非参数贝叶斯方法

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），以及半参数模型（Semi-parametric Model）。本节首先介绍非参数模型，然后介绍非参数贝叶斯方法。

1. 参数模型与非参数模型

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。例如 LDA 模型就是参数模型，它首先对文档的单词生成建立假设并构建模型，然后利用已有数据进行参数学习，对新的样本进行预测的时候，只需要利用学习到的参数即可。KNN 就是非参数模型，它并没有像 LDA 一样构建一个模型，然后利用数据学习出模型的参数，而是直接利用已有数据，对新的数据进行预测。以分类问题为例，下图 1 给出了参数模型和非参数模型的区别。

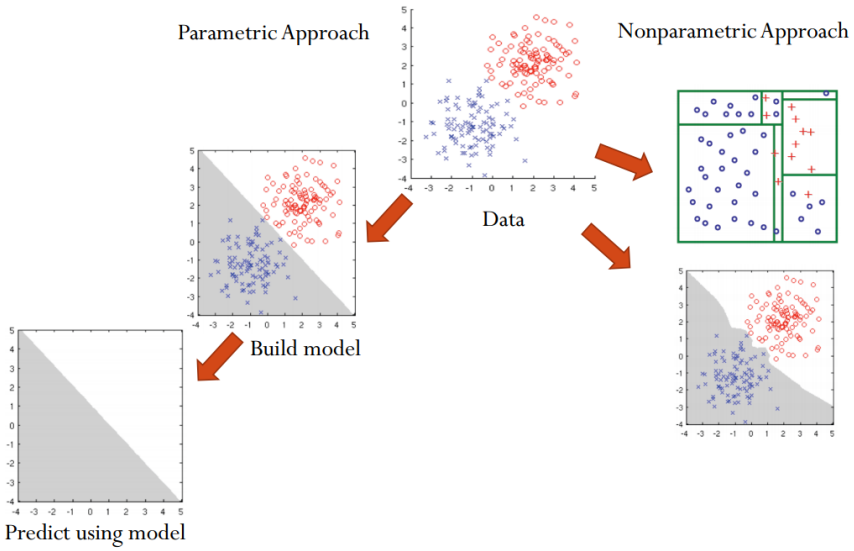

图 1：参数模型与非参数模型

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。对于有参数模型，如果基本的假设是正确的，那么学习到的模型就能比较简单，并且易于解释；

如果基本假设不正确，则对新数据的预测效果就会大打折扣。对于非参数模型，优点是避免了要求严格的假设，比较灵活，然而非参数模型往往难以解释，并且精度并不高。半参数模型参数由有限维度的部分和无限维度的部分两部分组成，即保留了参数模型易于解释的优点，又保留了模型的灵活性。图 2 给出了半参数模型和其它模型的关系。

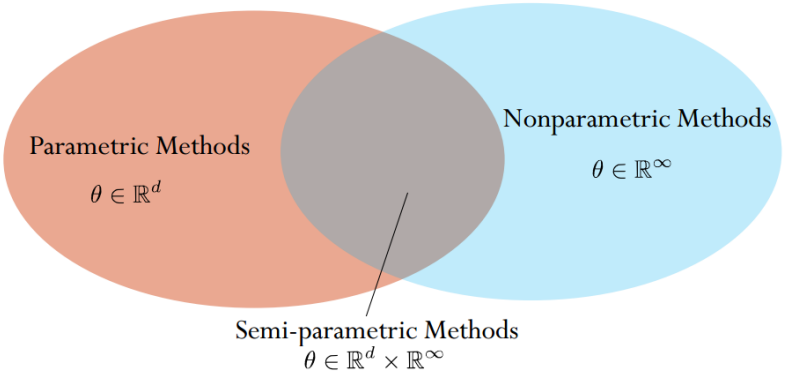

图 2：半参数模型

2. 贝叶斯非参数方法

非参数贝叶斯方法结合非参数模型和贝叶斯统计两大特点，首先希望模型的复杂度随着数据量的增加而逐渐上升，其次希望给属于无穷维空间（unbounded）的参数增加先验分布。我们以聚类为例，首先给出聚类的贝叶斯参数模型，然后介绍聚类的贝叶斯非参数模型。

2.1 贝叶斯统计

贝叶斯方法把统计问题中的参数作为随机变量，并给参数赋予不同的先验分布（Prior Distribution），联合数据的似然函数（Likelihood Function），利用贝叶斯定理可以得到参数的后验概率分布（Posterior Distribution），作为我们对参数的信度（Belief）。其中先验分布可以分为客观先验（Objective Priors），主观分布（Subjective Priors），分级先验（Hierarchical Priors），经验先验（Empirical Priors）。客观先验一般是均匀分布，拥有较好的频率论性质，并且能够防止出现黑天鹅悖论；主观先验一般是非均匀分布，把我们对参数的信度编码到先验分布当中，一般使用共轭分布作为先验分布；分层先验把先验分布的分布参数仍然当做随机变量，并给予先验分布；经验先验从观测数据中学习得到先验分布的分布参数，优点是全部让数据说话，防止人为增加错误的先验

知识，缺点是使用了两次数据，可能会发生过拟合。

2.2 贝叶斯参数模型

高斯混合模型，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$\phi_k=(\mu_k,\Sigma_k)$（可将参数$\boldsymbol{\pi}$和$\{\phi_k\}_{k=1}^K$联合记作$\theta$，高斯混合模型的参数），则数据的联合分布为：

$$p(\mathbf{x})=\sum_{k=1}^{K}\pi_k N(\mathbf{x}\mid\mu_k,\Sigma_k) \tag{1}$$

如果不对参数增加先验分布，那么我们就可以使用 EM 算法进行参数估计。如果我们把参数$\boldsymbol{\pi}$和ϕ_k当做随机变量，并赋予先验分布（取共轭分布），就变成了贝叶斯参数模型。我们可以令：

$$\boldsymbol{\pi}\sim Dirichlet\left(\alpha/K,...,\alpha/K\right) \tag{2}$$

$$\phi_k=(\mu_k,\Sigma_k)\sim G_0=Normal\text{-}Inverse\text{-}Wishart(\nu) \tag{3}$$

并令：

$$G=\sum_{k=1}^{K}\pi_k\delta_{\phi_k},\quad \theta_i\sim G \tag{4}$$

其中δ_{ϕ_k}表示在ϕ_k处的冲激函数。其中G是一个随机分布函数表示随机变量$\boldsymbol{\pi},\Phi$的联合分布函数，并且各自对应的边缘分布函数分别为式（2）和式（3）。以上的贝叶斯参数模型对应图 3 所示的概率生成模型（其中θ_i表示样本$\mathbf{x}_i$的分布参数）。

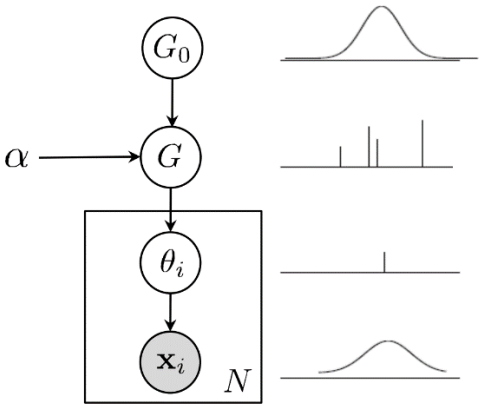

图 3：贝叶斯高斯混合模型

得到以上模型之后，我们就可以计算参数 $\boldsymbol{\pi}$ 和 ϕ_k 的后验概率分布：

$$p(\boldsymbol{\pi},\Phi|D)\propto p(D|\boldsymbol{\pi},\Phi)p(\boldsymbol{\pi},\Phi) \tag{5}$$

上式即我们最终想得到的结果：参数 $\boldsymbol{\pi},\Phi$ 的后验概率分布，如果我们能够得到该后验概率的解析表达式，那么我们可以利用最大后验概率得到参数 $\boldsymbol{\pi},\Phi$ 的一个点估计，或者将后验概率的期望作为参数 $\boldsymbol{\pi},\Phi$ 的点估计等等。然而显然很难得到上式的解析表达式。因此我们就可以使用马尔科夫链蒙塔卡罗方法（MCMC）来对后验概率分布进行采样，用样本来近似后验概率分布。或者用变分推断，通过一个函数来近似后验概率分布。

下面介绍一个可行的采样过程。引入随机变量 z_i 来表示数据 $\mathbf{x}_i$ 由哪一个聚类生成，则图 3 所示的生成模型可以等价于图 4。此时后验概率分布变成了：

$$p(\boldsymbol{\pi},Z,\Phi|D)\propto p(D|Z,\Phi)p(Z|\boldsymbol{\pi})p(\boldsymbol{\pi},\Phi) \tag{6}$$

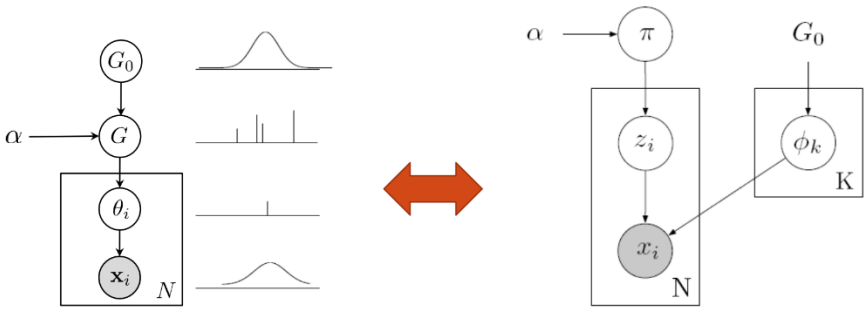

图 4：贝叶斯参数模型的等价表示

我们可以随机的初始化参数 $\boldsymbol{\pi},Z,\Phi$，然后首先从分布（直接利用贝叶斯公式，

忽略无关变量即可得到）

$$p(z_i \mid Z_{-i}, \boldsymbol{\pi}, \Phi, D) \propto \sum_{k=1}^{K} \pi_k p(\mathbf{x}_i \mid \phi_k) \delta_{z_i,k} \tag{7}$$

中采样 z_i，循环实现对 Z 的采样；其次从分布

$$p(\boldsymbol{\pi} \mid Z, \Phi, D) = Dirichlet(n_1 + \alpha / K, ..., n_K + \alpha / K) \tag{8}$$

中采样 $\boldsymbol{\pi}$，从分布 $NIW\left(\phi_k \mid \Phi_{-k} Z, D\right)$ 中采样 ϕ_k。遍历一遍上述过程，实现对 $\boldsymbol{\pi}, Z, \Phi$ 向量的一次采样。重复上述过程，直到采样收敛，利用收敛之后的样本来估计参数的取值。**由于参数 $\boldsymbol{\pi}$ 和参数 Z 的分布具有共轭的关系**，所以也可以把参数 $\boldsymbol{\pi}$ 通过积分消除，得到边缘分布 $p\left(Z, \Phi \mid D\right)$，然后就可以利用 Gibbs 采样对该边缘后验概率分布进行采样，即坍缩 Gibbs 采样（Collapsed Gibbs Sampling）。显然在得到 Z 的采样样本之后，由于存在共轭关系，通过简单的统计计算，就能够得到参数 $\boldsymbol{\pi}$ 的分布参数：$p(\boldsymbol{\pi} \mid Z) \propto p(Z \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}) = Dirichlet(n_1 + \alpha / K, ..., n_K + \alpha / K)$。

2.3 贝叶斯非参数模型

贝叶斯参数模型存在一个问题，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。我们可以用狄利克雷过程作为聚类个数 K 的先验分布。我们独立成立一节，来介绍一个典型的贝叶斯非参数模型：狄利克雷过程混合模型。

3. 狄利克雷过程混合模型

3.1 狄利克雷过程

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。聚焦参数的作用就是来指定分布的离散程度，极限情况下，当$\alpha \to 0$时，对应的分布是一个单点分布，当$\alpha \to \infty$时，对应的分布趋近于连续分布。

遗憾的是我们一般并不能显示的写出狄利克雷过程任意样本G的表达式，但是我们可以给出分布G在可测空间Θ（G为定义在可空间Θ上的概率测度）的有限划分上的概率分布值。假设$(A_1,\cdots,A_m)$是可空间Θ的的一个划分，则：

$$\left(G(A_1),\cdots,G(A_m)\right) \sim Dirichlet\left(\alpha G_0(A_1),\cdots,\ \alpha G_0(A_m)\right) \tag{9}$$

即分布函数虽无有解析的表达式，但是它在这些划分上的概率分布服从狄利克雷分布，分布参数由划分集合$(A_1,\cdots,A_m)$、基分布G_0和聚焦参数α共同决定。并且有$E[G(A)]=G_0(A)$，$Var[G(A)]=G_0(A)(1-G_0(A))/(\alpha+1)$。

假设按照狄利克雷过程$DP(G_0,\alpha)$采样一个分布G（一个离散的随机分布），那么我们按照分布G对可测空间Θ进行采样，得到独立同分布的样本$\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_n \sim G$，则在给定这些样本的条件下，$G$的后验概率分布仍然是一个狄利克雷过程：

$$G\,|\,\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_n \sim DP\left(\alpha+n,\frac{\alpha}{\alpha+n}G_0+\frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{\boldsymbol{\theta}_i}\right) \tag{10}$$

由于狄利克雷过程的采样一般是离散分布，G 依概率 1 具有以下形式：

$$G=\sum_{k=1}^{\infty}\pi_k\delta_{\boldsymbol{\theta}_k} \tag{11}$$

因而狄利克雷过程可以用来聚类，如图 5 所示为使用了狄利克雷过程作为先验分布的贝叶斯参数聚类模型的概率图表示。

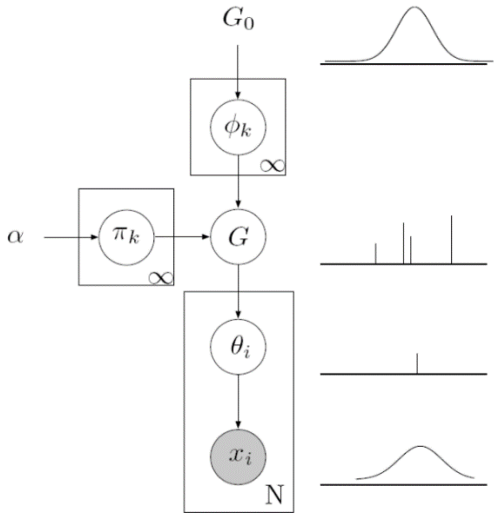

图 5：DP 用来聚类的生成模型

由聚焦参数α和基分布G_0决定的 *DP* 过程生成一个随机分布 G，作为样本数据参数服从的先验分布。**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**。

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）。

3.2 折断筷子模型

假定给定聚焦参数α和基分布G_0，我们可以按照如下过程构造一个狄利克雷过程：

$$\pi_k^{'} \mid \alpha_0, G_0 \sim Beta(1, \alpha_0) \qquad \phi_k \mid \alpha_0, G_0 \sim G_0 \tag{12}$$

根据以上采样得到的随机参数，构造如下概率测度：

$$\pi_k = \pi_k^{'} \prod_{l=1}^{k-1} \left(1 - \pi_l^{'}\right) \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \tag{13}$$

值得注意的是依概率 1，$\sum_{k=1}^{\infty} \pi_k = 1$；上式左式也可以写成$\boldsymbol{\pi} \mid \alpha_0 \sim GEM\left(\alpha_0\right)$。参数$\pi_k$和$\phi_k$根据参数$\alpha$和基分布$G_0$采样得到，所以分布$G$对应着狄利克雷过程$DP(\alpha, G_0)$的一个采样样本。

如果以折断筷子模型表示狄利克雷过程，引入示性变量z_i来表示样本 i 属于

哪一个聚类，则根据图 5 所示的模型，则以该狄利克雷过程为先验分布的非参数混合模型可以表示为：

$$\boldsymbol{\pi} \mid \alpha_0 \sim GEM\left(\alpha_0\right) \qquad z_i \mid \boldsymbol{\pi} \sim \boldsymbol{\pi} \tag{14}$$

$$\phi_k \mid G_0 \sim G_0 \qquad x_i \mid z_i, \left(\phi_k\right)_{k=1}^{\infty} \sim F\left(\phi_{z_i}\right) \tag{15}$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \qquad \theta_i = \phi_{z_i} \tag{16}$$

3.3 Blackwell-MacQueen 罐子模型

我们假定已经从狄利克雷过程 $DP(\alpha, G_0)$ 中采样得到了一个分布 G，在给定 G 的情况下我们采样一组独立同分布数据 $\theta_1, \theta_2, ... \mid G \sim G$，Blackwell-MacQueen 证明了，如果我们通过积分把分布 G 消除，则条件分布具有如下形式：

$$\theta_i \mid \theta_1, ..., \theta_{i-1}, \alpha, G_0 \sim \sum_{l=1}^{i-1} \frac{1}{i-1+\alpha} \delta_{\theta_l} + \frac{\alpha}{i-1+\alpha} G_0 \tag{17}$$

上式等号右侧的 G_0 表示按照基分布 G_0 采样得到的一个新的样本。上式的物理意义非常明确，可以解释成罐子模型：以正比于 $i-1$ 的概率从罐子内等概抽球，抽到球之后放回，并同时放回一个颜色相同的球；以正比于 α 的概率从分布 G_0 中产生一个新颜色的球，放入罐子中。

如果我们把 $\theta_1, ..., \theta_{i-1}$ 中所有不同的取值记为 $\phi_1, ... \phi_K$，则上式可以重新写成：

$$\theta_i \mid \theta_1, ..., \theta_{i-1}, \alpha, G_0 \sim \sum_{k=1}^{K} \frac{m_k}{i-1+\alpha} \delta_{\phi_k} + \frac{\alpha}{i-1+\alpha} G_0 \tag{18}$$

其中 m_k 表示属于第 k 个取值的样本个数。

3.4 中国餐厅过程

中国餐厅过程（Chinese Restaurant Process, CRP）是另一个表示狄利克雷过程有效方式，也是 Blackwell-MacQueen 罐子模型的另一种表达方式。基本流程是：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。

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 5 所示的模型，其对应的数学描述如下：

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

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 服从的分布为 $F\left(\boldsymbol{\theta}_i\right)$；因此任意样本由混合概率模型 $F\left(\bullet\right)=\sum_{k=1}^{\infty}\pi_k F\left(\bullet\mid\delta_{\boldsymbol{\theta}_k}\right)$ 产生（k 表示聚类的下标，由中国餐厅过程中的分析可知由 n 个样本决定的聚类个数 k_n 一般小于 n，因此求和不需要到正无穷）。

3.5 狄利克雷过程混合模型的参数学习

我们以中国餐桌过程表示狄利克雷过程，并引入示性参数 z_i 来表示样本 x_i

所属的聚类，参数 ϕ_k 表示聚类 k 的分布参数。其中 $p(Z)$ 由中国餐桌过程定义，$p(\Phi)$ 为分布参数的先验分布，可以取共轭分布， $p(D|Z,\Phi)$ 为似然函数。不妨假设在每一个聚类当中，数据 $\mathbf{x}_i$ 都服从高斯分布 $p(\mathbf{x}_i|\Phi,Z)\sim N(\mu_{z_i},\Sigma_{z_i})$，简单起见假设所有聚类的协方差矩阵 Σ_k 相同，且等于单位阵 I，区别在于均值 $\boldsymbol{\mu}_k$ 不同。共轭分布 $p(\phi_k)\sim N(0,\sigma^2 I)$，简单起见可以假设 $\sigma^2=1$。对于 $p(Z)$，可以用中国餐桌过程表示，则有：

$$p(z_i=k|Z_{-i})=\frac{n_k}{n-1+\alpha},\ if\ k\in\{1,...,K\} \tag{20}$$

$$p(z_i=K+1|Z_{-i})=\frac{\alpha}{n-1+\alpha} \tag{21}$$

其中 K 表示当前聚类的个数，$n_k=\sum_{j\in\{1,...,n\}\backslash\{i\}}\delta_{z_j,k}$ 表示属于第 k 个聚类的数据个数。

应用 Gibbs 采样学习参数 Z,Φ。首先随机初始化 Z,Φ（Z 的初始化涉及到聚类个数，也是随机的），然后对变量 z_i 进行采样，假设当前聚类为 $\{1,...,K\}$，则有（注意 z_i 之间不独立）：

$$\begin{aligned}p(z_i=k|D,\Phi,Z_{-i})&\propto p(z_i=k|Z_{-i})p(\mathbf{x}_i|Z,\Phi)\\&=\frac{n_k}{n-1+\alpha}\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\}\end{aligned} \tag{22}$$

$$\begin{aligned}p(z_i=K+1|D,\Phi,Z_{-i})&\propto p(z_i=K+1|Z_{-i})p(\mathbf{x}_i|Z,\Phi)\\&=p(z_i=K+1|Z_{-i})\int p(\phi_{K+1})p(\mathbf{x}_i|z_i,\Phi,\phi_{K+1})d\phi_{K+1}\\&=\frac{\alpha}{n-1+\alpha}\frac{1}{(2\pi)^{d/2}\sigma^d}\frac{|\Sigma'|^{1/2}}{|\Sigma|^{1/2}}\exp\{\frac{1}{2}\mathbf{x}_i^T(\Sigma^{-1}\Sigma'\Sigma^{-1}-\Sigma^{-1})\mathbf{x}_i\}\end{aligned} \tag{23}$$

其中 $\Sigma'=(1/\sigma^2+\Sigma^{-1})^{-1}$。得到参数 Z 的采样之后再对参数 Φ 进行采样，有：

$$
\begin{aligned}
p(\phi_k \mid D, Z, \Phi_{-k}) &\propto p(D \mid Z, \Phi)\, p(Z, \phi_k, \Phi_{-k}) \\
&= p(D \mid Z, \Phi)\, p(\phi_k)\, p(Z, \Phi_{-k} \mid \phi_k) \\
&\propto p(\phi_k) \prod_{i=1}^{n} p(\mathbf{x}_i \mid Z, \Phi) = p(\phi_k) \prod_{i \mid z_i = k}^{n} p(\mathbf{x}_i \mid Z, \Phi) \\
&\propto \exp\left\{ -\frac{1}{2\sigma^2} \boldsymbol{\mu}_k{}^T \boldsymbol{\mu}_k - \frac{1}{2} \sum_{i \mid z_i = k} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \\
&\sim N\left(\boldsymbol{\mu}_k^{'}, \Sigma_k^{'} \right)
\end{aligned}
\tag{24}
$$

上式第二行最后一项可以忽略是因为条件概率分布和 ϕ_k 无关，即 $p(Z, \Phi_{-k} \mid \phi_k) = p(Z, \Phi_{-k})$。