因子分析[1]

简介

之前的notes介绍了许多的学习模型，这些模型要想正常工作通常隐含了一个条件——训练数据足够充分，也就是说需满足$m \gg n$，其中m表示数据集大小，n表示输入空间维度。当训练数据不充分时，即当$m \approx n$甚至$m \ll n$时，让我们来看看如果我们用多元高斯分布来拟合数据会出现什么问题。多元高斯分布的参数估计为

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m} \left(x^{(i)} - \mu\right)\left(x^{(i)} - \mu\right)^T$$

数据不充分时协方差矩阵是一个奇异矩阵，这意味着Σ^{-1}不存在，且$1/|\Sigma| = 1/0$。在这种情况下，我们无法计算多元高斯分布的概率值。如果我们仍想将数据拟合成合理的高斯模型，我们该怎么做呢？

接下来我们将讨论两种解决方案，其一是限制协方差矩阵，令在模型在少量的训练数据下仍然能够工作，但这种方案并不能让人十分满意；其二是利用因子分析模型找到输入的隐含因子，大大降低输入的维度。

1 受限协方差矩阵

当我们缺乏足够数据去拟合完整的协方差矩阵时，我们可以考虑对协方差矩阵做出一些限制。例如，我们可以限制其为对角矩阵，此时易得

$$\Sigma_{jj} = \frac{1}{m}\sum_{i=1}^{m}\left(x_j^{(i)} - \mu_j\right)^2$$

其中，Σ_{jj}就是输入的第j个坐标的方差的经验估计。回想二元高斯分布的几何特性，它在平面上的投影是一个椭圆。如果Σ是一个对角矩阵，那么该椭圆的两个轴与坐标轴都平行。

如果我们继续加强对协方差矩阵的限制，令其不仅为对角矩阵，且对角线上所有元素相等，那么我们将得到$\Sigma = \sigma^2 I$。其中，

$$\sigma^2 = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\left(x_j^{(i)} - \mu_j\right)^2$$

也就是上一个矩阵的对角线元素的平均值，反映到二维高斯分布图上就是椭圆变成了圆。

如果我们想根据训练数据拟合出一个完整、不受限的协方差矩阵，我们至少需要$m \geq n + 1$条数据，否则就会得到一个奇异Σ；如果协方差矩阵取以上两种受限形式，在$m \geq 2$时我们就可以得到非奇异Σ。然而，对协方差矩阵做出这种限制意味着我们认为输入的每个位置是相互独立的，这个假设太强。

2 边缘和条件高斯分布

在进入因子分析模型之前，先让我们看看边缘和条件高斯分布怎么求。

假设我们有一个向量形式的随机变量

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

其中，$x_1 \in \mathbb{R}^r$，$x_2 \in \mathbb{R}^s$。如果$x \sim \mathcal{N}(\mu, \Sigma)$，那么$x$的期望和协方差矩阵分别是

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

注意到协方差矩阵是对称的，因此$\Sigma_{12} = \Sigma_{21}^T$。在我们以上的假设中，$x_1, x_2$是联合多元高斯分布，我们该如何求出$x_1$的边缘分布？并不难发现，
$E[x_1] = \mu_1, Cov(x_1) = E[(x_1 - \mu_1)(x_1 - \mu_1)^T] = \Sigma_{11}$。下面我们来验证第二个结果，由联合协方差（joint covariance）的定义可得

$$\begin{aligned} Cov(x) &= \Sigma \\ &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \\ &= E[(x - \mu)(x - \mu)^T] \\ &= E\left[\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \right] \\ &= E\begin{bmatrix} (x_1 - \mu_1)(x_1 - \mu_1)^T & (x_1 - \mu_1)(x_2 - \mu_2)^T \\ (x_2 - \mu_2)(x_1 - \mu_1)^T & (x_1 - \mu_2)(x_1 - \mu_2)^T \end{bmatrix} \end{aligned}$$

由此可见多元高斯分布的边缘分布仍是多元高斯分布，所以$x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$。接下来我们该如何得到给定$x_2$下$x_1$的条件分布$x_1|x_2$？事实上$x_1|x_2 \sim \mathcal{N}(\mu_{1|2}, \mu_{1|2})$，其中

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \qquad (1)$$
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \qquad (2)$$

3 因子分析模型

在因子分析模型中，我们构造了一个联合分布(x, z)如下所示

$$z \sim \mathcal{N}(0, I)$$
$$x|z \sim \mathcal{N}(\mu + \Lambda z, \Psi)$$

其中，$z \in \mathbb{R}^k$是一个隐含随机变量。我们模型的参数包括向量$\mu \in \mathbb{R}^n$，矩阵$\Lambda \in \mathbb{R}^{n \times k}$和对角矩阵$\Psi \in \mathbb{R}^{n \times n}$，且k的值一般选择为小于n。

我们认为每个n维输入$x^{(i)}$都是由通过采样k维服从多元高斯分布的$z^{(i)}$，然后经过一系列变化得

来的——首先，我们通过$\mu + \Lambda z^{(i)}$将$z^{(i)}$从k维空间映射到n维输入空间；接着，我们给$\mu + \Lambda z^{(i)}$加上噪音Ψ生成输入$x^{(i)}$。

这个过程用数学语言表示即

$$\begin{aligned} z &\sim \mathcal{N}(0, I) \\ \epsilon &\sim \mathcal{N}(0, \Psi) \\ x &= \mu + \Lambda z + \epsilon \end{aligned}$$

这时随机变量z和输入x的联合服从多元高斯分布

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(\mu_{zx}, \Sigma)$$

接下了我们要求出参数μ_{zx}和Σ。因为

$$\begin{aligned} E[x] &= E[\mu + \Lambda z + \epsilon] \\ &= \mu + \Lambda E[z] + E[\epsilon] \\ &= \mu \end{aligned}$$

所以有

$$\mu_{zx} = \begin{bmatrix} E[z] \\ E[x] \end{bmatrix} = \begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}$$

对于Σ，我们需要计算三个部分，即$\Sigma_{zz} = E\left[(z - E[z])\,(z - E[z])^T\right]$，$\Sigma_{zx} = E\left[(z - E[z])\,(x - E[x])^T\right]$和$\Sigma_{xx} = E\left[(x - E[x])\,(x - E[x])^T\right]$。由于$z \sim \mathcal{N}(0, I)$，易得$\Sigma_{zz} = I$。此外

$$\begin{aligned} E\left[(z - E[z])\,(x - E[x])^T\right] &= E[z(\mu + \Lambda z + \epsilon - \mu)^T] \\ &= E[zz^T]\Lambda^T + E[z\epsilon^T] \\ &= \Lambda^T \end{aligned}$$

在以上推导中，我们用到了$E[zz^T] = Cov(z) + (E[z])^2 = I$和$E[z\epsilon^T] = E[z]E[\epsilon^T] = 0$（$z, \epsilon$之间相互独立）。相似地，

$$\begin{aligned} E\left[(x - E[x])\,(x - E[x])^T\right] &= E[(\mu + \Lambda z + \epsilon - \mu)(\mu + \Lambda z + \epsilon - \mu)^T] \\ &= E[\Lambda z z^T \Lambda^T + \epsilon z^T \Lambda^T + \Lambda z \epsilon^T + \epsilon\epsilon^T] \\ &= \Lambda E[zz^T]\Lambda^T + E[\epsilon\epsilon^T] \\ &= \Lambda\Lambda^T + \Psi \end{aligned}$$

将前面的结论放在一起，可得

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}\right) \quad (3)$$

从上式不难看出x的边缘分布为$x \sim \mathcal{N}(\mu, \Lambda\Lambda^T + \Psi)$。给定训练集$\left\{x^{(i)}; i = 1, \cdots, m\right\}$，我们可写出log-likelihood

$$l(\mu, \Lambda, \Psi) = \log \prod_{i=1}^{m} \frac{1}{(2\pi)^{n/2}|\Lambda\Lambda^T + \Psi|^{1/2}} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu\right)^T (\Lambda\Lambda^T + \Psi)^{-1} \left(x^{(i)} - \mu\right)\right)$$

接着令对所有参数的偏导数为0不就可以得到各参数的估计了么？可惜的我们无法得到closed-form（也就是说无法得到解析解），参考前面存在隐含随机变量时的解法，我们可以利用EM算法来估计。

4 用EM算法进行因子分析

E-step相对简单，我们只需计算$Q_i\left(z^{(i)}\right) = p\left(z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi\right)$即可。将在式(3)得到的分布应用到式(1-2)中，我们可以得到条件高斯分布$z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi \sim \mathcal{N}\left(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}}\right)$，其中

$$\begin{aligned}
\mu_{z^{(i)}|x^{(i)}} &= \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\left(x^{(i)} - \mu\right) \\
\Sigma_{z^{(i)}|x^{(i)}} &= I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda
\end{aligned}$$

根据多元高斯分布公式可得

$$Q_i\left(z^{(i)}\right) = \frac{1}{(2\pi)^{k/2}|\Sigma_{z^{(i)}|x^{(i)}}|^{1/2}} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu_{z^{(i)}|x^{(i)}}\right)^T \Sigma_{z^{(i)}|x^{(i)}}^{-1}\left(x^{(i)} - \mu_{z^{(i)}|x^{(i)}}\right)\right)$$

在接下来的M-step中，我们的目标函数[2]是

$$\begin{aligned}
&\sum_{i=1}^{m} \int_{z^{(i)}} Q_i\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi\right)}{Q_i\left(z^{(i)}\right)} dz^{(i)} \\
&= \int_{z^{(i)}} Q_i\left(z^{(i)}\right) \left[\log p\left(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi\right) + \log p\left(z^{(i)}\right) - \log Q_i\left(z^{(i)}\right)\right] dz^{(i)} \\
&= \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\left[\log p\left(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi\right) + \log p\left(z^{(i)}\right) - \log Q_i\left(z^{(i)}\right)\right]
\end{aligned}$$

待估参数是μ, Λ, Ψ，其中$z^{(i)} \sim Q_i$表示$z^{(i)}$服从于分布Q_i[3]。上式中，去掉与参数无关的项，我们可以得到更精简的目标函数

$$\sum_{i=1}^{m} E\left[\log p\left(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi\right)\right] \qquad (4)$$

令式(4)[4]对各参数的偏导等于0便可得出参数的更新法则，由于太过复杂这里不给出。

总结

因子分析是一种数据简化技术，他通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并利用少量几个假想变量来表示基本的数据结构。这几个变量能够反映原来众多变量的主要信息。原始的变量是可观测的显在变量，而假想变量是不可观测的潜在变量，称为因

子。

---

1. Written by Jimmy on 2016/03/14.↩
2. 在目标函数中，$z^{(i)}$ 是随机连续变量，所以我们用积分符号代替了上一notes中EM算法的求和符号。↩
3. Q_i 表示以$\mu_{z^{(i)}|x^{(i)}}$ 为均值，以$\Sigma_{z^{(i)}|x^{(i)}}$ 为协方差的高斯分布，这两个参数在每一轮E-step是固定的，因而Q_i 分布是固定的。↩
4. 式(4)中$x^{(i)}|z^{(i)}$ 的分布可由式(1-2)给出。↩