第三次仿真作业

背景 1：参数模型与非参数模型。机器学习的模型主要分为参数模型（Parametric Model）和非参数模型（Nonparametric Model）。非参数模型并不意味着模型没有参数。划分参数模型和非参数模型的原则是“让数据说话”，参数模型往往利用已有数据，在一些假设的基础上建立一个模型，然后利用该模型进行预测；非参数模型直接利用数据建立模型，并作进一步预测。非参数模型不像参数模型一样构建一个模型，并利用数据拟合出该模型的参数，而是直接利用已有数据，对新的数据进行预测。由于参数模型一般建立在人为的假设上，因此参数属于固定的空间，而非参数学习并没有固定的参数空间，有可能属于无限维空间。图 1 给出了参数模型和非参数模型的一个简单示例说明。

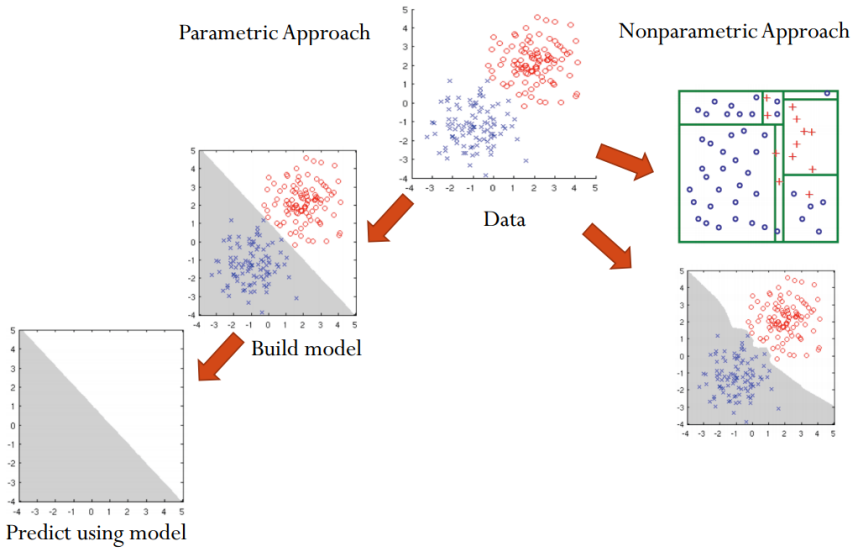


图 1：参数模型与非参数模型

背景 2：高斯混合模型与贝叶斯高斯混合模型。高斯混合模型是一种常用的聚类方法。假设有 K 个聚类，假设关于聚类的分布参数为 $\boldsymbol{\pi}=(\pi_1,...,\pi_K)$，每一个聚类的分布参数为 $\phi_k=(\mu_k,\Sigma_k)$（对应高斯分布的均值和方差），则任意一个数据样本所服从的概率密度为：

$$p(x)=\sum_{k=1}^{K}\pi_k N(x\,|\,\mu_k,\Sigma_k) \qquad (1)$$

其中参数 $\boldsymbol{\pi}$ 和 $\{\phi_k\}_{k=1}^{K}$（联合记作 $\boldsymbol{\theta}$）为高斯混合模型的待估计参数。

在贝叶斯统计中，往往把待估计的参数看做随机变量，并给随机变量赋予不同的先验分

布，通过贝叶斯公式得到随机变量的后验概率分布：

$$p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \tag{2}$$

其中 D 表示数据样本集合，$p(\boldsymbol{\theta})$ 为参数随机变量的先验分布，$p(D|\boldsymbol{\theta})$ 为似然函数。

对于高斯混合模型，同样可以利用贝叶斯方法进行参数估计，即将高斯混合模型的待估计参数 $\boldsymbol{\pi}$ 和 $\{\phi_k\}_{k=1}^{K}$（或者 $\boldsymbol{\theta}$）看作是随机变量，并给予先验分布，然后利用公式（2）即可得到后验概率分布。对于参数 $\boldsymbol{\pi}=(\pi_1,...,\pi_K)$，对应的随机变量的应该是一个 K 维的随机向量，并且各个维度满足 $\pi_k \geq 0, 1 \leq k \leq K; \sum_{k=1}^{K}\pi_k = 1$。对于这样的随机变量，其共轭分布为狄利克雷分布，即：

$$\boldsymbol{\pi} \sim Dirichlet\left(\alpha/K,...,\alpha/K\right) \tag{3}$$

其中参数 α 为该分布的分布参数，即先验分布的分布参数。对于参数 $\phi_k=(\mu_k,\Sigma_k)$，它的共轭分布为：

$$\phi_k \sim Normal\text{-}Inverse\text{-}Wishart(v) \tag{4}$$

其中 NIW 为正泰逆威沙特分布，v 为该分布大的分布参数，即先验分布的分布参数。

至此，高斯混合模型的参数 $\boldsymbol{\theta}$ 服从的先验分布可表示为：

$$G=\sum_{k=1}^{K}\pi_k\delta_{\phi_k}, \quad \theta_i \sim G \tag{5}$$

其中 δ_{ϕ_k} 表示在 ϕ_k 处的冲激函数。该混合模型对应的概率图模型如图 2 所示。

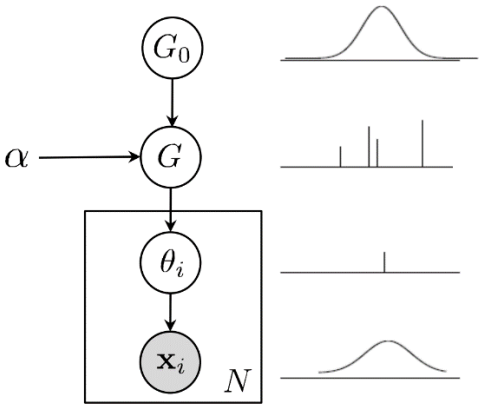


图 2：贝叶斯高斯混合模型的概率图模型表示

背景 3：狄利克雷过程混合高斯模型。贝叶斯高斯混合模型存在一个问题，就是我们事先要指定聚类的个数 K，为了得到合适的聚类个数，往往需要试验多个 K 的取值，取其中相对最好的一个。显然当问题比较复杂的时候比较困难，因此我们需要让观测数据“自己决定”聚类的个数。

我们在背景 2 中提高的狄利克雷分布，被称为分布的分布，因为该分布表达的意思是某一个 N 维分布出现的概率大小（实际上就是一个 N 为随机向量出现的概率，只不过这个 N 维随机向量的各维度取值恰好非负，并且求和等于 1）而狄利克雷过程（Dirichlet Process）类似于狄利克雷分布，它仍然是分布的分布，只不过狄利克雷过程将分布的维数从固定的 N 维扩展到了任意维。狄利克雷过程 $G \sim DP(G_0, \alpha)$ 由两个参数指定，一个是基分布 G_0（Base Distribution），一个是聚焦参数 α（Concentration Parameter）。基分布是狄利克雷过程的期望分布，即狄利克雷过程在期望分布 G_0 附近产生分布 G，类似于正态分布在均值附近取值。然而**即便基分布是一个连续的分布**，狄利克雷过程产生的分布（确定时间点，得到一个分布）一般都是离散分布。因此它可以被用来作为聚类个数的先验分布。

如果我们将狄利克雷过程作为高斯混合模型聚类个数参数的先验分布，那么对应的概率图模型可以表示为图 3。**贝叶斯非参数模型和贝叶斯参数模型的区别在于：聚类的个数不再是一个确定的值，而可以取值任何正整数**。

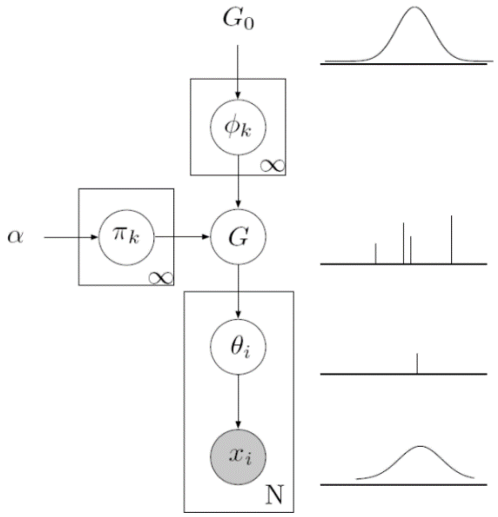


图 3：狄利克雷过程贝叶斯高斯混合模型

狄利克雷过程并没有确定的表达形式，但是有很多构造狄利克雷过程的方法，比如折断筷子模型（Stick-Breaking Model），Blackwell-MacQueen 罐子模型和中国餐厅过程（Chinese Restaurant Process）。

中国餐厅过程（Chinese Restaurant Process, CRP）是表示狄利克雷过程有效方式。基本流程为：假设一个中国餐厅里有无穷多张圆桌（桌子无穷大，可以坐的下任意多的顾客），顾客进入餐厅，按照一定的概率，随机的选择坐在已经有人坐的圆桌周围，或者还没有人做的圆桌周围。因此中国餐厅过程首先定义了一个正整数集合的划分（不同的桌子周围坐了不同数量的人，大部分桌子附近没人坐），并在这些划分上赋予了不同的概率。

引入示性变量 $z_i \in \{1,\cdots,K\}$ 表示样本 x_i 属于第几类的随机变量。中国餐厅过程可通过以下步骤实现：(I)假设刚开始餐厅没有人；(II)第一个顾客随机的选取一个桌子坐下；(III)第 $n+1$（$n \geq 0$）个顾客有两种选择，以概率 $\alpha/(\alpha+n)$ 坐在一个空桌子周围，以概率 $p(z_{n+1}=k \mid z_{1:n}) = n_k/(n+\alpha)$ 随机坐在已有人坐的第 k 个桌子周围。隐变量 z_i 保存着第 i 个顾客选择的桌子编号，$i \in \{1,2,\cdots k_n\}$，k_n 表示 n 个人所占桌子总数。注意到 $p(z_1,\cdots,z_n) = p(z_1)\cdots p(z_n \mid z_{1:n-1})$ 的取值与 z_i 的顺序没有关系，意味着如果两张桌子拥有相同的顾客，则它们发生的概率相同。

从上述过程可以看出中国餐桌模型是一种在给定观测数据下划分空间并赋予概率分布的方法（Specify a distribution over partitions of n points）。n 位顾客可以看作是观测变量，所有餐桌可以看作是整个空间，顾客们对餐桌的选择可以看作是对空间的划分。由于 n 位顾客选择的餐桌总数 k_n 一般小于等于 n（k_n 的期望值为 $O(\alpha \log n)$），因此表现出聚类的特性，这就是为什么狄利克雷过程被用来作为聚类算法先验分布的原因，它不仅可以表现出聚类属

性，而且不需要指定聚类的个数，聚类的个数完全由数据自身决定。

如果使用狄利克雷过程作为聚类参数的先验分布，图 3 所示的模型，其对应的数学描述如下：

$$\begin{aligned} x_i \mid \boldsymbol{\theta}_i &\sim F\left(\boldsymbol{\theta}_i\right) \\ \boldsymbol{\theta}_i \mid G &\sim G \\ G &\sim DP\left(\alpha, G_0\right) \end{aligned} \tag{19}$$

其中 G 就是聚类参数的先验分布，并且 G 也具有先验分布，即 $G=\sum_{k=1}^{\infty}\pi_k\delta_{\boldsymbol{\theta}_k}\sim DP\left(\alpha,G_0\right)$；$\boldsymbol{\theta}_i$ 是第 i 个数据 x_i 的聚类参数，由先验分布 G 产生；在知道样本 x_i 的聚类参数 $\boldsymbol{\theta}_i$ 的条件下，样本 x_i 服从的高斯分布。

题目：利用 Gibbs 采样实现狄利克雷过程高斯混合模型。假设有数据集：$D=\left\{x_i\right\}_{i=1}^{n}$；待估计参数为：

——示性变量 $\mathbf{z}=\left(z_1,\cdots,z_n\right), z_i\in\left\{1,\cdots,K\right\}$（表示样本 x_i 属于第几类的随机变量）；

——类别参数变量：$\boldsymbol{\phi}=\left(\phi_1,\cdots,\phi_K\right)$（表示第不同类别数据所服从分布的分布参数）。

利用 Gibbs 采样推断待估计参数变量的后验概率分布，从而实现聚类：

$$p\left(\boldsymbol{\phi},\mathbf{z}\middle|D\right)\propto p\left(\boldsymbol{\phi}\right)p\left(\mathbf{z}\right)p\left(D\middle|\boldsymbol{\phi},\mathbf{z}\right)$$

其中先验分布 $p\left(\boldsymbol{\phi}\right)$ 选取共轭分布；先验分布 $p\left(\mathbf{z}\right)$ 由中国餐厅模型得到。

具体地，假设不同类数据具有相同的协方差矩阵 Σ，区别在于均值不同，即：$\phi_k=\mu_k$，$p\left(x_i\middle|\boldsymbol{\phi},\mathbf{z}\right)\sim N\left(\mu_{z_i},\Sigma\right)$，简单起见，令 $\Sigma=I$。对于先验分布 $p\left(\boldsymbol{\phi}\right)$，我们采用共轭分布 $p\left(\phi_k\right)\sim N\left(0,\sigma^2 I\right)$，简单起见令 $\sigma^2=1$。对于先验分布 $p\left(\mathbf{z}\right)$，我们用中国餐厅模型来实现狄利克雷过程，有：

—— $p\left(z_i=k\middle|\mathbf{z}_{-i}\right)=\dfrac{n_k}{n-1+\alpha}$，如果 $k\in\left\{1,\cdots,K\right\}$

—— $p\left(z_i=K+1\middle|\mathbf{z}_{-i}\right)=\dfrac{\alpha}{n-1+\alpha}$，其它

其中 $n_k = \sum_{j \in \{1,\cdots,n\} \setminus \{i\}} \delta_{z_j,k}$ 表示已经在第 k 张桌子上的客人数量。

下面是 Gibbs 采样的具体实现过程：

- Randomly initialize $\mathbf{z}, \boldsymbol{\phi}$
- Sample each z_i from
  - old component ($k \in \{1, \ldots, K\}$):

$$p(z_i = k|\mathcal{D}, \boldsymbol{\phi}, \mathbf{z}_{-i}) \propto p(z_i = k|\mathbf{z}_{-i}) p(\mathbf{x}_i|\mathbf{z}, \boldsymbol{\phi})$$
$$= \frac{n_k}{n-1+\alpha} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^\top \Sigma^{-1}(\mathbf{x}_i - \mu_k)\}$$

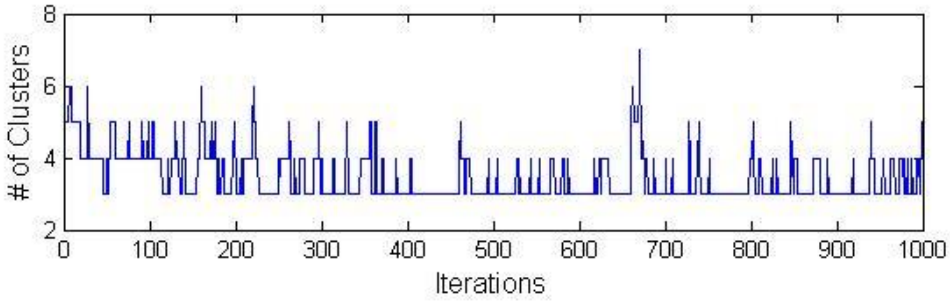
  - new component:

$$p(z_i = K+1|\mathcal{D}, \boldsymbol{\phi}, \mathbf{z}_{-i}) \propto p(z_i = K+1|\mathbf{z}_{-i}) p(\mathbf{x}_i|\mathbf{z}, \boldsymbol{\phi})$$
$$= p(z_i = K+1|\mathbf{z}_{-i}) \int p_0(\phi_{K+1}) p(\mathbf{x}_i|z_i, \boldsymbol{\phi}, \phi_{K+1}) d\phi_{K+1}$$
$$= \frac{\alpha}{n-1+\alpha} \frac{1}{(2\pi)^{d/2}\sigma^d} \frac{|\Sigma'|^{1/2}}{|\Sigma|^{1/2}} \exp\left\{\frac{1}{2}\mathbf{x}_i^\top (\Sigma^{-1}\Sigma'\Sigma^{-1} - \Sigma^{-1})\mathbf{x}_i\right\}$$
$$(\text{where } \Sigma' = \left(\frac{1}{\sigma^2} I + \Sigma^{-1}\right)^{-1})$$

- Sample each ϕ_k from

$$p(\phi_k|\mathcal{D}, \mathbf{z}) \propto p_0(\phi_k) \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{z}, \boldsymbol{\phi})$$
$$= p_0(\phi_k) \prod_{i|z_i=k} p(\mathbf{x}_i|\mathbf{z}, \boldsymbol{\phi})$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}\mu_k^\top \mu_k - \frac{1}{2}\sum_{i|z_i=k} (\mathbf{x}_i - \mu_k)^\top \Sigma^{-1}(\mathbf{x}_i - \mu_k)\right\}$$
$$\sim \mathcal{N}(\mu_k', \Sigma_k')$$
$$(\text{where } \mu_k' = \Sigma_k'\left(\Sigma^{-1}\sum_{i|z_i=k} \mathbf{x}_i\right),\ \Sigma_k' = \left(\frac{1}{\sigma^2} I + c_k \Sigma^{-1}\right)^{-1},\ c_k = \sum_{i=1}^{n} \delta_{z_i,k})$$

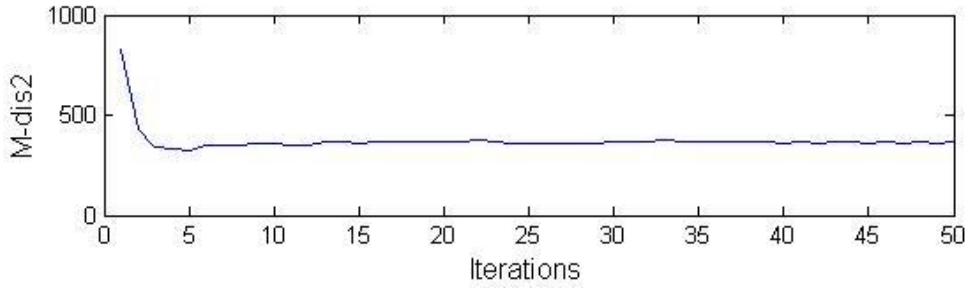
要求：1）数据获取：自行产生 100 个协方差为单位矩阵，均值为 $(2.4, 2)^T$ 的数据样本，100 个协方差为单位矩阵，均值为 $(-1.8, 1.4)^T$ 的数据样本和 100 个协方差为单位矩阵，均值为 $(-0.2, -2.6)^T$ 的数据样本作为实验所用数据。2）记录在学习过程中（迭代过程中），聚类个数的变化曲线，下图是一个示例：



3）记录在学习过程中（迭代过程中）所有数据样本距离它们所属类别的均值向量的距离，即：

$$D_M\left(D;\boldsymbol{z},\boldsymbol{\phi}\right)=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left|\left(x_i-\mu_{z_i}\right)^T\left(x_i-\mu_{z_i}\right)\right|^{\frac{1}{2}}$$

下图是一个示例结果：



4）给出采样算法收敛之后样本数量最多的前三个聚类的聚类参数和对应的样本个数，比较它们和生成数据时所用的参数（均值向量）有何不同？4）中国餐厅过程的参数 α 默认为 1，另外请尝试不同的 α 取值，研究该参数对聚类过程和结果的影响；5）算法可通过 python 或者 MATLAB 实现；6）自学 Gibbs 采样方法。

说明：1）要求中所说的迭代过程的任意一次迭代指的是对参数 $\boldsymbol{z}$ 和 $\boldsymbol{\phi}$ 的一次完整采样；2）采样算法收敛指的是 Gibbs 采样算法所得到的采样样本所服从的概率分布最终收敛到了其对应的真实的后验概率分布（采样初期得到的样本并不服从真实的后验概率分布）；3）对参数 $\boldsymbol{z}$ 进行初始化时，可以先假定有 m 个类别（例如 m=20），然后给所有样本数据随机赋予一个 1~m 的值，作为初始化结果；在给定初始类别个数之后，就可以随机初始化每个类别对应的分布参数了；4）作业附件 1，附件 2 和附件 3 为关于狄利克雷过程高斯混合模型的介绍；附件 4 为 Gibbs 采样的介绍。关于 Gibbs 采样方法，也可以参考 bishop 书 11.3 小结介绍（注：附件 1 和附件 4 仅供参考）。

提交：1）简明的实验报告，包括要求中的所有结果；2）代码，并有详细的注释说明。

致谢：朱军老师 PPT
朱军老师个人主页：http://bigml.cs.tsinghua.edu.cn/~jun/index.shtml