量 Y 在该类别 z 下的条件概率 $P(Y = y|Z = z)$。在条件独立假设下，变量 Y=$(y_1, y_2, ...y_n)$ 在潜在分类 K = k 下的概率为:

$$P(Y_1 = y_1, Y_2 = y_2, Y_n = y_n|K = k) = P(Y_1 = y_1|K = k)...P(Y_n = y_n|K = k) \tag{7}$$

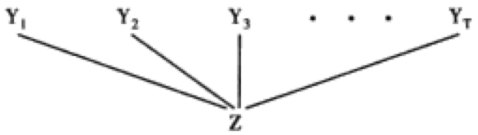


图 13.1　潜类模型的关联图

如果 $(Y_1, Y_2, ...Y_n)$ 构造一个 $n \times n$ 列联表，则加上潜分类 k 形成 $(n+1) \times (n+1)$ 列联表，对于每个单元格发生概率为:

$$\begin{aligned} P_{y_1, y_2 ... y_n} &= \sum_{k=1} P(Y_1 = y_1, Y_2 = y_2, Y_n = y_n | K = k) P(K = k) \\ &= \sum_{k=1} [\prod_{i=1}^{n} P(Y_i = y_i | K = k)] P(K = k) \end{aligned} \tag{8}$$

3.1.2　潜类别模型拟合

设$\lambda y_1, y_2 ... y_n$ 为所观测单元格计数，对于所有单元格求和则多项分布的对数似然函数的核函数为：（Kernel function 是将低维下无法分类或回归的特征经过非线性变换映射到高维空间进行分类、于高维空间非线性变换的内积相等的函数，详见支持向量机相关机器学习理论）$\sum \lambda_{y_1, y_2 ... y_n} log P_{y_1, y_2 ... y_n}$，利用 EM 算法使该似然函数最大化。

3.2　嵌套 logit 模型 (the Nested Logit Regression)

Multinomial，Conditional 和 Mixed Logit 成立的前提是 IIA，“Blue Bus, Red Bus”的例子已经体现了 IIA 假设局限性：否定了备选项间可能存在的相关性。显然”Blue Bus, Red Bus”例子是不满足 IIA 假设的。对此引入嵌套 logit 模型 (the Nested Logit Regression，GEV model)（Ben-Akiva and McFadden），该模型适用于无法观测对象的备择选项很相似或者有很强的相关性的情况，解决方法是构造一个决策树，分层次构建集合，集合下又可以有多个数量不同的分支并且逐层嵌套。同层的子集不相关但同一子集内的选项相关。假设一个深度为 2 的决策树，第一层的备选有 J 个，对应的第二层的备选有 K_j 个，最终的选项表示为 $j_1, j_2, ..., J_{k_j}$，个体 i 效用函数定义为：

$$P_{jk} = \beta X_{jk} + z_j \alpha_j + \epsilon_{jk} \tag{9}$$

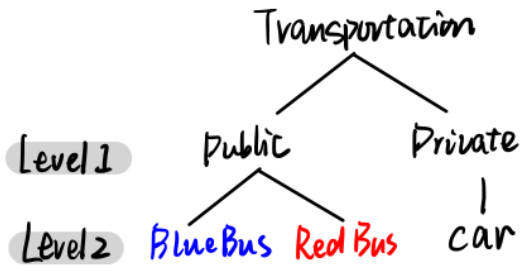
假设干扰项 ϵ_{jk} 服从广义极值分布（GEV），累积分布表示如下。$\lambda_k = 1$ 时说明 ϵ_{jk} 互不相干。

$$F(e) = exp(-\sum_{k=1)}^{K} (\sum e^{-\epsilon_{nj}/\lambda_k})^{\lambda_k}) \tag{10}$$

在第一层选择 j 的情况下，第二层 k 方案的概率为

$$P_{jk} = Prob(choose\ j\ at\ level1) \times Prob(choose\ k\ at\ level2 \mid j) \tag{11}$$

对于“Blue Bus, Red Bus”的例子，其对应的树状图为：



3.3 Logit 模型其他分类模型的比较

3.3.1 Logit 模型和朴素贝叶斯模型的比较

朴素贝叶斯模型 朴素贝叶斯模型是根据贝叶斯原理（Assignment 1: Brief Introduction of Bayesian Estimation），通过条件独立假设简化估计参数的模型（在某个条件下另一个变量贡献的信息失效）。朴素贝叶斯模型和 Logit 模型区别在于：

1. 估计的量不同：朴素贝叶斯模型估计的时 P(X) 和 P(Y|X) 的分布，Logit 模型估计的是 P(Y|X)。
2. 当 GNB(Guass Naive Bayes) 成立时，logit 和 NB 收敛结果相似；GNB 不成立，但数据量更大时 logit 精确度更大。
3. Biase-Variance Tradeoff 方面，NB 的方差更小但偏差更大，因此在小数据样本 NB 的精确度更大。
4. NB 需要基于严格的条件独立假设，并非所有数据都适用；而且需要先验概率假设，具有很强的主观性。

3.3.2 Logit 模型和 Probit 模型的比较

1. 从形式上的区别在于干扰项的分布，logit 为独立同分布的 Gumble：μ 是 Gumbel 分布的众数，$\frac{\pi^2}{6}\beta^2$ 是 Gumbel 的方差，Gumbel 分布的 PDF

$$f(x;\mu,\beta) = e^{-z-e^{-z}}, where z = \frac{x-\mu}{\beta} \tag{12}$$

Gumbel 分布的 CDF

$$F(x;\mu,\beta) = e^{-e^{-(x-\beta)/\beta}} \tag{13}$$

，probit 为标准正态分布，因此 probit 不受独立同分布的限制，也不存在 IIA 效应问题，可以解决变量的相关性和个体随机偏好差异；同样地可以构建混合嵌套 Logit 模型 (mixed nested logit model) 达到同样的效果。

2. Lecture 里总结 binary choice 两者相同，multinomial 时 probit 的协方差的参数比较多，运算量比较大；另外不合理的模型设定让 probit 的识别性和显著性明显降低。

4 学习回顾与总结（参考 TA Session）

变量	自变量 Y: 定量变量 ┬ 连续 └ 离散 ┬ 二值 (Cropland; Dose Response) page24-40 └ 多值 (Crop Choice) page105-115 定性变量 ┬ 定序（Ordinal） └ 名义 ┬ 二值 (Voting) page4-23 └ 多值 (Transportation, Ketchup) page71-103 解释变量 X: ┬ Individual specific: Multinomial Logit ├ Alternative specific: Conditional Logit └ Individual and Alternative specific: Mixed Logit
统计分类器（不包括 KNN 等结构分类器）	1. Logit: ┬ 1.1 Binomial /Multinomial Logit Regression ├ 1.2 Nested Logit Regression: 解决 MML 假设各个选择相互独立的问题（IIA） └ 1.3 Mixed Logit Regression: 解决 MML 里忽略个体异质性、无法处理随机差异 2. Probit: 没有 MNL 模型中独立同分布的限制，解决相关性、随机偏好差异等 3. 朴素贝叶斯模型：根据贝叶斯原理，通过条件独立假设简化估计参数的模型
模型判断	基本指标：confusion matrix，基于 confusion matrix 的其他指标有：ROC, AUC 等

5 参考文献

[1]《高级计量经济学及 Stata 应用》第二版，陈强编著.

[2]《分类数据分析》,（美）阿格莱斯蒂著.

[3]《统计学习方法》, 李航著.

[4] 离散选择模型研究进展. 王灿, 王德, 朱玮, 宋姗. 地理科学进展, 34 (10): 1275- 1287.

[5] 离散选择模型的基本原理及其发展演讲评价. 聂冲, 贾生华. 浙江大学管理学院. 数量经济技术经济研究：2005 年第 11 期.

[6] *Nested Logit Model*,Asif Khan. IRE, Georg-August University Goettingen.

[7]*Discrete Choice Analysis Theory and Application to Travel Demand.* Moshe Ben-Akiva, Steven R. Lerman.

[8]*Generative and Discriminative Classifiers: Naive Bayse and Logistic Regression*.Chapter3. Mitchell.