第 6 章 多项逻辑回归

本章将逻辑回归推广至“多项逻辑回归” (Multinomial Logit)，应用于多分类问题。

多分类问题也很常见。比如，在识别手写数字时，响应变量 $y \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$分别为从 0-9 的十个整数,共分为十类。

又比如，在使用数据集 `iris`，根据花瓣与花萼的长度与宽度判断鸢尾花的品种时，响应变量$y \in \{setosa, versicolor, virginica\}$，共分为三类。

6.1 多项逻辑回归

假设响应变量y的取值可分为K类，即$y \in \{1, 2, \cdots, K\}$。

给定特征向量$\mathbf{x}_i$，假设事件“$y_i = k$”$(k = 1, \cdots, K)$的条件概率为

$$\mathrm{P}(y_i = k \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta}_k)}{\sum_{l=1}^{K} \exp(\mathbf{x}_i' \boldsymbol{\beta}_l)} \quad (k = 1, \cdots, K) \tag{6.1}$$

这就是**多项逻辑回归**(Multinomial Logit)。其中，参数向量$\boldsymbol{\beta}_k$为对应于第k类的回归系数，$k = 1, \cdots, K$。

在机器学习中，方程(6.1)的右边也称为**软极值函数**(softmax function)，广泛用于分类问题的神经网络模型。

各类别的条件概率之和为 1，即

$$\sum_{k=1}^{K} \mathrm{P}(y_i = k \mid \mathbf{x}_i) = 1 \qquad (6.2)$$

在方程(6.1)中，无法同时识别所有的系数$\boldsymbol{\beta}_l \ (l = 1, \cdots, K)$。

这是因为，如果将$\boldsymbol{\beta}_l$变为$\boldsymbol{\beta}_l + \boldsymbol{\alpha}$(其中，$\boldsymbol{\alpha}$为某常数向量)，方程(6.1)的右边依然不变，并不影响此模型的拟合效果：

$$\frac{\exp\left[\mathbf{x}_i'(\boldsymbol{\beta}_k+\boldsymbol{\alpha})\right]}{\sum_{l=1}^{K}\exp\left[\mathbf{x}_i'(\boldsymbol{\beta}_l+\boldsymbol{\alpha})\right]}=\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_k)\cdot\cancel{\exp(\mathbf{x}_i'\boldsymbol{\alpha})}}{\cancel{\exp(\mathbf{x}_i'\boldsymbol{\alpha})}\cdot\sum_{l=1}^{K}\exp(\mathbf{x}_i'\boldsymbol{\beta}_l)}=\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_k)}{\sum_{l=1}^{K}\exp(\mathbf{x}_i'\boldsymbol{\beta}_l)} \tag{6.3}$$

为此，通常将某类(比如，第 1 类)作为**参照类别**(base category)，然后令其相应系数$\boldsymbol{\beta}_1=\mathbf{0}$。

由此，类别k的条件概率可写为

$$\mathrm{P}(y_i = k \mid \mathbf{x}_i) = \begin{cases} \dfrac{1}{1+\sum_{l=2}^{K}\exp(\mathbf{x}_i'\boldsymbol{\beta}_l)} & (k=1) \\ \dfrac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_k)}{1+\sum_{l=2}^{K}\exp(\mathbf{x}_i'\boldsymbol{\beta}_l)} & (k=2,\cdots,K) \end{cases} \tag{6.4}$$

其中，“$k=1$”所对应的类别为参照类别，故$\boldsymbol{\beta}_1 = \mathbf{0}$。

显然，当$K=2$时，多项逻辑模型就是逻辑回归：

$$\mathrm{P}(y_i = k \mid \mathbf{x}_i) = \begin{cases} \dfrac{1}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_2)} & (k=1) \\ \dfrac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_2)}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_2)} & (k=2) \end{cases} \tag{6.5}$$

方程(6.5)与第 5 章所介绍的 Logit 模型并无实质区别。

6.2 最大似然估计

假设样本数据为“独立同分布”(independently and identically distribution，简记 iid)，则第i个观测值的似然函数为

$$L_i(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_K)=\prod_{k=1}^{K}\mathrm{P}(y_i=k\,|\,\mathbf{x}_i)^{I(y_i=k)} \qquad (6.6)$$

其中，$\prod_{k=1}^{K}(\cdot)$表示连乘；而$I(\cdot)$为示性函数(indicator function)，即当 $y_i=k$时，$I(y_i=k)=1$，反之则为 0。

将上式取对数，可得第i个观测值的对数似然函数：

$$\ln L_i(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_K)=\sum_{k=1}^{K}\left[I(y_i=k)\cdot\ln \mathrm{P}(y_i=k\,|\,\mathbf{x}_i)\right] \tag{6.7}$$

将所有观测值的对数似然函数加总，即得到整个样本的对数似然函数：

$$\max_{\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_K}\ \ln L(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_K)=\sum_{i=1}^{n}\sum_{k=1}^{K}\left[I(y_i=k)\cdot\ln \mathrm{P}(y_i=k\,|\,\mathbf{x}_i)\right] \tag{6.8}$$

最大化此目标函数，即可得到系数估计值$\hat{\boldsymbol{\beta}}_1,\cdots,\hat{\boldsymbol{\beta}}_K$。

对于多项逻辑模型，也可根据对数似然函数，定义“准R^2”(Pseudo R^2) 与“残差偏离度”(residual deviance)，与 Logit 模型类似。

6.3 多项逻辑回归的解释

如果响应变量y分为K类，则多项逻辑模型有$(K-1)$个参数向量$\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_K$。

假设将第 1 类作为参照类别，故$\boldsymbol{\beta}_1=\mathbf{0}$。应如何解释这些参数向量$\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_K$呢？

在多项 Logit 模型中，对于系数$\hat{\boldsymbol{\beta}}_k\ (k=2,\cdots,K)$的解释，依赖于参照方案的设定。

由方程(6.4)可知，响应变量y归属第k类$(k=2,\cdots,K)$的条件概率，与y归属第 1 类(参照类别)的条件概率之比为

$$\frac{\mathrm{P}(y_i=k\,|\,\mathbf{x}_i)}{\mathrm{P}(y_i=1\,|\,\mathbf{x}_i)}=\exp(\mathbf{x}_i'\boldsymbol{\beta}_k) \qquad (6.9)$$

这就是事件“$y_i=k$”与“$y_i=1$”发生的**几率**(odds)，也称为**相对风险**(Relative Risk)。

进一步，如果某变量x_j为离散变量(比如，性别、子女数)，则可通过几率比来解释该变量对y的作用。

假设x_j增加 1 单位，从x_j变为x_j+1，记条件概率$\mathrm{P}(y_i=1|\mathbf{x}_i)$与$\mathrm{P}(y_i=k|\mathbf{x}_i)$的新值分别为$\mathrm{P}^*(y_i=1|\mathbf{x}_i)$与$\mathrm{P}^*(y_i=k|\mathbf{x}_i)$。

可计算新几率与原几率的比率，即**几率比**(odds ratio)，也称为**相对风险比率**(Relative Risk Ratio，简记 RRR)：

$$\text{RRR} \equiv \frac{\dfrac{\mathrm{P}^*(y_i = k \mid \mathbf{x}_i)}{\mathrm{P}^*(y_i = 1 \mid \mathbf{x}_i)}}{\dfrac{\mathrm{P}(y_i = k \mid \mathbf{x}_i)}{\mathrm{P}(y_i = 1 \mid \mathbf{x}_i)}} = \frac{\exp[\beta_1 x_1 + \cdots + \beta_j (x_j + 1) + \cdots + \beta_p x_p]}{\exp(\beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_p x_p)} = \exp(\beta_j) \tag{6.10}$$

其中，$\beta_1, \cdots, \beta_p$为参数向量$\boldsymbol{\beta}_k$的p个分量。

若$\beta_j = 0.12$，则几率比$\exp(\beta_j) = e^{0.12} = 1.13$。

这意味着，当x_j增加 1 单位时，则(相对于参照方案的)新几率变为原几率的 1.13 倍，即增加 13%。

6.4 多项逻辑回归的 Python 案例

我们使用 Glass 数据集演示多项逻辑回归。

该数据集最初来自 UCI Machine Learning Repository。

响应变量 Type 表示 7 种玻璃的类别(但样本中仅包含 6 种玻璃)，包括

1：建筑窗户的浮法玻璃 (building windows, float processed)
2：建筑窗户的非浮法玻璃 (building windows, non float processed)
3：车辆窗户的浮法玻璃 (vehicle windows, float processed)
4：车辆窗户的非浮法玻璃 (vehicle windows, non float processed，未在样本中出现)

5：容器(containers)的玻璃
6：餐具(tableware)的玻璃
7：车前灯(headlamps)的玻璃

为了法医学(forensic science)的目的，有时需根据玻璃碎片(glass fragments)的折射率以及不同化学元素的含量，预测犯罪现场的玻璃类别。

特征变量包括 RI(Refractive Index，折射率)，以及 8 种不同元素在相应氧化物(oxides)中的重量占比(weight percent)：Na(Sodium，钠)，Mg(Magnesium，镁)，Al(Aluminum，铝)，Si(Silicon，硅)，K(Potassium，钾)，Ca(Calcium，钙)，Ba(Barium，钡)与 Fe(Iron，铁)。

* 详见教材，以及配套 Python 程序（现场演示）。