Dong Wang

现代机器学习技术导论

2018 年 1 月 29 日

Springer

Contents

Chapter 1
机器学习概述

Chapter 2
线性模型

Chapter 3
神经模型

Chapter 4
深度学习

Chapter 5
核方法

Chapter 6
图模型

Chapter 7
非监督学习

Chapter 8
非参数模型

在前面章节中，我们已经讨论了各种模型，这些模型有些是线性的，有些是非线性的，有些是神经网络的，有些是概率图的，有些是描述性的，有些是区分性的，有些是监督学习的，有些是非监督学习的... 这些模型的构造过程都遵循如下步骤：设计好一个模型形式M，收集训练数据集D，依某种优化方法对M进行优化，使之对某一目标函数L最大化（或某一损失函数C最小化）。特别的，绝大多数模型是由一组参数P决定的，如在概率图模型中概率函数的尺度变换参数，神经模型中的连接权重等。因此，参数P决定了模型M，对模型M进行优化的过程即是对P的选择过程。这种模型参数化极大简化的学习过程：它相当于预先定义了一个知识表达形式，并用可学习的参数确定这一形式的最终表达。这事实上提供了一种将先验知识（模型形式）和后验学习相结合的方法。被固定的参数完全定义的模型称为‘参数模型’。参数模型的一个重要特点是：这一模型的知识表达形式是确定的，因此模型的规模也是确定的，不会随着训练数据量的多少改变而改变。

参数模型的优势在于其对知识的抽象能力，但这一方法也存在一些问题。例如，当先验知识不足时，对模型参数形式的设计可能是不合理的；其次，当训练数据较丰富时，我们希望模型可以相应增大规模，以描述更多细节，而参数模型不具有这种扩展能力；最后，训练数据在不同区域分布可能是不均衡的，我们希望在训练数据较多的区域有更好的模型，但参数模型通常是全局的，无法实现依数据分布情况的合理调节。总而言之，在一些建模任务中，我们希望模型由训练数据本身来确定，而不是预先设计的固定形式。这种数据驱动的建模方式称为‘非参数模型’（No-Parametric Model）。

本章我们将讨论非参数模型的基本概念，并集中讨论两种非参数模型：高斯过程和狄利克雷过程。这两种模型基于概率图模型，通过设计无限维空

间上的先验概率，实现依训练数据规模对模型复杂度进行调整，同时保证这种调整不致因过度依赖数据而产生太大偏差。

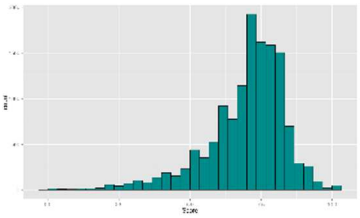
8.1 简单非参数模型

直观地说，非参数模型是不被参数形式限制的模型。例如当我们统计一个一维数据的分布规律时，经常会假设一个高斯分布，通过拟合这一分布的均值和方差对数据的实际分布进行近似。这一方法的优点在于引入了一个高斯假设，因而只要少数几个样本点就可以得到一个很好的估计，但当实际数据本身并不是高斯分布时，由于模型本身形式已经固定（高斯的），因此再多训练数据也无法提高模型的描述能力。换一个思路，如果我们不假设高斯分布，而是在数轴上设计若干个区间，并统计在这些区间上的数据分布比例，形成如图 8.1（a）所示的直方图模型。当训练数据较少时，直方图模型给出的概率显然是粗糙的，但当数据量增多时，通过对区间进行细致划分，直方图模型可以完美逼近真实数据分布。注意，直方图模型的复杂度是随着数据量增长而增加的：更多数据支持更细致的区间划分，使得模型的规模增大，从而提供更细致的数据分布描述。

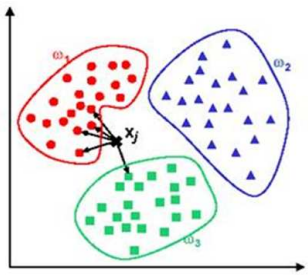
K近邻（K-Nearest Neighbour, KNN）是另一种用于分类的非参数模型。这一方法的基本假设是数据空间的分类是有连续性的，因而一个待考察点周围的点应具有相似的分类。因此，可以通过考察目标点周围的训练样本所属的类别，通过投票等方式来决定待考察目标点的类别。在KNN中，选择和目标点最近的K个训练样，用这些样本的分类来判断目标点的分类，如图 8.1（b）所示。和典型的参数模型（如GMM）相比，KNN不对数据分布做任何假设。这使得该方法在训练数据较少时性能较低，但当数据量增大，可以充满测试数据所在的空间时，KNN会超过任何一种参数模型。这是因为任何参数模型都对数据的性质做了某些假设（高斯或线性等），这些假设在数据量有限时是一种有价值的先验，但当数据足够多时，将成为模型表达能力的限制。

另一种典型的非参数模型是支持向量机（SVM）。SVM在再生核希尔伯特空间设计一个线性分类器，该分类器不是以参数形式设定的，而是通过保存训练数据中的支持向量来实现。基于这些支持向量，计算未知数据到这些支持向量的距离（用核函数表示），以这些距离为权重对每个支持向量的分类标记进行平均，即得到未知数据的分类标记。和KNN类似，当训练数据较多时，特别是分类面处的数据较多时，SVM中包含的支持向量数会显著增

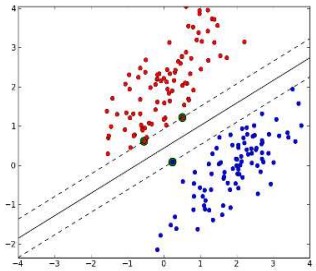
长。增长的支持向量集合可以提供更细节的分类面，从而提高模型的表达能力和分类能力。



(a)



(b)



(c)

Fig. 8.1 几种典型的非参数模型。(a)：直方图模型；(b)：KNN 模型；(c)：SVM模型。

通过上述三个典型非参数模型，我们看到这类模型具有如下特征：（1）对数据的分布状态不做过强的假设，让数据自己表达；（2）保留全部或部分训练数据进行推理，因此模型随训练数据的增长而增长。这两点显然是相关的，正因为不做过强假设，因此需要保留训练集来做推理。

如果将参模型看作抽象学习，非参数模型更接近联想学习；前者可以看作是对知识的抽象，后者更多是对经验的记忆。和参数模型相比，非参数模型多用在相对复杂的任务中。在这些任务中，人们的知识相对缺乏，无法设计合理的参数形式，但数据量相对较大。这时可利用非参数模型，将这些数据记录下来，用联想方式进行推理。

非参数模型并非没有参数，也并非没有受限形式。例如SVM，对数据的表达能力依赖核函数的选择，也依赖目标函数中错分样本的权重参数。非参数模型是有参数的，只不过很大一部分参数由训练数据决定，如SVM中的支

持向量。同时，一个模型也未必是全部非参数的，可以是参数模型和非参数模型的混合。SVM即是这种混合模型，这一模型在核函数映射部分是参数的，但在支持向量学习时是非参数的。

朴素的非参数模型（如直方图或KNN）只能依靠训练数据的增长来提高预测能力。SVM将参数模型和非参数模型结合在一起，因而不再单纯依靠数据覆盖度。事实上，非参数模型在现代机器学习中的意义不在于取代参数模型做简单的联想学习，而在于帮助参数模型打破形式化限制，使得当数据量增长的时候可以自动调整模型复杂度，以提高模型的精度。贝叶斯方法提供了一种非常优美的框架，在这一框架中可以通过设计非参数的先验概率来打破传统贝叶斯模型的固化形式，实现模型复杂度随数据的自动调整。这一方法称为贝叶斯非参数模型方法（Bayesian No-Parametric），是本章要介绍的重点内容。

下面我们将讨论两种贝叶斯非参数模型，一种用于预测任务，另一种用于聚类任务。对预测任务，传统贝叶斯参数模型方法定义一个参数结构（如线性回归），并对参数赋以一个高斯先验，因此所有先验都是基于该参数结构所代表的映射函数集的先验。高斯过程定义了另一种先验，这些先验不是基于某种模型形式的参数上的先验，而是某类映射函数的先验 [9]。这类映射函数只需符合非常宽泛的假设，因而比传统参数形式具有更广泛的覆盖性。高斯过程在第 5章中已经有讨论，当时我们关注的是高斯过程中的核函数意义，本章我们将着重讨论高斯过程的非参数性质。

对于聚类任务，非参数模型将帮助我们解决聚类中的一个重要问题，即聚类数的不确定性。绝大多数聚类算法都需要事先定义好聚类数，或设定某些相关参数（如AF中相关矩阵对角值的相对大小，DBSCAN中的到达半径）。事实上，我们应该让数据自己决定应该聚成几类，特别是当数据量较大时，算法应该有能力聚出更多类。狄利克雷过程通过对所有可能的聚类方式赋予一个先验概率来解决这一问题 [13]。通过给定这一先验，应用概率图模型的推理方法即可得到数据应该如何聚类的后验概率。我们从高斯过程开始讨论。

8.2 高斯过程

第 5章中我们已经讨论过高斯过程。本节对这一方法做一回顾，并着重强调其非参数模型的意义。

8.2.1 高斯过程

考察如下线性高斯模型

$$y = \mathbf{w} \cdot \mathbf{x}; \qquad \mathbf{w} \sim N(\mathbf{0}, \alpha^{-1} I).$$

因为$\mathbf{w}$具有随机性，因而由$\mathbf{x}$到y的映射函数$y(\mathbf{x})$是随机的。我们可以用另一种非参数形式来表达这种随机性。考虑任意一个点集$\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]$和对应的预测$\mathbf{y} = [y_1, ..., y_N]^T$，由于$\mathbf{w}$ 是高斯的，因此$\mathbf{y}$ 也是一个高斯分布。且有:

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\mathbb{E}(\mathbf{w}) = \mathbf{0},$$

$$\mathrm{cov}(\mathbf{y}) = \mathbb{E}(\mathbf{y}^T \cdot \mathbf{y}) = \mathbf{X}^T \mathbb{E}(\mathbf{w} \cdot \mathbf{w})\mathbf{X} = \alpha^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{K},$$

其中$\mathbf{K}$为Gram矩阵，且：

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \alpha^{-1}\mathbf{x}_i \cdot \mathbf{x}_j.$$

从K的表达式我们可以看出，任意两个样本点$\mathbf{x}_i$和$\mathbf{x}_j$的预测值y_i和y_j是相关的，其协方差取决于$\mathbf{x}_i$和$\mathbf{x}_j$ 之间以$k(\mathbf{x}_i, \mathbf{x}_j)$描述的‘距离’：这两个样本离得越近，他们对应的预测点$y(\mathbf{x}_i), y(\mathbf{x}_j)$越相关。

通过上述讨论，我们看到一个以随机变量$\mathbf{w}$为参数的随机函数$y(\mathbf{x})$可以表达为这一函数在任意点集X上取值的分布规律。如果$\mathbf{w}$为高斯分布，则对应的随机函数在任意点集上的取值亦符合高斯分布，且这一分布的性质由协方差函数$k(\mathbf{x}_i, \mathbf{x}_j)$ 描述。注意上述分布性质在任意点集上都成立，这事实上定义了一个随机过程，称为*高斯过程* [16, 9]。

具体而言，高斯过程是定义在映射函数$y(\mathbf{x})$上的一个概率分布（即映射函数的概率分布），该概率分布使得在任意有限点集$\mathbf{x}_1, \cdots, \mathbf{x}_N$处计算的$y(\mathbf{x})$服从如下多元高斯分布:

$$\begin{bmatrix} y(\mathbf{x}_1) \\ \vdots \\ y(\mathbf{x}_N) \end{bmatrix} \sim N\left(\begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_N) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix} \right),$$

其中$m(\cdot)$代表随机函数的均值函数，$k(\cdot, \cdot)$代表随机函数的协方差函数。记上述高斯过程为$y(\cdot) \sim G(m(\cdot), k(\cdot, \cdot))$。

换句话说，高斯过程本质上是一个随机函数，描述函数的不确定性。然而，将随机函数表达成随机变量的话将有无限维（每一个$\mathbf{x}_i$看作一维）。为了

描述这一随机函数，我们考察其中任意有限维空间上的取值并确认其具有一致的高斯分布特性，即不论$\mathbf{x}_1,\cdots,\mathbf{x}_N$如何选择，其所对应的$y(\mathbf{x}_1),...,y(\mathbf{x}_N)$ 都应满足由均值函数$m(\cdot)$和协方差函数$k(\cdot,\cdot)$定义的多元高斯分布。事实上，更广义的随机过程是由这种一致性定义的：一个由给定的联合概率分布定义的随机过程$y(\mathbf{x})$ 是指对任意有限个取值$y(\mathbf{x}_1),\cdots,y(\mathbf{x}_N)$，其联合分布都一致地服从某一给定的分布。

在前述线性高斯模型的例子中，对应的协方差函数$k(\mathbf{x}_j,\mathbf{x}_j)$为$\alpha^{-1}\mathbf{x}_i\cdot\mathbf{x}_j$。事实上，如果我们将模型取为更复杂的形式，如$y(\mathbf{x})=\mathbf{w}\cdot\boldsymbol{\phi}(\mathbf{x})$，其中$\boldsymbol{\phi}$为映射函数，则有$k(\mathbf{x}_i,\mathbf{x}_j)=\alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_i)\cdot\boldsymbol{\phi}(\mathbf{x}_j)$。在第 5 章中讲过，我们可以直接定义$k(\mathbf{x}_i,\mathbf{x}_j)$，使之对应复杂的特征映射，实现更复杂的相关性建模（如在原始空间中不满足高斯过程的任务）。这事实上即是我们在第 5章中所讨论的核函数方法。进一步，我们可以不关心$y(\mathbf{x})$的具体形式，只需定义$m(\cdot)$和$k(\cdot,\cdot)$即可完整定义高斯过程的性质。这意味着高斯过程并不局限于一个明确定义的预测函数，它是由$m(\cdot)$和$k(\cdot,\cdot)$定义的非常广泛的一类随机函数。一般取$m(\cdot)=0$，这时高斯过程由协方差函数$k(\cdot,\cdot)$唯一确定。

在讨论高斯过程的应用之前，我们首先对高斯过程有个直观印象。注意高斯过程是函数的概率分布，或随机函数，其每一个采样是一个确定的函数。基于高斯过程的定义，我们可以通过采样得到这些函数。首先确定协方差函数$k(\cdot,\cdot)$，然后确定若干个函数取值点X，高斯过程保证在这些点上的取值$\mathbf{y}$符合高斯分布，其协方差矩阵由$k(\cdot,\cdot)$确定。由此，对高斯过程的采样转化为对$\mathbf{y}$的采样。由于$p(\mathbf{y})$已知，原则上这一采样不存在困难。一种比较简单的方式是利用高斯过程的性质，从y_1开始，依次对y_2,y_3,...采样。当对y_i进行采样时，将既有采样值$\{y_1,y_2,...,y_{i-1}\}$作为条件，即依$p(y_i|y_1,...,y_{i-1})$中进行采样。由于$p(y_1,...,y_i)$是高斯的，$p(y_i|y_1,...,y_{i-1})$也是一个高斯分布，因此采样很容易实现。

图8.2给出从两个具有不同协方差函数的高斯过程采样得到的两簇函数。可见，不同协方差函数决定了采样函数的性质。一般来说，协方差越大，则不同样本间的相关性越强，函数取值越倾向于一致，表现出来函数的曲线越平坦。反之，协方差越小，函数曲线越不受相关性影响，表现的越起伏。

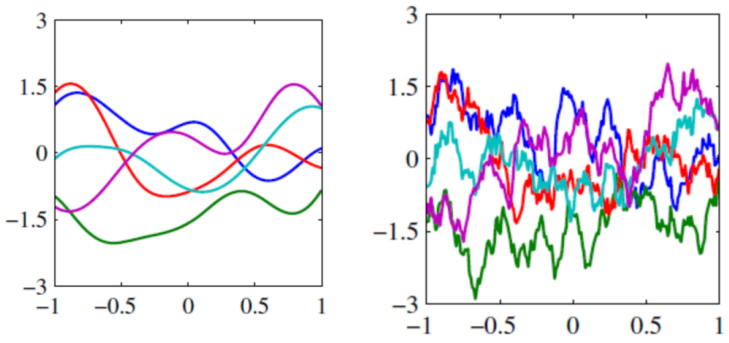


Fig. 8.2 从两个不同协方差函数的高斯过程采样出的两簇函数。图片来源于 [1]，Figure 6.4。

8.2.2 高斯过程回归

在上一小节中，我们提到高斯过程描述了定义在映射函数上的概率分布，即随机函数。在本小节中，我们将把这一随机函数作为先验应用到贝叶斯回归分析中。

8.2.2.1 贝叶斯线性回归

首先简要回顾贝叶斯线性回归模型。该模型假设数据$\mathbf{x}$的目标值t之间具有如下关系：

$$t = \mathbf{w} \cdot \mathbf{x} + \varepsilon \tag{8.1}$$

其中$\mathbf{w}$是模型参数，该参数是一个高斯随机变量，满足：

$$\mathbf{w} \sim N(\mathbf{0}, \alpha^{-1}\mathbf{I}).$$

ε是一个高斯噪声，满足:

$$\varepsilon \sim N(0, \beta^{-1}).$$

设训练数据集$\{(\mathbf{x}_i, t_i); i = 1, \cdots, N\}$。记$\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]$，对应的标记为$\mathbf{t} = [t_1, t_2, ..., t_N]^T$。利用贝叶斯公式，可以得到参数$\mathbf{w}$ 的后验概率形式为一个高斯分布：

$$\mathbf{w}|\mathbf{X},\mathbf{t} \sim N(\mathbf{m}_N, \mathbf{S}_N),$$

其中，

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{X}\mathbf{X}^T,$$

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{X}\mathbf{t}.$$

对任意一个测试数据$\mathbf{x}_*$，可以计算其输出t_* 的概率分布。因为$\mathbf{w}$具有随机性，因此需要对所有可能的$\mathbf{w}$做边缘化：

$$p(t_*|\mathbf{x}_*,\mathbf{t},\mathbf{X}) = \int_{\mathbf{w}} p(t_*|\mathbf{x}_*,\mathbf{w})p(\mathbf{w}|\mathbf{t},\mathbf{X}).$$

注意上式右侧积分中两个概率分布都是高斯的，因此t_*的分布也是一个高斯分布:

$$t_*|\mathbf{x}_*,\mathbf{X},\mathbf{t} \sim N(\mathbf{m}_N \cdot \mathbf{x}_*, \sigma_N^2(\mathbf{x}_*)),$$

其中

$$\sigma_N^2(\mathbf{x}_*) = \beta^{-1} + \mathbf{x}_*^T\mathbf{S}_N\mathbf{x}_*.$$

上述贝叶斯线性回归是典型的参数模型：设计一个线性模型形式，定义模型中每个变量的随机性，基于一个训练集，得到参数的后验概率，由此完成模型训练。训练完成后，得到的模型即是对训练数据中知识的抽象，基于该模型即可进行推理，原来的训练数据可以丢弃。注意，这一知识抽象基于预先设计的模型形式，这一形式不会因训练数据的多少而改变。

8.2.2.2 高斯过程回归建模

现在我们从另一个角度来理解贝叶斯线性回归。由上节对高斯过程的讨论可知，式 8.1中的回归函数$y = \mathbf{w}\cdot\mathbf{x}$是一个高斯过程，因而对训练数据中的样本点$\{\mathbf{x}_i\}$，其预测值$\{y(\mathbf{x}_i)\}$ 的联合分布是一个高斯分布$G(\mathbf{0},\mathbf{K})$，其中$\mathbf{K} = \alpha^{-1}\mathbf{X}^T\mathbf{X}$)。加入一个随机高斯噪声$\varepsilon$后，$t(\mathbf{x})$依然是一个高斯过程。对训练样本集$\mathbf{X}$及其对应的观测值$\mathbf{t}$，容易验证$\mathbf{t}$是一个多元高斯变量，且：

$$\mathbb{E}(\mathbf{t}) = \mathbf{0},$$

$$cov(\mathbf{t}) = \mathbf{K} + \beta^{-1}\mathbf{I}.$$

和$y(\mathbf{x})$相比，$t(\mathbf{x})$是一个协方差更大的高斯过程，但这一协方差的增加仅表现在对角元素上，说明加入噪声仅增加了预测时的不确定性，但不改变不同样本点处的相关性。

特别重要的是，上式中对$\mathbf{K}$的定义形式完全来源于式8.1中的线性形式$\mathbf{w}\cdot\mathbf{x}$以及对$\mathbf{w}$的先验概率。事实上，我们完全可以抛开这些限制，让$\mathbf{K}$自由定义一个高斯过程$y(\mathbf{x})$，用预测函数本身的随机性（更确切地说，高斯随机性）来代替基于先验和模型形式衍生出的随机性，这将极大摆脱线性模型定义的形式束缚。

8.2.2.3 高斯过程回归预测

因为高斯过程的性质由其采样点上取值的分布特性（联合高斯分布）决定，因此对新采样点的预测也由训练数据的采样点取值决定。具体而言，如果给定训练数据集$\{(\mathbf{x}_i, t_i); i = 1, \cdots, N\}$，对一个测试数据$\mathbf{x}_*$，其预测值为$t_*$。依高斯过程定义，联合向量$[\mathbf{t}^T \quad t]$符合如下高斯分布：

$$\begin{bmatrix} \mathbf{t} \\ t_* \end{bmatrix} | \mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{x}_* \sim N(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \beta^{-1}\mathbf{I} & \mathbf{k} \\ \mathbf{k}^T & c + \beta^{-1} \end{bmatrix})$$

其中$\mathbf{K} = cov(\mathbf{y})$，$\mathbf{k} \in R^{N\times 1}$的元素为:

$$k_i = \alpha^{-1}\mathbf{x}_i \cdot \mathbf{x}_*, \quad i = 1, 2, \cdots, N$$

$$c = \alpha^{-1}\mathbf{x}_* \cdot \mathbf{x}_*$$

由这一联合高斯分布形式可求t_*的后验概率如下：

$$t_* | \mathbf{t}, \mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{x}_* \sim N(m(\mathbf{x}_{N+1}), \sigma^2(\mathbf{x}_*)) \tag{8.2}$$

其中:

$$m(\mathbf{x}_{N+1}) = \mathbf{k}^T(\mathbf{K} + \beta^{-1}\mathbf{I})^{-1}\mathbf{t},$$

$$\sigma^2(\mathbf{x}_*) = c - \mathbf{k}^T(\mathbf{K} + \beta^{-1}\mathbf{I})^{-1}\mathbf{k}.$$

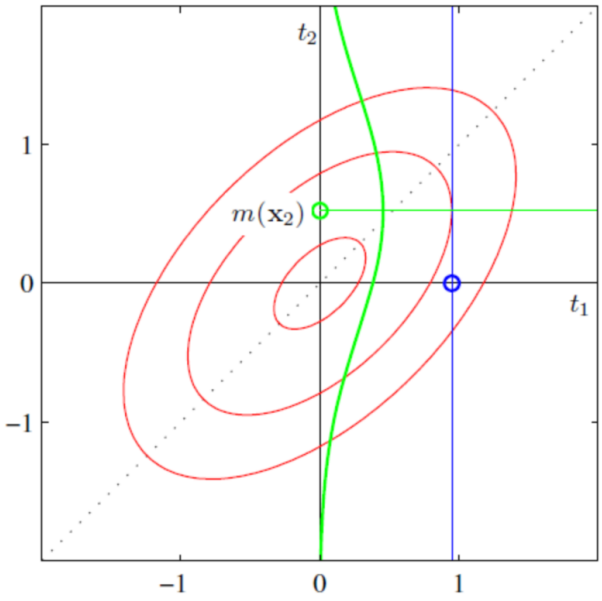


Fig. 8.3 高斯过程回归的后验预测概率分布。$(\mathbf{x}_1, t_1)$为训练数据，$(\mathbf{x}_2)$为预测数据，t_2为预测数据取值。红色等概率曲线表示t_1, t_2的联合概率分布，绿色等概率曲线表示给定t_1后t_2的后验概率分布。图片来源于 [1]，Figure 6.7.

图8.3给出上述高斯过程回归的预测概率分布。其中t_1为训练数据输出，t_2为测试数据的输出。依高斯过程，我们知道(t_1, t_2)的联合分布$p(t_1, t_2)$服从二元高斯分布，如图中红色椭圆等概率曲线所示。在给定具体的t_1值（对应于图中蓝色圈）的情况下，依式8.2可知，t_2的后验概率$p(t_2|t_1)$服从另一高斯分布，如图中绿线所示，其中分布的均值$m(\mathbf{x}_2)$（即出现概率最大的t_2的取值）为红色等高线$p(t_1, t_2)$与取值为t_1的直线相切点对应的t_2 的值。

图 8.4给出预测的置信度，其中绿线为一个产生数据的正弦函数，蓝色点为基于这一函数产生并加入高斯噪声的数据。红色曲线表示以这些点为观测数据，基于高斯过程回归建模得到的后验概率（用于预测）的均值函数$m(\mathbf{x})$，阴影部分的宽度表示每个取值点$\mathbf{x}$ 上的预测方差$\pm 2\sigma(\mathbf{x})$。可以看出，在训练数据比较多的地方，模型的置信程度更高。

仔细考察上述基于高斯过程的预测方法，我们可以发现这一方法与前面所述的贝叶斯线性回归有本质不同。首先，我们并未定义$t(\mathbf{x})$的具体形式，

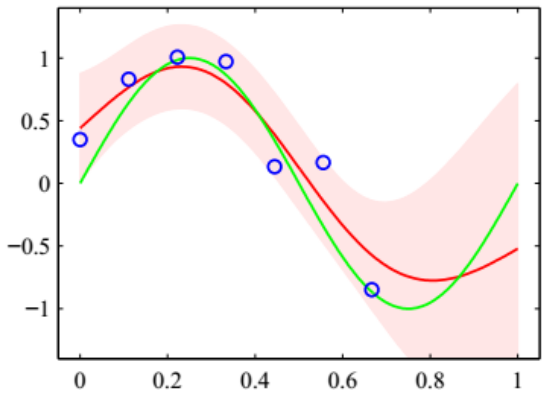


Fig. 8.4 高斯过程回归的置信区间，其中绿线为真实映射函数，蓝点为从这一映射函数采样并加入高斯噪声后的训练数据，红色曲线为后验概率的均值函数，阴影部分的宽度表示预测方差$\pm 2\sigma(\mathbf{x})$。图片来源于 [1]，Figure 6.8.

只是确认了$y(\mathbf{x})$ 是一个以$k(\cdot,\cdot)$为协方差函数的高斯过程；第二，在对测试样本进行预测时，高斯过程回归利用了训练集中的所有数据。这意味着训练数据越多，模型的复杂度越高。因此，高斯过程回归显然是一个无参数模型。

事实上，我们可以将任何模型设计过程认为是一种先验预设过程，从而确定了一个有效函数空间。例如线性回归模型确定了预测函数只能是线性的，且只能有一个函数；贝叶斯线性回归允许多个预测函数，这些函数依概率进行预测，但这些预测函数必须是线性的。高斯过程提供了另外一种先验，这种先验允许所有符合高斯分布特性的预测函数参与，并为这些函数赋予不同的先验概率。这些函数有可能是线性的，也可能不是线性的，有可能是局部的，也有可能是全局的。高斯过程给这些广泛的可选函数赋予如下先验，即依这一先验，对任意有限样本点，这些函数依此概率在这一样本点上进行取值，这些取值一致地符合确定的高斯分布即可。基于这一先验分布，给定训练数据，即可得到在这些函数上的后验概率，并在预测时基于这一后验概率对所有这些函数进行边缘化。乍一看这似乎是不可能的，因为这样的函数有无穷多个且形式未知，因而无法对它们进行边缘化。然而，我们可以把函数的概率转化为特定样本上取值的联合概率，因为这些样本点（训练样本和测试样本）是有限的，因而可实现前面所述的预测过程。

8.2.3 高斯过程用于分类任务

高斯过程的本质是对提供一簇函数的先验概率，基于这一先验概率，可实现对任何预测任务的非参数贝叶斯学习。前面介绍的高斯过程回归是一个典型的例子，同样的方法也可用于分类任务。在分类任务中，预测函数为

$$y = \sigma(\mathbf{w} \cdot \mathbf{x}),$$

其中σ 为一个Sigmoid函数。和回归任务类似，我们希望高斯过程给预测函数$y(\mathbf{x})$赋予一个先验概率，使得分类任务摆脱线性形式的束缚。然而，这一思路在实现时遇到一个问题，即yy的取值是受限的，只能取值在$[0,1]$区间，而高斯过程中假设采样点取值是高斯的，因而取值是没有限制的。因此，直接在$y(\mathbf{x})$上设计高斯过程先验并不合理。一个解决方法是将分类任务的预测函数分成两部分：

$$y = \sigma(a(\mathbf{x})),$$

并在$a(\mathbf{x})$上设计高斯过程先验，即对任何一组数据$\mathbf{X}$，有：

$$[a_1, ..., a_N] \sim N(\mathbf{0}, \mathbf{K}).$$

预测时，测试数据$\mathbf{x}_*$ 的预测值t_*的后验预测分布形式为:

$$p(t_*|t_1, ..., t_N) = \int_{a_*} p(t_*|a_*) p(a_*|t_1, ..., t_N) \mathrm{d}a_*,$$

上式中后验概率部分$p(a_*|t_1, ..., t_N)$无法直接求出，一般通过变分推断或者Laplace近似给出近似解。关于Laplace近似方法的细节，可参考 [1]第六章6.4.6节。

8.3 狄利克雷过程

高斯过程$G(m,k)$事实上定义了一个在函数上的先验概率，这一先验概率可用在预测任务建模上，包括回归任务和分类任务，这些任都属于监督学习范畴。对于非监督学习，特别是聚类任务，我们需要定义另一种先验。

在聚类任务中，我们一般要定义一个贝叶斯生成模型，并确定模型中的聚类数K，对该模型参数进行优化使其对训练数据的生成概率最大。这一方

式显然是参数的。如果我们不是定义某一个聚类模型，而是定义一个在所有可能聚类方式上的先验概率，并基于训练数据得到在每个聚类方式上的后验，则可让数据自动选择出最合理的模型复杂度。狄利克雷过程正是一种定义在所有聚类方式上的先验概率。所谓聚类方式，我们是指N个样本任何一种可能的分组设计。

8.3.1 回顾高斯混合模型

一个高斯混合模型是K个高斯概率密度的叠加。对任意一个数据集$\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N$，其联合概率密度函数如下：

$$\ln p(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N)=\sum_{i=1}^{N}\ln\sum_{k=1}^{K}\pi_k N(\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),$$

其中参数π_k为混合系数，满足:

$$\sum_{k=1}^{K}=1,\quad 0\le\pi_k\le 1.$$

这一模型中包含如下参数：每个高斯成分的权重$\boldsymbol{\pi}=\{\pi_k\}$和每个高斯成分的参数$\boldsymbol{\theta}_k=\{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}$。贝叶斯方法将这些参数视为随机变量。我们假设$\boldsymbol{\pi}$由一个以$\boldsymbol{\alpha}$为参数的狄利克雷分布$Dir(\boldsymbol{\alpha})$得到:

$$p(\boldsymbol{\pi})=C(\boldsymbol{\alpha})\prod_{k=1}^{K}\pi_k^{\alpha_k-1},$$

其中$C(\boldsymbol{\alpha})$为归一化系数。进一步假设每个高斯成分的参数$\boldsymbol{\theta}_k$由一个连续分布H得到。基于上述假设，贝叶斯高斯混合模型可表述为如图 8.5所示的生成模型，并可形式化为如下过程：

$$\boldsymbol{\mu}_k^*\sim H\quad for\ \ k=1,2,...,K \tag{8.3}$$

$$\boldsymbol{\pi}\sim Dir(\boldsymbol{\alpha}) \tag{8.4}$$

$$z_i|\pi\sim Multi(\boldsymbol{\pi}) \tag{8.5}$$

$$\mathbf{x}_i|\boldsymbol{\theta}_{z_i}^*\sim N(\boldsymbol{\theta}_{z_i}^*), \tag{8.6}$$

其中$Multi(\boldsymbol{\pi})$是以$\boldsymbol{\pi}$为参数的多项式分布。

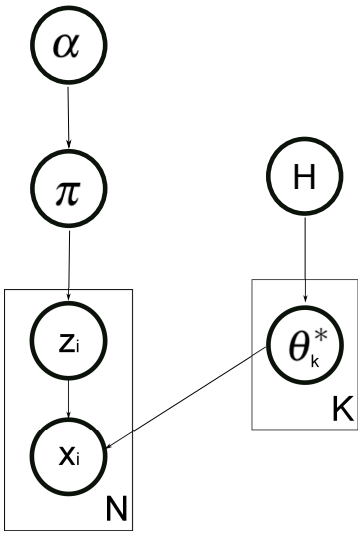


Fig. 8.5 贝叶斯高斯混合模型的有向图模型表示。首先由H中生成K个高斯成分的参数$\{\boldsymbol{\theta}_k^*\}$，并基于$Dir(\boldsymbol{\alpha})$生成高斯成分的权重参数$\boldsymbol{\pi}$。每一个数据样本$\mathbf{x}_i$，首先由$Multi(\boldsymbol{\pi})$生成高斯成分指示变量$z_i$，再由$N(\boldsymbol{\theta}_{z_i}^*)$生成$\mathbf{x}_i$。

与混合高斯模型类似的是Latent Dirichlet Allocation （LDA）主题模型。在这一模型中，数据是离散的，且由一个多项式分布生成，如图 8.6 所示。LDA模型被广泛应用在自然语言处理中，用于对文档的主题进行建模。注意LDA中每篇文档都有一个独立的主题分布概率，这一分布概率由一个狄利克雷分布采样得到。在GMM中没有文档的概率，所有数据共享同一个高斯成分上的先验概率。

不论是GMM还是LDA，都需要设定模型中高斯成分或主题的个数。实际应用中，我们很难判断一参数应该如何设置，而且一旦设定好后，新增训练数据也无法修改。我们希望得到一个模型，它能自动地根据训练数据的多少选取合理的聚类数或主题个数，并随数据量的增加做合理调整。一种思路是对数据集的不同聚类方式赋予一个先验概率，从而可以利用概率图模型的推理方法得到聚类方式的后验概率。下面将要介绍狄利克雷过程（Dirichlet Process，DP）即是这种先验概率。

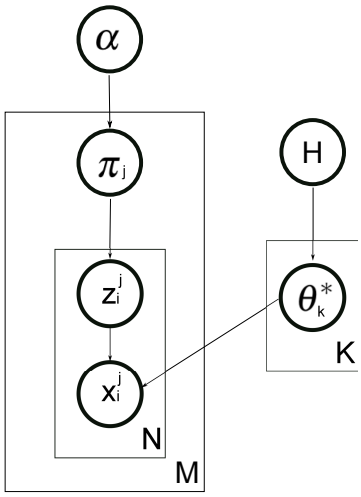


Fig. 8.6 LDA主题模型的有向图模型表示。对第j篇文档由狄利克雷分布生成一个主题分布概率$\boldsymbol{\pi}_j$，对该文档中的每个词，由$\boldsymbol{\pi}_j$生成一个主题变量z_i^j，基于该主题变量，由$Multi(\boldsymbol{\theta}_{z_i^j})$生成词$\mathbf{x}_i^j$。主题$\boldsymbol{\theta}_k^*$由一个以$H$为基础变量的狄利克雷分布生成。

8.3.2 中国餐馆问题

我们的目的是对N个数据的所有可能聚类方式赋予一个先验概率，从而可设计一个贝叶斯模型，推理得到聚类方式的后验概率。得到这一后验概率后，我们可以（1）基于最大后验准则，得到最合理的聚类方式；（2）对任何一个测试数据样本，可以基于该后验概率进行贝叶斯推断，计算该数据样本的生成概率。

中国餐馆问题（CRP）问题给出这一先验概率的直观设计方法。CRP可想象成如下过程：假设一个中国餐馆有无限张桌子，每张桌子提供不同的菜，可记为$\boldsymbol{\theta}_k^*$。每个顾客可选择和其他已经就座的顾客同坐，也可以选择坐一张新桌子。假设第N 顾客到来的时候，已经有K张桌子上有顾客了，这些桌子上分别坐了$n_1, n_2, ...n_K$个顾客。那么第N个顾客将以概率$\frac{n_k}{\alpha+N-1}$ 坐在第k张桌子上，或以概率$\frac{\alpha}{\alpha+N-1}$选择一张新的桌子坐下。这样在第$N$个顾客坐定之后，这$N$个顾客聚成为$K$ 类或$K+1$类。图 8.7 给出了一个CRP过程的实例。

设第N个顾客坐到了第c_N桌，则上述过程可写成如下公式：

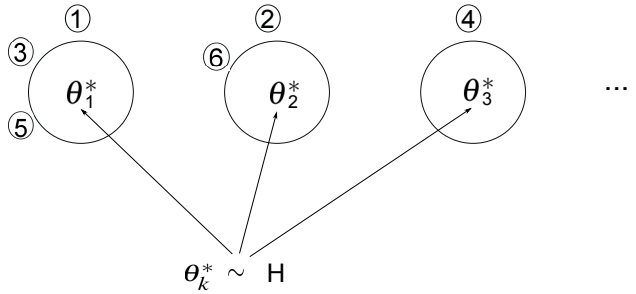


Fig. 8.7 一个CRP过程的实例。顾客1、2、4各选择了一个新桌，其对应的概率分别为1、$\frac{\alpha}{\alpha+1}$、$\frac{\alpha}{\alpha+3}$；顾客3和5选择和顾客1坐到一桌，对应的概率分别为$\frac{1}{\alpha+2}$、$\frac{2}{\alpha+4}$；顾客6选择和顾客2一起坐在第二桌，其对应的概率为$\frac{1}{\alpha+5}$。每一桌的参数由分布H随机生成。

$$c_N|c_{1,2,...,N-1} = \begin{cases} k & \text{with probability} \frac{n_k}{\alpha+N-1} \\ K+1 & \text{with probability} \frac{\alpha}{\alpha+N-1} \end{cases}$$

上式说明CRP过程具有聚类效应，即当某一桌上的人越多的时候，下一个顾客越倾向于哪一桌。基于上述公式，我们可以得到:

$$\begin{aligned} p(c_1,c_2,...,c_N) &= p(c_1)p(c_2|c_1)...p(c_n|c_1,c_2,...,c_{N-1}) \\ &= \frac{\alpha^K \prod_{k=1}^K (n_k-1)!}{\alpha(1+\alpha)...(N-1+\alpha)}. \end{aligned}$$

上式说明，对N个顾客的CRP过程，每一位顾客坐哪一桌的联合概率只与每桌的顾客数相关，与顾客的顺序和桌子的顺序都是无关的。换句话说，这一概率表示的是将无差别的客户进行无差别聚类后，每种聚类方式的概率分布。这事实上提供了一种在聚类方式上的先验分布，基于这一分布，我们将可以通过观测数据得到这些数据的后验分布。

值得说明的是，上述CRP过程提供的先验分布是无界但有限的。无界意味着这一分布可以为无限多的数据提供聚类先验，有限是因为对一个特定数据集，样本数总是有限的，CRP只需对有限数据的有限聚类方式提供先验。可以计算CRP对N个样本的聚类数的期望值为：

$$\mathbb{E}[K|N] = \sum_{i=1}^{N} \frac{\alpha}{\alpha+i-1} \approx \alpha \log(1+\frac{n}{\alpha}).$$

可见，CRP倾向的聚类方式，既非过于分散（如每个样本自成一类），也非过于聚拢（如所有类自成一类），而是一个与样本数成对数关系的适中聚类方

式。随着N的增长，CRP先验给出的预期聚类数也会跟着增长，但增长速度是对数性的，要落后于样本的增长，这显然是我们希望的。

基于CRP给出的先验，可以得到如下描述聚类任务的生成模型：设对$i-1$个样本点完成聚类后，已经有K类存在，则对第i个样本$\mathbf{x}_i$，以$\frac{n_k}{\alpha+i-1}$为概率选择第k个类，依该类的参数$\boldsymbol{\theta}_k^*$得到一个数据生成模型$F(\boldsymbol{\theta}_k^*)$，由此生成$x_i$; 或以$\frac{\alpha}{\alpha+i-1}$为概率生成一个新的类，由$H$随机得到该类的参数$\boldsymbol{\theta}_{K+1}^*$，并依数据生成模型$F(\boldsymbol{\theta}_{K+1}^*)$生成$\mathbf{x}_i$。记第$i$个样本所属类别的参数为$\boldsymbol{\theta}_i$（不同样本$\boldsymbol{\theta}_i$可能共享同一个参数$\boldsymbol{\theta}_k^*$），这一过程可形式化为：

$$\boldsymbol{\theta}_i|\boldsymbol{\theta}_{1,2,\ldots,i-1} \sim \frac{\alpha}{\alpha+i-1}H + \sum_{k=1}^{K}\frac{n_k}{\alpha+i-1}\delta_{\boldsymbol{\theta}_k^*} \tag{8.7}$$

$$\mathbf{x}_i \sim F(\boldsymbol{\theta}_i) \tag{8.8}$$

如果数据x是连续的且生成模型F是高斯的，我们即得到了一个基于无参数先验概率的贝叶斯高斯混合模型；如果数据$\mathbf{x}$是离散的且生成模型F是多项式分布，即得到一个基于无参数先验概率贝叶斯主题模型。基于上述生成模型，可以通过推理得到聚类形式的后验概率。基于这一后验概率，我们可以知道（1）数据应该聚成几类；（2）对任何一个数据样本的生成概率。然而，由于聚类形式是组合增长的，精确推理很困难，一般采用近似推理方法，如MCMC [7]或变分法 [2]。后面我们会详细讨论基于这一无参数先验的近似推理方法。

8.3.3 狄利克雷分布及性质

通过CRP，我们设计了一个聚类方式的先验概率，并依此得到了后验概率的推理方法和对数据样本概率的预测方法。现在我们将CRP放到一个更通用的理论框架里讨论，说明这一方法的合理性。

首先我们讨论狄利克雷分布。一个狄利克雷分布可表示为：

$$(\pi_1,\ldots,\pi_K) \sim Dir(\alpha_1,\ldots,\alpha_K),$$

其中$\alpha=(\alpha_1,\ldots,\alpha_K)$为参数。狄利克雷分布的概率密度函数为：

$$p(\pi_1,\ldots,\pi_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)}\prod_{k=1}^{K}\pi_k^{\alpha_k-1},$$

其中$\Gamma(\cdot)$为Gamma函数:

$$\Gamma(z) = \int_0^{\infty} \frac{t^{z-1}}{e^t} dt.$$

值得注意的是，$\pi = (\pi_1, ..., \pi_K)$本身是一个概率分布，并满足：

$$\sum_k \pi_k = 1; \quad \pi_k \geq 0.$$

因此，$p(\pi_1, ..., \pi_K)$是一个概率分布的概率分布，或随机概率。这意味着狄利克雷分布的每个采样点是一个离散概率分布，这些采样点分布在一个受限$K-1$维空间中，如图 8.8所示。

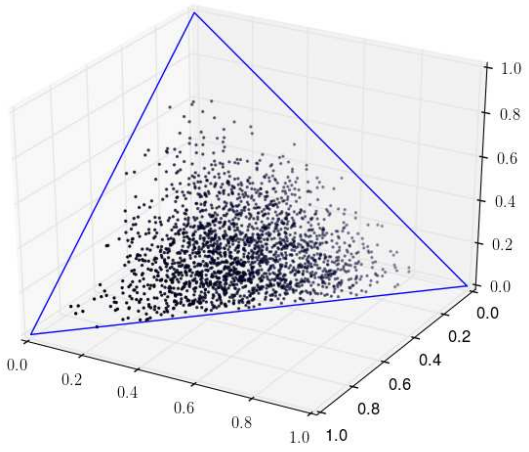


Fig. 8.8 三维空间中的Dirichlet分布Dir(4,4,2)，样本点个数为2000。

狄利克雷分布具有交换性，即：

$$(\pi_{e(1)}, ..., \pi_{e(K)}) \sim Dir(\alpha_{e(1)}, ..., \alpha_{e(K)}).$$

其中$e(\cdot)$是$1, 2, ..., K$上的交换函数。

狄利克雷分布具有累加性，即将该分布中的任意两维做加和，依然是一个狄利克雷分布。即如果$(\pi_1, ..., \pi_K) \sim Dir(\alpha_1, ..., \alpha_K)$，则有：

$$(\pi_1 + \pi_2, ..., \pi_K) \sim Dir(\alpha_1 + \alpha_2, ..., \alpha_K).$$

写成更一般的形式，有：

$$(\sum_{i\in l_1}\pi_i,...,\sum_{iinl_j}\pi_i)\sim Dir(\sum_{i\in l_1}\alpha_i,...,\sum_{i\in l_j}\alpha_i),$$

其中$l_1,...,l_j$是对$1,2,...,K$的一个任意划分。

最后，狄利克雷分布最有可分性，即一个狄利雷克分布可由部分狄利雷克分布嵌套实现。具体来说，如果:

$$(\pi_1,...,\pi_K)\sim Dir(\alpha_1,...,\alpha_K),$$

且:

$$(\tau_1,\tau_2)\sim Dir(\alpha_1\beta_1,\alpha_1\beta_2)\quad \beta_1+\beta_2=1,$$

则有：

$$(\pi_1\tau_1,\pi_1\tau_2,\pi_2,...,\pi_K)\sim Dir(\alpha_1\beta_1,\alpha_1\beta_2,\alpha_2,...,\alpha_K).$$

上式可通过狄利克雷分布的概率密度形式证明。

狄利克雷分布可写成更明确的随机分布形式如下：

$$(\pi_1,...,\pi_K)\sim Dir(\alpha,H),$$

其中$\alpha=\sum_i^K\alpha_k$, H是一个在离散空间X上的多项式分布：

$$H=Multi(\frac{\alpha_1}{\alpha},...,\frac{\alpha_K}{\alpha}).$$

可以证明，如果：

$$\pi\sim Dir(\alpha,H);\quad \mathbf{x}\sim Multi(\boldsymbol{\pi}),$$

则有：

$$p(\mathbf{x})=\sum_{\pi}p(\mathbf{x}|\boldsymbol{\pi})p(\boldsymbol{\pi}|H)=H(\mathbf{x}),$$

上式说明狄利克雷分布$Dir(\alpha,H)$是一个以基础分布H为中心的随机分布。α是控制随机性的参数，可称为中心因子（Contraction Factor）。α越大，π的随机性越小，越接近H。

8.3.4 狄利克雷过程

现在我们设想一个K非常大的狄利克雷分布$Dir(\alpha,H)$。依累加性，对H的支持空间X的任意划分都将是一个一致的狄利克雷分布。所谓一致，是指其基础分布都由同一个基础分布H衍生（累加）得到，中心因子都是同一个参数α。

当我们把K扩展到无限维，甚至X变成连续空间，依Kolmogorov一致性定理，狄利克雷分布$Dir(\alpha,H)$被扩展成一个随机过程，通常称为狄利克雷过程，表示为$DP(\alpha,H)$。由于无限维空间上的分布很难形式化表示，可以用前面所述的累加过程中的概率一致性来定义狄利克雷过程：所谓狄利克雷过程，是指一个分布在X上的随机分布G，使得在X的任何有限划分$(A_1,A_2,...,A_K)$下，每个子空间A_k的累加概率值一致地符合如下狄利克雷分布:

$$\begin{aligned}(G(A_1),G(A_2),...,G(A_K)) &\sim Dir(\alpha H(A_1),\alpha H(A_2),...,\alpha H(A_K))\\ &= Dir(\alpha,[H(A_1),H(A_2),...,H(A_K)])\\ &= Dir(\alpha,H(A_1,A_2,...,A_K)),\end{aligned}$$

这时我们说G符合狄利克雷过程$DP(\alpha,H)$，或简称DP，记为:

$$G\sim DP(\alpha,H).$$

注意上述用概率一致性来定义无限维上随机过程的方法和高斯过程的定义是一样的：在高斯过程中我们定义无限维映射函数上任意有限个样本集上的分布具有一致性（高斯分布），在狄利克雷过程中我们定义无限维分布函数上任意有限个划分上的分布具有一致性（狄利克雷分布）。Kolmogorov一致性定理保证具有上述一致性的无限维随机分布扩展为一个随机过程。

设G是一个DP: $G\sim DP(\alpha,H)$，A是X的任意一个子集，则在A上的概率$G(A)$是一个随机变量，其均值和方差为：

$$\begin{aligned}\mathbb{E}[G(A)] &= H(A)\\ Var(G(A)) &= \frac{H(A)(1-H(A))}{\alpha+1}.\end{aligned}$$

上述均值和方差事实上通过G在任意一个子集上的特性考察了G的随机性。

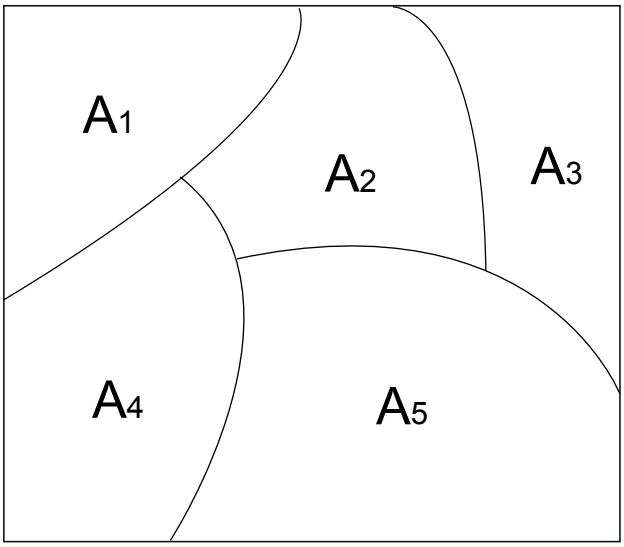


Fig. 8.9 狄利克雷过程是无限维空间X上随机分布，该分布满足如下概率一致性：对X的任何划分，其累积概率符合狄利克雷分布。图中A_1-A_5是X一个划分，$G(A_i) = \int_{\mathbf{x}\in A_i} G(\mathbf{x})\mathrm{d}\mathbf{x}$. $G(A_1,A_2,A_3,A_4,A_5) = Dir(\alpha H(A_1),\alpha H(A_2),\alpha H(A_3),\alpha H(A_4),\alpha H(A_5))$。

8.3.5 狄利克雷过程的表示

考虑如下一个基于狄利克雷分布的采样过程:

$$\boldsymbol{\pi} \sim Dir(\alpha, H),$$
$$z|\boldsymbol{\pi} \sim Multi(\boldsymbol{\pi})$$

可以计算z的边缘概率和$\boldsymbol{\pi}$的后验概率如下：

$$z \sim Multi(H),$$
$$\pi|z \sim Dir(1+\alpha, \frac{\alpha H + \delta_k(z)}{1+\alpha}),$$

其中$\delta_k(z)$为指示函数，当z取k时为1，否则为0。

现考察一个狄利克雷过程$G \sim DP(\alpha, H)$。对任一个划分$(A_1,A_2,...,A_K)$，有:

$$(G(A_1),G(A_2),...,G(A_K)) \sim Dir(\alpha, H(A_1,A_2,...,A_K)).$$

依狄利克雷分布的边缘概率，有：

$$P(\boldsymbol{\theta} \in A_i) = H(A_i),$$

特别的，有：

$$P(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}|G)dG = H(\boldsymbol{\theta}). \tag{8.9}$$

依狄利克雷分布的后验概率，有：

$$(G(A_1), G(A_2), ..., G(A_K))|\boldsymbol{\theta} \sim Dir(1+\alpha, \frac{\alpha H(A_1, ..., A_K) + (\delta_{\boldsymbol{\theta}}(A_1), ..., \delta_{\boldsymbol{\theta}}(A_K))}{1+\alpha}).$$

由于上式对任何一个划分都成立，依DP的定义，可知$G|\boldsymbol{\theta}$也是一个狄利克雷过程，且有：

$$G|\boldsymbol{\theta} \sim DP(1+\alpha, \frac{\alpha H + \delta_{\boldsymbol{\theta}}}{1+\alpha}). \tag{8.10}$$

狄利克雷过程的后验概率依然是一个狄利克雷过程，这是个非常重要的结论，为我们更深入理解DP的性质并对其进行实际建模提供了基础 [5]。

首先我们考察一个重要结论，任何从DP中抽取出的采样都是一个离散分布，即便H是连续的。以一个连续分布作基础函数采样出离散分布，而这一采样的期望又是一个连续函数，这看似是非常不符合常识的，但Ferguson [3]证明这确实如此。事实上，如果我们考察DP的后验概率形式，就会发现给定一个观察值$\boldsymbol{\theta}$后，$G|\boldsymbol{\theta}$的基础函数已经是非连续的了，这说明DP本身确实具有很强的非连续性。下面我们说明这种非连续性如何导致了DP采样的离散性。

考虑如下采样过程：

$$G \sim DP(\alpha, H)$$

$$\boldsymbol{\theta} \sim G,$$

我们得到了$(G, \boldsymbol{\theta})$的一个联合采样。然而，因为分布函数G是未知的，因此也无法由G得到采样$\boldsymbol{\theta}$，因而上述采样过程事实上只是理论的，无法实用。幸运的是，基于DP 的后验概率，我们可以得到采样过程的一个等价过程：

$$\boldsymbol{\theta} \sim H$$

$$G \sim G|\boldsymbol{\theta}$$

通过上式得到的$(G,\boldsymbol{\theta})$具有和原始采样相同的联合概率（虽然这里边的G依然是不可见的）。我们已经知道，后验概率$G|\boldsymbol{\theta}$是一个如式8.10所示的新的DP。

对上述采样进行扩，依G得到两个独立同分布采样，即:

$$G \sim DP(\alpha, H)$$

$$\boldsymbol{\theta_1}, \boldsymbol{\theta}_2 \sim G,$$

其等价采样为：

$$\boldsymbol{\theta}_1 \sim H$$

$$\boldsymbol{\theta}_2 \sim \boldsymbol{\theta}|\boldsymbol{\theta}_1$$

$$G \sim G|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2.$$

其中$\boldsymbol{\theta}|\boldsymbol{\theta}_1$是已知$\boldsymbol{\theta}_1$后系统的概率分布。因为$G|\boldsymbol{\theta}_1$是一个DP:

$$G|\boldsymbol{\theta}_1 \sim DP(1+\alpha, \frac{\alpha H + \delta_{\boldsymbol{\theta}_1}}{1+\alpha}),$$

依DP采样的边缘概率公式 8.9，因而$\boldsymbol{\theta}|\boldsymbol{\theta}_1$ 是该$G|\boldsymbol{\theta}_1$的基础分布$\frac{\alpha H+\delta_{\boldsymbol{\theta}_1}}{1+\alpha}$。上述过程重复进行，可得$G$的$N$个采样如下：

$$\boldsymbol{\theta}_n|\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\boldsymbol{\theta}_i}}{n-1+\alpha} \quad n = 1, 2, ..., N \tag{8.11}$$

基于这N个采样点，G的后验概率为：

$$G|\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N \sim DP(N+\alpha, \frac{\alpha H + \sum_{i=1}^{N} \delta_{\boldsymbol{\theta}_i}}{N+\alpha}) \tag{8.12}$$

式8.11给出了一个从DP的某一采样G中随机抽取采样$\boldsymbol{\theta}$的过程。注意的是，我们事先并不知道我们正在采样的是哪个G，但是我们知道通过这一采样方法得到的$\{\boldsymbol{\theta}_i\}$是基于$G$的独立同分布采样，因而$\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N$本身即代表了$G$的分布特性。同时，我们也知道$N$个采样点是有限的，$G$中还有一些分布性质无法由这些采样点得到，而这些不确定性即是由式8.12所代表的后验概率。由该式可知，当得到越来越多的采样点，G的特性越来越清晰地表现出来。让人惊讶的是，随着N的增加，G的绝大部分概率值越来越多地分布在若干离散点上，而这些离散点正是已经得到的采样值（不同采样可能有相同采样值）。极限情况下，当N趋近于无穷时，采样点的概率分布可以准确描述G的分布，这时对一个新$\boldsymbol{\theta}$的采样分布即为G所代表的概率分布:

$$[\boldsymbol{\theta}|\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_N]_{N\to\infty} \sim G = \frac{\sum_{i=1}^{N}\delta_{\boldsymbol{\theta}_i}}{N+\alpha} \tag{8.13}$$

注意上述概率分布正是当N取无穷大时，G的后验概率8.12所趋近的极限值（注意当N取无穷大时，G无限趋近其基础分布）。式8.13清晰地表明，G是一个离散分布，该分布有无穷多个取值点，但这些点是离散的。图 8.10 给出一个由$DP(\alpha,H)$的采样过程，可见虽然H是个连续分布，G却是个离散分布。值得注意的是，G中每个采样点的高度$p(\boldsymbol{\theta}|G) = G(\boldsymbol{\theta})$ 与$H(\boldsymbol{\theta})$没有直接关系，更多取决于哪个$\boldsymbol{\theta}$ 先被选中。依采样过程，越先被随机到的$\boldsymbol{\theta}$ 其$G(\boldsymbol{\theta})$越大。由于$\boldsymbol{\theta}$的采样依赖$H(\boldsymbol{\theta})$，因而$H(\boldsymbol{\theta})$较大的$\boldsymbol{\theta}$ 容易被选中，因此G会倾向在$H(\boldsymbol{\theta})$ 较大处表现出更大的$G(\boldsymbol{\theta})$。

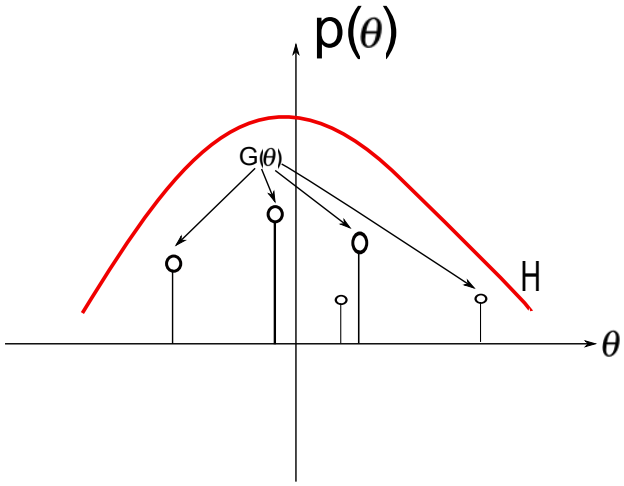


Fig. 8.10 狄利克雷过程$DP(\alpha,H)$的一个采样G。

上述采样过程可用如下假想实验来模拟：假设我们有颜色分布H，现在我们拿到一个空袋子，从下面的两种方式中随机选择一种往袋子里放球：

方式一：从H中抽样出一种颜色，并将一个该种颜色的球放入袋中;

方式二：从袋中随机抽取出一个球，将球放回袋中，并往袋中放入一个同样颜色的球。

上述两种方式的比例为$\alpha : N$，其中N为袋中已有球数。当袋中球的个数无限大时，袋中不同颜色球的分布即是一个狄利克雷过程的抽样。上述假想实验称为BlackWell-MacQueen Urn。图 8.11给出这一过程。

进一步，如果我们观察BlackWell-MacQueen Urn过程，可以发现放入一个新球时，所选颜色的概率与当前袋中该颜色球的个数成正比。以$\boldsymbol{\theta}_i$代表

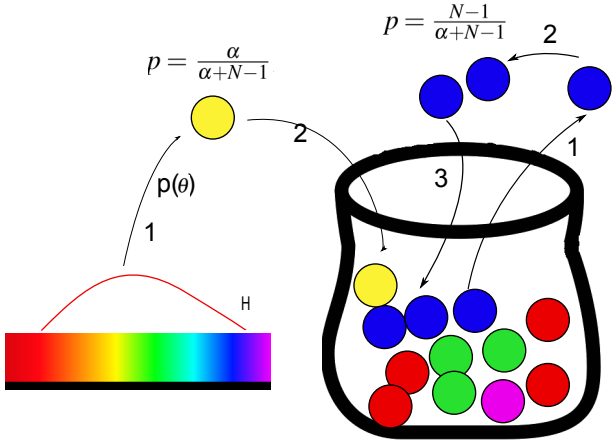


Fig. 8.11 狄利克雷过程的BlackWell-MacQueen Urn表示。进行第N次采样时，以$p=\frac{N-1}{\alpha+N-1}$为概率选择从袋中取出一个球，按该球颜色'复制'一个同颜色的球放入袋中；或以$p=\frac{\alpha}{\alpha+N-1}$为概率选择一种新颜色，制作一个该种颜色的球放入袋中。颜色选择依分布H。

第i个球的颜色，$\boldsymbol{\theta}_k^*$为第k种颜色，K为已有颜色总数，n_k为第k个颜色的个数，则有：

$$\boldsymbol{\theta}_N|c_{1,2,\ldots,N-1}=\begin{cases}\boldsymbol{\theta}_k^* & \text{with probability}\ \frac{n_k}{\alpha+N-1}\\ \boldsymbol{\theta}_{K+1}^* & \text{with probability}\ \frac{\alpha}{\alpha+N-1}.\end{cases}$$

上式事实上正是CRP过程中的选座位方式，$\boldsymbol{\theta}_k^*$相当于第k桌（或桌上的菜），K为已经被占的总桌数，n_k为第k桌的顾客数。因此，CRP所定义的先验概率事实上正是DP定义在概率分布函数上的先验，每一次CRP过程得到DP的一个采样G。

由BlackWell-MacQueen Urn得到的$\boldsymbol{\theta}_i$其联合概率与顺序无关（见CRP的概率分布公式），且是依G条件独立同分布的。依de Finetti 定理，必然存在一个在G上的概率分布来保证这一独立同分布属性。这一G上的概率分布即是狄利克雷过程。

8.3.6 狄利克雷过程的构造

前面所述的CRP过程和BlackWell-MacQueen Urn从数据生成角度描述了DP的性质，该过程通过采样$\boldsymbol{\theta}$来间接生成G，其中每一步采样出的$\boldsymbol{\theta}$有很大概率与前面得到的$\boldsymbol{\theta}$相等。现在我们讨论一种通过采样各异值$\boldsymbol{\theta}^*$直接生成DP采样G的方法，即每次采样都得到一个新的$\boldsymbol{\theta}^*$。这一方法称为Stick-breaking。

回到$(G,\boldsymbol{\theta})$的联合采样，G可由下述采样过程得到：

$$\boldsymbol{\theta} \sim H$$

$$G|\boldsymbol{\theta} \sim DP(\alpha+1, \frac{\alpha H+\delta_{\boldsymbol{\theta}}}{1+\alpha}).$$

考虑将X分成两部分：$A_1=\boldsymbol{\theta}, A_2=X\backslash\boldsymbol{\theta}$。依DP定义，有：

$$\begin{aligned}(G(\boldsymbol{\theta}), G(X\backslash\boldsymbol{\theta})) &\sim Dir((1+\alpha)\frac{\alpha H+\delta_{\boldsymbol{\theta}}}{1+\alpha}(\boldsymbol{\theta}), (1+\alpha)\frac{\alpha H+\delta_{\boldsymbol{\theta}}}{1+\alpha}(X\backslash\boldsymbol{\theta})) \\ &= Dir(1,\alpha)\end{aligned}$$

注意$(\beta, 1-\beta) \sim Dir(1,\alpha)$等价于$\beta \sim Beta(1,\alpha)$，因此，$G$包含两部分，或者以$Beta(1,\alpha)$ 为概率取$\boldsymbol{\theta}$，或者以$1-Beta(1,\alpha)$为概率取$X\backslash\boldsymbol{\theta}$ 上的值，因此有:

$$G=\beta\delta_{\boldsymbol{\theta}}+(1-\beta)G';\quad \beta\sim Beta(1,\alpha)\quad \boldsymbol{\theta}\sim H$$

进一步，将X做如下划分$X=\{\boldsymbol{\theta}\}\cup A_1\cup A_2...\cup A_K$，依DP定义，有：

$$(G(\boldsymbol{\theta}), G(A_1), ..., G(A_K)) = (\beta, (1-\beta)G'(A_1), ..., (1-\beta)G'(A_K)) \sim Dir(1, \alpha H(A_1), ..., \alpha H(A_K))$$

因此有：

$$G' \sim DP(\alpha, H).$$

重复上述过程，即有：

$$
\begin{aligned}
&G \sim DP(\alpha, H) \\
&G = \beta_1 \delta_{\boldsymbol{\theta}_1^*} + (1-\beta_1) G_1 \\
&G = \beta_1 \delta_{\boldsymbol{\theta}_1^*} + (1-\beta_1)(\beta_2 \delta_{\boldsymbol{\theta}_2^*} + (1-\beta_2) G_2) \\
&\quad \cdots \\
&G = \sum_{k=1}^{\infty} \beta_k \prod_{i=1}^{k-1} (1-\beta_i) \delta_{\boldsymbol{\theta}_k^*}
\end{aligned} \tag{8.14}
$$

其中:

$$\beta_k \sim Beta(1, \alpha); \ \boldsymbol{\theta}_k^* \sim H.$$

注意上式中β_k和$\boldsymbol{\theta}_k^*$的随机性是重要条件，否则G的分解形式不能成立。

式8.14提供了一种非常简便的DP采样方法：每次从H中随机采样出一个值$\boldsymbol{\theta}_k^*$，由$Beta(1,\alpha)$生成一个β_k，依式8.14计算$\boldsymbol{\theta}_k^*$上的概率$\beta_k \prod_{i=1}^{k-1}(1-\beta_i)$。这一过程可用一个折筷子游戏模拟，称为Stick-breaking过程。假设一根单位长度的筷子，密度分布均匀。首先采样$\beta_1 \sim Beta(1,\alpha)$，按$\beta_1 : 1-\beta_1$ 将筷子折成两半，取第一半为第一个采样点$\boldsymbol{\theta}_1^*$的概率；再将剩余的$1-\beta_1$部分作为整体，采样$\beta_2 \sim Beta(1,\alpha)$，取第一部分作为第二个采样点$\boldsymbol{\theta}_2^*$的概率...如此往复进行，每次都对前次折后的剩余部分依$Beta(1,\alpha)$进行再次拆分，作为新一个采样点的概率。经过上述拆分过程后，每一小段的长度即等于对应采样点的概率，剩余部分即为所有其它采样值的概率和。Stick-Breaking过程首先由Sethuraman于1994年提出 [11]，并在文章中证明该过程得到的随机概率确实是DP先验。由Stick-Breaking得到的对单位筷子的长度分布亦称为Griffiths-Engen-McCloskey（GEM）分布 [8]。

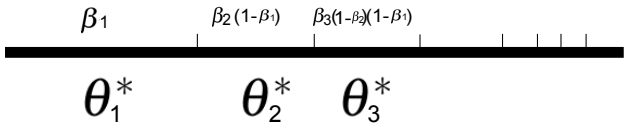


Fig. 8.12 Stick-breaking过程。一根单位长度的筷子被分割成无数小段，每一个小段长度代表一个概率值。这一分割由无数次二分过程得到，其中第k个二分过程包含如下步骤：$\beta_k \sim Beta(1,\alpha), \boldsymbol{\theta}_k^* \sim H$，则第$i$次分割的比例为$\beta_k : 1-\beta_k$，对应的采样为$\boldsymbol{\theta}_k^*$。

8.3.7 推理方法

本节以DPGMM为例，讨论贝叶斯非参数模型中基于采样的推理方法。另一种常用的推理方法是变分法，读者可参考 [2]。关于采样和变分两种推理方法的基本原理，我们在第 **??**章中已有细致讨论，本节只关注将采样法应用于DPGMM的具体做法。

DPGMM是对高斯成分$\boldsymbol{\theta}_k^*$和高斯成分的先验概率π_k都引入随机性的GMM模型，其中k可以无穷大。从生成模型角度看，DPGMM可以有三种表示，如图 8.13所示。

在表示（A）中，首先由CRP生成z_i，再由$\boldsymbol{\theta}_{z_i}$生成数据：

$$z_i|z_1,...z_{i-1} \sim CRP(\alpha) \tag{8.15}$$

$$\boldsymbol{\theta}_k^* \sim H \quad if \;\; z_i \neq z_j \;\; \forall j < i \tag{8.16}$$

$$\mathbf{x}_i \sim F(\boldsymbol{\theta}_{z_i}^*). \tag{8.17}$$

在表示（B）中，首先由BlackWell-MacQueen Urn过程生成$\boldsymbol{\theta}_i$，再由该参数生成数据：

$$\boldsymbol{\theta}_i \sim DP(\alpha, H) \tag{8.18}$$

$$\mathbf{x}_i \sim F(\boldsymbol{\theta}_i). \tag{8.19}$$

在表示（C）中，首先由Stick-breaking过程（即GEM过程）生成$\boldsymbol{\theta}_k^*$和$p(k)$，由此计算出n_k，再由这些参数对每个高斯成分生成n_k个数据：

$$n_k \sim GEM(\alpha) \;\; k = 1,2.,, \tag{8.20}$$

$$\boldsymbol{\theta}_k^* \sim H \;\; k = 1,2,... \tag{8.21}$$

$$\mathbf{x}_i^k \sim F(\boldsymbol{\theta}_k^*) \;\; k = 1,2,...; i = 1,2,...,n_k \;\; for \;\; each \;\; k \tag{8.22}$$

研究表明，基于表示（A）来设计采样推理方法相对简单直观。在这一模型中，观测变量为数据$\{\mathbf{x}_i\}$，隐藏变量为指示变量$\{z_i\}$和模型参数变量$\{\boldsymbol{\theta}_k^*\}$。注意上述观测数据和模型参数一般是向量而非数值，但这对下述推理算法没有影响。模型推理的任务是基于观测变量求隐藏变量的后验概率，即$p(\mathbf{z}, \{\boldsymbol{\theta}_k^*\}|\mathbf{X})$，其中$\mathbf{z} = [z_1,...,z_N]^T$, $\mathbf{X} = [\mathbf{x}_1,...,\mathbf{x}_N]$。

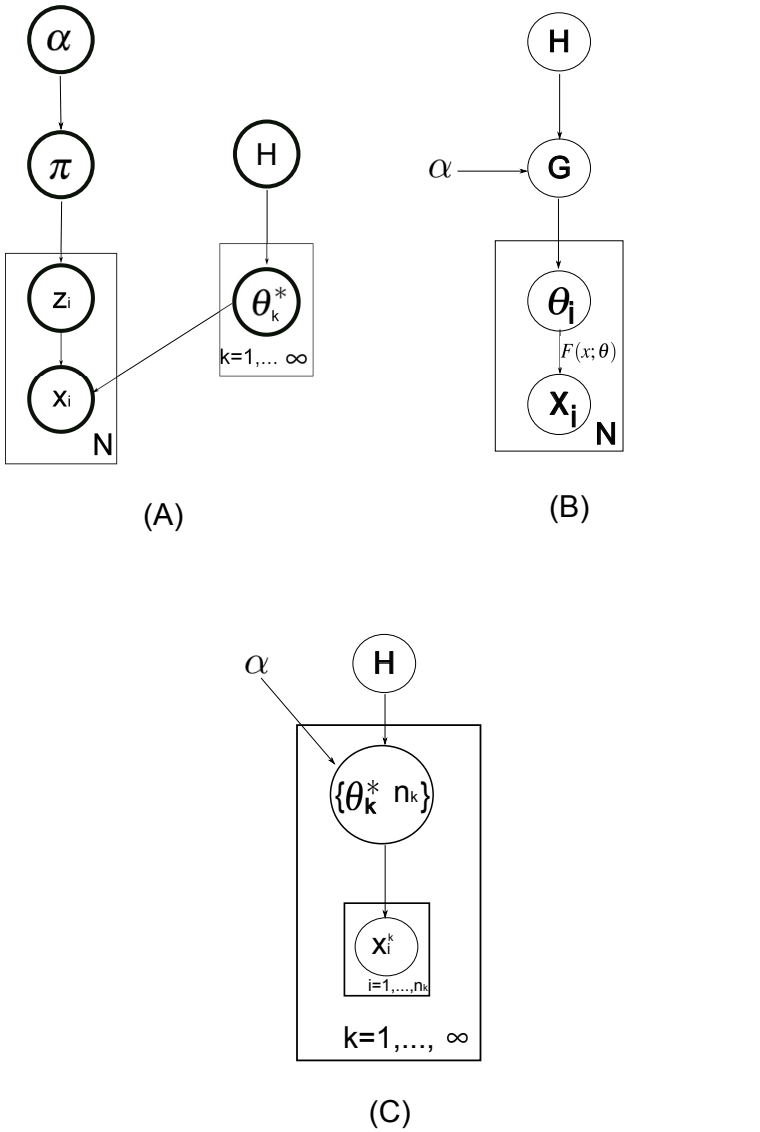


Fig. 8.13 三种DPGMM的表示方法。（A）基于CRP生成指示变量z_i和参数$\boldsymbol{\theta}_{z_i}$，再由$F(\boldsymbol{\theta}_{z_i})$生成数据；（B）基于DP生成每个样本的参数$\boldsymbol{\theta}_i$，再由$F(\boldsymbol{\theta}_i)$生成数据样本；（C）基于Stick-Breaking生成取类参数$\boldsymbol{\theta}_k^*$和n_k，再对每个类k依$F(\boldsymbol{\theta}_k^*)$生成$n_k$个数据样本。

我们用Gibbs采样法来进行近似推理。第 **??**一章中介绍过，该方法的基本原理是通过构造一个马尔可夫过程，该过程运行到稳定状态时得到采样符合要推理的后验概率。这一马尔可夫过程可由Gibs采样得到。该采样过程的采样点为后验概率中的目标变量，包括每个数据样本的指示变量z_i和每个高斯成分的参数$\boldsymbol{\theta}_k^*$，即$\{z_i\},\{\boldsymbol{\theta}_k^*\}$。Gibbs采样时，每次采样选择一个变量，保持其它变量不变，从而使简化采样过程。

z_i采样

我们首先对每个z_i采样。记除z_i外的所有指示变量为$\mathbf{z}^{-i}$，则可计算对z_i采样所需的条件概率如下：

$$\begin{aligned} P(z_i|\mathbf{z}^{-i},\mathbf{X},\{\boldsymbol{\theta}_k^*\}) &\propto P(z_i,\mathbf{z}^{-i},\mathbf{X},\{\boldsymbol{\theta}_k^*\}) \\ &\propto p(\mathbf{X}|\mathbf{z},\{\boldsymbol{\theta}_k^*\})p(z_i|\mathbf{z}^{-i}) \\ &\propto p(\mathbf{x}_i|z_i,\{\boldsymbol{\theta}_k^*\})p(z_i|\mathbf{z}^{-i}). \end{aligned}$$

其中，后验概率$p(z_i|\mathbf{z}^{-i})$依CRP过程计算得到：

$$p(z_i|\mathbf{z}^{-i}) = \begin{cases} \frac{n_k}{\alpha+N-1} & z_i = k \\ \frac{\alpha}{\alpha+N-1} & z_i = K+1 \end{cases}$$

其中K是当前$\mathbf{z}^{-i}$中不同值的个数。

下面我们计算条件概率$p(\mathbf{x}_i|z_i;\{\boldsymbol{\theta}_k^*\})$。如果$z_i \le K$，可计算$p(\mathbf{x}_i|z_i;\{\boldsymbol{\theta}_k^*\}) = F(\mathbf{x}_i,\boldsymbol{\theta}_{z_i}^*)$; 如果$z_i = K+1$，$p(\mathbf{x}_i|z_i;\{\boldsymbol{\theta}_k^*\})$需要考虑所有可能的新参数$\boldsymbol{\theta}_{K+1}^*$，因而有：

$$p(\mathbf{x}_i|z_i;\{\boldsymbol{\theta}_k^*\}) = \int p(\mathbf{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int F(\mathbf{x}_i,\boldsymbol{\theta})H(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{8.23}$$

如果$H(\boldsymbol{\theta})$是$F(\mathbf{x},\boldsymbol{\theta})$的共轭先验，则上式可较方便计算出来。

综合上述后验概率和条件概率的计算过程，可得对z_i采样的过程如下：

1. 如果z_i和所有$\mathbf{z}^{-1}$中的值都不同，则说明第z_i个高斯只包含$\mathbf{x}_i$这一个点。因为要重新随机z_i，这意味着当前这第z_i个高斯里已经不包含任何数据，因此将$\boldsymbol{\theta}_{z_i}^*$去掉，将$K$减1.

2. 依如下概率采样z_i:

$$P(z_i|\mathbf{z}^{-i},\mathbf{X},\{\boldsymbol{\theta}_k^*\}) = \begin{cases} \frac{n_k}{\boldsymbol{\alpha}+N-1}F(\mathbf{x}_i,\boldsymbol{\theta}_{z_i}^*) & z_i = k \\ \frac{\alpha}{\alpha+N-1}\int F(\mathbf{x}_i,\boldsymbol{\theta})H(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} & z_i = K+1 \end{cases}$$

3. 如果$z_i = K+1$，则将K加1，并随机抽取一个新的$\boldsymbol{\theta}_K^*$如下：

$$P(\boldsymbol{\theta}^*|\mathbf{x}_i) \propto F(\mathbf{x}_i,\boldsymbol{\theta}^*)H(\boldsymbol{\theta}^*).$$

$\boldsymbol{\theta}_k^*$采样

对$\{z_i\}$做完采样后，我们将这些变量固定，接下来对每个高斯成分的参数$\boldsymbol{\theta}_k^*$采样。记除第k个高斯外的所有高斯成分的参数为$\boldsymbol{\theta}_{-k}^*$，这一采样的概率为：

$$\begin{aligned} p(\boldsymbol{\theta}_k^*|\mathbf{X},\mathbf{z},\boldsymbol{\theta}_{-k}^*) &\propto p(\boldsymbol{\theta}_k^*,\mathbf{X},\mathbf{z},\boldsymbol{\theta}_{-k}^*) \\ &\propto \prod_{i:z_i=k} p(\mathbf{x}_i|\boldsymbol{\theta}_k^*)p(\boldsymbol{\theta}_k^*) \\ &= \prod_{i:z_i=k} F(\mathbf{x}_i,\boldsymbol{\theta}_k^*)H(\boldsymbol{\theta}_k^*) \end{aligned} \tag{8.24}$$

值得说明的是，上述Gibbs采样算法需要H和F是共轭的，否则式 8.23可能无法计算。如果确实要选择非共轭先验，可采用第 **??**章中所讨论的Metropolis-Hastings算法，这一算法不需要计算条件概率。Neal 2000年的论文给出了DPGMM MCMC 近似推理算法的细节 [7]。

8.3.8 Hierarchical DP (HDP)

DP可以对相似环境下数据进行有效学习，但有时数据比较复杂，包括在不同环境下的不同分布。例如在对互联网文章进行聚类时，不同类别的文章其分布规律是不同的，如果不把这些文章分类处理，分布可能过于复杂，用DP做先验可能得不到好的效果。然而，对不同类别的文章完全分开处理又会降低统计显著性，毕竟不同类间具有很多相似性。

研究者将DP扩展成层次结构来处理这个问题 [15, 14, 12]。其基本思路是不同环境下的数据分布共享相同的DP作为‘先验分布的先验’，之后不同环境下基于该共享DP生成独特的DP作为先验。基于这些环境相关的DP进

行建模，即可实现对不同环境下数据相关性的学习。这一层次性DP 模型称为Hierarchical DP（HDP），如图 8.14所示。在HDP中，首先以$DP(\gamma,H)$抽样出一个概率函数G_0，对每一类j，以$DP(\alpha,G_0)$抽样出一个概率函数G_j。由G_j可抽样出每个样本对应的$\boldsymbol{\theta}_i^j$，最后由数据分布概率$F(\boldsymbol{\theta}_i^j)$抽样出数据样本$\mathbf{x}_i^j$。

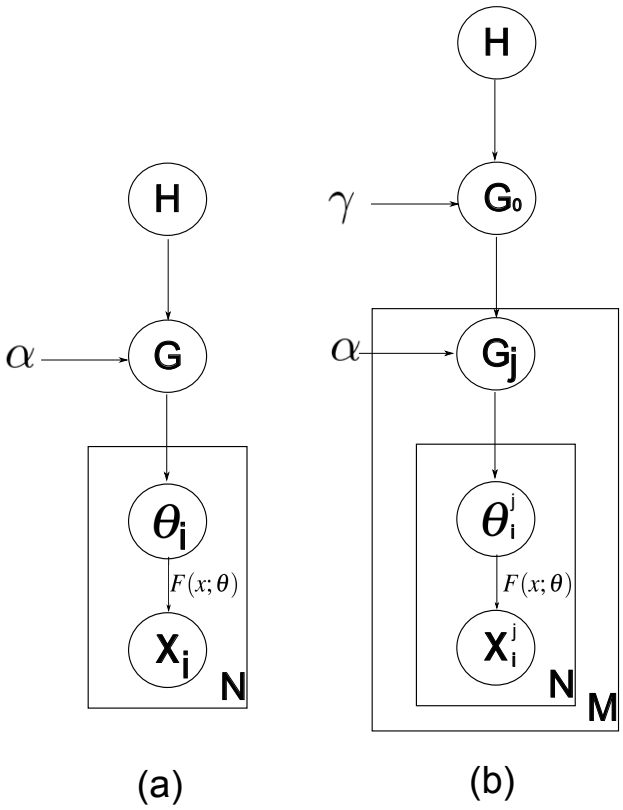


Fig. 8.14 基于DP(左）与HDP（右）的数据生成模型。$F(x;\theta)$为数据分布概率。在HDP中，每个场景j由共享的$DP(\gamma,H)$中得到G_0，再由$DP(\alpha,G_0)$得到不同场景下的概率G_j。由G_j可得到每个数据样本对应的模型参数$\boldsymbol{\theta}_i^j$，进而生成数据$\mathbf{x}_i^j$。

由HDP的数据生成过程可知，不同类的数据事实上共享一个基础分布G_0。由于G_0是离散的，且具有很强的聚类性，G_j只能取G_0所确定的参数点，且有较大概率集中在G_0中概率最大的参数，由此实现类间的参数取值及参数分布形式的共享。用于DP的近似推理方法可同样用于HDP。

8.4 本章小结

本章讨论了非参数模型的基本概念和方法。我们讨论了几种朴素的非参数模型，包括用于概率描述的直方图方法和用于分类的KNN方法。这些方法不设计模型的具体形式，依靠大量数据对空间的覆盖实现对测试数据的描述和分类。因为没有模型结构的限制，只要训练数据量足够大，这些非参数模型可实现对未知数据非常精确的描述和预测。

SVM是另一种典型的非参数模型。这一模型并非完全没有模型设计，而是将参数模型和非参数模型结合起来，在特征映射上采用参数模型（即核函数），但在预测是采用非参数模型，通过计算测试数据和部分训练数据（即支持向量）间的距离，用训练数据的分类来预测测试数据的分类。这种将参数模型和非参数模型结合起来的方法既可以避免非参数模型对数据量的过度依赖，也可以摆脱参数模型对预测函数形式的限制，从而极大提高建模的精确性，实现模型依数据的调整。从另一个角度看，这一方法也极大提高了对训练数据的利用效率。

基于这一思路，我们着重讨论了贝叶斯框架下的非参数模型。贝叶斯框架提供了非常强大的建模能力，但基于参数的贝叶斯方法其模型结构是固化的，这极大限制了模型对数据的描述上限。贝叶斯非参数模型通过提供非参数的先验概率，打破了固化模型带来的限制，使得贝叶斯模型的建模能力随数据的增长而大幅提高。

我们讨论了两种贝叶斯非参数模型方法。在预测任务中，我们引入对预测函数的先验概率，即高斯过程。这一先验概率不依赖预测函数的具体形式，而是对所有值域在$(-\infty,+\infty)$上的预测函数赋予适当的先验概率，使得基于该先验概率，任何有限采样点上的取值是高斯分布的。这一普适的先验概率大大丰富了贝叶斯方法中可选的预测函数范围，使得当训练数据足够多时，其后验概率更加灵活地趋近于真实预测函数。

在聚类任务中，我们已经讨论了狄利克雷过程。总体来说，DP可以认为是狄利克雷分布向无限维空间的扩展。在贝叶斯框架下，这使得我们得以在无限参数空间上设计概率密度函数的先验概率，从而实现依训练数据选择最合适的聚类形式。特别是，这一方法考虑所有可能的聚类形式，因而不需对聚类数进行人为指定。通过推理在所有聚类形式上的后验概率，可以让数据自动选择最合理的聚类数和在每个类上的分布概率。将狄利克雷分布扩展到狄利克雷过程具有可行性，不论是Kolmogorov一致性定理还是de Finetti 定理都表明DP是存在的；同时，BlackWell-MacQueen Urn和CRP给出了基于数据

生成过程的构造方法，Stick-breaking给出了基于参数生成过程的构造方法。利用这些构造方法，我们可以设计基于MCMC的推理算法。

8.5 相关资源

- 本章对高斯过程的讨论，参考了Bishop的‘Pattern recognition and machine Learning’第六章。
- 本章对狄利克雷过程的讨论，参考了Yee Whye Teh的‘Dirichlet Processes: Tutorial and Practical Course’[1].
- 关于高斯过程，可参考Seeger的综述文章 [10]，Williams等人的‘Gaussian Processes for Regression’[16]，Rasmussen等人的‘Gaussian processes for machine learning’ [9]。
- 关于贝叶斯非参数模型，请参考Gershman等的Review论文 [4]以及Muller等的论文 [6]。

[1] https://www.stats.ox.ac.uk/ teh/teaching/npbayes/mlss2007.pdf

Chapter 9
遗传学习

Chapter 10
强化学习

Chapter 11
优化方法

References

[1] Bishop CM (2006) Pattern recognition and machine Learning. Springer

[2] Blei DM, Jordan MI, et al (2006) Variational inference for dirichlet process mixtures. Bayesian analysis 1(1):121–143

[3] Ferguson TS (1973) A bayesian analysis of some nonparametric problems. The annals of statistics pp 209–230

[4] Gershman SJ, Blei DM (2012) A tutorial on bayesian nonparametric models. Journal of Mathematical Psychology 56(1):1–12

[5] Ghosal S (2010) The Dirichlet process, related priors and posterior asymptotics, vol 2. Chapter

[6] Müller P, Mitra R (2013) Bayesian nonparametric inference–why and how. Bayesian analysis (Online) 8(2)

[7] Neal RM (2000) Markov chain sampling methods for dirichlet process mixture models. Journal of computational and graphical statistics 9(2):249–265

[8] Pitman J, et al (2002) Combinatorial stochastic processes

[9] Rasmussen CE (2006) Gaussian processes for machine learning

[10] Seeger MW (2004) Gaussian processes for machine learning. International Journal of Neural Systems 14(2):69–106

[11] Sethuraman J (1994) A constructive definition of dirichlet priors. Statistica sinica pp 639–650

[12] Teh YW (2006) A hierarchical bayesian language model based on pitman-yor processes. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp 985–992

[13] Teh YW (2011) Dirichlet process. In: Encyclopedia of machine learning, Springer, pp 280–287

[14] Teh YW, Jordan MI (2010) Hierarchical bayesian nonparametric models with applications. Bayesian nonparametrics 1

[15] Teh YW, Jordan MI, Beal MJ, Blei DM (2005) Sharing clusters among related groups: Hierarchical dirichlet processes. In: Advances in neural information processing systems, pp 1385–1392

[16] Williams CKI, Rasmussen CE (1996) Gaussian processes for regression pp 514–520