关于零膨胀负二项分布的介绍

汪翔

2023/5/20

1 零膨胀负二项分布

零膨胀负二项 (Zero-Inflated Negative Binomial, ZINB) 分布[1] 能够很好地建模 scRNA-seq 数据的特点[2], 并且已经被验证能够取得较好的结果, 本文主要介绍 ZINB 分布.

1.1 二项分布和 Poisson 分布

考虑到 ZINB 分布的实际含义, 本节首先介绍两种常用的离散分布, 二项分布和 Poisson 分布[3], 在此之前首先介绍 Bernoulli 试验, 其定义如定义1.1所示[3].

定义 1.1. *考虑一个试验, 其结果分为两类, 或者成功, 或者失败, 该试验称为 Bernoulli 试验. 令随机变量*

$$X=\begin{cases}1, \text{试验结果为成功},\\ 0, \text{试验结果为失败},\end{cases} \tag{1}$$

假设试验结果为成功的概率为 p, 则此时称随机变量 X 服从 Bernoulli 分布, 记为 $X\sim \mathrm{B}(x;1,p)$, 其概率函数为:

$$f(x=k;1,p)=p^k(1-p)^{n-k}, k=0,1. \tag{2}$$

当我们考虑 n 次独立 Bernoulli 试验时即可以得到二项分布, 其定义如定义1.2所示[3].

定义 1.2. *考虑 n 次独立 Bernoulli 试验中试验成功的次数 X, 其中每次 Bernoulli 试验成功的概率为 p, 此时试验成功次数 X 服从二项分布, 记为 $X\sim \mathrm{B}(x;n,p)$, 此时:*

$$f(x=k;n,p)=\binom{n}{k}p^k(1-p)^{n-k}, k=0,1,\cdots,n. \tag{3}$$

由 Bernoulli 试验定义可知, 二项分布通常用来建模只有两种试验结果的离散分布. 例如, 考虑 n 次独立重复抛硬币的结果, 我们考虑硬币正面朝上的次数 X. 显然在硬币质地均匀的条件下, 第 i 次抛硬币得到正面朝上的概率为 $p = 0.5$, 以 X_i 表示这一事件, 则 $X_i \sim \mathrm{B}(x_i; 1, 0.5)$. 由此可知 $X = \sum_{i=1}^{n} X_i \sim \mathrm{B}(x; n, 0.5)$. 类似的, 若每次试验的结果不再是两种, 而是变为 K 种, 且每种结果发生概率为 $(p_1, p_2, \cdots, p_K)$, 那么此时我们称 X 服从多类别分布[4], 记作 $X \sim \mathrm{Cat}(\boldsymbol{x}; 1, p_1, \cdots, p_K)$. 当该事件独立重复发生 n 次时可以得到多项分布[4], 记作 $X \sim \mathrm{Multi}(\boldsymbol{x}; n, p_1, \cdots, p_K)$, 其中 $\boldsymbol{x} \in \mathbb{R}^K, \sum_{i=1}^{K} x_i = n, \sum_{i=1}^{K} p_i = 1$, x_i 表示试验中第 i 种结果发生的次数. 两种分布的概率函数分别如式 (4) 和式 (5) 所示.

$$f(\boldsymbol{x} = \boldsymbol{e}_k; 1, p_1, \cdots, p_K) = p_i, \ k = 1, \cdots, K, \tag{4}$$

$$f(\boldsymbol{x} = \boldsymbol{n}; n, p_1, \cdots, p_K) = n! \prod_{i=1}^{K} \frac{p_i^{n_i}}{n_i!}, \ \sum_{i=1}^{K} n_i = n, \tag{5}$$

其中,$\boldsymbol{e}_k \in \mathbb{R}^K$ 为第 k 维为 1 的指示向量, 即 $\boldsymbol{e}_k$ 满足:

$$\sum_{i=1}^{K} e_{ki} = 1, \quad e_{kk} = 1.$$

当我们考虑掷骰子这一事件时, 显然结果有六种可能, 则对于质地均匀的骰子, 可知其结果服从多类别分布, 即 $X \sim \mathrm{Cat}(\boldsymbol{x}; 1, \frac{1}{6}, \cdots, \frac{1}{6})$, 而当我们考虑 n 次独立掷骰子这一事件时, 其结果服从多项分布, 即 $X \sim \mathrm{Multi}(\boldsymbol{x}; n, \frac{1}{6}, \cdots, \frac{1}{6})$.

除了二项分布和多项分布, Poisson 分布也是常用的离散分布, 其定义如定义1.3所示.

定义 1.3. *离散随机变量 X 服从参数为 $\lambda > 0$ 的 Poisson 分布 (记作 $X \sim \mathrm{Poisson}(\lambda)$), 当且仅当其概率函数为:*

$$f(x = k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, \cdots. \tag{6}$$

Poisson 分布是一种常用的计数分布, 通常可以用 Poisson 分布来建模离散数据. 根据 Poisson 分布的概率函数可计算出其均值和方差为:

$$\mu = \lambda, \quad \sigma^2 = \lambda, \tag{7}$$

可以看出 Poisson 分布均值和方差相等, 因此对于某些离散数据来说, Poisson 分布并不能很好的建模数据特征, 此时可以考虑用负二项分布进行建模.

1.2 负二项分布

本节介绍负二项 (Negative Binomial, NB) 分布[3], 其原始定义如定义[?] 所示.

定义 1.4. 若每次 Bernoulli 试验有两种可能结果, 分别为成功或者失败. 在每次试验中, 成功的概率为 p, 失败的概率为 $(1-p)$. 反复进行该 Bernoulli, 直到观察到第 r 次成功发生, 此时试验失败次数 X 的分布即为负二项分布, 记为 $X \sim \mathrm{NB}(x; r, p)$, 概率函数为:

$$p(x=k) = \binom{k+r-1}{r-1} p^r (1-p)^k, \quad k = 0, 1, 2, \cdots. \tag{8}$$

特别地, 此处我们考虑的是 $r \in \mathbb{Z}$ 的情况, 此时负二项分布又称为 Pascal 分布.

根据概率函数可求出 NB 分布的均值和方差[3], 均值和方差为:

$$\mu = r\frac{1-p}{p}, \quad \sigma^2 = r\frac{1-p}{p^2}. \tag{9}$$

可以看出, NB 分布的均值和方差不相等, 因此相比于 Poisson 分布, NB 分布能够对离散数据给出更精确的建模.

由定义1.4并不能很好地看出 NB 分布与离散数据之间的关系, 命题1.1可以很好地反映 NB 分布和 Poisson 分布的关系[5], 命题描述如下:

命题 1.1. 设随机变量 $X \sim \mathrm{Poisson}(\Lambda)$, $\Lambda \geq 0$, Λ 为 Poisson 分布的参数, 且 $\Lambda \sim \mathrm{Gamma}(\lambda; \theta, \beta)$, 即 Λ 的概率分布为:

$$\mathrm{Gamma}(\lambda; \theta, \beta) = \frac{1}{\beta^\theta \Gamma(\theta)} \lambda^{\theta-1} e^{-\lambda/\beta},$$

θ, β 均为 Gamma 分布的参数, $\Gamma(\cdot)$ 表示 gamma 函数, 其定义为:

$$\Gamma(x) = \int_0^{+\infty} t^{x+1} e^{-t} \mathrm{d}t,$$

则此时有 $X \sim \mathrm{NB}(\theta, \beta)$.

我们在此给出命题1.1的证明[5].

证明. 由定义可知, 当给定 $\Lambda = \lambda$ 时, $X \sim \mathrm{Poisson}(\lambda)$, 即:

$$P(X = k | \Lambda = \lambda) = \frac{\lambda^k}{k!} e^{-\lambda},$$

从而有:

$$\begin{aligned} P(X=k) &= \int_0^{+\infty} P(X=k|\Lambda=\lambda) p(\lambda) \mathrm{d}\lambda \\ &= \int_0^{+\infty} \frac{\lambda^k}{k!} e^{-\lambda} \cdot \frac{1}{\beta^\theta \Gamma(\theta)} \lambda^{\theta-1} e^{-\lambda/\beta} \mathrm{d}\lambda \\ &= \frac{1}{\beta^\theta \Gamma(\theta) \cdot k!} \int_0^{+\infty} \lambda^{\theta+k-1} e^{-\lambda(\beta+1)/\beta} \mathrm{d}\lambda, \end{aligned}$$

考虑计算积分项 $I=\int_0^{+\infty}\lambda^{\theta+k-1}e^{-\lambda(\beta+1)/\beta}\mathrm{d}\lambda$, 令 $\varphi=\frac{1+\beta}{\beta}\lambda$, 即 $\lambda=\frac{\beta}{1+\beta}\varphi$, 代入积分项可得:

$$\begin{aligned}I&=\int_0^{+\infty}\lambda^{\theta+k-1}e^{-\lambda(\beta+1)/\beta}\mathrm{d}\lambda\\&=\int_0^{+\infty}\left(\frac{\beta}{1+\beta}\right)\left(\frac{\beta}{1+\beta}\varphi\right)^{\theta+k-1}e^{-\varphi}\mathrm{d}\varphi\\&=\left(\frac{\beta}{1+\beta}\right)^{\theta+k}\int_0^{+\infty}\varphi^{\theta+k-1}e^{-\varphi}\mathrm{d}\varphi\\&=\left(\frac{\beta}{1+\beta}\right)^{\theta+k}\Gamma(\theta+k),\end{aligned}$$

进而有:

$$\begin{aligned}P(X=k)&=\frac{1}{\beta^\theta\Gamma(\theta)\cdot k!}\left(\frac{\beta}{1+\beta}\right)^{\theta+k}\Gamma(\theta+k)\\&=\frac{\Gamma(\theta+k)}{\Gamma(k+1)\Gamma(\theta)}\left(\frac{\beta}{1+\beta}\right)^k\left(\frac{1}{1+\beta}\right)^\theta,\end{aligned}$$

若此时取 $p=\frac{1}{1+\beta}$, 则可得:

$$P(X=k)=\frac{\Gamma(\theta+k)}{\Gamma(k+1)\Gamma(\theta)}(1-p)^kp^\theta=\binom{k+\theta-1}{\theta-1}p^\theta(1-p)^k,\tag{10}$$

这即表示 $X\sim\mathrm{NB}(x;\theta,p)$. □

由此我们得到 NB 分布的第二个定义, 其概率函数为:

$$P(X=k)=\frac{\Gamma(\theta+k)}{\Gamma(k+1)\Gamma(\theta)}\left(\frac{\beta}{1+\beta}\right)^k\left(\frac{1}{1+\beta}\right)^\theta,\tag{11}$$

由式(9)可算出此时 NB 分布的均值和方差为:

$$\mu=\theta\frac{1-p}{p}=\theta\beta,\quad\sigma^2=\theta\frac{1-p}{p^2}=\theta\beta(1+\beta),\tag{12}$$

将均值 $\mu=\theta\beta$ 代入式(11), 此时 $\beta=\frac{\mu}{\theta}$, 由此概率函数可化简为:

$$\begin{aligned}P(X=k)&=\frac{\Gamma(\theta+k)}{\Gamma(k+1)\Gamma(\theta)}\left(\frac{\mu/\theta}{1+\mu/\theta}\right)^k\left(\frac{1}{1+\mu/\theta}\right)^\theta\\&=\frac{\Gamma(\theta+k)}{\Gamma(k+1)\Gamma(\theta)}\left(\frac{\mu}{\theta+\mu}\right)^k\left(\frac{\theta}{\theta+\mu}\right)^\theta,\end{aligned}$$

综上, 对于给定参数 μ,θ, NB 分布的概率函数为:

$$\mathrm{NB}(x;\mu,\theta)=\frac{\Gamma(x+\theta)}{\Gamma(x+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^\theta\left(\frac{\mu}{\theta+\mu}\right)^x.\tag{13}$$

1.3 零膨胀负二项分布

关于 scRNA-seq 数据的建模, 假设某条 RNA 被测序捕捉到的概率为 p, 将这一事件记为 X_i, 则可知 $X_i \sim \mathrm{B}(x;1,p)$. 当我们考虑从总数为 n 的所有 RNA 中捕捉到基因 g 对应的 RNA 这一事件时, 记这一事件为 X, 则可知:

$$X = \sum_{i=1}^{n} X_i, \quad X_i \text{ i.i.d} \sim \mathrm{B}(x;1,p),$$

由定义[?] 可知, 这时 X 服从总数为 n, 概率为 p 的二项分布, 即 $X \sim \mathrm{B}(x;n,p)$. 同时, 对 scRNA-seq 数据, RNA 总数往往非常大, 这时, 当 $n \to +\infty$ 时, 若我们假设 $np = \lambda = \mathrm{Const}$, 此时二项分布就近似成为 Poisson 分布, 即:

$$\begin{aligned}
\lim_{n\to+\infty} P(X=k;n,p) &= \lim_{n\to+\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to+\infty} \frac{k!}{n^k}\binom{n}{k} \cdot \frac{(np)^k}{k!} \cdot \left(1-\frac{\lambda}{n}\right)^{-n/\lambda\cdot(-\lambda)} \cdot (1-p)^{-k} \\
&= \frac{\lambda^k e^{-\lambda}}{k!},
\end{aligned}$$

这即是 Poisson 分布的概率函数, 由此可知:

$$X \sim \mathrm{Poisson}(\lambda = np). \tag{14}$$

研究者最开始就是通过 Poisson 分布来对 scRNA-seq 进行建模, 但后来发现 scRNA-seq 数据具有 “过散度” 性质[6], 即随着表达值的均值的增加, 方差相对于均值越来越大. 而根据 Poisson 分布的性质可知, Poisson 分布的均值和方差相等, 这时 Poisson 分布已经不能很好地完成对 scRNA-seq 数据的建模.

由命题1.1可知, NB 分布是对 Poisson 分布的改进, 同时由式(12)可知, 当 $\beta > 0$ 时有:

$$\sigma^2 = \theta\beta(1+\beta) > \theta\beta = \mu, \tag{15}$$

此时 NB 分布的方差大于均值, 可以用于表示 “过散度” 性质, 因此考虑利用 NB 分布建模 scRNA-seq 数据. 同时, 在 scRNA-seq 数据中, 由于测序技术等原因, scRNA-seq 数据包含大量零值, 因此可以考虑引入参数来单独建模零值分布, 即考虑用混合模型来建模 scRNA-seq 数据分布, 我们在此给出混合模型的数学描述:

给定一个事件 X, 假设其产生的结果受 K 个因素 $X_1, \cdots, X_K$ 的影响, 每个因素发生的先验概率为 $p(X_i) = \pi_i$, 且在已知 X_i 发生的条件下, X 服从某一特定分布, 该分布的参数记作 θ_i, 即 $p(X|X_i) = f_{X_i}(x;\theta)$, 则此时 X 的概率函数为:

$$f_X(x;\theta) = \sum_{i=1}^{K} p(X_i)p(X|X_i) = \sum_{i=1}^{K} \pi_i f_{X_i}(x;\theta_i), \tag{16}$$

此时我们称 X 的分布模型为 K 因素混合模型.

为建模 scRNA-seq 数据的高稀疏性, 考虑用 $K = 2$ 的混合分布来进行建模. 以 Dirac 分布来额外建模 0 值发生的情况, 则我们可以定义两个随机事件

$$X_1 := \{X\text{来自 Dirac 分布}\}, \quad X_2 := \{X\text{来自 NB 分布}\},$$

其中 Dirac 分布表示事件只在 0 值发生, 即:

$$P(X = 0) = 1, \quad P(X \neq 0) = 0,$$

若假设 X_1 发生的先验概率为 $P(X_1) = \pi$, 则此时我们得到混合分布的概率函数:

$$f(x; \mu, \theta, \pi) = p(X_1)p(X|X_1) + p(X_2)p(X|X_2) = \pi\delta(x) + (1 - \pi)f_{\text{NB}}(x; \mu, \theta), \quad (17)$$

其中, μ, θ 为 NB 分布的参数, π 为 Dirac 分布的发生先验, 可以理解为 0 值的额外发生概率, $\delta(x)$ 为示性函数, 即:

$$\delta(x) = \begin{cases} 1, & x = 0, \\ 0, & x \neq 0. \end{cases}$$

我们称这一混合模型为 ZINB 分布模型, 由式(17)可以知道 ZINB 分布的概率函数为:

$$f_{\text{ZINB}}(x; \mu, \theta, \pi) = \pi\delta(x) + (1 - \pi)f_{\text{NB}}(x; \mu, \theta), \quad (18)$$

类似地, 若考虑其他分布和 Dirac 分布的混合, 则可以得到其他的零膨胀分布模型, 例如零膨胀 Poisson(Zero-Inflated Poisson, ZIP) 分布[7] 等.

综上, 我们已经详细介绍了 ZINB 分布. ZINB 分布可以理解为在 NB 分布的基础上引入混合系数 π 来额外建模零值分布, 这能够处理某些 scRNA-seq 数据的高度稀疏性. 需要注意的是, 在实际问题上, NB 分布和 ZINB 分布都能够很好地建模 scRNA-seq 数据, 具体的分布选择依赖于具体数据.

2 神经网络实现

在深度神经网络中, 通常用自编码器模型来处理 scRNA-seq 数据, 基于 ZINB 分布建模的自编码器模型如图1所示.

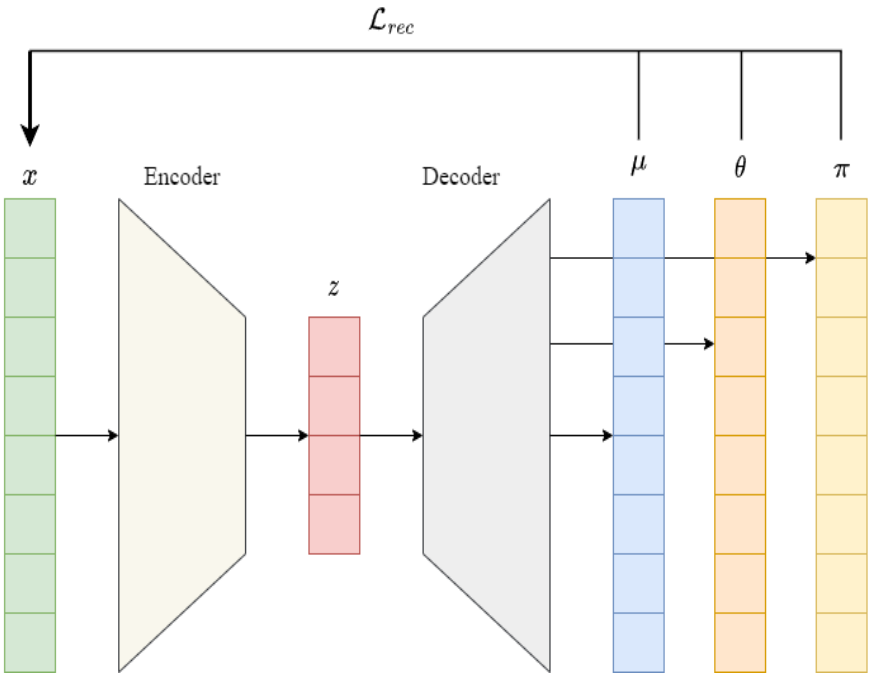

图 1: ZINB 自编码器网络结构图

利用神经网络 Θ 来拟合函数 $f : \boldsymbol{x} \to (\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\pi})$, 从而对给定 $\boldsymbol{x}$, 可以通过神经网络 f_Θ 算出其服从的 ZINB 分布的参数 $(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\pi})$. 其中 $\boldsymbol{x} = (x_1, \cdots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$, 假设 $\boldsymbol{x}$ 的各个分量相互独立且都服从 ZINB 分布, 即 $x_i \sim \mathrm{ZINB}(x_i; \mu_i, \theta_i, \pi_i)$.

为得到合适的拟合参数 $(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\pi})$, 我们以最小化负对数似然函数来作为网络的损失函数, 即:

$$\mathcal{L}_{rec} = -\log \mathrm{P}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\pi}) = -\sum_{i=1}^{n} \log \mathrm{P}(x_i; \mu_i, \theta_i, \pi_i), \tag{19}$$

由式(13)和(18)可知, 对每个分量 x, 概率函数为:

$$\log \mathrm{P}(x; \theta, \mu, \theta) = \begin{cases} \log\left((1-\pi) \cdot \mathrm{NB}(x; \mu, \theta)\right), & x > 0, \\ \log\left(\left(\dfrac{\theta}{\theta+\mu}\right)^{\theta}(1-\pi) + \pi\right), & x = 0, \end{cases} \tag{20}$$

结合式(20)和式(19)可以算出最终的网络损失, 网络训练完成之后, 可以取出隐藏变量 $\mathbf{Z}$ 来完成细胞聚类, 细胞注释等下游分析任务.

参考文献

[1] Yau K K, Wang K, Lee A H. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros[J]. Biometrical Journal,

2003, 45(4): 437-452.

[2] Eraslan G, Simon L M, Mircea M, et al. Single-cell rna-seq denoising using a deep count autoencoder[J]. Nature Communications, 2019, 10(1): 1-14.

[3] 李贤平. 概率论基础[M]. 北京：高等教育出版社, 1997.

[4] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[5] Anscombe F J. The transformation of poisson, binomial and negative-binomial data[J]. Biometrika, 1948, 35(3/4): 246-254.

[6] Hinde J, Demétrio C G. Overdispersion: models and estimation[J]. Computational Statistics & Data Analysis, 1998, 27(2): 151-170.

[7] Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing[J]. Technometrics, 1992, 34(1): 1-14.

A Code

```
import torch
from torch import nn

class ZINBLoss(nn.Module):
    def __init__(self):
        super(ZINBLoss, self).__init__()

    def forward(self, x, mean, disp, pi, scale_factor=1.0,
        ridge_lambda=0.0):
        eps = 1e-10
        mean = (mean.T * scale_factor).T

        t1 = torch.lgamma(disp+eps) + torch.lgamma(x+1.0) - \
            torch.lgamma(x+disp+eps)
        t2 = (disp+x) * torch.log(1.0 + (mean/(disp+eps))) + \
            (x * (torch.log(disp+eps) - torch.log(mean+eps)))
        nb_final = t1 + t2

        nb_case = nb_final - torch.log(1.0 - pi + eps)
        zero_nb = torch.pow(disp / (disp + mean + eps), disp)
        zero_case = -torch.log(pi + ((1.0 - pi) * zero_nb) + eps)
```

```
        result = torch.where(torch.le(x, 1e-8), zero_case, nb_case)

        if ridge_lambda > 0:
            ridge = ridge_lambda * torch.square(pi)
            result += ridge

        result = torch.mean(result)
        return result
```