判别式概率隐层空间模型学习与复杂度问题研究

（申请清华大学工学博士学位论文）

培 养 单 位：计算机科学与技术系

学　　科：计算机科学与技术

研 究 生：陈 宁

指 导 教 师：孙 富 春 教 授

二〇一二年五月

Discriminative Probabilistic Latent Space Learning and Model Complexity

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Computer Science and Technology

by

Chen Ning

Dissertation Supervisor : Professor Sun Fuchun

May, 2012

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘 要

从复杂数据中学习隐层空间表示可以揭示数据隐含的本质特征，提高文本挖掘、图像分类、多模态数据融合与推理等理论方法的性能。包含隐变量的概率图模型作为隐空间学习的基本框架之一，可以有效提高模型表示的紧致性、计算的高效性以及学习的鲁棒性。虽然传统的隐层空间表示的学习/分析方法已获得了一定成功，但随着互联网及数字化技术的飞速发展，近年来多种大规模有监督信息可几乎“免费”地获取，如何有效利用这些大规模的有监督信息，学习判别式的隐层空间表示、提高模型的预测性能、同时自动确定模型的复杂度，已成为机器学习领域亟待解决的几大关键问题。本文面向非结构化信息的分类与预测问题，系统地研究隐层空间学习的模型表示、判别式学习/推理方法、以及模型复杂度问题。本文的主要创新点包括：

1. 针对多模态复杂数据的高维、互补特性，提出一种基于无向马尔可夫网络的参数化有监督多模态隐层空间分类与回归模型，为发现复杂多模态数据的判别性、紧致、低维的隐层空间表示，提高分类、回归、图像检索及标注等任务的性能提供了一条新的途径；

2. 为了提高传统最大似然估计方法的预测性能，提出了一种无向图有监督隐层空间模型的判别式最大间隔学习方法，使用线性期望算子定义隐层空间模型的判别函数与损失函数，系统地将概率推理与确定性的最大间隔学习原理融合在一起，形成了一种通用的高效学习方法，与传统的基于最大似然估计方法相比，显著提高各种有监督概率隐层空间模型的预测性能；

3. 为了研究非参数化的无向图隐层空间模型的模型复杂度问题，克服参数化隐层空间模型需要时间代价很高的模型选择的缺陷，提出了一种基于链图及经验贝叶斯模型的正则化贝叶斯推理理论框架，创新性地将非参数贝叶斯方法成功应用于无向图隐层空间模型，并将非参数贝叶斯方法与最大间隔学习这两个互相分离的经典机器学习方法有机结合于统一的无向图隐层空间模型中，为深入研究其他有监督无向图隐层空间模型的模型复杂度问题提供了新的理论方法。

关键词： 隐层空间模型、多模态数据分析、最大间隔学习、非参数化贝叶斯推理

Abstract

Learning latent space representations from complex data can discover underlying intrinsic structures and improve the performance in various tasks, including text mining, image categorization, multi-view data fusion and inference, etc. For probabilistic graphical models, one basic modeling framework for learning latent space representations, introducing latent variables could effectively improve the compactness of model representation, the efficiency of computation, and the robustness of learning. Although classic latent representation learning/analyzing methods have obtained huge success, new challenges have arisen as the fast growing of Internet and digital techniques. We can easily obtain large-scale of supervising side information almost "for free". But how to effectively use such supervising information to learn discriminative latent space representations, how to improve the prediction performance, and meanwhile how to automatically resolve the unknown model complexity have become key problems that call for urgent attention in machine learning. To address these key issues, this thesis presents a systematic study on several important aspects of learning latent space representations, including model representation, discriminative learning/inference methods, and model complexity. In summary, the major novel contributions are as follows:

1. To effectively deal with complex, high-dimensional and complementary multi-view data, this thesis proposes multi-view latent space Markov networks for both classification and regression. The model offers a novel approach to discovering discriminative, compact and low dimensional latent space representations and to dealing with classification/regression, image retrieval and annotation in an efficient manner;

2. To overcome the shortcomings of the extant maximum likelihood estimation (MLE), this thesis proposes a discriminative max-margin learning method for learning supervised latent space models; presents a novel method to use the linear expectation operator to define the discriminant function as well as loss function for probabilistic latent space models; and offers a systematic way to integrate probabilistic inference and the discriminative max-margin learning principle into one coherent framework for learning predictive latent representations and performing probabilistic inference on missing data. Compared to the MLE-based methods, the new method dramatically improves prediction

performance;

3. To resolve the unknown model complexity of undirected latent space models, and bypass the hard and computationally expensive model selection step of a parametric discriminative latent space model, this thesis extends the Regularized Bayesian Inference theory to chain graphs and the case of empirical Bayesian inference. Under this framework, this thesis successfully applies nonparametric Bayesian techniques to resolve the model complexity of undirected latent space Markov networks for the first time and integrates the two important fields of nonparametric Bayesian inference and max-margin learning that have been largely treated as isolated in the machine learning community for about 20 years since their births. The novel framework of Regularized Bayesian Inference for undirected Markov networks could provide insightful theoretical guidance for developing new latent space models and present a promising new approach to automatically resolving the model complexity of undirected latent space models.

Key words: Latent Subspace Models; Multi-view Data Analysis; Max-margin Learning; Nonparametric Bayesian Inference.

目 录

主要符号对照表

$\mathbb{R}$	实数空间
$\mathbf{X}$	**x**的随机变量
$\mathbf{Y}$	所有响应变量值的集合，由若干y组成
$\mathbf{H}$	隐式随机变量
$\mathbf{W}$	参数矩阵
C	正则化常数
F	判别函数
$\mathcal{L}$	变分上界
$\mathcal{H}$	熵（Entropy）
KL($p\|\|q$)	概率分布p和q之间的KL-散度（Kullback-Leibler Divergence）
I	单位矩阵（identity matrix）
$\mathcal{N}$	高斯分布
$\mathcal{P}_{\text{prob}}$	概率分布的集合
$\mathcal{P}_{\text{post}}$	满足一定后验约束的概率分布的集合
$\|\|\cdot\|\|_2^2$	ℓ_2-范数
$\mathbb{I}(\cdot)$	指示（Indicator）函数，当变量为真时取值为1，否则取值为0
$\mathcal{G}$	图
ξ	松弛变量
ϵ	回归模型精确度参数
$\mathcal{R}$	风险函数
$\mathbb{E}$	期望
Beta	贝塔分布
Mult	多项式分布
Bernoulli	伯努利分布
$\mathcal{M}$	概率图模型
$\mathcal{D}$	数据样本
$\mathcal{I}_{\text{tr}}$	训练数据集合
$\mathcal{I}_{\text{tst}}$	测试数据集合
$\mathcal{DP}$	狄利克雷随机过程
EFH	指数族Harmonium
DWH	双模态Harmonium

MMH	最大间隔Harmonium
LDA	潜在狄利克雷分配
iLSVM	无限维隐支持向量机
MT-iLSVM	多任务无限维支持向量机
iEFH	无限维指数族Harmonium
iMMH	无限维最大间隔Harmonium
MedLDA	最大熵判别式潜在狄利克雷分配
BOW	词袋（Bag-of-Words）

第 1 章 引言

互联网及数字化技术的飞速发展为用户提供了数以亿计的复杂数据，如文本、图像、视频、社会关系网络等。这些数据的复杂性主要体现在高维、多模态、异构、动态等方面。探索实时处理与理解这些丰富多样的海量、高维数据的方法已成为机器学习及相关应用领域所关注的核心问题之一。而隐层特征空间（简称隐空间或隐特征）学习作为揭示复杂数据隐含规律信息的重要方法，是近年来机器学习领域一大研究热点。隐层空间学习涉及机器学习的多个重要子领域，包括概率图模型方法与推理、有监督学习、核方法、非参数化贝叶斯方法、大规模分布式优化与学习等。随着互联网共享技术的飞速发展，大规模有监督信息（如样本标签、用户评价分数等）都可几乎“免费”地获取。如何有效利用这些有监督信息，学习判别式（即区分性强、预测性强、更适于提高模型预测性能）的有监督隐层空间模型，是本文的根本研究动机。本文面向互联网上带有监督信息的多模态文本及图像数据等，拟系统解决判别式隐层空间学习中存在的模型表示、学习、以及模型复杂度等方面的若干基础性关键问题。

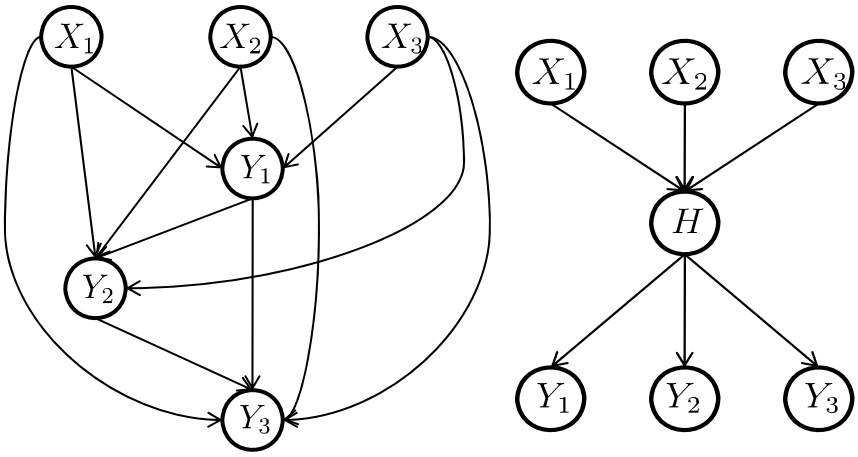
本章首先概述隐层空间学习的研究背景；通过回顾国内外研究现状，提出面临的问题和挑战；进而阐述本文的研究内容、创新点；最后介绍本文的组织结构。

1.1 研究背景及意义

隐层空间学习的核心思想是通过训练数据建立或者学习一个映射（通常可用一个包含隐变量的统计模型表示），将高维复杂的数据（如文本、图像、网页、DNA结构等）映射到一个低维的隐层子空间。这种用于学习隐层特征空间表示的隐变量模型（简称为隐层空间模型，或者隐特征模型）可以简化复杂的数据依赖关系，显著提高模型的灵活性，发现数据的潜在结构，可用于解决多种应用问题，例如文本分析[1]、社会网络分析[2]、图像语义理解[3]、语音识别[4]，以及生物信息学[5]等，具有重要的应用背景。无论从机器学习理论角度，或是相关应用角度，隐层空间模型都具有重要的价值。下面从不同方面简要介绍隐层空间学习。

1.1.1 理论研究价值

（1）简化复杂的数据依赖关系[6,7]



(a) 无隐变量的模型依赖关系　　(b) 有隐变量的模型依赖关系

图 1.1　隐变量模型可以简化复杂的数据依赖关系示意图[6]

真实世界中观测到的数据通常存在复杂的统计依赖关系，例如文本段落间的因果关系或图像像素间的空间位置关系等。然而，用于产生观测数据的物理过程可能具有很简单的隐含结构[8]。在这种情况下，如果只考虑观测数据，可能造成模型过于复杂，学习效率低下等问题。举例说明，假设已经得到变量$X1$，$X2$，$X3$以及$Y1$，$Y2$，$Y3$的观测数据，采用无隐变量的贝叶斯网络对其进行建模，很可能出现如图 1.1(a)中所示的复杂结构，**Y**变量间全连接，每个**X**变量与所有**Y**变量连接。图中共有12条边，此时需要定义模型中各变量的先验分布及含有依赖关系的所有变量的条件概率分布，假设**X** 和 **Y** 都是二值变量，则此模型中共有59个参数需要学习，其复杂程度不言而喻。由于观测数据有限，这种复杂网络很容易忽略潜在的依赖关系，同时易导致“Data Fragmentation”的问题（即通常所说的数据稀疏问题），很大程度上影响参数估计的鲁棒性[6]。

相比之下，当适当引入隐变量**H**后，模型可简化为如图1.1(b)所示的简单结构。这个含有隐变量的模型1.1(b)中只有6条边，不需要在具有复杂结构的观测变量上定义复杂的概率分布，只需在简单的隐变量模型结构定义更加容易计算的联合分布（如树状结构模型的概率分布）[9]，变量的边缘分布可以通过变量消元去除隐变量得到。假设**X**、**Y**及**H** 都是二值变量，此时需要学习的参数只有17个。可见，含有隐变量的模型无论从模型的表示或是学习角度，都可以显著降低模型的复杂程度，提高模型的鲁棒性和效率。

（2）发现数据的潜在结构

隐变量模型将高维数据映射到低维紧致的子空间，不仅可以发现数据潜在的结构，还尽可能有效地保留输入数据的有用信息。虽然至今为止尚未有理论保证

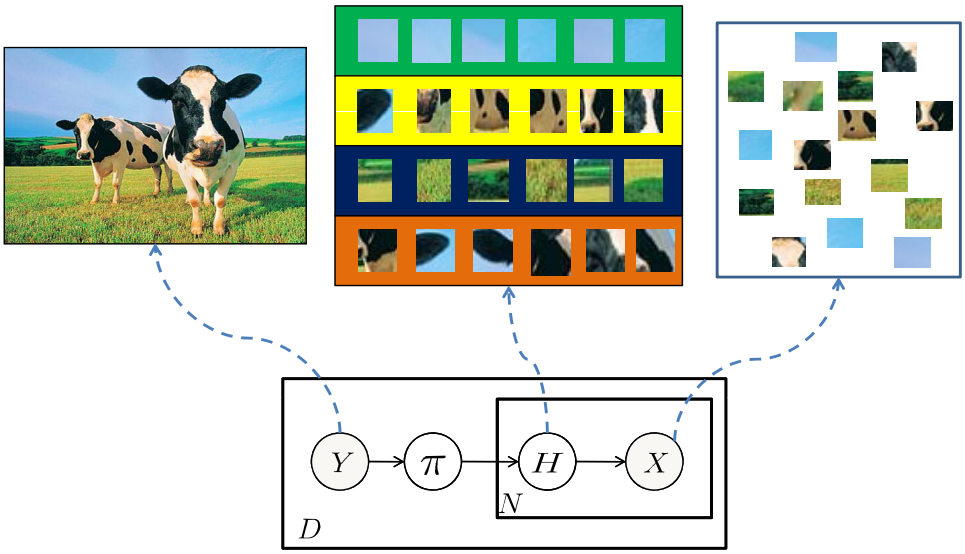


图 1.2　有监督隐层空间模型在图像分类任务中的应用。

使用隐变量模型一定能够得到可解释的聚类，或其他有用信息，已有的大量研究工作[1,10,11] 表明当这些模型很好地描述了数据真正的产生过程，许多数据源都可用高效的隐变量模型表示[11]，例如利用深度概率网络（Deep Networks）[12]，学习稀疏的隐层空间表示来解释人类大脑V2区域的认知原理等。

1.1.2　应用研究价值

基于上述原因，隐层空间学习已经在很多重要应用问题中得到了广泛使用，下面从文本信息检索、自然场景图像分类、社会网络分析、多模态数据分析等方面简单介绍。

（1）文本信息检索

文本信息检索需要计算不同文本之间的相似性及相关性，由于文本的维度较高并且数据通常存在噪声，学习一个有效的低维表示对于去除噪声、提高检索效率具有重要意义。使用隐层空间模型发现对高维复杂数据的低维描述，可以保留观测数据中用于分类、摘要抽取、相关与相近性判决等任务的信息，实现对高维数据的降维。通过计算不同文本的低维表示之间的相似性及相关性，提高检索效率。典型模型包括基于概率的潜在语义索引（Probabilistic Latent Semantic Indexing，简称pLSI）[13]、潜在狄利克雷分配模型（Latent Dirichlet Allocation，简称LDA）[1]，以及本文主要研究对象之一的Harmonium模型[14]等。

（2）自然场景图像分类

在计算机视觉领域，隐层空间模型可用于图像分割和自然场景分类[15,16]。如

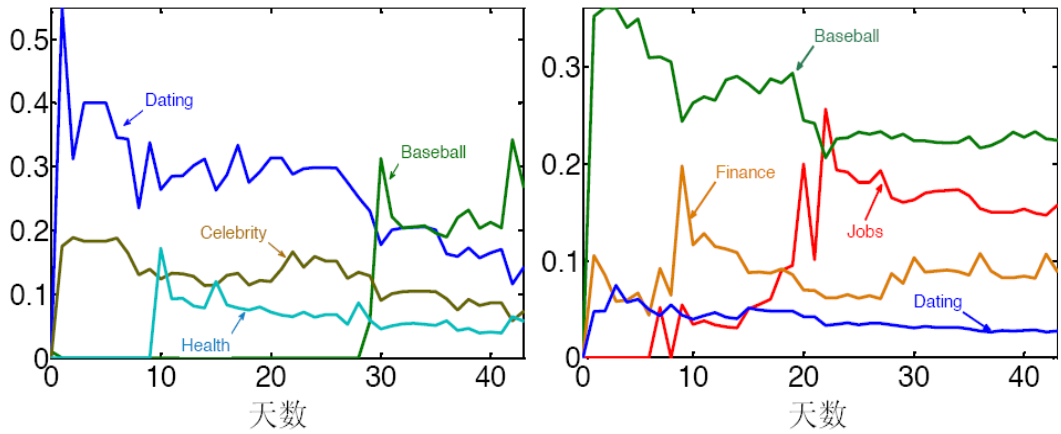


图 1.3　隐空间话题模型自动发现的两位用户40天内在社会网络上的行为表现[11]。

图1.2所示为一个用于自然场景图像分类的有向隐特征模型，将一幅图像分成不同大小的区域用观测变量X表示；H表示隐变量，用于学习更适于分类的区分式隐特征，来表示不同对象（Object）的集合（如天空、草地、动物等）；Y表示图像类别（如图1.2中为动物）；π为选择不同目标的多项式分布的参数，于是，通过对训练数据的学习，不仅可以得到具有区分性的隐层空间表示H，同时可以预测响应变量Y，实现对自然场景图像的分类。

（3）社会关系网络挖掘

隐层空间模型在社会关系网络领域也得到越来越多的关注[11,17]。特别是当处理大规模的弱监督社会关系网络数据时，通常需要掌握事件与事件之间的相互作用。例如，将文本分配到不同的类别中，将用户行为按照其兴趣不同进行分组，或在社会关系网络中为不同用户推荐其感兴趣人物和内容。引用文献[11]中的例子，图1.3 所示为由话题模型[1]得到的两位用户40天内在社会关系网络上的感兴趣活动的变化，这些活动的变化曲线是模型自动产生出来的，而活动的名字是由人工标记的。可以发现，这些关于活动的曲线很好地解释了用户的感兴趣行为，此信息对于向不同用户提供不同的广告推荐具有重要价值。

（4）多模态数据分析

信息技术的飞速发展为用户提供了包含大量多媒体信息的数据（包括相关的文本、图像、音频、视频以及其他多模态信息）。隐空间模型可对多模态数据建模，通过学习隐层空间表示来捕捉多模态数据之间的潜在关系。例如，有向图狄利克雷分配（Correspondence Latent Dirichlet Allocation）隐特征模型[18] 以及无向图的多模态Harmonium模型[3]可以表示相同图像的不同区域以及边缘文本信息的潜在相关性，并能自动进行图像标注。关于多模态数据的分析与理解，本文将会在第3，4，5章详细介绍。

1.2 相关研究现状

按照隐层空间模型的表示、学习、以及模型复杂度问题三个方面，本小节对隐层空间模型的研究进行文献综述。

1.2.1 隐层空间模型的表示问题

如前所述，这些隐变量模型被成功应用于文本挖掘、计算机视觉、社会网络分析等多种应用中。目前隐层空间学习的主流方法可分为如下两种。

（1）基于观测数据统计量的确定性优化方法。包括经典主成份分析（Principal Component Analysis，简称PCA）[19,20]，独立成份分析（Independent Component Analysis，简称ICA）[21]，典型相关分析（Canonical Correlation Analysis，简称CCA）[22]、Fisher线性判别分析（Linear Discriminant Analysis）[23]等。这种确定性方法的优点是简单、高效，但是缺点也很明显：不能处理数据的缺失问题。

（2）基于概率模型的不确定性推理方法。这种方法通常可用概率图来直观地表示。按照图结构的不同可分为：基于有向图的贝叶斯网络，包括有向图潜在狄利克雷分配模型（ LDA）[1]；以及基于无向图的马尔可夫网络，包括无向图受限波尔兹曼机（RBM：Restricted Boltzmann Machine）[10]等。相比之下，基于统计理论的概率模型使用概率分布形式表示可观测数据和隐变量，它具有很好的统计理论基础，可以根据不同数据采用不同的概率假设，在学习隐层空间表示的同时，处理数据缺失问题，已成为隐层空间学习的主要框架之一。

本文着重研究基于概率图的隐层空间模型。下面分别对基于有向图和基于无向图的概率隐层空间模型稍做更详细的介绍：

（1）基于有向图的贝叶斯网络，包括混合高斯模型（Mixture of Gaussians，简称MOG）[24]、概率主成份分析（Probabilistic PCA）[25]、指数族主成份分析（Exponential Family PCA，简称ePCA）[26]、概率潜在语义索引（Probabilistic Latent Semantic Indexing，简称pLSI）[13]、潜在狄利克雷分配（LDA）[1]等；这种有向图的特点是各互相关联节点之间存在具有因果倾向的有向依赖关系，通常需要引入领域知识等先验信息，并建立似然模型，通过贝叶斯推理，从而得到研究者感兴趣的后验分布结果。有向图模型的优点是数据之间的依赖关系明确，可以使用多种方法（例如变量消元等精确推理方法，以及基于均值场假设（Mean Field Assumption）的置信度传播算法（Belief Propagation）等近似推理方法）对其进行推理；而缺点是有向图中可能存在V-结构（V-Structure），会导致“Explain Away”

效应①，影响模型的后验推理效率。但是近年来高效的推理算法（如均值场假设等变分方法）的提出有望提高模型中的推理效率。

（2）相比之下，基于无向图的马尔可夫网络隐空间模型的代表性工作包括联合机器（Combination Machine）[27]，受限波尔兹曼机[10]，指数族 Harmoniums模型[14]等。其特点是各节点之间是成对的相互影响关系，而没有明显的因果依赖关系，因此只需定义模型的联合分布，不需定义先验分布和似然函数。特别的，以受限波尔兹曼机和Harmonium模型为代表的无向图隐特征模型具有简单的二部图（Bipartite Graph）结构（如本文第2章图2.1所示）。当已知一组变量时，另一组变量间满足条件独立性。因此这种无向图隐层空间模型的推理简单、高效。同时，它可以对多模态数据建模。但这些模型的缺点是似然函数中存在很难计算的归一化因子（又叫配分函数），近年来研究界提出多种近似的推理算法[10][14]，用以避免对复杂配分函数的计算。其中，Contrastive Divergence方法使用两个KL散度的差值近似似然函数，在无向隐特征模型中得到广泛应用[10,14,28]。

本论文的主要内容将围绕着基于无向图的隐层空间模型，研究其在典型的多模态数据分析与融合任务中的模型表示、判别式学习以及模型复杂度自动确定等若干关键问题。

1.2.2 隐层空间模型的学习问题

对于无监督概率图模型，最大似然估计是最常用的参数学习方法。但是由于没有考虑任务的额外的有监督信息（如分类问题中的类别标注等），无监督方法不能独立完成预测任务（如分类和回归），学习得到的隐层空间表示一般没有显著的判别性和区分性，对于预测任务通常是次优的。

随着互联网共享技术的飞速发展，互联网上许多有监督数据集如Flickr②，ImageNet③， LabelMe④，TripAdvisor⑤ 等为用户提供大量的样本标签、用户评价分数等多种有监督信息，都可几乎“免费”地获取，如图1.4所示。如何有效利用这些大规模的有监督信息，学习判别式隐层空间表示，提高模型的预测性能，是研究者面临的挑战性问题。有必要提出新的模型和训练方法以期提高模型的预测能力，得到更加具有判别性/区分性的隐层空间表示。

① 当概率图中存在V-结构，即两个变量X和Y分别有边同时指向一个隐变量H时，可以得到$X \perp Y$，但不满足条件独立$X \perp Y|H$，这种效应即为“Explain Away”效应。
② http://www.flickr.com/
③ http://www.image-net.org/
④ http://labelme.csail.mit.edu/
⑤ http://www.tripadvisor.com/

图 1.4　有监督隐层空间学习示意图。

于是，研究者相继提出多种有监督的概率图隐层空间模型。例如，有监督的有向图模型包括：用于分类和回归分析的产生式有监督LDA[29]和判别式LDA[30]，用于同时分类和标注的贝叶斯模型[31]，联合文本和有监督评价分数进行情感分析的贝叶斯模型[32]等；有监督的无向图模型包括：判别式受限波尔兹曼机[33]，分层的Harmonium模型[34]等。与无监督模型相比，这些模型可以在学习隐层空间表示的同时实现预测任务。然而，上述所有有监督概率模型都基于最大似然估计的学习方法，得到的预测性能通常不如使用判别式最大间隔分类器的预测结果[35,36]，因此，如何改进有监督隐层空间模型的学习方法、提高模型的预测性能，是本文研究内容的根本出发点之一。

最大间隔的学习方法是通过计算每个样本相对于决策平面的距离，从而选取间隔最大的决策平面的一种确定性学习方法，这种“间隔”可被作为广义分类误差的上界。由于其高效、简单的原理，已被成功应用于分类、回归等预测问题中[37–39]。如何将最大间隔的学习方法用于不确定性的概率隐层空间模型中，是一个有趣且有挑战性的问题。于是，文献[40,41]中提出将最大间隔的方法分别用于有向图话题模型和用于理解自然场景图像的话题模型，与基于最大似然估计的学习方法相比，显著提高了模型的预测性能。这些成功为本文进一步研究隐层空间模型的判别式学习方法带来了希望。本文将在第3，4，5章中分别研究参数化的判别式隐层空间模型的基于最大间隔准则的学习方法，以及非参数化的判别式隐层空间模型的基于最大间隔准则的推理方法。

1.2.3 隐层空间模型的复杂度问题

以上内容中介绍的所有隐层空间模型都是参数化模型。即当给定观测数据时，模型的复杂度固定。但所有参数化隐层空间模型都存在一个公认难题：无法事先确定模型的隐层特征的维度（这里指隐变量的数目）。最常用的策略是使用“模型选择”，例如交叉验证（Cross Validation），似然比测试（Likelihood Ratio Test）等，即设置不同数量的隐变量进行多次的模型学习，从中选取在特定数据集上性能最好的隐特征个数。但这种方法由于需要比较多组模型的实验结果，时间代价通常很高；同时由于只能比较（较少的）有限多个可能的模型，因此无法保证一定得到最优解。近年来，非参数化方法在隐变量模型中的成功运用为研究者带来了新契机。非参数方法将模型复杂度考虑在模型推理过程中，可以自动地确定一定意义下最优的模型复杂度。它允许在例如无限深度的树[42]、划分（Partitions）[43]、隐特征[44]、测度（Measures）[45]和函数（Functions）[46] 等对象空间上建立概率分布，将传统的参数化贝叶斯先验替换为随机过程先验。非参数化方法假设模型（如隐类别模型/隐特征模型）中有无限多个成份或特征，从候选的无限维隐变量中，通过非参数贝叶斯推理确定一个能够充分表示有限观测数据的候选变量子集。

非参数化模型的代表性工作包括：用于聚类[47]或预测问题的[48]狄利克雷过程混合模型（Dirichlet Process Mixture，简称DP Mixture），基于印度自助餐随机过程（Indian Buffet Process，简称IBP）先验的隐特征模型[49]，以及用于预测任务的无限高斯过程混合模型[50]等。但是，标准的非参数贝叶斯模型一般局限于对观测数据做严格但不实际的假设。例如，假设观测值为“同质”（Homogeneous）或 “可互换”（Exchangeable）的。随着智能处理复杂数据的需求不断增加，很多最新的关于非参数贝叶斯的研究试图放宽这些约束，以更好地适应复杂数据处理的要求。例如，为了处理异质（Heterogenous）的观测数据，文献[51]提出属性依赖的狄利克雷（Predictor-Dependent）随机过程；为了松弛“可互换”假设，最近的研究工作提出不同的方法将多种关联结构（Correlation Structures）引入到非参数化随机过程中，其中包括层次结构（Hierarchical Structures）[52]，时域/空域依赖（Temporal or Spatial Dependencies）结构[53]，以及随机排序依赖（Stochastic Ordering Dependencies）结构[54,55]等。但所有这些方法都是单纯通过设计一些具有特殊结构的非参数化贝叶斯先验分布，通过结合似然模型间接地影响模型的后验分布。事实上，“后验分布”才是研究者真正关心的对象，它包含描述问题本质的隐含结构信息。此外，这种间接的推理方法得到的隐空间表示通常判别性不强，例如在分类任务上的性能一般比不上直接使用判别式学习的方法（例如支持向量

表 1.1 各章节研究内容在模型表示、学习、复杂度三方面的具体呈现。

任务 / 章节	模型表示	学习方法	模型复杂度
第3章	参数化多模态无向图	最大间隔学习	固定/模型选择
第4章	参数化多模态无向图	最大间隔学习	固定/模型选择
第5章	非参数化多模态无向图	最大间隔学习	自动

机等)。根据判别式学习的思想，本文期望直接在后验分布中引入正则化约束来控制隐层空间表示的性质。使用后验正则化因子的另一个原因是：在很多情况下，在正则化的框架下引入领域知识其实更加自然和简单，例如直接在后验分布（而非先验分布）中引入最大间隔约束[40,56–58]或者流形约束[59]等。但是，经典的贝叶斯推理是基于贝叶斯准则的，它并没有一个显示的优化问题，不能直接在贝叶斯公式中引入后验正则化约束。于是，针对已有的非参数化模型中存在的问题，本文将在第5章提出相应的解决办法。

综上所述，鲁棒地学习判别式隐空间表示需要从模型表示、学习方法以及模型复杂度等三个关键方面进行研究。虽然，目前存在的方法在每个方面分别做了一些探讨，但是，如上述文献调研显示，已有的方法还存在许多问题。为了有效利用和复杂数据相关的标注信息，亟需一套系统的理论与方法，有效地学习具有强判别性的隐层空间表示，同时避免模型选择的难题。

1.3 本文研究内容和主要贡献

为了解决上述隐层空间模型的表示、学习方法以及模型复杂度等方面的关键性基础问题，本文系统地研究了如何学习具有强判别性的隐空间表示以及自动确定模型复杂度的问题。具体来说，本文研究基于无向图的判别式有监督隐层空间模型。本着重点突出、循序渐进的原则，本文主要研究内容分为两大部分。第一部分：基于无向图的判别式参数化隐层空间模型。当模型的复杂度固定时，研究参数化有监督隐层空间模型的表示和判别式学习问题；第二部分：基于无向图的判别式非参数化隐层空间模型。研究如何自动确定基于无向图的判别式隐空间模型复杂度的非参数化贝叶斯方法以及如何有效地直接控制后验概率分布的判别性。各章节研究的主要内容在模型表示、学习方法、模型复杂度三个方面的具体呈现请见表1.1所示，可以看出**“基于无向图的模型表示”和“基于最大间隔的学习准则”是串联本文所有研究内容的主线，因此，本文所有内容形成一个完整的体系**。

在研究内容和研究思路上，本文的创新性及核心贡献可用图1.5来阐述。首

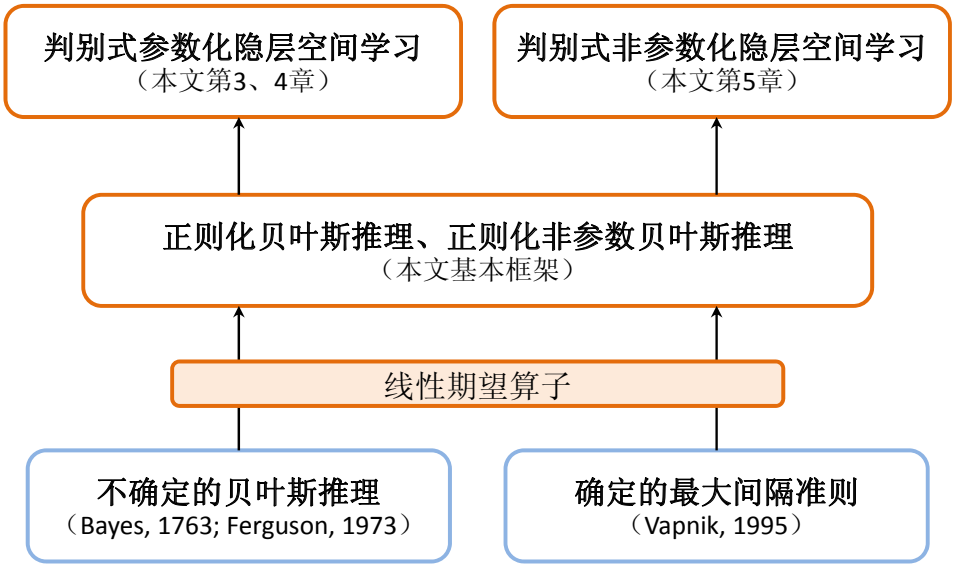


图 1.5　本文的基本框架及主要贡献。

先，本文提出使用线性期望算子将基于无向图概率模型的不确定贝叶斯推理与确定的最大间隔准则有机地融合在一个框架下，即本文提出的适用于链图的正则化贝叶斯推理与正则化非参数贝叶斯推理；其次，本文系统研究了基于最大间隔准则的判别式参数化及非参数化无向图隐空间学习中的若干关键问题。下面具体介绍两部分的研究内容。

第一部分，包括第3，4章。对于参数化有监督隐层空间模型，本文首先对多模态有监督（如包含图像的类别标签）的文本与图像数据建模，考虑多模态观测数据的不同特性及输入数据的互补特性。在模型表示方面，本文提出有监督多模态隐层空间马尔可夫网络（Multi-view Latent Space Markov Network）。与有向的隐层空间概率图模型不同，该模型建立在弱条件独立性假设基础上：即当已知一系列隐层变量时，多模态输入变量之间以及与响应变量满足条件独立。由于此弱条件独立性假设，模型的推理过程简单、高效。在模型学习方面，为了克服传统的基于最大似然估计方法的缺陷，在无向图隐层空间概率模型中引入判别式的最大间隔学习方法，提出基于最大间隔的有监督隐层空间分类和回归模型。为了填补概率推理和最大间隔的确定性学习方法之间的鸿沟，创新性地提出使用线性的期望算子来定义判别函数和预测准则，并且在分类模型中采用基于模型期望的铰链损失函数（Hinge Loss），在回归模型中采用基于期望的ϵ-不敏感损失函数（ϵ-insensitive loss），通过联合最大数据似然与最小化训练数据的预测损失函数进行优化，学习具有强判别性的隐层空间表示。为了解决马尔可夫网络中难以计算归一化因子的问题，本文采用Contrastive Divergence的变分推理方法进行学习

与推理，通过用两项0状态与1状态下的KL散度的差值来近似目标函数，这样可以有效地近似目标函数，同时避免计算复杂归一化因子。鉴于概率图模型可以处理缺失数据的能力，有监督多模态隐层空间马尔可夫网络还可以实现输入变量级别的预测：如图像自动标注功能。本文将多模态隐层空间马尔可夫网络运用在图像、文本等不同数据集上：（1）TRECVID video，（2）Flickr web image，（3）Hotel review数据集。分别用于图像分类、检索、自动标注，以及文本点评分数的预测。实验结果表明这种基于最大间隔的隐层空间马尔可夫网络与基于最大似然估计的隐层空间学习方法相比，不仅预测（如分类、回归等）性能得到显著提高，同时还可以学习判别性的隐层空间表示。该部分主要研究成果已发表于机器学习顶级期刊IEEE Trans. on PAMI 2012[60]及顶级会议NIPS 2010[36]和ICML 2010[61]上，并且该部分工作是优秀973项目“基于视觉认知的多模态信息融合与交互”（项目号2007CB311003）的代表性成果之一。

第二部分，包括第5章。研究基于无向图的非参数化判别式隐层空间模型中的学习与复杂度问题。基于统计学家Zellner教授于1988年提出的贝叶斯对偶理论[62]，即由贝叶斯推理得到模型后验分布等价于求解一个优化问题，本文作者合作提出带有后验约束的“正则化贝叶斯推理”理论框架[57]。在该框架下，成功地将近20年来互相独立的机器学习子领域非参数贝叶斯推理与最大间隔判别式学习有机地融合在一起。不仅可以避开隐层空间模型的模型选择问题，还可以提高模型的预测性能。但是，对于本文主要研究的对于——基于无向图的隐空间模型，如何有效地利用非参数化贝叶斯方法的长处自动确定其模型复杂度仍然是一个开放的问题。本文的主要贡献之一是将用于有向图贝叶斯网络的贝叶斯推理及正则化贝叶斯推理推广到无向图，自动确定第3，4章提出的隐空间马尔可夫网络的模型复杂度。因此，本文所有研究内容形成一个完整的体系，如图1.5示。该部分工作发表于机器学习顶级会议NIPS 2011[57,63]和ICML 2011[58]。

具体地说，至今为止，绝大多数非参数化隐层空间模型都无一例外地在有向图贝叶斯网络框架下提出的[53,64,65]。尚未有人使用非参数贝叶斯方法解决无向图隐变量马尔可夫网络（如指数族Harmonium模型(EFH)[14]）中的模型选择问题，原因是基于贝叶斯推理的无向图马尔可夫网络模型是一个链图，它具有和贝叶斯网络不同的马尔可夫性质，无法直接运用贝叶斯定理和正则化贝叶斯推理。为了扩大非参数贝叶斯方法在无向隐变量模型中的使用，本文将 Zellner教授提出的贝叶斯对偶理论推广至链图以及含有经验参数的模型中。在此理论框架下，提出非参数化的无向图隐空间模型，即无限指数族Harmonium模型（Infinite Exponential Family Harmoniums，简称iEFH模型），用于自动确定无向的隐层空间模型的模型

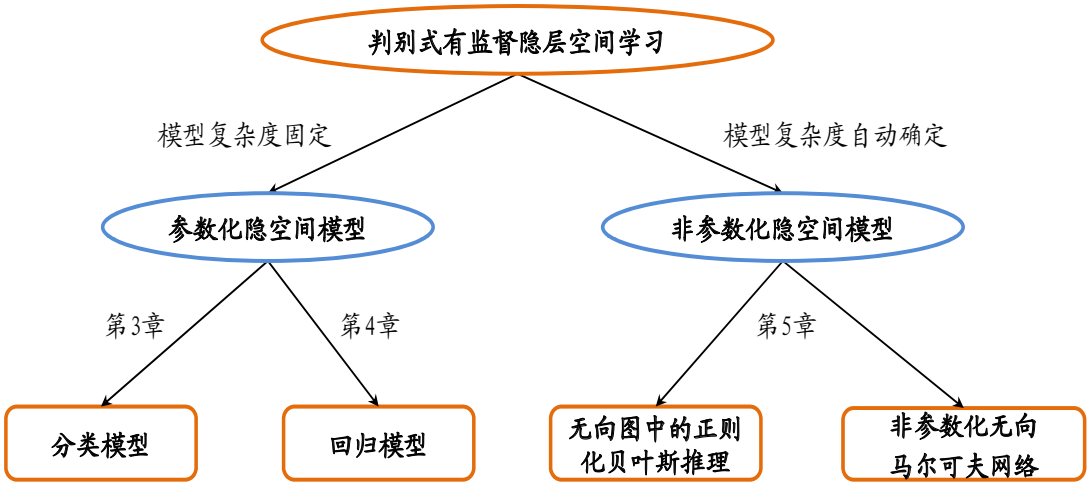


图 1.6　本文各章节关系。

复杂度。更具体一点说，本文将每个隐特征与二值化的指示变量相关联，在这些二值变量（即二值矩阵）基础上假设一个稀疏的印度自助餐随机过程（Indian Buffet Process，简称IBP）先验[49]。由此得到的模型为一个链图模型[66]。此外，本文将论文[57]提出的正则化贝叶斯推理推广至链图及有经验参数的模型中。在此基础上，本文提出有监督无限维最大间隔Harmonium模型（Infinite Max-margin Harmonium，简称iMMH模型），此模型不仅可以自动确定模型复杂度，还可以通过在隐变量的后验分布上引入最大间隔约束，发现判别式的隐层空间表示。

1.4　本文各章节的结构安排

本文的组织结构如下，各章节之间的关系，请见图1.6所示。

第1章为引言部分，概括本文的主要研究内容和贡献。

第2章介绍本文后续章节所需的基础理论背景知识。

第3章提出基于最大间隔的参数化多模态隐层空间马尔可夫网络分类模型。利用有监督信息（如图像的类别标签等），学习判别式的隐层空间表示，同时提高模型的分类性能。

第4章提出基于最大间隔的参数化多模态隐层空间马尔可夫网络回归模型，以及有结构化输入的隐层空间分类模型。不仅可以用于Hotel Review文本评价分数回归分析问题，还可对有段落依赖关系的Review文本进行建模，学习判别式隐层空间表示的同时，提高模型的预测性能。

第5章提出基于无向马尔可夫网络的非参数化隐特征模型。我们将正则化贝叶斯推理推广至链图以及含有经验参数的正则化贝叶斯推理。在此框架下，提出

无限指数族Harmonium模型。根据推广的贝叶斯推理对偶理论及正则化贝叶斯推理，将iEFH推广到有监督无限维最大间隔Harmonium模型，不仅可以自动确定模型的复杂度，同时通过引入最大间隔约束来正则化隐变量，发现判别式的隐层空间表示，

第6章对本文的研究内容和结果进行总结，并展望未来的可行工作。

1.5 本章小结

本章首先概述了判别式隐层空间学习的研究背景及意义；通过对国内外研究现状的综述，提出判别式隐层空间模型相关研究面临的关键问题和挑战；进而阐述本文的研究内容及创新点；最后介绍本文各章节的组织结构。

第 2 章 基础知识

第1章介绍了本文的基本研究内容，本章将从概率隐层空间模型的表示、模型的学习与推理方法、以及模型复杂度等三方面分别介绍本论文的背景基础知识。其中，模型的表示包括：经典无向图隐空间模型即受限波尔兹曼机、指数族Harmonium模型等；学习与推理方法部分包括：最大似然估计和本文后续章节将要系统研究的最大间隔学习准则；隐层空间的模型复杂度部分包括本文第5章将要系统研究的非参数化方法及若干典型的随机过程。所有这些内容构成了本文后续章节的根本性研究基础。

2.1 概率隐层空间模型的表示

这里主要介绍基于无向图的隐层空间模型的两个经典例子，即受限波尔兹曼机和指数族Harmonium模型。

2.1.1 受限波尔兹曼机

受限波尔兹曼机（Restricted Boltzmann Machines，简称RBM）① 是一种特殊的对数线性马尔可夫随机场（Markov Random Field，简称MRF），其能量函数与未知的自由参数成线性关系。RBM中包含很多隐变量，可以表示复杂的数据分布（例如从有限的参数设置到无限的非参数设置等）。通过引入更多的隐变量（也称为隐单元），研究者可以增加波尔兹曼机模型的表示能力。RBM的一个基本假设是模型中只有可观测变量和隐变量间存在对应边相关联，而观测变量之间或者隐变量之间没有边相关联，因此叫做受限的波尔兹曼机模型。图2.1所示为RBM模型的典型图结构，图中$\mathbf{X}$表示N维可观测变量， $\mathbf{H}$表示K维隐变量。

具体地说，RBM的能量函数可以定义为

$$E(\mathbf{x},\mathbf{h}) = -\boldsymbol{\alpha}^{\top}\mathbf{x} - \boldsymbol{\beta}^{\top}\mathbf{h} - \mathbf{h}^{\top}\mathbf{W}\mathbf{x} \tag{2-1}$$

其中$\mathbf{W}$表示连接输入变量与隐变量的权值矩阵，$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$分别表示与特征函数对应的权值向量。由于RBM的特殊结构，各观测变量间，以及各隐变量间在分别已知隐

① http://deeplearning.net/tutorial/rbm.html

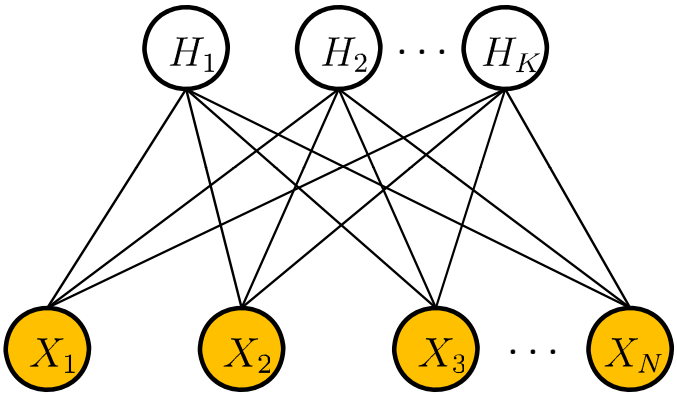


图 2.1　受限波尔兹曼机的典型图结构。

变量和观测变量时满足条件独立性。这个性质可以写为

$$p(\mathbf{h}|\mathbf{x}) = \prod_k p(h_k|\mathbf{x})$$
$$p(\mathbf{x}|\mathbf{h}) = \prod_n p(x_n|\mathbf{h})$$

经典RBM中观测变量与隐变量都是离散的二值变量，即$x_n(n = 1, ..., N)$ 以及 $h_k(k = 1, ..., K) \in \{0,\ 1\}$。于是，可以得到

$$p(h_k = 1|\mathbf{x}) = \text{Sigmoid}(\beta_k + \mathbf{W}_{.k}\mathbf{x})$$
$$p(x_n = 1|\mathbf{h}) = \text{Sigmoid}(\alpha_n + \mathbf{W}_{n.}\mathbf{h})$$

其中$\text{Sigmoid}(x) = \frac{1}{1+\exp(-x)}$。基于以上定义，二值RBM的能量函数可简化为

$$E(\mathbf{x}) = -\boldsymbol{\alpha}^\top\mathbf{x} - \sum_k \log(1 + \exp\{\beta_k + \mathbf{W}_{.k}\mathbf{x}\}) \tag{2-2}$$

当给定一组训练样本$\mathcal{D} = \{\mathbf{x}_d\}_{d=1}^D$时，可以使用经典的最大似然估计方法对模型进行参数学习 ①。但是，即使学习一个具有很少隐变量的RBM也是一个有挑战性的问题，其困难主要在于计算模型分布的配分函数（Partition Function），又称归一化因子。当模型的隐变量增多时，此配分函数将会非常复杂无法计算。于是研究者相继提出多种近似的方法，其中一个重要的工作是Contrastive Divergence，这是由Goeffrey Hinton 等人在2001年提出的[10,28]，后经若干学者进一步发展成为学习包含隐变量马尔可夫网络的最重要工具之一。

① 关于RBM模型的参数学习，读者可参考如下链接：
http://www.iro.umontreal.ca/ lisa/twiki/bin/view.cgi/Public/DBNEquations

2.1.2 指数族Harmonium模型

在离散的受限波尔兹曼机基础上，Max Welling教授等人于2004年[14]提出指数族Harmonium模型（Exponential Family Harmoniums，简称EFH）。EFH是RBM的推广，即允许输入变量和隐变量服从更广义的指数族分布① 。与RBM一样，EFH模型也是一个具有两层结构的马尔可夫网络，包括两层变量（即输入变量$\mathbf{X}$ 和隐层变量 $\mathbf{H}$）。对于每一个数据$d \in \{1, \cdots, D\}$，$\mathbf{x}_d = \{x_{d1}, \cdots, x_{dN}\}$表示观测变量集合。此时，模型的联合概率为

$$p(\mathbf{x}, \mathbf{h}) \propto \exp\Big\{ \sum_n \alpha_n f(x_n) + \sum_k \beta_k g(h_k) + \sum_{n<k} W_{nk} f(x_n) g(h_k) \Big\} \tag{2-3}$$

其中$f(\cdot)$和$g(\cdot)$表示定义在一个单独顶点的特征函数②，即充分统计量；$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$为与特征函数对应的权值向量。这个联合概率中其实还有一项对数配分函数（归一化因子）没有明确地写出来，因为这一项通常比较复杂，不能写出解析表达式。与RBM类似，由于EFH同样满足一些条件独立性（如给定$\mathbf{X}$，$\mathbf{H}$之间相互独立等）。观测变量与隐变量的条件概率分布即可写成因子化形式 $p(\mathbf{x}|\mathbf{h}) = \prod_n p(x_n|\mathbf{h})$，　$p(\mathbf{h}|\mathbf{x}) = \prod_k p(h_k|\mathbf{x})$，且

$$p(x_n|\mathbf{h}) = \exp\{\sum_a \hat{\alpha}_{na} f_{na}(x_n) - A_n(\hat{\alpha}_{na})\}$$
$$p(h_k|\mathbf{x}) = \exp\{\sum_b \hat{\beta}_{kb} g_{kb}(h_k) - B_k(\hat{\beta}_{kb})\}$$

其中$A(\cdot)$和$B(\cdot)$表示对数配分函数（Log-partition Function），带有偏移量的参数$\hat{\alpha}$及$\hat{\beta}$为

$$\hat{\alpha}_{na} = \alpha_{na} + \sum_{kb} W_{na} g_{kb}(h_k), \quad \hat{\beta}_{kb} = \beta_{kb} + \sum_{na} W_{kb} f_{na}(x_n) \tag{2-4}$$

因此在输入特征给定的情况下，推理隐层空间表示（即$p(\mathbf{h}|\mathbf{x})$）很简单，效率很高。特别的，当模型中$\mathbf{X}$与$\mathbf{H}$都取连续值时，EFH的联合高斯分布可定义为

$$p(\mathbf{x}_d, \mathbf{h}_d) \propto \exp\Big\{ \boldsymbol{\alpha}^\top \mathbf{x}_d + \boldsymbol{\beta}^\top \mathbf{h}_d - \frac{\mathbf{x}_d^\top \mathbf{x}_d}{2\sigma_{d1}^2} - \frac{\mathbf{h}_d^\top \mathbf{h}_d}{2\sigma_{d2}^2} + \mathbf{x}_d^\top \mathbf{W} \mathbf{h}_d \Big\}, \tag{2-5}$$

其中 $\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}$（$N \times K$ 矩阵），$\{\sigma_{d1}^2, \sigma_{d2}^2\}$为模型参数。此模型本文接下来将在第3，4，6章用到。

① 指数族包括多种常用的概率分布家族，包括二值伯努利分布、连续的正态分布、指数分布、Gamma分布、chi-squared分布、Beta分布、狄利克雷（Dirichlet）分布、泊松分布等。

② 当$f(\cdot)$和$g(\cdot)$分别定义为$f(x_n) = x_n$，$g(h_k) = h_k$时，模型将直接使用输入变量和隐变量，其能量函数与RBM中的相似。

值得说明的是，EFH中所有的输入变量和隐变量并不是边缘独立的，但是满足条件独立性，这是有向图模型中没有的优点。如上所述，这种条件独立性显著降低了模型的推理代价。但是与此同时，由于很难计算的全局化配分函数的存在，Harmonium模型中的学习问题比较复杂。同样地，Contrastive Divergence方法是有效地学习Harmonium模型最成功的方法之一。

2.2 有监督隐层空间模型的学习方法

下面介绍两种隐层空间模型的参数学习方法，包括面向概率模型的最大似然估计，以及面向确定性判别式学习的最大间隔学习准则。

2.2.1 最大似然估计的近似推理方法

最传统的隐层空间模型的学习方法是最大似然估计。下面本文介绍有向图隐层空间模型与无向图隐层空间模型的最大似然估计方法。

2.2.1.1 有向图隐层空间模型中的最大似然估计

含有隐变量的有向图隐层空间模型中的最常用学习方法是期望-最大化（Expectation-Maximization，简称EM）算法[24,67–69]。

下面举个简单的例子说明。已知一个统计模型，其中包含可观测数据$\mathbf{X}$、一系列隐变量$\mathbf{H}$，以及未知参数Θ。此时模型的对数似然函数为$L(\Theta;\mathbf{X}) = \log p(\mathbf{X}|\Theta) = \log \int_{\mathbf{H}} p(\mathbf{X},\mathbf{H}|\Theta)d\mathbf{H}$。于是，最大似然估计就是通过最大化$L(\Theta;\mathbf{X})$来估计参数$\Theta$。然而，由于对数运算里面包含积分运算（当$\mathbf{H}$为离散变量时，积分运算为加运算），所以这个边缘似然函数通常非常复杂，很难计算。为了解决上述困难，EM算法的原理可以理解为一种变分的方法。具体地说，这里引入分布$q(\mathbf{H})$，然后通过Jensen不等式可以得到似然函数的一个下界

$$\begin{aligned} L(\Theta;\mathbf{X}) &= \log \int_{\mathbf{H}} q(\mathbf{H}) \frac{p(\mathbf{X},\mathbf{H}|\Theta)}{q(\mathbf{H})} d\mathbf{H} \\ &\geq \int_{\mathbf{H}} q(\mathbf{H}) \log \frac{p(\mathbf{X},\mathbf{H}|\Theta)}{q(\mathbf{H})} d\mathbf{H} \\ &= \mathbb{E}_{\mathbf{H}}[\log p(\mathbf{X},\mathbf{H}|\Theta)] + \mathcal{H}(q(\mathbf{H})), \end{aligned} \tag{2-6}$$

其中，$\mathcal{H}(p)$表示概率分布p的熵。然后通过两步迭代的方法可以推导出EM算法的基本步骤:

（1） E步（Expectaton step）：固定参数$\Theta^{(t)}$，最大化下界可以得到$q(\mathbf{H}) = p(\mathbf{H}|\mathbf{X}, \Theta^{(t)})$，同时可以计算

$$Q(\Theta|\Theta^{(t)}) = \mathbb{E}_{\mathbf{H}|\mathbf{X},\Theta^{(t)}}[\log L(\Theta; \mathbf{X}, \mathbf{H})]$$

（2） M步（Maximization step）：通过最大化$Q(\Theta|\Theta^{(t)})$进行参数估计

$$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}}\, Q(\Theta|\Theta^{(t)})$$

值得注意的是，在没有对$q(\mathbf{H})$做任何限制的情况下，其最优解等于真实的后验分布$p(\mathbf{H}|\mathbf{X}, \Theta^{(t)})$。但是对于复杂的模型，通常是不能精确求解后验分布的，此时，需要对$q(\mathbf{H})$做一些约束，如均值场假设[68]等，以得到尽量精确的近似解。

2.2.1.2 无向图隐层空间模型中的最大似然估计

相比于有向图隐层空间模型，无向图隐层空间马尔可夫网络的最大似然估计一般更具有挑战性，其困难主要在于计算模型分布的配分函数。当模型的隐变量较多时，此配分函数将会变得非常复杂，导致无法计算。

为了避开很难计算的对数似然函数，研究者相继提出多种近似的方法，例如均值场变分近似[70]、Contrastive Divergence方法[28]、Gibbs 采样法[71]、Langevin方法[72]等。这里详细介绍Contrastive Divergence（CD）方法，如前所述，CD方法由Geoffrey Hinton等人在2001年提出[28]，后经若干学者进一步发展[10,73–75]，它已成为学习无向图隐变量模型中最常用的变分近似方法之一[76–78]。

假设p表示模型的真实分布，q表示模型的近似分布。为了避开很难计算的负对数似然函数$-L(\Theta)$，这里使用与第2.2.1.1节中类似的方法推导一个上界$L'(\Theta)$来近似负对数似然函数，这个上界实际上是一个KL散度，可以被描述为两个自由能量函数的差值

$$L(\Theta)' = F_0 - F_\infty$$

上式第一项F_0中的可观测变量取观测到的真实值，第二项F_∞是自由能量函数，其中所有变量的值都是未知的。由于F_∞仍然很难计算， Hinton教授提出可以使用马尔可夫链蒙特卡罗（Markov chain Monte Carlo，简称MCMC）方法，通过n步迭代来重建一个近似的分布，这就是Contrastive Divergence学习的基本原理，即

$$\begin{aligned} L(\Theta)' &= F_0 - F_\infty \\ &\approx F_0 - F_i = \mathrm{KL}(q_0 \| p) - \mathrm{KL}(q_n \| p) \end{aligned} \tag{2-7}$$

其中p表示模型分布，q_0表示当可观测变量取可观测数据值时的变分分布，而q_n中所有变量都是未知的。在Contrastive Divergence变分推理过程中，从变分分布q_0开始使用马尔可夫链，然后只需运行很少的n步（如$n = 1$）即可得到q_n。这种方法显著地降低了模型的计算复杂性，诸多文献[10,73–75]证明此方法可以得到关于负对数似然$L(\Theta)$的理想的近似。

下面以本章2.1.2节介绍的指数族Harmonium模型为例，介绍Contrastive Divergence的变分近似方法。EFH模型中的负对数似然函数为[14]

$$L(\Theta) = -\sum_n \log p(x_n) = -\sum_n \log\Big\{\sum_a \alpha_{na} f_{na}(x_n) + B_k(\beta_{kb} + \sum_{na} W_{na}^{kb} f_{na}(x_i))\Big\} - \log Z$$

其中Z表示很难计算的配分函数。于是，在EFH模型中使用CD方法，令$\mathcal{L}(q_0, q_1)$近似负对数似然函数$L(\Theta)$：

$$\mathcal{L}(q_0, q_1) = \mathrm{KL}(q_0(\mathbf{x}, \mathbf{h}) \| p(\mathbf{x}, \mathbf{h})) - \mathrm{KL}(q_1(\mathbf{x}, \mathbf{h}) \| p(\mathbf{x}, \mathbf{h})),$$

其中$\mathrm{KL}(q\|p)$为变分分布q（q_0或 q_1）与模型分布p的KL散度。q_0为当$\mathbf{x}$取可观测数据值时的变分分布，而q_1中所有变量都是未知的。于是通过优化这个近似的目标函数$\mathcal{L}(q_0, q_1)$，可以进行参数估计。这里可以使用梯度下降的方法，其中模型参数的梯度公式为

$$\begin{aligned}
\partial\alpha_{na} &= \mathbb{E}_{q_0}[x_n] - \mathbb{E}_{q_1}[x_n] \\
\partial\beta_{kb} &= \mathbb{E}_{q_0}[B'_{kb}(\hat{\beta}_{kb})] - \mathbb{E}_{q_1}[B'_{kb}(\hat{\beta}_{kb})] \\
\partial W_{nk}^{ab} &= \mathbb{E}_{q_0}[f_{na}(x_n) B'_{kb}(\hat{\beta}_{kb})] - \mathbb{E}_{q_1}[f_{na}(x_n) B'_{kb}(\hat{\beta}_{kb})]
\end{aligned}$$

其中$B'_{kb} = \partial B_k(\hat{\beta}_{kb})/\partial\hat{\beta}_{kb}$，$\hat{\beta}_{kb}$的定义如公式（2-4）所示。

2.2.2　最大间隔学习方法

目前已有的有监督隐变量模型的学习方法大部分是基于如上所述的最大似然估计的，例如用于目标识别和自然场景分割的条件随机场[79,80]、用于网页抽取的多层次马尔可夫网络[81]，用于发现隐层语义结构的有监督话题模型[29,30]，以及用于自然语言处理的含有隐变量的对数线性文法[82]等。然而这些方法的预测性能通常比不上使用判别式最大间隔分类器（如支持向量机）的预测性能[36]。虽然最大间隔学习准则已在很多预测任务中得到广泛成功应用，但在包含隐变量的概率模型中如何有效运用最大间隔学习准则，仍然是个比较有挑战性的问题。主要原因是最大间隔准则缺乏对随机性隐变量的直接的概率描述[83]，所以不确定性的概率图模型与确定性的最大间隔学习准则之间存在一个鸿沟。虽然近年来一些文献中

提出了在隐变量模型中使用类似于最大间隔准则的方法，例如文献[84,85]中忽略隐变量不确定性，使用最大后验概率估计来近似代替隐变量的真实分布。但这种使用最大后验概率估计的方法对于比较平缓的后验概率分布，很可能丢失很多有用信息[83]。相比之下，一个更合理的概率描述方法是考虑隐变量的所有可能取值并且进行线性加权平均，例如使用隐变量的线性期望值等。文献[85]提出一种可以用于预测结构化输出的支持向量机模型，但需要CCCP（Constrained Concave-convex Procedure）进行优化。而且文献[83]中指出这种方法还存在一个明显的缺陷：其结构预测准则的定义与预测准则中损失函数的定义不一致。

最大熵判别式理论[56,83,86,87]的提出为将贝叶斯学习与最大间隔准则巧妙地集成在一起提供了非常好的框架。近年来，相继有一些代表性工作将最大间隔思想用于有向图的包含隐变量概率模型中。具体包括：文献[87]中提出的部分可观测的最大熵判别式马尔可夫网络，用于实现结构化的输出变量的预测。文献[40]中提出了最大熵判别式话题模型，使用最大间隔方法学习判别式的有向图隐层空间模型，以及文献[41]中将最大间隔学习用于自然场景理解的有监督隐变量模型，提高模型的预测性能。本文主要研究如何在基于无向图的隐变量模型以及非参数化贝叶斯推理中考虑最大间隔准则。

下面首先介绍用于多类别分类的最大间隔支持向量机，进而介绍最大熵判别式学习方法。

2.2.2.1 最大间隔多类别支持向量机

最早提出的基于最大间隔的支持向量机分类模型是针对二分类问题的，对于多类别的分类问题，有很多种不同的策略，包括基于二分类支持向量机的“1对1”，“1对其他”等方式，以及Crammer学者等人于2001年提出的多类别的最大间隔学习[88]。下面具体以Crammer学者等人的工作为例介绍多类别分类的最大间隔学习。本文的后续工作将用到这种定义方式。

设输入数据为$\mathcal{D} = \{(\mathbf{x}_d, y_d)\}_{d=1}^{D}$，输入变量$\mathbf{x}_d$的维度为$N$, $y_d \in \{1, ..., T\}$。定义$\mathbf{f}(y, \mathbf{x}_d)$为$NT$维的特征向量，其中从$(y-1)N+1$到$yN$的元素为$\mathbf{x}_d$，而其他元素为0。$\boldsymbol{\eta}$是由$T$个子向量$\boldsymbol{\eta}_y$拼接的向量，其中每一个$\boldsymbol{\eta}_y$与类别标签$y$相对应。这里使用线性的判别函数$F(y, \mathbf{x}_d; \boldsymbol{\eta}) = \boldsymbol{\eta}^{\top} f(y, \mathbf{x}_d)$。此时，模型的预测准则为

$$y = \underset{y}{\operatorname{argmax}}\, F(y, \mathbf{x}_d; \boldsymbol{\eta}) \tag{2-8}$$

于是含有松弛变量ξ的最大间隔分类器求解下面受约束的优化问题

$$\begin{aligned} &\min_{\boldsymbol{\eta},\xi} \ \frac{1}{2}\|\boldsymbol{\eta}\|_2^2 + \frac{C}{D}\sum_d \xi_d \\ &\forall d, \ \text{s.t.}: F(y_d, \mathbf{x}_d; \boldsymbol{\eta}) - F(y, \mathbf{x}_d; \boldsymbol{\eta}) \geq \Delta\ell_d^{0/1}(y) - \xi_d \end{aligned} \tag{2-9}$$

其中$\Delta\ell_d^{0/1}(y) = \ell^{0/1}(y, y_d)$是预测值$y$与真实值$y_d$的0/1代价函数。问题（2-9）中的约束即为最大间隔约束。以上问题等价于求解下面带有铰链损失函数（Hinge Loss）$\mathcal{R}_{hinge}(\mathbf{v})$的无约束优化问题

$$\min_{\boldsymbol{\eta}} \ \frac{1}{2}\|\boldsymbol{\eta}\|_2^2 + \frac{C}{D}\mathcal{R}_{hinge}(\boldsymbol{\eta}), \tag{2-10}$$

其中$\mathcal{R}_{hinge}(\boldsymbol{\eta}) = \sum_d \max_y[\Delta\ell_d^{0/1}(y) - [F(y_d, \mathbf{x}_d; \boldsymbol{\eta}) - F(y, \mathbf{x}_d; \boldsymbol{\eta})]]$。此问题可以用凸优化技术求解，在原问题中对于任意的d与y，引入拉格朗日乘子$\mu_d(y)$，于是模型的对偶问题为

$$\begin{aligned} &\max_{\mu} \ -\frac{1}{2}\|\sum_{d,y}\mu_d(y)(f(y_d, \mathbf{x}_d) - f(y, \mathbf{x}_d))\|_2^2 + \sum_{d,y}\mu_d(y)\Delta\ell_d^{0/1}(y) \\ &\forall d, \forall y, \ \text{s.t.}: \sum_y \mu_d(y) = \frac{C}{D}; \ \mu_d(y) \geq 0 \end{aligned}$$

根据KKT条件，可以得到对偶问题的最优解$\mu^\star$与原问题最优解$\boldsymbol{\eta}^\star$的关系

$$\mu_d^\star(y)\{F(y_d, \mathbf{x}_d; \boldsymbol{\eta}^\star) - F(y, \mathbf{x}_d; \boldsymbol{\eta}^\star) - \Delta\ell_d^{0/1}(y) + \xi_d\} = 0$$

当$\mu_d(y) \neq 0$时，对应的$\mathbf{x}_d$满足原问题的等式约束，此时$\mathbf{x}$在决策边界上，称为支持向量。最大间隔学习中只需要得到并保存少数支持向量即可确定判决边界。目前常用的最大间隔学习方法包括次梯度下降法[83,88]，以及切平面法[89]等。本文将在下面章节的最大间隔学习中使用。

2.2.2.2 最大熵判别式方法

和上述最大间隔多类别分类支持向量机中通过优化某个目标函数学习一个特定的模型参数$\boldsymbol{\eta}^\star$不同，最大熵判别式（Maximum Entropy Discrimination，简称MED）[56,86]学习一种基于平均模型的最大间隔学习方法，它通过最小化相对熵（即KL散度）来学习所有可能模型最优的后验概率分布$p(\boldsymbol{\eta})$。对于多类别的分类问题，最大熵判别式模型定义为

$$\min_{p(\boldsymbol{\eta}),\xi} \ \text{KL}(p(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta})) + \frac{C}{D}\sum_{d=1}^{D}\xi_d \tag{2-11}$$

$$\forall d, \forall y, \text{s.t.}: \int p(\boldsymbol{\eta})[F(y_d, \mathbf{x}_d; \boldsymbol{\eta}) - F(y, \mathbf{x}_d; \boldsymbol{\eta})]d\boldsymbol{\eta} \geq \Delta\ell_d^{0/1}(y) - \xi_d$$

其中$\mathrm{KL}(p\|p_0) = \int p \log(p/q)dp$表示KL散度。

可以证明上小节介绍的线性支持向量机是MED的一个特例[83]。具体地说，若只考虑满足正态分布的先验概率分布情况，即$p_0(\eta) = \mathcal{N}(0,\ I)$，可以很容易地得到$p(\eta)$也是一个正态分布$q(\eta) = \mathcal{N}(\mu_\eta, I)$，此时引入对偶变量$\mu_d(y)$，可以得到下面的对偶问题，于是可用多种支持向量机分类器工具求解此对偶问题。

$$\begin{aligned}
&\max_{\mu} \ -\frac{1}{2}\|\sum_{d,y}\mu_d(y)(f(y_d, \mathbf{x}_d) - f(y, \mathbf{x}_d))\|_2^2 + \sum_{d,y}\mu_d(y)\Delta\ell_d^{0/1}(y) \\
&\forall d, \forall y, \ \text{s.t.}: \sum_{y}\mu_d(y)\Delta\ell_d^{0/1}(y) \in [0, C].
\end{aligned}$$

如前所述，MED的优点是可以系统地考虑模型的不确定性，如包含隐含变量等，在之前的工作中，已将MED用于学习不同的基于有向图的参数化隐变量模型，本文的主要工作将MED的思想系统地推广到无向图隐变量模型以及非参数化贝叶斯推理。

2.3　非参数化贝叶斯推理

非参数化贝叶斯方法是建立在无限维参数空间上的贝叶斯方法，其中参数空间包含学习问题的所有可能解的集合，例如隐类别模型中的无限维成份类别，或者是隐特征模型中的无限维隐特征的集合。通过引入不同的随机过程先验（例如狄利克雷随机过程先验、印度自助餐随机过程先验）可以从无限维的候选解中推理选择一个充分表达观测数据的有限子集。非参数化方法另外一个特点是模型的复杂度可以随数据集的变化而自适应改变，例如在观测数据增加时，非参数化方法通过推理可以适当增加模型复杂度以达到充分表达数据的目标。下面分别介绍经典的狄利克雷随机过程以及印度自助餐随机过程。

2.3.1　狄利克雷随机过程

狄利克雷随机过程（Dirichlet process，简称DP）是1973年由Ferguson[45] 首先提出的用于密度估计的无限维随机过程。后来， Sethuraman教授[90] 提出其Stick-breaking表示。具体来说，一个服从狄利克雷过程的随机测度G可写成

$$G = \sum_{i=1}^{\infty}\pi_i(\mathbf{v})\delta_{\theta,\theta_i},\ \pi_i(\mathbf{v}) = v_i\prod_{j=1}^{i-1}(1-v_j)$$

其中 $\theta_i \sim G_0$, $v_i \sim \text{Beta}(1, \alpha)$，而$\delta_{a,b}$ 为 Kronecker Delta 函数（即若$a = b$，$\delta_{a,b} = 1$；否则$\delta_{a,b} = 0$)。从这个表示框架下，读者可以很清晰地看出，G 几乎确定是离散的[91]，即 G的支撑集（support）包含可数多无限个元素，这些元素是从G_0中独立采样而产生的。

Antoniak教授[47]首先将DP作为混合模型（DP Mixtures）中的先验分布。由于从DP采样的概率分布几乎确定是离散的，所以DP混合模型产生的数据可以通过采样分布的不同值来划分。因此，DP混合模型是一种灵活的混合模型，模型中混合成份的数量是随机的，且随着观察到的新数据的增多而增长。 DP混合模型已成功运用于处理无监督的聚类和密度估计（Density Estimation）等问题中来自动确定未知聚类的数量等。

最近，也有一些工作提出运用狄利克雷过程混合模型进行预测，如分类等，其中代表性工作包括：无限维高斯过程混合模型（Infinite Mixture of Gaussian Processes）[50]和广义线性分类器狄利克雷过程混合模型（DP Mixture of Generalized Linear Models，简称GLMs）[48,92]。但是，这些方法很大程度上局限于以下两方面：第一，虽然基于线性分类器的混合模型（Mixture of Linear Experts）可以产生全局非线性的分类器[48]，但为了捕获复杂数据的局部非线性，它可能导致不必要的多余成份产生；第二，这些模型的分类器都是基于概率模型的分类器，为了实现贝叶斯推理，模型中必须为响应变量定义一个似然模型（例如广义线性模型），这其中通常包含一个很难计算的归一化因子（又叫配分函数），从本质上将后验推理复杂化。

为了解决上述问题，本文的主要贡献之一是将判别式的最大间隔准则与非参数化贝叶斯推理方法有机地融合在一起，并且提出自适应的具有强区分性的隐层空间学习方法。这部分工作将在本文第5章详细阐述。

2.3.2 印度自助餐随机过程

另外一个典型的非参数化随机过程是印度自助餐随机过程（Indian Buffet Process，简称IBP），它是由Thomas L. Griffiths等人在文献[93]中提出。至今为止，IBP已被成功运用于诸多机器学习相关任务（例如社会网络链接分析[94]，多任务学习[95]等）中。IBP是定义在无限维二值特征矩阵上的随机过程。它的产生过程可以形象地用印度自助餐过程描述为

（1） 第一位顾客选择K_1种食物，$K_1 \sim Poisson(\alpha)$;

（2） 第i位顾客：以概率$\frac{m_k}{i}$选择已被之前顾客选过的食物；并且还选择了K_i种其

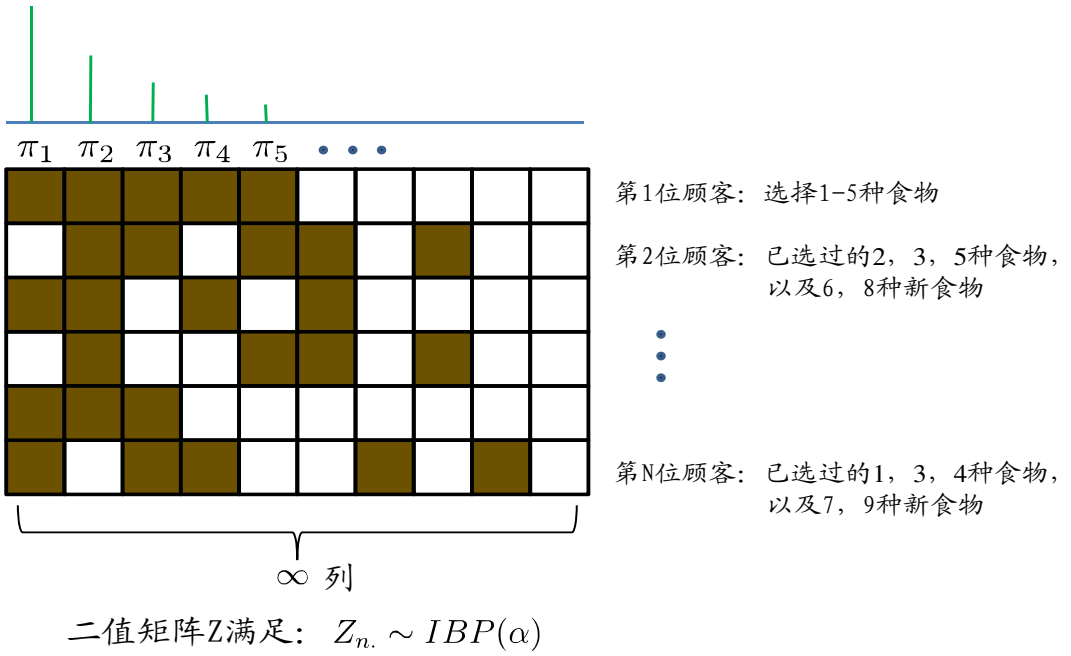


图 2.2　印度自助餐随机过程的示意图。

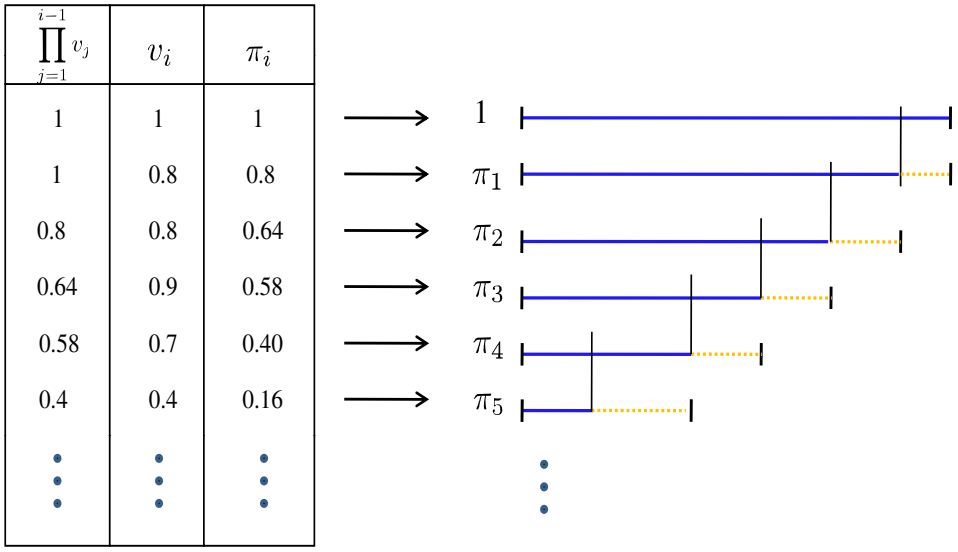


图 2.3　印度自助餐随机过程的Stick-breaking表示。

他的食物，其中$K_i \sim Poisson(\frac{\alpha}{i})$；以此类推；$m_k$表示$k$种食物被顾客选择的次数。

如图2.2所示，图中的方格表示二值矩阵$\mathbf{Z}$的每个元素，此时矩阵$\mathbf{Z}$理论上可以有无限列。

这里着重介绍IBP的Stick-breaking 表示[96]，这种表示对于本文提出高效的推理方法至关重要。令$\pi_k \in (0,1)$表示二值矩阵$\mathbf{Z}$中对应于第k列的参数。当π_k已知，

每一个z_{nk}在第k列独立地从Bernoulli(π_k)中采样得到。参数$\boldsymbol{\pi}$从一个Stick-breaking过程中产生，可表示为

$$\pi_1 = \nu_1, \text{ and } \pi_k = \nu_k \pi_{k-1} = \prod_{i=1}^{k} \nu_i, \tag{2-12}$$

其中$\nu_i \sim \text{Beta}(\alpha, 1)$。如图2.3所示为IBP的stick-breaking表示，如图可以看出这个随机过程将产生一个递减的概率序列π_k。具体地说，当观测到一个有限的数据集，特征k的概率随着k的增长成指数级下降。本文将在第5章研究非参数化无向马尔可夫网络中详细介绍。

2.4　本章小结

本章从概率隐层空间模型的表示、学习与推理方法、以及模型复杂度三方面介绍了本文的基础知识，包括：经典无向图隐空间模型即受限波尔兹曼机、指数族Harmonium模型、最大似然估计，与最大间隔学习，以及经典的非参数学习方法如狄利克雷随机过程、印度自助餐随机过程等。所有这些内容构成了本文后续章节的研究基础。

第 3 章　参数化隐层空间分类模型

本章研究参数化的隐层空间分类模型，并将以多模态复杂数据融合与分析为典型应用场景，具体介绍隐层空间学习的模型表示和学习问题。具体地说，在模型表示方面，为了考虑多模态特征，本章提出一种有监督的无向图机器学习框架，即多模态的隐层空间马尔可夫网络（Multi-view Latent Space Markov Network）。与有向的隐层空间概率图模型不同，该模型建立在弱条件独立性假设基础上：即当已知一系列隐层变量时，多模态输入变量与响应变量之间是条件独立的。该模型可以利用有监督信息（如图像的类别标签等），学习判别式的隐层空间表示。鉴于概率图模型可以处理缺失数据的能力，模型还可以实现对输入变量的预测（如图像自动标注功能）。在模型学习方面，本章克服最大似然估计容易过拟合以及判别性不强的缺点，创新性地在无向隐空间模型中采用基于线性期望算子的最大间隔学习方法，通过联合最小化负对数似然与训练数据的预测损失函数的方法学习模型参数。实验结果表明本文提出的基于最大间隔的隐层空间马尔可夫网络与基于最大似然估计的隐层空间学习方法相比，不仅在预测性能（如分类正确率、F1-score等）方面有很大程度的提高，同时得到更加具有判别性和可解释性的隐层空间表示。

3.1　研究动机

自然界中许多问题都包含来自多个数据源的数据集，或从不同的领域抽取出来的数据。本文把描述同一对象的多种来源的信息叫做多模态数据①。例如，网络环境下描述同一对象的数据可以是包含不同类型的文本网页以及网页之间的链接信息[97]；一段视频数据可以包含多种特征（如关键帧图像的颜色、形状等视觉信息[98] 以及视频的文本字幕信息等），已经被学术界广泛关注和研究的还有很多其他例子[99–103]，在此不一一列举。然而，传统的预测模型方法，例如支持向量机[104]，boosting算法[105]等，通常都没有考虑这些多模态数据的特点。这些方法需要建立在当所有变量都是完全可观测的基础上，而不考虑非同源数据的存在性及不同特性；或者对每一个数据源的数据建立一个独立的弱分类器，而不考虑非同源信息之间的关联。这些方法通常会在一定程度上影响模型的预测性能[106]和计算性能[102]，而且不能实现不同模态输入数据的分析和预测[103]（例如已知图像

① 有时也称多视角数据，在不引起歧义的情况下，本文将不作深入比较并统一称为多模态数据。

的视觉特征，推理未知的文本特征，从而实现图像自动标注任务，或者学习模态之间的隐层关系结构等）。

同时，如前面章节所述，随着互联网共享技术的飞速发展，互联网上许多有监督数据集如Flickr①，ImageNet②， LabelMe③，TripAdvisor④ 等为用户提供大量的样本标签、用户评价分数等多种有监督信息，都可几乎“免费”地获取，如图1.4所示。如何有效利用这些大规模的有监督信息，学习判别式隐层空间表示，提高模型的预测性能，是研究者面临的挑战性问题。有必要提出新的模型和训练方法以期提高模型的预测能力，得到更加具有判别性/区分性的隐层空间表示。

本章的研究目标是运用有监督信息，提出从多模态数据中学习具有强判别性隐层空间表示的统计模型，同时实现输入特征级别的分析预测（如图像标注等）以及响应变量级别的预测（如分类等）。本文的方法与使用多源信息进行半监督学习[97,103,107–109]、非监督聚类[110]以及结构化输出预测问题[111]的工作等有着本质的区别。学习多模态数据的隐层空间表示的代表性工作包括：非监督的模型如典型相关分析（CCA）[22,109]，核化典型相关分析CCA[112]等，但这些模型都忽略了目前广泛存在的有监督信息。因此，得到的隐层空间表示通常不具有很强的区分性，在分类任务上一般不能显著提高模型预测性能。多模态Fisher判别式分析（Fisher Discriminant Analysis，简称FDA）[23]提出一种有监督的方法来学习一个子空间。但是，这种确定性方法不能完成“输入数据的分析与预测”的推理任务，同时，可能需要依赖于密度估计的方法以便于使用信息准则来检测不同模态之间的不一致[100]。相反，概率隐层空间模型以概率分布的形式表示可观测的输入数据（例如图像）和隐层的概念（例如潜在话题[1]），而且概率图模型可以很灵活而方便地使用高效的计算方法进行统计因果推理。概率隐层空间模型已被广泛应用于信息表示和处理系统，可高效地从多源数据中探索潜在隐层表示[18,34,98]。

因此，本章主要研究基于概率图模型的多模态判别式隐层空间表示的学习框架。首先，在模型表示方面，提出一种处理大规模文本和多媒体数据的概率隐层空间模型。此模型建立在一个基于无向图的隐层空间模型基础之上，可用于多种应用（如文本检索、视频分类等）中[14,98,113]。虽然基于有向图的贝叶斯网络（例如用来处理单模态数据的隐狄利特列分配模型（LDA）[1]以及其推广模型[29,40]）等都可以扩展至处理多模态数据的问题。但是，由于贝叶斯网络中通常会存在“V-结构”，导致观测的变量已知时所有的隐变量都互相依赖地关联在一

① http://www.flickr.com/
② http://www.image-net.org/
③ http://labelme.csail.mit.edu/
④ http://www.tripadvisor.com/

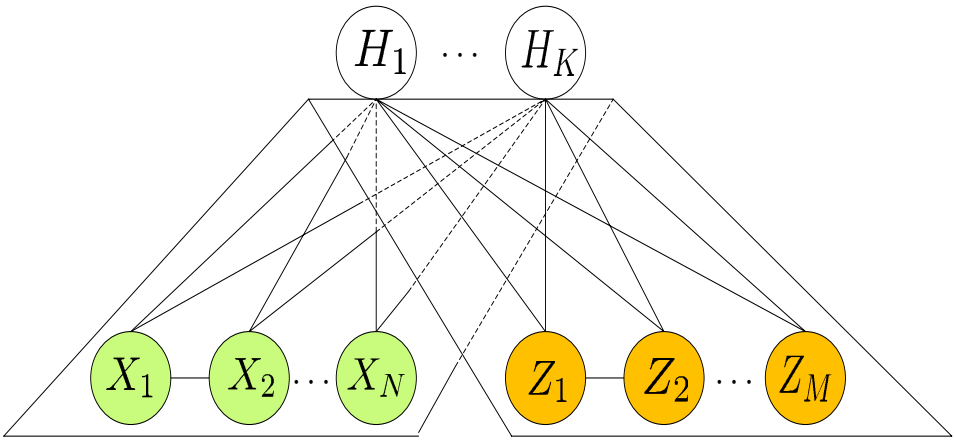


图 3.1　无监督两模态隐层空间马尔可夫网络。

起[14]，从而使得有向贝叶斯网络的后验推理很复杂。其次，在模型学习方面，目前学习概率隐层空间模型的方法，大都局限于最大似然估计的框架下[14,34,98,114]，很可能产生不理想的后果（如过拟合等）。这些限制都制约了模型在多种复杂学习场景（如高维稀疏和有噪声环境）下的效用。当处理带监督信息的数据时，其预测性能有时不如传统分类器（例如基于最大间隔准则的支持向量机等）的性能。为此，本章提出一种新的判别式学习方法用于无向图隐层空间模型中。

3.2　基于最大似然估计的隐层空间马尔可夫网络

下面，本章首先介绍一种无监督的隐层空间马尔可夫网络，在此基础上介绍基于传统最大似然估计的有监督隐层空间马尔可夫网络。

3.2.1　无监督多模态隐层空间马尔可夫网络

图3.1为一个两模态的隐层空间马尔可夫网络的图结构①。它包括两种输入数据 $\mathbf{X} \overset{\text{def}}{=} \{X_i\}_{i=1}^{N}$，$\mathbf{Z} \overset{\text{def}}{=} \{Z_j\}_{j=1}^{M}$（分别对应不同的模态），以及一组隐变量$\mathbf{H} \overset{\text{def}}{=} \{H_k\}_{k=1}^{K}$，其中隐变量$\mathbf{H}$对应着需要推理得到的隐层空间表示。对不同模态对应的输入变量，分别定义一个马尔可夫网络表示变量间互相依赖关系。为了方便描述，这里只考虑每一种模态中输入变量间的一阶依赖关系。这种假设也可很容易地扩展到更加通用的情况，如含有高阶依赖关系等。令E_x表示输入变量$\mathbf{X}$之间的边②，同理，E_z表示输入变量$\mathbf{Z}$之间的边。这里用e表示某个特定的边，而$\mathbf{X}_e$表示与边e相关联的变量。

① 对于两个以上模态模型的结构类似

② 本文将一个单独的顶点看作一条退化的边。

多模态隐层空间马尔可夫网络的的联合分布可以通过如下方式来构造性地定义。首先，分别定义每一种模态输入变量的概率分布，以及隐变量的概率分布。根据随机场理论，对于每一种模态数据定义它的边缘分布为一个指数族分布

$$p(\mathbf{x}) = r(\mathbf{x}) \exp\Big\{ \sum_{e \in E_x} \theta_e^\top \phi(\mathbf{x}_e) - A(\theta) \Big\},$$
$$p(\mathbf{z}) = s(\mathbf{z}) \exp\Big\{ \sum_{e \in E_z} \eta_e^\top \psi(\mathbf{z}_e) - B(\eta) \Big\}, \tag{3-1}$$

其中ϕ 和ψ 是特征函数的向量，θ和η是相应的权值参数，A和B是对数配分函数（log-partition function）。与文献[14]相似，这里把log($r(\mathbf{x})$)和log($s(\mathbf{z})$)当做额外的特征，且其权重为常数1。对于隐变量**H**，每一个成份H_k 都有一个指数族分布

$$p(\mathbf{h}) = \prod_k p(h_k) = \prod_k \exp\Big\{ \lambda_k^\top \varphi(h_k) - C_k(\lambda_k) \Big\},$$

其中$\varphi(h_k)$是定义在h_k的特征函数向量，C_k 是另一个对数配分函数。

有了上述定义的边缘概率分布，还需要引入一些其他项将随机变量**X**与**H**，以及**Z**与**H**关联在一起。综合以上讨论，可以定义该模型的联合概率分布

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}, \mathbf{h}) \propto \exp\Big\{ & \sum_{e \in E_x} \theta_e^\top \phi(\mathbf{x}_e) + \sum_{e \in E_z} \eta_e^\top \psi(\mathbf{z}_e) + \sum_k \lambda_k^\top \varphi(h_k) \\ & + \sum_{e \in E_x, k} \phi(\mathbf{x}_e)^\top \mathbf{W}_e^k \varphi(h_k) + \sum_{e \in E_z, k} \psi(\mathbf{z}_e)^\top \mathbf{U}_e^k \varphi(h_k) \Big\}, \end{aligned} \tag{3-2}$$

其中**W**和**U**分别表示**X**与**H**，以及**Z**与**H**之间的关联强度。从联合概率分布，可以推导出每一模态变量的条件概率分布

$$p(\mathbf{x}|\mathbf{h}) = \exp\{ \textstyle\sum_{e \in E_x} \hat{\theta}_e^\top \phi(\mathbf{x}_e) - A(\hat{\theta}) \}$$
$$p(\mathbf{z}|\mathbf{h}) = \exp\{ \textstyle\sum_{e \in E_z} \hat{\eta}_e^\top \psi(\mathbf{z}_e) - B(\hat{\eta}) \}$$
$$p(\mathbf{h}|\mathbf{x}, \mathbf{z}) = \textstyle\prod_k \exp\{ \hat{\lambda}_k^\top \varphi(h_k) - C_k(\hat{\lambda}_k) \},$$

其中

$$\hat{\theta}_e = \theta_e + \sum_k \mathbf{W}_e^k \varphi(h_k),$$

$$\hat{\eta}_e = \eta_e + \sum_k \mathbf{U}_e^k \varphi(h_k),$$

$$\hat{\lambda}_k = \lambda_k + (\sum_{e \in E_x} \phi(\mathbf{x}_e)^\top \mathbf{W}_e^k + \sum_{e \in E_z} \psi(\mathbf{z}_e)^\top \mathbf{U}_e^k)^\top$$

为原来参数通过平移得到的新参数。可以看到，在隐变量$\mathbf{H}$已知的情况下，$p(\mathbf{x}|\mathbf{h})$和$p(\mathbf{z}|\mathbf{h})$定义了一个条件随机场（Conditional Random Fields，简称CRF）模型[115]，这里$\mathbf{h}$对应于CRF中的全局条件输入，而$\mathbf{x}$ 和$\mathbf{z}$对应于CRF中的有结构的输出。同样地，从联合分布，可以推导出观测数据的边缘似然$p(\mathbf{x},\mathbf{z})$。

相反的过程，这里还可以先定义上面的局部条件分布，然后直接写出与公式(3-2)中的对数线性形式相兼容/相一致的联合概率分布。本章用Θ表示所有的模型参数$(\theta,\eta,\lambda,\mathbf{W},\mathbf{U})$。值得一提的是，指数族Harmonium模型 (EFH)[14]以及它的推广模型两模态Harmonium（Dual-wing Harmonium，简称DWH）[98]都是多模态隐层空间马尔可夫模型的特例，此时广义边的集合E_x和E_z只包含单独的顶点。因此可见，多模态隐层空间马尔可夫网络继承了EFH 的多种优良属性[14]，例如，如上所述的模型的联合分布可以通过在每一个模态上的局部条件概率而构造性地定义。

这里简要介绍DWH模型，因为它建立了本章3.4节实验部分的基础。在[98]中，DWH被定义为一个有两个模态（two-view）的图结构，这里$\mathbf{X}$表示一个离散的文字特征向量（如图像里面的标签），$\mathbf{Z}$表示一个连续的实值特征向量（例如描述图像的归一化的颜色直方图）。假设X_i服从伯努利（Bernoulli）分布，表示这幅图像的标签在词典中的第i项是否出现；Z_j 为一个实值变量，表示一副图像的归一化的颜色直方图在第j维的值。每一实值的H_k都服从单变量的高斯分布。于是，这些变量的条件分布可以定义为

$$
\begin{aligned}
p(x_i = 1|\mathbf{h}) &= \text{Sigmoid}(\alpha_i + \mathbf{W}_{i\cdot}\mathbf{h}),\\
p(z_j|\mathbf{h}) &= \mathcal{N}(z_j|\sigma_j^2(\beta_j + \mathbf{U}_{j\cdot}\mathbf{h}), \sigma_j^2),\\
p(h_k|\mathbf{x},\mathbf{z}) &= \mathcal{N}(h_k|\mathbf{x}^\top\mathbf{W}_{\cdot k} + \mathbf{z}^\top\mathbf{U}_{\cdot k}, 1),
\end{aligned}
$$

其中$\mathbf{W}_{i\cdot}$ 和$\mathbf{W}_{\cdot k}$ 分别表示权值矩阵$\mathbf{W}$的第i行和第k列，$\mathbf{U}_{j\cdot}$和$\mathbf{U}_{\cdot k}$分别表示权值矩阵$\mathbf{U}$的第j行和第k列。

上面介绍了一种用于多模态数据分析的隐层空间统计模型的表示。下面讨论该隐层空间模型的学习问题。学习无监督的多模态隐层空间马尔可夫网络的最自然的方法是最大似然估计（Maximum Likelihood Estimation，简称MLE），MLE已被广泛用于学习有向[31,116]和无向隐变量模型[14,34,98,113]。为了处理很难计算的对数似然$\log p(\mathbf{x},\mathbf{z})$，多种近似推理的方法（如均值场方法[68]或第2章介绍的Contrastive Divergence[98]方法）已被广泛采用。关于近似算法的更多细节，本文将会在下面研究最大间隔有监督学习中详细介绍。

由于在建模和学习过程中没有考虑有监督信息，上述无监督的多模态隐层空

间马尔可夫网络不能独立实现预测任务（如分类或回归分析等），因此，为了实现预测任务，最常用的解决办法可以描述为以下两步：

（1） 学习一个无监督的隐层空间表示；

（2） 在隐层空间表示基础上建立一个预测模型（例如用于分类的SVM模型，或用于回归的SVR模型[38]）。

但是，由于没有在学习隐层空间表示的过程中考虑有监督的信息，这样的两步方法对于预测问题来说通常不是最优的，学习得到的隐层空间表示通常不具备较强的区分性。正如之前所述，大量有监督的额外信息可以在互联网上几乎免费地获取，因此研究新的可以有效利用这些有监督信息的模型表示和学习方法有望大大提高模型的预测性能以及提高隐层空间表示的判别性。下面，本章研究有监督的隐层空间马尔可夫网络，在学习隐层空间表示的同时完成预测任务。读者将会看到，与MLE方法相比，若采用合适的训练方法，有监督的隐层空间模型不仅可以学习得到预测性、区分能力强的隐层空间表示，同时可以大大提高模型的预测性能。

3.2.2　基于最大似然估计的有监督隐层空间马尔可夫网络

与无监督模型相似，MLE也是学习一个有监督的隐层空间马尔可夫网络的最常用方法。下面介绍基于MLE的有监督隐层空间马尔可夫网络。

为了使用MLE, 需要在可观测的输入数据（包括输入变量和响应变量）上定义似然模型。令Y表示响应变量，$\mathbf{V}$ 表示预测模型的参数。然后定义联合概率分布$p(\mathbf{x}, \mathbf{z}, \mathbf{h}, y)$。这里只考虑单个响应变量的预测模型，其中$Y$可以为分类模型里面的离散变量，或回归模型里面的连续变量。本章只考虑离散的分类模型，下一章将介绍连续的回归模型。根据前面介绍的结构化定义方法，对于有监督的隐层空间马尔可夫网络，需要额外指定已知$\mathbf{H}$的情况下Y 的条件概率分布，进而定义联合分布$p(\mathbf{x}, \mathbf{z}, \mathbf{h}, y)$。对于多类别分类问题，$y \in \{1, \cdots, T\}$，一种最常用的策略是使用softmax 函数① 来定义条件分布

$$p(y|\mathbf{h}) = \frac{\exp\{\mathbf{V}^\top \mathbf{f}(y, \mathbf{h})\}}{\sum_{y'} \exp\{\mathbf{V}^\top \mathbf{f}(y', \mathbf{h})\}}, \tag{3-3}$$

其中$\mathbf{f}(y, \mathbf{h})$为特征向量，其中从$(y-1)K+1$到$yK$的元素为$\mathbf{h}$，而其他元素为0。$\mathbf{V}$是由$T$个子向量 $\mathbf{V}_y$拼接的向量，其中每一个$\mathbf{V}_y$与类别标签y相对应。而联合分布

① 为了表示方便，在分类模型和下一章的回归模型中都忽略偏移量参数（Offset parameters）。偏移量参数可以通过在$\mathbf{h}$中加入一个维度取值为1的特征，很容易地包含进来。

$p(\mathbf{x}, \mathbf{z}, \mathbf{h}, y)$与公式 (3-2)有相同的形式，唯一的区别在于指数函数里有一个额外项$\mathbf{V}^\top \mathbf{f}(y, \mathbf{h}) = \mathbf{V}_y^\top \mathbf{h}$。

DWH模型是无监督隐层空间马尔可夫网络的一个特例，相应的，有监督层次Harmonium （在本文中记为Tri-wing Harmonium，简称TWH）[35]模型是上述有监督隐层空间马尔可夫网络分类模型的一个特例。在联合似然函数基础上，采用均值场或Contrastive Divergence等近似推理方法，可以使用标准的MLE来学习模型参数①。由于推导过程与学习TWH[34]模型很相似，在此不再赘述，主要的区别在后验推理。在本章介绍最大间隔学习之后，读者会有更加深入的了解。

3.3　最大间隔多模态隐层空间马尔可夫网络

如之前所述，基于MLE的有监督隐层空间马尔可夫网络需要定义如公式(3-3)所示的归一化概率分布，其中的归一化因子通常很难计算，使模型推理变得更加复杂，特别是在有向图模型中[31,117]，计算的困难将会变得更加突出。此外，如文献[34] 和本文实验结果所显示，基于MLE的有监督隐层空间马尔可夫网络并不比第3.2.1节介绍的最简单的两步方法（即首先推理得到隐层空间表示；然后基于隐层空间表示，学习一个预测模型）得到更好的预测结果。因此，提出一种更合适的学习有监督隐层空间马尔可夫网络的方法，是一个非常有意义的研究问题。为此，本章提出最大间隔有监督隐层空间马尔可夫网路。在模型表示和学习方法上都与基于MLE的有监督隐层空间马尔可夫网络不同，但都是希望通过考虑有监督信息（如分类问题中离散的类别标签或回归问题中连续的评价分数等）从多模态数据中学习具有预测性和区分性的隐层空间表示。

3.3.1　分类模型

为了避免赘述，这里考虑更加广义的多类别分类问题，二分类问题可以用相似方法推导求解。

3.3.1.1　模型表示/定义

与公式(3-3)中的对数线性模型相似，当隐变量$\mathbf{H}$已知时，定义线性的判别函数为$F(y, \mathbf{h}; \mathbf{V})$

$$F(y, \mathbf{h}; \mathbf{V}) \stackrel{\text{def}}{=} \mathbf{V}^\top \mathbf{f}(y, \mathbf{h}), \tag{3-4}$$

① 一种最大化条件似然$p(y|\mathbf{x}, \mathbf{z})$的判别式方法[114]，但此方法不如混合产生式/判别式的方法。

其中$\mathbf{f}$和$\mathbf{V}$与公式 (3-3)中定义相同。但是这里将采用和上述基于定义Y的条件概率完全不同的策略来考虑有监督的信息。现在的问题是：如何在确定性的最大间隔准则里面考虑隐变量$\mathbf{H}$? 为了去掉$\mathbf{H}$的不确定性，这里对隐变量$\mathbf{H}$求期望（即一阶矩），并且定义有效判别函数为

$$F(y, \mathbf{h}; \mathbf{V}) \stackrel{\text{def}}{=} \mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[\mathbf{f}(y, \mathbf{h}; \mathbf{V})]$$

相应地，用于多分类任务的预测准则为

$$\begin{aligned} y^* &\stackrel{\text{def}}{=} \operatorname*{argmax}_y \mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[F(y, \mathbf{h}; \mathbf{V})] \\ &= \operatorname*{argmax}_y \mathbf{V}^\top \mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[\mathbf{f}(y, \mathbf{h})]. \end{aligned} \tag{3-5}$$

当$\mathbf{x}$和$\mathbf{z}$都完全可观测时，上述期望可以利用$p(\mathbf{h}|\mathbf{x}, \mathbf{z})$的因子化表示高效地计算。如果$\mathbf{x}$或 $\mathbf{z}$中存在缺失信息时，可以通过推理得到缺失元素的期望，正如下面要介绍的公式 (3-10)中所示。

有了上述定义，很显然，学习一个预测模型，就是找到满足最小化某种损失函数的一个最优的参数$\mathbf{V}^*$。正如最大间隔支持向量机模型使用的，本文中采用最小化铰链损失函数（hinge loss）。当训练数据$\mathcal{D} = \{(\mathbf{x}_d, \mathbf{z}_d, y_d)\}_{d=1}^D$已知，期望预测准则 (3-5)的铰链损失函数可以定义为

$$\mathcal{R}_{hinge}(\mathbf{V}) \stackrel{\text{def}}{=} \sum_d \max_y [\Delta\ell_d(y) - \mathbf{V}^\top \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\Delta\mathbf{f}_d(y)]],$$

其中$\Delta\ell_d(y)$为代价函数（如 0/1 loss），用来度量当一个候选的预测值y与真实的类别标签y_d不同时，需要付出的代价；$\mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\Delta\mathbf{f}_d(y)] = \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\mathbf{f}(y_d, \mathbf{h})] - \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\mathbf{f}(y, \mathbf{h})]$。可以证明上述铰链损失函数是经验错误率

$$\mathcal{R}_{emp}(\mathbf{V}) \stackrel{\text{def}}{=} \sum_d \Delta\ell_d(y_d^*)$$

的上界[①]，即

$$\forall \mathbf{V},\ \mathcal{R}_{emp}(\mathbf{V}) \le \mathcal{R}_{hinge}(\mathbf{V}).$$

① 根据定义，

$$\begin{aligned} \max_y [\Delta\ell_d(y) - \mathbf{V}^\top \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\Delta\mathbf{f}_d(y)]] &\ge \Delta\ell_d(y^*) - \mathbf{V}^\top \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\Delta\mathbf{f}_d(y^*)] \\ &= \Delta\ell_d(y^*) + (\mathbf{V}^\top \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\mathbf{f}(y^*, \mathbf{h})] - \mathbf{V}^\top \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\mathbf{f}(y_d, \mathbf{h})]) \\ &\ge \Delta\ell_d(y^*), \end{aligned}$$

其中，最后一个不等式的成立是因为y^*的定义。

根据正则化的经验风险最小化（Empirical Risk Minimization）准则[39,118]，同时考虑输入数据的拟合程度，可以定义一个联合学习预测模型$\mathbf{V}$和似然模型Θ的目标函数

$$\mathrm{P1}:\ \min_{\Theta,\mathbf{V}}\ L(\Theta)+\frac{1}{2}C_1\|\mathbf{V}\|_2^2+C_2\mathcal{R}_{hinge}(\mathbf{V}), \tag{3-6}$$

其中，$L(\Theta)\overset{\text{def}}{=}-\sum_d \log p(\mathbf{x}_d,\mathbf{z}_d)$为负对数似然函数，$C_1$和$C_2$为非负正则化常量，可以使用交叉验证（Cross Validation）来确定它们的值。值得注意的是$\mathcal{R}_{hinge}$是关于Θ的函数。该问题P1可以等价的写成受约束的优化问题

$$\begin{aligned}&\mathrm{P1}':\ \min_{\Theta,\mathbf{V},\xi}\ L(\Theta)+\frac{1}{2}C_1\|\mathbf{V}\|_2^2+C_2\sum_d\xi_d\\ &\forall d,\ \forall y,\ \text{s.t.}:\ \begin{cases}\mathbf{V}^\top\mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\Delta\mathbf{f}_d(y)]\geq\Delta\ell_d(y)-\xi_d\\ \xi_d\geq 0\end{cases}\end{aligned} \tag{3-7}$$

这是因为，由问题P1′，我们可以得到松弛变量的最优解为

$$\forall d:\ \xi_d^*=\max_y[\Delta\ell_d(y)-\mathbf{V}^\top\mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\Delta\mathbf{f}_d(y)]]. \tag{3-8}$$

同时，从上述最优解中我们可以看出ξ_d^*是恒大约等于0的，因此，问题P1′中的对松弛变量的非负性约束也可以省略掉。

由上面的定义可以看出，问题（3-6）联合最小化负对数似然和最小化训练损失函数，因此，通过求解问题（3-6），可以期望同时学习预测性的隐层空间表示（即$\mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[\mathbf{h}]$）以及一个预测模型（用参数$\mathbf{V}$来表示），一方面在训练数据集上保证尽可能高的预测准确率，另一方面更好地解释输入数据。在下一节介绍算法的过程中，本章将讨论更多更深层次的理解。

3.3.1.2　基于Contrastive Divergence的优化方法

和一般的EFH以及DWH模型相似，由于归一化因子的存在，模型的数据似然$L(\Theta)$很难计算。因此，本章采用第2章介绍过的高效的变分推理方法（即Contrastive Divergence变分推理技术）[10,14,28,98]来近似地表示联合概率似然，它是学习受限波尔兹曼机等无向图隐变量模型的有效方法。具体地说，使用变分目标函数$\mathcal{L}^v(q_0,q_1)$来近似负对数似然函数$L(\Theta)$

$$\mathcal{L}^v(q_0,q_1)\overset{\text{def}}{=}\mathrm{KL}(q_0(\mathbf{x},\mathbf{z},\mathbf{h})\|p(\mathbf{x},\mathbf{z},\mathbf{h}))-\mathrm{KL}(q_1(\mathbf{x},\mathbf{z},\mathbf{h}),\|p(\mathbf{x},\mathbf{z},\mathbf{h})),$$

Algorithm 1 最大间隔隐空间马尔可夫网络分类模型的学习算法

输入: 数据$\mathcal{D} = \{(\mathbf{x}_d, \mathbf{z}_d, y_d)\}_{d=1}^{D}$，常量$C_1$和$C_2$
输出: 模型参数Θ和$\mathbf{V}$
初始化$\mathbf{V} = 0$，Θ设为对输入矩阵$M = [\mathbf{x}_1, \cdots, \mathbf{x}_D; \mathbf{z}_1, \cdots, \mathbf{z}_D]$做SVD分解得到的前$K$个的左特征向量。
repeat
  for $d = 1$ **to** D **do**
    使用公式（3-10）计算隐空间表示$q_0(\mathbf{h}_d)$;
    初始化$q_1(\mathbf{h}_d) = q_0(\mathbf{h}_d)$
    for $t = 1$ **to** T（实验中设为5） **do**
      使用公式（3-10）计算$q_1(\mathbf{x}_d)$和$q_1(\mathbf{z}_d)$
      使用公式（3-10）计算$q_1(\mathbf{h}_d)$
    end for
  end for
  通过求解多个类别的SVM分类器（3-11）计算参数$\mathbf{V}$;
  通过次梯度下降法求解参数Θ)
until 问题P1的目标函数相对变化小于τ (如$1e^{-4}$) 或者迭代次数超过某个阈值（如20）。

其中KL($q\|p$)为变分分布q[①]（q_0或 q_1）与模型分布p的L散度。 q_0为当$\mathbf{x}$和$\mathbf{z}$取可观测数据值时的变分分布，而q_1中所有变量都是未知的。对于q，应用结构化均值场（Structured Mean Field）假设[119] [②]

$$q(\mathbf{x}, \mathbf{z}, \mathbf{h}) = q(\mathbf{x})q(\mathbf{z})q(\mathbf{h}) \tag{3-9}$$

将变分近似似然值$\mathcal{L}^v(q_0, q_1)$ 代入问题 (3-6)，得到近似的目标函数$\mathcal{L}(\Theta, \mathbf{V}, q_0, q_1)$。然后，如算法1所示，交替地对目标函数$\mathcal{L}(\Theta, \mathbf{V}, q_0, q_1)$做分别关于（$q_0$，$q_1$）和（$\Theta$，$V$）的最小化运算，其中计算$q_0$和$q_1$的问题是后验推理。具体地说，当$(\Theta, \mathbf{V})$固定时，根据均值场的一般理论[24]，对于一个变分分布q（q_0或者q_1），可以得到更新每一个边缘分布为

$$q(\mathbf{x}) = p(\mathbf{x}|\mathbb{E}_{q(\mathbf{h})}[\mathbf{h}])$$
$$q(\mathbf{z}) = p(\mathbf{z}|\mathbb{E}_{q(\mathbf{h})}[\mathbf{h}])$$

① 这里都用q统一表示。
② 之前的相关工作[34,98]对q做了更多假设，这里可以发现，那些假设是不需要的。

$$q(\mathbf{h}) = \prod_k p(h_k | \mathbb{E}_{q(\mathbf{x})}[\mathbf{x}], \mathbb{E}_{q(\mathbf{z})}[\mathbf{z}]) \tag{3-10}$$

对于q_0, $(\mathbf{x}, \mathbf{z})$ 固定在它们的输入观测值，所以只需推理$q_0(\mathbf{h})$。同时，由于$q_0(\mathbf{h})$的因子化表示，这一步可以很高效地进行推理。而分布q_1 需要从q_0开始进行更新，通常需要几步迭代（在本文的实验中选用5步），以得到一个较好的q_1。更详细的迭代步骤读者可以参考文献[10]。同样，读者可以发现$q(\mathbf{x})$和 $q(\mathbf{z})$都是CRF模型，其中期望$\mathbf{H}$ 可以看作全局条件。因此，对于线性链式结构（Linear-Chain）的模型，可以使用消息传递（Message Passing）的方法[115]推理边缘分布，这些边缘分布是参数估计以及在输入多模态数据的预测任务（如图像标注）中需要的，本文会在接下来的内容以及第4章中介绍。对于更加广义的结构化模型，可以应用近似推理技术[120]进行模型的参数学习。

在推理得到q_0和q_1之后，可以使用坐标下降（Coordinate Descent）方法来进行参数学习，具体步骤分为

（1） 固定Θ，估计$\mathbf{V}$，这个问题可通过学习一个多类别SVM[88]求解；

$$\begin{aligned} &\min_{\mathbf{V},\xi} \ \frac{1}{2}\|\mathbf{V}\|_2^2 + \frac{C_2}{C_1}\sum_d \xi_d \\ &\forall d, \ \forall y, \ \text{s.t.}: \ \mathbf{V}^\top \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\Delta \mathbf{f}_d(y)] \geq \Delta \ell_d(y) - \xi_d, \end{aligned} \tag{3-11}$$

注意，相对于问题P1′，这里我们忽略了松弛变量的非负性约束，由前面的讨论可以知道，这种忽略将对结果没有影响。

（2） 固定$\mathbf{V}$，估计Θ，此步可用次梯度下降（Sub-Gradient Descent）方法求解。

为了简化符号表示，定义

$$\Delta\mathbb{E}[\cdot] \stackrel{\text{def}}{=} \mathbb{E}_{q_1}[\cdot] - \mathbb{E}_{q_0}[\cdot].$$

我们可以得到参数θ，η以及λ的梯度计算公式分别为

$$\begin{aligned} \forall e \in E_x, \partial\theta_e &= \Delta\mathbb{E}[\phi(\mathbf{x}_e)] \\ \forall e \in E_z, \partial\eta_e &= \Delta\mathbb{E}[\psi(\mathbf{z}_e)] \\ \forall k, \partial\lambda_k &= \Delta\mathbb{E}[\varphi(h_k)]; \end{aligned}$$

相应的，矩阵$\mathbf{W}$和$\mathbf{U}$的梯度计算公式分别为

$$\partial\mathbf{W}_e^k = \Delta\mathbb{E}[\phi(\mathbf{x}_e)\varphi(h_k)^\top] - C_2 \sum_d (\mathbf{V}_{y_d k} - \mathbf{V}_{\bar{y}_d k}) \frac{\partial \mathbb{E}_{q_0}[h_k]}{\partial \mathbf{W}_e^k}$$

$$\partial \mathbf{U}_e^k = \Delta\mathbb{E}[\psi(\mathbf{z}_e)\varphi(h_k)^\top] - C_2 \sum_d (\mathbf{V}_{y_d k} - \mathbf{V}_{\bar{y}_d k}) \frac{\partial \mathbb{E}_{q_0}[h_k]}{\partial \mathbf{U}_e^k},$$

其中

$$\bar{y}_d = \arg\max_y [\Delta\ell_d(y) + \mathbf{V}^\top \mathbb{E}_{q_0}[\mathbf{f}(y, \mathbf{h})]$$

是在考虑代价损失的情况下的预测类别。而期望$\mathbb{E}_{q_0}[\phi(\mathbf{x}_e)]$实际上是$\phi(\mathbf{x}_e)$在训练数据集D上出现的频率。同理，$\mathbb{E}_{q_0}[\psi(\mathbf{z}_e)]$也为经验期望。在这些次梯度的基础上，可以很容易地应用LBFGS拟牛顿方法[121]来迭代地更新模型的参数，直到收敛或者迭代次数大于某个阈值。

值得一提的是，在本章提出的最大间隔的框架下，与标准的基于MLE的DWH模型[98]相比，**W**和**U** 的次梯度中包含额外项（即公式中的最后一项）。这一项对学习隐层空间表示起到了正则化的作用，即如果在考虑代价损失函数情况下的预测值 $\bar{y}_d$与真实的类别标签y_d不同，这一项将为非零，从而它会使得模型朝着更有利于发现预测性的隐层空间表示（即更有利于分类）的方向改进。这个“偏差项”将使基于最大间隔的多模态隐层空间模型更趋于发现具有强预测性/区分性的隐层空间表示。

3.3.2　最大间隔Harmonium模型

上面提出了在通用的多模态隐层空间马尔可夫网络上的最大间隔学习框架，该模型不仅可以解决分类问题，还可以解决回归问题（详见第4章）。为了更加深入地研究基于最大间隔的基本学习准则的有效性，并与之前其他工作相比较，下面介绍一个关于有监督隐层空间马尔可夫网络的特例：假设模型中每一种模态的输入变量间不存在依赖结构，即各输入变量满足条件独立性，这时模型退化为Harmonium模型。这种模型在文献[14,34,98]中得到广泛的研究。本章我们提出最大间隔的Harmonium模型（即Max Margin Harmonium，简称MMH）。值得一提的是，这个特例并没有制约本文展示提出框架的通用性，因为前述的问题定义和优化方法适用于所有结构化或非结构化的输入数据。特别的， MMH是基于之前3.2.1章节所述的DWH模型，DWH模型是一种建立在输入数据$(\mathbf{x}, \mathbf{z})$之上的概率似然模型，其中**x**为一种离散的文本特征向量（如图像标签），而**z** 为实数值的特征向量（如归一化的颜色直方图特征）。可使用前述方法进行参数估计。为了推理q_0和q_1, **x**, **z**和 **h**的分布都完全地满足因子化。因此，可以很容易地计算得到分类模型中的次梯度。具体的内容请参考附录A。

3.3.3　时间复杂度

下面讨论有监督的隐层空间马尔可夫网络在多种应用（包括分类、图像检索和图像标注）中理论上的时间复杂度。

隐层空间马尔可夫网络用于分类和检索任务的共同之处在于这些应用都只需要推理隐层空间表示（即$\mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[\mathbf{h}]$），而输入数据都是完全可观测的。对于最大间隔隐层空间马尔可夫网络，由于它的似然函数只定义在输入数据$(\mathbf{x},\mathbf{z})$上，所以不需要迭代，一步就可以推理得到这些隐表示。更加精确地说，隐层表示是$\mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[\mathbf{h}] = \Upsilon$，其中$\Upsilon_k = \sum_{e\in E_x} \phi(\mathbf{x}_e)^\top \mathbf{W}_e^k + \sum_{e\in E_z} \psi(\mathbf{z}_e)^\top \mathbf{U}_e^k$, $\forall 1 \le k \le K$为输入数据的线性组合，因此可以非常高效地计算，其复杂度关于输入特征的维度是线性的。相反，对于基于MLE的有监督隐层空间马尔可夫网络，它定义了一个关于输入变量$(\mathbf{x},\mathbf{z})$和响应变量y的完全的似然函数$p(\mathbf{x},\mathbf{z},y)$。在测试过程中，因为$Y$是未知的，因此，这时推理需要一个迭代过程，即（1）给定$\mathbf{H}$，推理得到Y的后验分布$p(Y|\mathbf{h})$；（2）对每一个可能的Y取值，推理得到隐层空间表示$p(\mathbf{h}|\mathbf{x},\mathbf{z},y)$。因此，基于MLE的有监督隐层空间马尔可夫网络的测试时间是用本文提出的最大间隔方法的一个恒定倍。但概括来说，这些无向图隐层空间模型的推理效率与有向图隐层特征模型（如MedLDA[40]等）相比更加高效。本文会在第4章第4.5.4节详细介绍。

对于图像标注问题，令$\mathbf{x}$表示标签，在训练过程中$\mathbf{x}$为可观测的，但在测试过程中是不可观测的，此时可以通过公式（3-10）计算后验分布$p(\mathbf{x}|\mathbf{z})$，其中$\mathbf{z}$为可观测值。推理得到的概率值高的标签就被选作标注结果。可以看出，这个推理过程与分类问题中的基于MLE的有监督隐层空间马尔可夫网络相似，都需要一个迭代的过程。因此，无监督、有监督的多模态隐层空间马尔可夫网络（无论是基于MLE还是基于最大间隔方法）在这个任务上的时间复杂度基本是相同的。

3.4　实验结果与分析

下面，本章通过在两个真实数据集上的实验来定量与定性地评估本章提出的最大间隔有监督多模态隐层空间马尔可夫网络模型的效果，包括隐层空间表示、预测性能、以及参数敏感度分析等。为了与以往的大量用于分类和图像标注与检索的隐层空间模型（参见3.4.3章节）进行比较，本章深入研究特例 MMH模型，验证学习方法的通用性和有效性，对于回归分析以及有结构的输入数据，将在下一章详细介绍。

3.4.1　数据集与特征

两个数据集分别为：TRECVID 2003视频数据集[98]以及13 class-animal Flickr图像数据集。这两个数据集从特征类型和维度来说，都非常的丰富和多样化。所有这些数据集都已公布于网页： http://www.cs.cmu.edu/ junzhu/data.htm，具体描述为

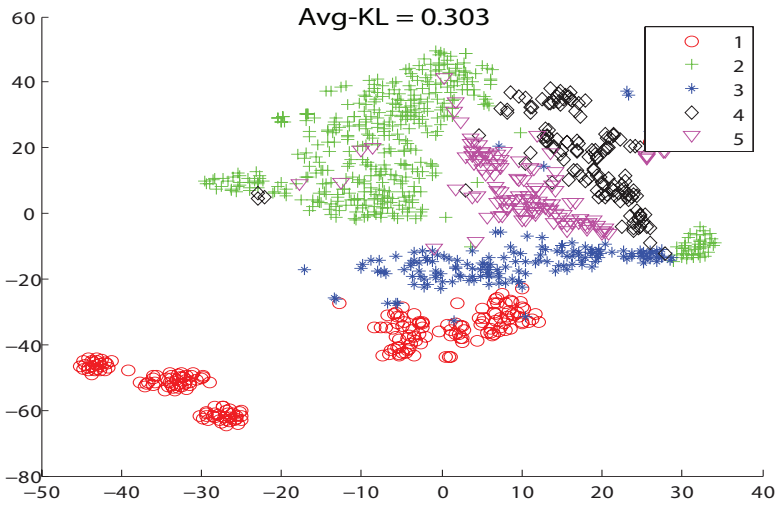
（1）TRECVID 2003数据集包含1078个分别属于5个类别且带有类别标签的视频片段。每一视频片段由1894维文本特征向量和关键帧图像的165维归一化的HSV 颜色直方图特征向量组成。将这个数据集平均分成两份：即539个样本作为训练数据，另外539个样本作为测试数据。

（2）13-class animal Flickr数据集是NUS-WIDE数据集[122]（由Flickr网页图像组成的一个大规模数据集）[122]的一个子集。这个数据集包含3411幅图像，属于13个类别，分别为 *squirrel*, *cow*, *cat*, *zebra*, *tiger*, *lion*, *elephant*, *whales*, *rabbit*, *snake*, *antlers*, *hawk* and *wolf*。图3.9所示为每一种类别的一部分示例图片。对于每一幅图片，都由634维实数值特征（即64维颜色直方图，144维颜色相关图，73维边方向的直方图， 128维小波纹理特征以及225维颜色矩等），以及500维SIFT[123]特征组成[122]。实验中随机选出2054幅图片作为训练数据，剩余的1357幅图片作为测试数据。同时下载1000维标签特征用于图像标注。

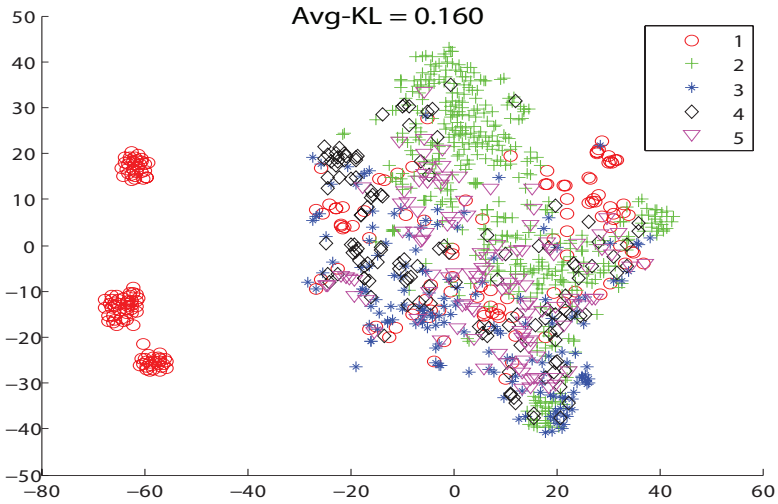
需要说明的是，在本章所有实验中，离散的词袋（Bag-of-Words，简称BOW）特征（如文本特征或SIFT特征），都被转化为0/1二值特征（即非零特征转化为“1”，其他特征转化为“0”），并假设它们服从伯努利分布。

3.4.2　判别性隐层空间表示

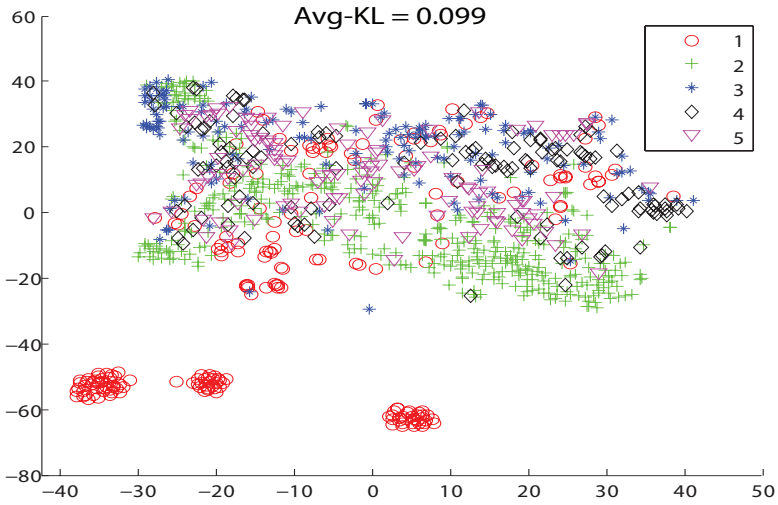
本小节首先研究隐层空间表示的整体效果。图3.2所示为MMH，DWH和TWH模型在TRECVID视频数据集上得到的10维隐层空间表示的2维可视化表示。这里使用t-SNE随机近邻投影算法[124]得到数据在2维平面空间中的表示。这个结果非常明确地说明使用最大间隔的MMH方法得到的隐层空间表示对于相同类别图像有很强的聚集模式，而对于不同类别的图像有很强区分性的特点（即不同种类的图像在2维投影平面上可以非常容易被区分）。相反，由基于MLE的DWH 和TWH模型得到的隐层空间表示除了第一种类别外，对其他四种类别的样本并没有一种清晰的聚集模式，而不同类别的图像都混淆在一起。这些观测结果说明基于最大间隔的MMH模型可以发现更加具有区分性/判别性的隐层空间表示，得到更好的预测性能。在Flickr数据集上可以观察到相似的结果。



(a) MMH隐空间



(b) DWH隐空间



(c) TWH隐空间

图 3.2　三种模型在TRECVID视频数据集上得到的隐层空间表示的 t-SNE 2维投影。

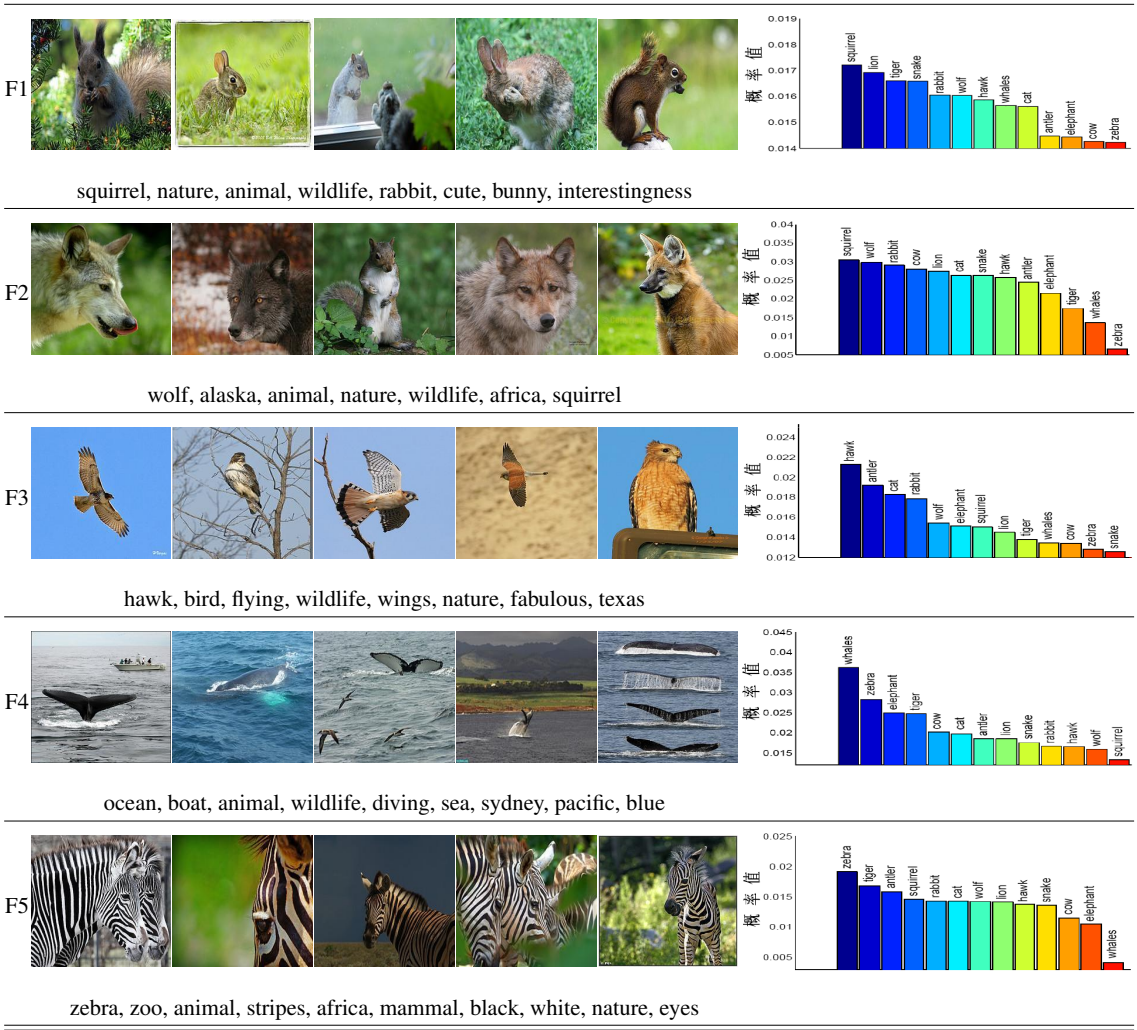


图 3.3　MMH模型在Flickr animal数据集上发现得到的60维隐变量的代表性图片。对于每一个隐变量（topic），这里展示该隐特征期望值最高的5幅图像，以及该隐特征对于表示13个类别的所有图像的平均概率分布值。

下面，更加仔细地研究每一维隐层空间特征上的特性。这里将Flickr数据集作为一个例子，图3.3展示了隐层空间中的5维特征（每一特征都与隐层空间中的一维相对应）。对于每一维特征T_k，将所有样本所得到的隐层空间的第k维特征H_k的期望按照从高至低进行排序，取前5幅图像以及其相应的标签显示在图3.3中。在图3.4中也显示了每一个特征对应所有排序中最后5幅图像及相关的标签。同时，为了展示每一特征对于13类样本的区分能力，还画出第k维隐特征在13类样本上的平均分布（如图3.3右侧或图3.4的右侧所示）①。

① 为了计算平均分布需要对其进行如下变换：首先每一隐变量期望值都减去最小值，然后将所有K维隐变量值进行归一化，按此方法将$\mathbf{H}$的期望值转化为非负的概率分布。每一类的平均值是指对该类别中所有样本的隐特征期望值进行平均。此外，图中还展示每一个隐特征在13类样本上的平均分布。需要说明的是，实验中本章所做的这种归一化变换并没有影响结果的区分性能表示，这也是对所有类别的平均隐特

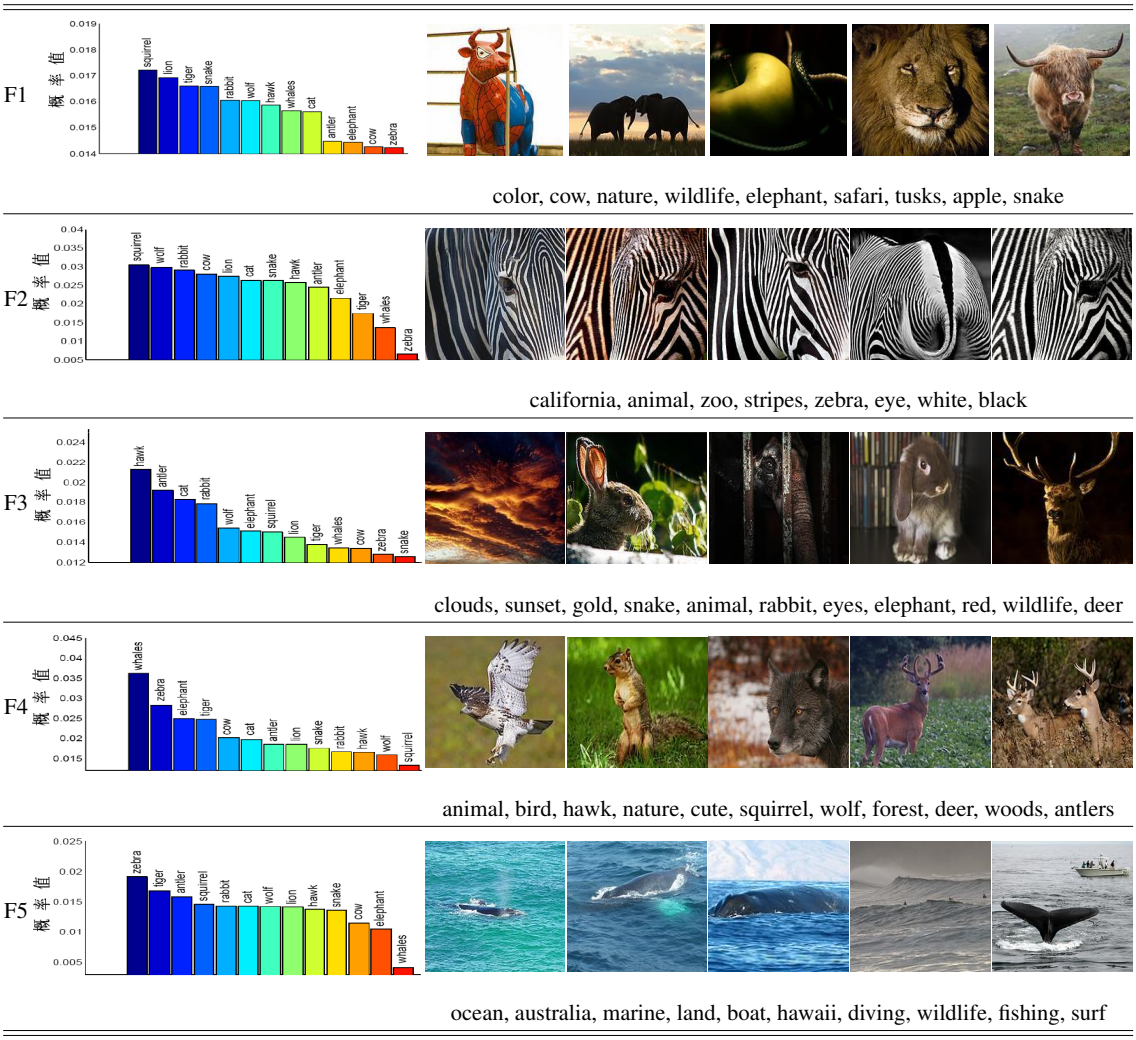


图 3.4　MMH模型在Flickr animal数据集上学习得到的60维隐变量的代表性图片。对于每一个隐变量，展示该隐特征期望值最低的5幅图像，以及该隐特征对于表示13个类别的所有图像的平均概率分布值。

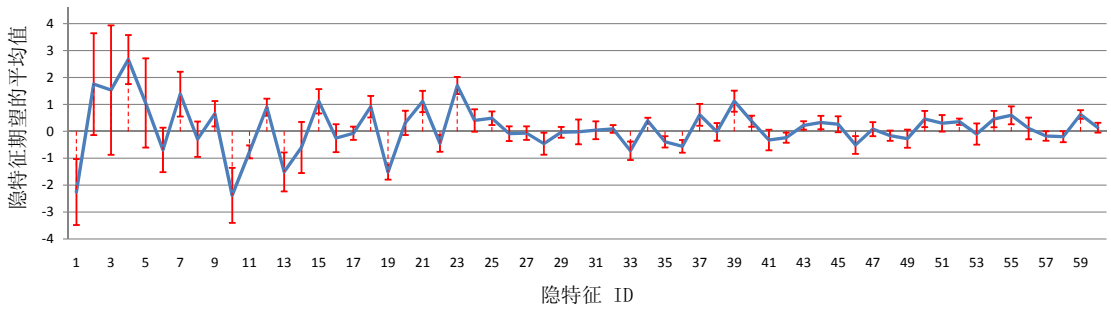


图 3.5　MMH在Flickr图像数据集上得到的60维隐特征的期望在所有13类样本上平均及方差。

可以观察到，MMH学习得到的隐层空间表示对于多种类别都具有较强的判别性。例如，F3和F4分别表示类别*hawk*和 *whales*。相似的， F1和F5分别表示类别*squirrel*和*zebra*。还有一些隐特征适于区分几种类别的组合。例如，F2适于区分类别组合{*squirrel*, *wolf*, *rabbit*}，和类别组合{*tiger*, *whales*, *zebra*}；而不能很好地区分类别*squirrel*和*wolf*。

为了进一步定量地研究判别式的隐层空间表示，本章通过计算所有类别间的$\mathbb{E}[\mathbf{H}]$分布的KL散度值① 衡量隐层空间的区分性程度。如图3.2所示，基于最大间隔的MMH与基于MLE的DWH和TWH方法相比，平均KL散度值更高。这又一次证明MMH模型得到的隐层空间表示更加具有区分性。在Flickr数据集上可得到相似的结果（如图3.3所示），对于隐特征数为60的MMH，DWH, TWH得到的平均KL散度值分别是 1.62，1.28，0.232。这与作者预期的结果相吻合，即MMH得到的隐层空间表示（图3.3中所示）更加具有区分性。为了与图3.3进行比较，图3.4右侧所示为与图3.3中相同的5维隐特征对应的$\mathbb{E}[H_k]$值最低的5幅示例图像。同样，图3.4左侧所示为每一维隐特征在所有13类样本上期望$\mathbb{E}[H_k]$值的平均分布。为了提供一个整体分布结果，图3.5所示为MMH得到的60维隐特征在所有13类样本上隐特征期望的平均值及方差。每一维特征对应的方差一定程度上表示特征k对于13种类别图像的区分能力（即方差越大，区分性越强）。

3.4.3　预测性能

下面将本章提出的模型用于图像分类、检索和标注等实际问题中，从而展示大量实验结果证明模型的有效性。

3.4.3.1　分类

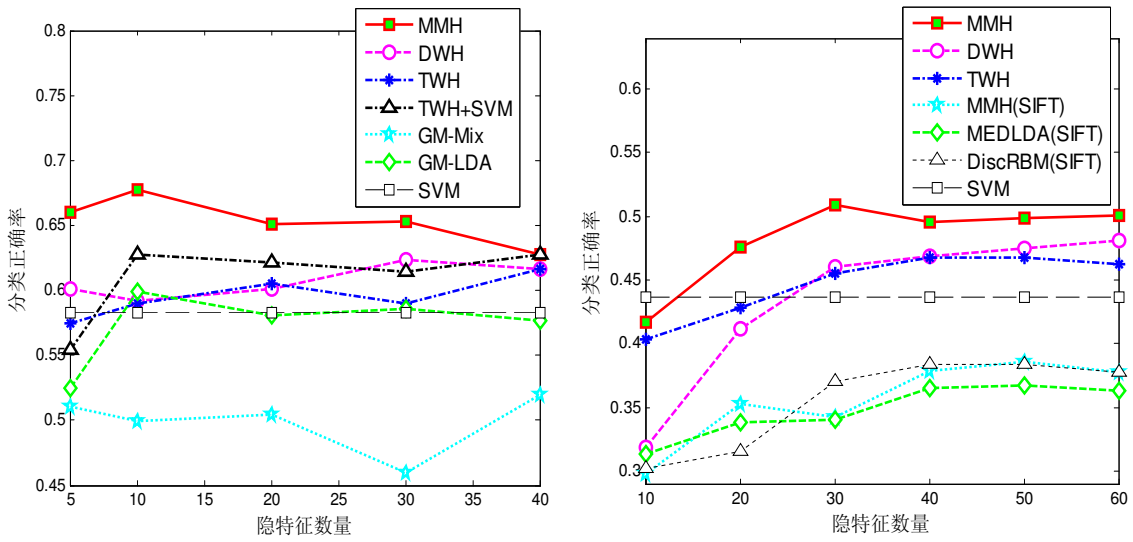
首先比较MMH模型与其他主流方法，包括多类别线性支持向量机SVM②，无监督DWH，基于MLE的有监督TWH，高斯混合模型（Gaussian Mixture，简称GM-Mix），Gaussian Mixture LDA（GM-LDA）和 Correspondence LDA（CorrLDA）在TRECVID 2003数据集上的分类结果。对于后三种模型的结果，读者请参考文献[18]。对于MMH模型，实验中使用SVM^{multiclass} ③ 求解MMH预测模型的参

征期望值的区别性度量。使用$\mathbb{E}[\mathbf{H}]$原始值可以得到相似的可视化结果。

① 首先将$\mathbf{H}$的期望值变换为K维的分布。每一类的平均值通过平均化相同类别图像的分布得到。对于每一对分布p和q，平均KL散度是$1/2(R(p,q)+R(q,p))$。

② 这里我们不考虑使用非线性核函数（如比较常用的RBF径向核函数等），主要原因是线性模型（尤其是MMH）具有更好的解释性。

③ http://svmlight.joachims.org/svm_multiclass.html



(a) TRECVID数据分类正确率　　(b) Flickr数据分类正确率

图 3.6　在两个图像数据集上的分类正确率结果比较。

数**V**。对于SVM分类器，这里将多模态的输入特征拼接在一起作为整体特征向量建立线性SVM分类器，同样地，使用SVM^{multiclass}进行训练和预测。对于无监督模型（即DWH，GM-Mix，GM-LDA，CorrLDA），在得到的隐层空间表示上建立一个线性SVM分类器。

图3.6(a)所示为不同模型在TRECVID数据集上的分类正确率，由于CorrLDA的结果比其他方法正确率低很多，在此忽略之。可以观察得到，基于最大间隔的MMH模型比其他方法得到更好的分类效果。相比之下，基于MLE的有监督TWH相对于无监督DWH并没有表现出优越性能。如果在TWH得到的隐层空间表示基础上再建立一个SVM分类器（记为TWH+SVM）① 这时模型的分类性能将会有所提高，但仍然比不上MMH的分类结果。这样的结果表明当模型在经过有效训练之后（如采用最大间隔学习方法），有监督的信息有助于发现判别式的隐层空间表示，且更适合提高模型预测性能。与传统SVM分类器相比，MMH的优异结果说明对多模态输入数据的建模有助于提高分类性能。其他模型（如CorrLDA，GM-Mix模型）的预测结果不理想的原因在文献[34,98]中有详细说明，在此不再赘述。

图3.6(b)所示为模型在Flickr 数据集上的分类正确率。为了简化描述，这里只将MMH与性能最佳的DWH，TWH，SVM相比较。实验中使用500维SIFT特

① 与MMH模型相比，这种最简单的结合两次使用监督信息，并不是一种严格意义上合理的做法。作者同样进行了 MMH+SVM的实验，这种方法理论上不能得到优于MMH的结果，因为两种方法的分类器建立在相同的隐特征基础上。读者可以参考[40]中相似的研究结果。

征和 634维实值特征作为多模态MMH，DWH，TWH模型的多种特征输入。同时，将MMH与单模态的MedLDA[40]模型（一种有向贝叶斯网络）和单模态的判别式受限波尔兹曼机（DiscRBM）模型[114]相比较。为了公平起见，将MMH和MedLDA、DiscRBM都仅使用一种SIFT特征进行比较，此时MMH模型记为MMH（SIFT）。同样地，读者可以发现基于最大间隔的多模态MMH与其他任何方法（包括忽略多模态特征的SVM）相比都得到更出色的预测性能。对于单模态MMH（SIFT），分类结果与最大间隔MedLDA和DiscRBM的结果不相上下。事实上，MMH可以被理解为MedLDA在无向图上的一种推广。对于DiscRBM，由于它是一种基于MLE的判别式模型（即已知输入特征时，最大化Y的条件似然），并不需要对输入特征建模，因此不能实现输入变量级别的分析和预测（如预测图像标签）。文献[114]中提出一种优化混合产生式/判别式似然函数的方法学习RBM模型，并得到比DiscRBM更好的结果。然而，MMH与这种方法有以下三方面不同：

（1） MMH的产生式似然函数$L(\Theta)$中并不包含Y；

（2） MMH的判别式部分不是条件对数似然，而是铰链损失函数；

（3） MMH可以使用一种明确的正则化规则，而非文献[114]中使用的不明确的正则化规则（如“提前停止迭代”等启发式方法）。

从上述分类结果中，我们也可以看到参数化隐空间模型的维度（即隐特征的个数K）对模型的性能影响很大，只有K设置较合适时，才能获得较满意的性能。如何自动有效地确定K是本文第5章要研究的主要问题。

3.4.3.2　图像检索

下面将MMH模型用于TRECVID和Flickr数据集的图像检索。具体方法为：将每一幅测试图片作为一个查询样本（Query），所有训练样本作为待检索图像。将所有训练样本按照其与查询样本在隐特征空间的Cosine相似度① 由高至低进行排序。实验中约定：若一幅图像与查询样本属于同一类别，那么它们相关（relevant）。按此方法将MMH模型用于图像检索。图3.7展示了不同模型的平均准确率（Average Precision Score，简称AP值）[34,98]。这里AP值按照下面的公式计算

$$AP = \frac{\sum_{T=1}^{N} Prec(T)Rec(T)}{average\ number\ of\ relevant\ documents} \tag{3-12}$$

① 定义向量$\mathbf{x}_1$和$\mathbf{x}_2$的cosine相似度为$\frac{\mathbf{x}_1^{\top}\mathbf{x}_2}{\|\mathbf{x}_1\|_2\|\mathbf{x}_2\|_2}$.

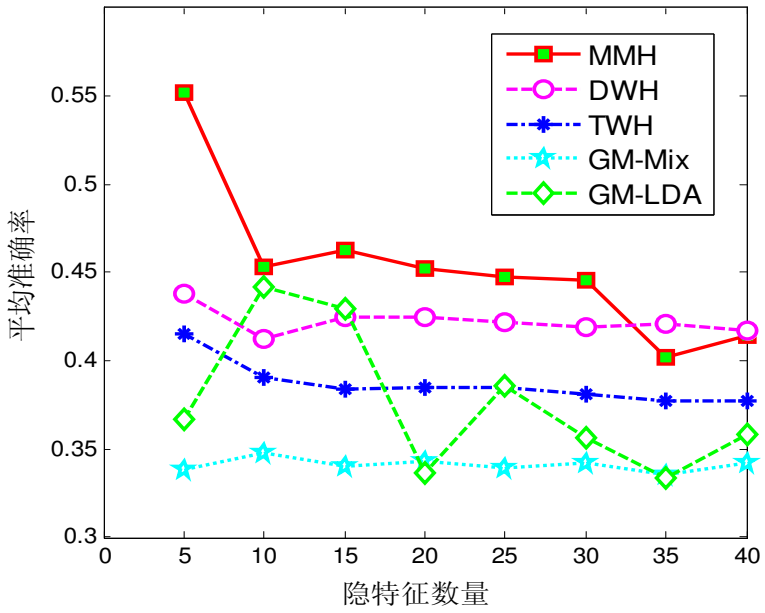
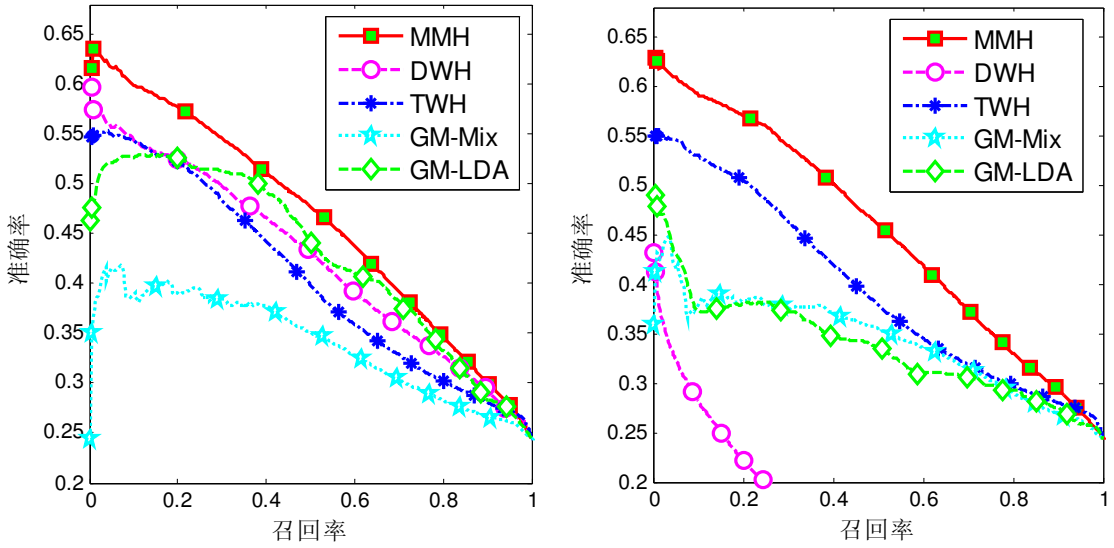


图 3.7　TRECVID 2003数据图像检索的平均准确率曲线。



(a) 15维隐特征　　(b) 20维隐特征

图 3.8　TRECVID 2003数据图像检索问题的准确率-召回率曲线。

其中$Prec(T)$和$Rec(T)$是在待检索图像排序表中截断级别为T的准确率（Precision）和召回率（Recall）结果。

同时，图3.8展示了MMH与其他四种模型在隐特征数量K取不同值（K = 15及K = 20）时的准确率-召回率（Precision-Recall）曲线。从上述结果可以看出，虽然MMH并没有直接优化一个基于排序的损失函数，但由MMH得到的隐层空间表示在大部分情况下，可以得到比其他隐层空间学习方法更高的检索性能。这

表 3.1　各模型在Flickr图像标注问题中的Top-N F1值。

	MMH	**DWH**	**TWH**	**sLDA**
$F1@3$	**0.245**	0.202	0.218	0.146
$F1@4$	**0.258**	0.208	0.228	0.159
$F1@5$	**0.262**	0.210	0.236	0.169
$F1@6$	**0.259**	0.208	0.240	0.171
$F1@7$	**0.256**	0.206	0.239	0.175

主要是因为MMH学习得到的隐层空间表示具有更强的区分特性。对于Flickr数据集，实验中有相似的观测结果。例如，60维隐特征MMH，DWH和TWH的AP值分别为0.163， 0.153 和0.158。

3.4.3.3　图像标注

下面将MMH用于Flickr数据集的图像标注问题。图像对应的标签的字典大小为1000维。每一幅图像的平均标签数是4.5。

将MMH与两模态DWH，TWH比较，其中$\mathbf{X}$表示1000维标签，$\mathbf{Z}$ 表示634维实数值图像特征。将MMH与 sLDA标注模型[31]比较（只使用SIFT特征）。使用top-N F1值[31]评估模型性能，记为$F1@N$。表3.1所示是使用60维隐特征MMH的top-N F1值。再次，可以观察得出：虽然MMH没有直接最小化关于标注的损失函数，基于最大间隔的MMH比其他的模型标注性能高的原因是其具有判别式的隐层空间表示。图3.9所示为13类别的标注结果，对于每一类别，左侧的图像是正确的标注结果，而右侧的图像是错误的标注结果。

3.4.3.4　参数敏感度分析

最后，分析MMH模型关于正则化参数C_2的敏感度。在所有上述实验中，固定C_1为0.5，在训练过程中通过交叉验证方法来选择C_2。图3.10所示为正则化常数C_2变化时，模型MMH的分类正确率变化。如图可见，当隐特征维数K适当选取时，MMH在TRECVID 2003和Flickr数据集上对于正则化常数C_2并不敏感。

3.5　本章小结

本章研究参数化的隐层空间分类模型。以多模态数据融合与分类为例，提出多模态隐层空间马尔可夫网络模型。该模型可发现判别式的隐层空间表示，同时实现分类任务。为了克服最大似然估计易于过拟合以及弱判别性的缺点，本章创新性地采用最大间隔学习，通过使用线性期望算子，联合最小化负对数似然与最

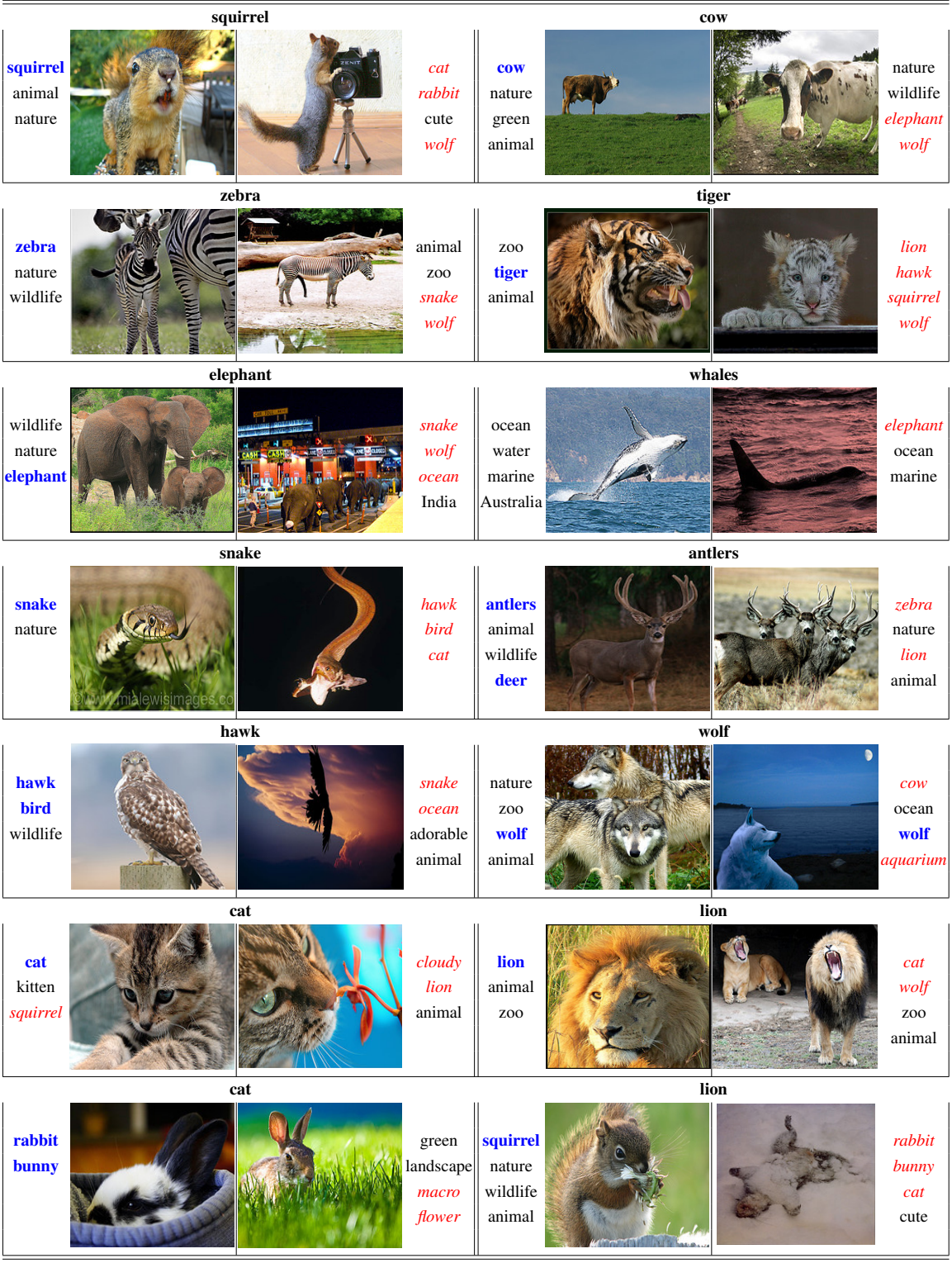
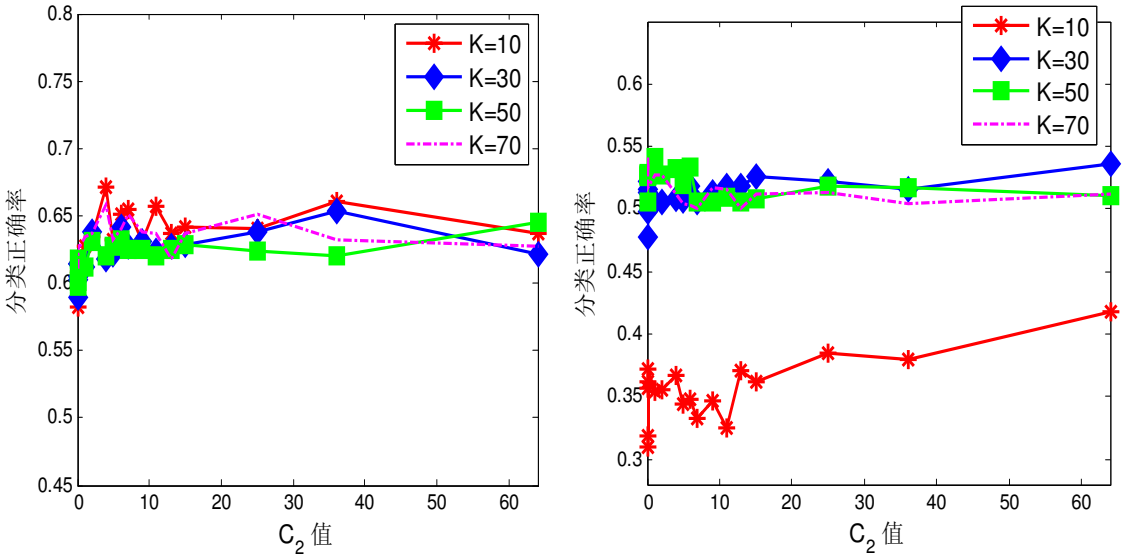


图 3.9　13类动物Flickr图像数据集的标注结果。带有蓝色加粗的标签是正确的标注结果，而红色斜体的标签为错误的标注结果。其他颜色的标注为中性的。

小化训练数据的预测损失函数方法学习模型参数。本章展示了模型在多种真实数据集上对于多种预测任务的评估结果（包括图像分类、检索、标注等问题中）。



(a) TRECVID数据上敏感度　　(b) Flickr数据上敏感度

图 3.10　MMH模型在两个数据集上关于参数C_2的敏感度分析。

本章提出的最大间隔学习方法，可用于广泛的无向隐层空间模型。本文将在下一章将其用于响应变量连续的回归模型。

第 4 章　参数化隐层空间回归及有结构化输入的隐层空间分类模型

第3章介绍了一种对多模态数据有效建模的表示框架，并提出一种新的基于最大间隔准则的学习方法。本章将上述结果扩展到处理回归分析，以及对有结构的输入数据进行统计建模。作为第3章提出的最大间隔隐层空间马尔可夫网络的扩展，在模型表示方面，本章首先提出用于预测连续响应变量的隐层空间马尔可夫网络回归模型；此外，研究输入数据满足一阶马尔可夫链式结构依赖关系的隐层空间模型表示与学习问题。该模型仍满足无向图隐层空间的弱条件独立的优良性质，高效地学习隐层空间表示的同时实现预测任务。在模型学习方面，为了克服传统最大似然估计方法存在容易过拟合以及判别性不强的缺点，采用和第3章相似的最大间隔学习方法，对隐变量的期望定义损失函数，通过最小化负对数似然和预测损失函数进行参数学习。最后，将提出的最大间隔马尔可夫网络回归模型和有结构输入的分类模型用于Hotel Review真实数据集上的预测问题。实验结果表明，最大间隔马尔可夫回归模型不仅可以发现判别式的隐层空间表示，还可高效地处理回归分析问题；同时，当在模型中考虑输入数据的结构依赖关系时，可进一步提高模型预测性能。

4.1　研究动机

与分类问题相似，回归分析是机器学习领域另一典型预测问题，它具有广泛的应用场景[29,40]。例如，近年来互联网为用户提供多种用于商品交易① 、旅游② 、娱乐信息的交流平台，其中包括大量的用户评论信息。基于这些大规模的评论文本，自动地预测被评论对象的优劣（如评价分数），是一个很有意义的应用问题。再比如，研究者希望根据Digg网站③ 中文本的描述，预测该网站中一篇文章的“diggs”（被点击收藏）数量，当dig（点击收藏）该页面的人数越多，说明该文章越受欢迎。上述两个例子都可以看作是预测连续值响应变量的回归分析问题，本章希望有监督的隐层空间模型也可在发现判别式隐层空间表示的同时，高效地

① www.amazon.com

② www.tripadvisor.com

③ www.digg.com，该网站完全由用户上传信息来源，然后让阅读者来决定文章是否有用，dig （点击收藏）页面的人数越多，说明该页面越受欢迎。当该文章的dig数达到一定数量时，该网页将被置于网站首页或其他重要页面上。

实现回归预测。

此外，虽然可简单地通过统计文本中词频，或者采用尺度不变特征变换提取图像的词袋（Bag-of-words）特征作为模型的输入，但无论是文本评论或图像数据都存在潜在的结构信息。例如，宾馆的评论信息通常是用户按照网站（如TripAdvisor等）提供的宾馆各方面（例如价格、房间、地理位置、整洁度、服务质量等）的评价标准而撰写的有序的满意度评价[60]；再比如，一幅图像中的各像素（如相邻像素点，或互相连通的像素点）间存在着空域上的关联[15]，如何引入这些数据本身存在的依赖信息，帮助模型发现更有价值的隐层空间表示，进一步提高模型预测性能，是本章的又一研究出发点。在第3章中，本文着重研究了隐层空间马尔可夫网络的特例，即不考虑输入变量间的结构信息情况下的隐空间模型。本章将对其进行进一步研究，考虑有结构的输入变量的有监督隐层空间马尔可夫网络分类问题。

下面，本章将分别介绍有监督隐层空间回归模型，以及考虑有结构输入数据的隐层空间分类模型。

4.2　有监督隐层空间回归模型

本节首先介绍经典的支持向量机回归分析问题，以及传统基于最大似然估计的有监督隐层空间回归模型，并在此基础上引出基于最大间隔的隐层空间回归模型，即最大间隔隐层空间马尔可夫网络回归模型。

4.2.1　支持向量回归分析模型

回归问题的定义：假设有一系列有监督信息的训练数据$\mathcal{D} = \{(\mathbf{x}_d, y_d)\}_{d=1}^{D}$，其中$\mathbf{X}$表示输入空间，$Y \in \mathbb{R}$表示连续的监督信息（如评价分数）。回归问题的任务[38]是：寻找一个尽可能平滑的函数$f(\mathbf{x})$，对于所有训练数据$\{\mathbf{x}_d\}_{d=1}^{D}$，满足$f(\mathbf{x}_d)$与真实目标$y_d$的差值尽量小，比如不多于一个阈值$\epsilon$。

下面考虑经典的线性支持向量机回归模型来解决上述回归问题。假设线性函数f满足

$$f(\mathbf{x}) = \mathbf{w}^{\top}\mathbf{x} + b$$

其中$\mathbf{w}$是预测参数向量，$b \in \mathbb{R}$是偏置量。与带有松弛变量的支持向量机分类模型类似，回归问题中求解支持向量机软约束目标函数

$$\min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} \ \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{d}(\xi_d + \xi_d^*)$$

$$\forall d,\ \text{s.t.}: \begin{cases} y_d - (\mathbf{w}^\top + b)\mathbf{x}_d \leqslant \epsilon + \xi_d \\ -y_d + (\mathbf{w}^\top + b)\mathbf{x}_d \leqslant \epsilon + \xi_d^* \\ \xi_d,\ \xi_d^* \geqslant 0 \end{cases} \tag{4-1}$$

其中常数$C > 0$是正则化常数，通常可以使用交叉验证选取；ϵ是一个准确度参数（Precision Parameter），它反映了模型能容忍预测误差的范围。上面的问题等价于最小化如下ϵ-不敏感（ϵ-insensitive）损失函数

$$\min_{\mathbf{w},b}\ \frac{1}{2}\|\mathbf{w}\|_2^2 + C\mathcal{R}_\epsilon(\mathbf{w}) \tag{4-2}$$

其中ϵ-不敏感损失函数定义为

$$\mathcal{R}_\epsilon(\mathbf{w}) \stackrel{\text{def}}{=} \sum_d \max(0, |y_d - (\mathbf{w}^\top \mathbf{x}_d + b)| - \epsilon),$$

即对于样本d，当预测值$(\mathbf{w}^\top \mathbf{x}_d + b)$与真实值$y_d$之间的差值小于$\epsilon$时，没有损失，否则，有一个线性的损失$|y_d - (\mathbf{w}^\top \mathbf{x}_d + b)| - \epsilon$。这是因为问题（4-2）中的松弛变量的最优解为

$$\forall d: \begin{cases} \xi_d = \max(0, y_d - (\mathbf{w}^\top + b)\mathbf{x}_d - \epsilon) \\ \xi_d^* = \max(0, -(y_d - (\mathbf{w}^\top + b)\mathbf{x}_d) - \epsilon) \end{cases} \tag{4-3}$$

通过考虑三种不同的情况，即：（1）$y_d - (\mathbf{w}^\top + b)\mathbf{x}_d > \epsilon$；（2）$-(y_d - (\mathbf{w}^\top + b)\mathbf{x}_d) > \epsilon$；（3）$-\epsilon \leq y_d - (\mathbf{w}^\top + b)\mathbf{x}_d \leq \epsilon$，可以得到

$$\forall d,\ \ \xi_d + \xi_d^* = \max(0, |y_d - (\mathbf{w}^\top + b)\mathbf{x}_d| - \epsilon).$$

直观上，ϵ-不敏感损失函数可用图4.1表示，其中左侧图中不在灰色区域内的红色数据点表示该样本的预测值$(\mathbf{w}^\top \mathbf{x}_d + b)$与真实值$y_d$之间的差值大于$\epsilon$，对应到右侧图中的损失值$\zeta$。对某个特定样本$d$，其损失函数为

$$\zeta_d \stackrel{\text{def}}{=} \begin{cases} 0 & e\ |y_d - \mathbf{w}^\top \mathbf{x}_d| \leq \epsilon \\ |y_d - \mathbf{w}^\top \mathbf{x}_d| - \epsilon & K \end{cases} \tag{4-4}$$

由于函数f是线性的，上述凸二次规划问题通常可以用已有高效工具包（如SVM-light等）求解其原问题或对偶问题，本章介绍最大间隔隐层空间回归模型的求解过程将会求解类似子问题。

为了简化，上述定义中的偏置量可以被吸收到$\mathbf{w}$中，即引用一维恒取1值的特征，b为该维特征对应的权重，下文中将不再显式地写出偏置量。

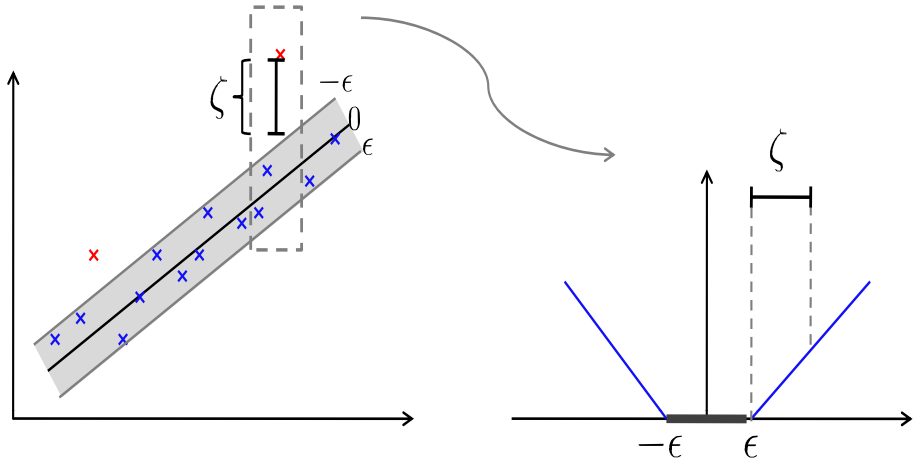


图 4.1　ϵ-不敏感损失函数示意图[38]。

4.2.2　基于最大似然估计的有监督隐层空间回归模型

图4.2为一个两模态的隐层空间马尔可夫网络回归模型的图结构。与第3章的分类模型相似，$\mathbf{X} \overset{\text{def}}{=} \{X_i\}_{i=1}^{N}$和$\mathbf{Z} \overset{\text{def}}{=} \{Z_j\}_{j=1}^{M}$分别表示两种不同模态的输入变量，$\mathbf{H} \overset{\text{def}}{=} \{H_k\}_{k=1}^{K}$表示一组需要推理得到的隐变量。回归模型与分类模型的区别在于，这里Y是连续的响应变量。根据多模态隐层空间马尔可夫网络的构造定义方法，需要定义Y的条件分布。对于连续变量Y，当已知隐变量$\mathbf{H}$的情况下，定义$y \in \mathbb{R}$的条件概率为正态分布

$$p(y|\mathbf{h}) = \mathcal{N}(y|\mathbf{V}^{\top}\mathbf{h}, \sigma^2), \tag{4-5}$$

其中$\mathbf{V}$为预测模型的参数。于是，可以定义该模型的联合分布$p(\mathbf{x}, \mathbf{z}, \mathbf{h}, y)$，与分类模型中的联合分布 (3-2)形式相似，唯一的区别在于指数函数中存在$-\frac{1}{2\sigma^2}(y^2 - y\mathbf{V}^{\top}\mathbf{h})$项。

如第3章所述，若不考虑输入变量间存在的结构信息，则无监督的多模态隐层空间马尔可夫网络模型退化为文献[98]中所述的两模态Harmonium（本文中记为Dual-wing Harmonium，简称DWH）回归模型。同样的，基于最大似然估计的隐层空间回归模型退化为和文献[35]中介绍的层次Harmonium（本文中记为Tri-wing Harmonium，简称TWH）模型相似，但是具有连续响应变量的TWH回归模型。对于所有这些基于最大似然估计的模型，都可以采用均值场或Contrastive Divergence等近似推理方法近似目标函数，使用最大似然估计方法学习模型参数。在本章实验中，将比较DWH、TWH在回归问题中的预测性能。

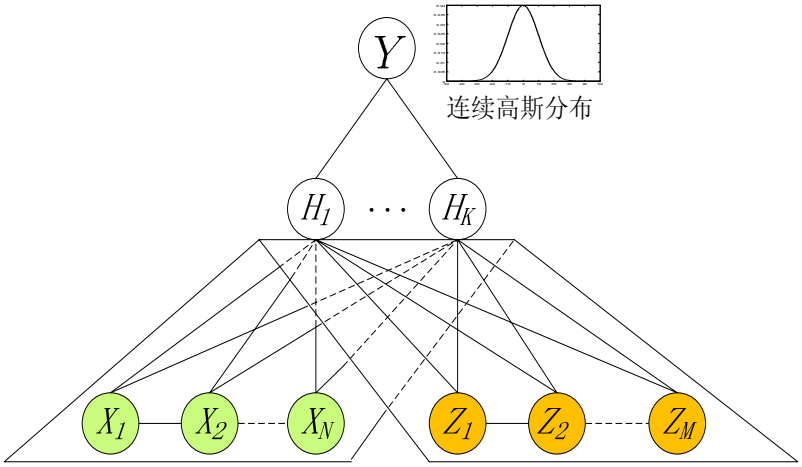


图 4.2　基于MLE的多模态隐层空间马尔可夫网络回归模型。

4.2.3　最大间隔有监督隐层空间回归模型

和上述基于最大似然估计的回归方法不同，本章提出一种判别式的最大间隔有监督隐层空间马尔可夫网路回归模型，考虑连续性的有监督信息（如回归问题中连续的评价分数等），学习判别式的隐层空间表示的同时，提高模型的回归分析性能。

4.2.3.1　回归问题定义

和第3章分类模型相似，这里要解决的一个关键问题是如何将确定性的最大间隔准则和基于概率的后验推理有机地融合在一起。这里的解决办法同上一章类似，即采用线性的期望算子。具体地说，当隐层空间表示$\mathbf{h}$给定时，可以定义线性的回归分析模型

$$F(\mathbf{h};\mathbf{V}) \stackrel{\text{def}}{=} \mathbf{V}^{\top}\mathbf{h}, \tag{4-6}$$

其中$\mathbf{V}$是一个K维向量。为了去掉隐变量$\mathbf{H}$的不确定性，这里对隐变量$\mathbf{H}$求期望，并定义适于回归问题的线性*期望*的预测准则

$$y^{*} \stackrel{\text{def}}{=} \mathbf{V}^{\top}\mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[\mathbf{h}]. \tag{4-7}$$

为了学习预测模型$\mathbf{V}$，需要设计一个损失函数，结合最大间隔准则解决回归预测问题。在此选择最小化 ϵ-不敏感损失函数, 该损失函数也被用于4.2.1节中所述的标准支持向量机回归模型[38]中。对于预测准则（4-7），其ϵ-不敏感损失函数为

$$\mathcal{R}_{\epsilon}(\mathbf{V}) \stackrel{\text{def}}{=} \sum_{d} \max(0, |y_d - \mathbf{V}^{\top}\mathbb{E}[\mathbf{h}_d]| - \epsilon), \tag{4-8}$$

其中，$\epsilon \in \mathbb{R}_+$为精确度参数，通常比较小。为了便于表示，上述定义将$\mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\mathbf{h}]$简记为$\mathbb{E}[\mathbf{h}_d]$。根据**正则化风险最小化**（Regularized Risk Minimization）准则，最大间隔隐层空间马尔可夫网络回归模型求解下面的联合优化问题

$$\text{P2}: \min_{\Theta,\mathbf{V}} \; L(\Theta) + \frac{1}{2}C_1\|\mathbf{V}\|_2^2 + C_2\mathcal{R}_\epsilon(\mathbf{V}), \tag{4-9}$$

其中，$L(\Theta)$是输入数据的负对数似然，它和最大间隔隐层空间分类模型的$L(\Theta)$相同。

同理，可以看出，上述问题$P2$通过联合最小化负对数似然和最小化回归问题的损失函数，可以同时学习隐层空间表示和一个预测模型，并且一方面试图更好的解释数据，另一方面希望得到更准确的预测结果。

4.2.3.2　基于Contrastive Divergence的优化方法

理论上对于回归问题$P2$，可以采用和第3章相似的次梯度下降方法学习参数Θ。但由于回归问题中的ϵ-不敏感损失函数 $\mathcal{R}_\epsilon$（如公式4-8）中存在一个不可导的绝对值项，这比分类问题中的铰链损失函数$\mathcal{R}_{hinge}$ 更加复杂。于是，这里求解P2的一个等价的有约束优化问题$P2'$

$$\begin{aligned}
&\text{P2}': \min_{\Theta,\mathbf{V},\xi,\xi^\star} \; L(\Theta) + \frac{1}{2}C_1\|\mathbf{V}\|_2^2 + C_2\sum_d(\xi_d + \xi_d^*) \\
&\forall d, \text{ s.t.}: \begin{cases} y_d - \mathbf{V}^\top\mathbb{E}[\mathbf{h}_d] \leqslant \epsilon + \xi_d \\ -y_d + \mathbf{V}^\top\mathbb{E}[\mathbf{h}_d] \leqslant \epsilon + \xi_d^* \\ \xi_d,\ \xi_d^* \geqslant 0 \end{cases}
\end{aligned} \tag{4-10}$$

其中ξ_d和ξ_d^* 都为松弛变量。然而，由于似然函数$L(\Theta)$中存在很难计算的正则化因子，所以有约束的问题$\text{P2}'$仍然很难求解。与分类问题相似，又一次采用Contrastive Divergence的变分推理方法，用$\mathcal{L}^v(q_0,q_1)$近似似然函数$L(\Theta)$。于是，得到近似的最大间隔回归问题$P2''$

$$\begin{aligned}
&\text{P2}'': \min_{\Theta,\mathbf{V},\xi,\xi^\star} \; \mathcal{L}^v(q_0,q_1) + \frac{1}{2}C_1\|\mathbf{V}\|_2^2 + C_2\sum_d(\xi_d + \xi_d^*) \\
&\forall d, \text{ s.t.}: \begin{cases} y_d - \mathbf{V}^\top\mathbb{E}[\mathbf{h}_d] \leqslant \epsilon + \xi_d \\ -y_d + \mathbf{V}^\top\mathbb{E}[\mathbf{h}_d] \leqslant \epsilon + \xi_d^* \\ \xi_d,\ \xi_d^* \geqslant 0 \end{cases}
\end{aligned} \tag{4-11}$$

采用拉格朗日方法，对于每一样本d，为四个约束分别引入拉格朗日乘

Algorithm 2 最大间隔隐空间马尔可夫网络回归模型的学习算法

输入: 数据$\mathcal{D} = \{(\mathbf{x}_d, \mathbf{z}_d, y_d)\}_{d=1}^{D}$，正则化常量$C_1$，$C_2$以及隐空间的维度$K$
输出: 模型参数Θ和$\mathbf{V}$
初始化$\mathbf{V} = 0$，Θ设为对输入矩阵$M = [\mathbf{x}_1, \cdots, \mathbf{x}_D; \mathbf{z}_1, \cdots, \mathbf{z}_D]$做SVD分解得到的前$K$个的左特征向量。
repeat
  for $d = 1$ **to** D **do**
    使用公式（3-10）计算隐空间表示$q_0(\mathbf{h}_d)$;
    初始化$q_1(\mathbf{h}_d) = q_0(\mathbf{h}_d)$
    for $t = 1$ **to** T（实验中设为5） **do**
      使用公式（3-10）计算$q_1(\mathbf{x}_d)$和$q_1(\mathbf{z}_d)$
      使用公式（3-10）计算$q_1(\mathbf{h}_d)$
    end for
  end for
  通过求解SVR回归分析模型（4-14）计算参数$\mathbf{V}$以及对偶变量μ和μ^*;
  通过梯度下降法求解参数Θ
until 目标函数L相对变化小于τ (如$1e^{-4}$) 或者迭代次数超过某个阈值（如20）。

子$\mu_d, \mu_d^*, v_d, v_d^*$，得到拉格朗日函数$L$为

$$
\begin{aligned}
L = \mathcal{L}^v(q_0, q_1) + \frac{1}{2}C_1\|\mathbf{V}\|_2^2 + C_2\sum_d(\xi_d + \xi_d^*) - \sum_d(v_d\xi_d + v_d^*\xi_d^*) \\
- \sum_d\Big\{\mu_d(\epsilon + \xi_d - y_d + \mathbf{V}^\top\mathbb{E}[\mathbf{h}_d]) + \mu_d^*(\epsilon + \xi_d^* + y_d - \mathbf{V}^\top\mathbb{E}[\mathbf{h}_d])\Big\}.
\end{aligned}
$$

于是，如算法2所示，可以通过交替地进行下面三步优化拉格朗日函数L

（1） 固定变量$\Theta, \mathbf{V}, \mu, \mu^*$，推理得到$q_0$和$q_1$: 这一步与分类模型的原理相同，在此不再赘述;

（2） 固定$\mathbf{V}$、μ_d和μ_d^*，估计Θ: 这一步可以使用梯度下降的方法（如分类模型中L-BFGS[121]方法）求解，其中参数 (θ, η, λ)的梯度计算公式与第3章所介绍的分类模型中的相应计算公式相同，在此不再赘述；而$\mathbf{W}$和$\mathbf{U}$的梯度计算公式分别为

$$
\begin{aligned}
\partial\mathbf{W}_e^k &= \Delta\mathbb{E}[\phi(\mathbf{x}_e)\varphi(h_k)^\top] - \sum_d(\mu_d - \mu_d^*)\mathbf{V}_k\frac{\partial\mathbb{E}_{q_0}[h_k]}{\partial\mathbf{W}_e^k} \\
\partial\mathbf{U}_e^k &= \Delta\mathbb{E}[\psi(\mathbf{z}_e)\varphi(h_k)^\top] - \sum_d(\mu_d - \mu_d^*)\mathbf{V}_k\frac{\partial\mathbb{E}_{q_0}[h_k]}{\partial\mathbf{U}_e^k}
\end{aligned}
\tag{4-12}
$$

（3） 估计参数$\mathbf{V}$和拉格朗日乘子$\{\mu_d, \mu_d^*\}$：令

$$\partial L/\partial \xi_d, \partial L/\partial \xi_d^*, \partial L/\partial \mathbf{V} = 0,$$

利用KKT条件，可以得到

$$\mathbf{V} = \frac{1}{C_1}\sum_d (\mu_d - \mu_d^*)\mathbb{E}[\mathbf{h}_d]. \tag{4-13}$$

将公式(4-13)代入拉格朗日函数L，得到$P2''$的对偶问题

$$\begin{aligned} &\max_{\mu,\mu^*} -\frac{1}{2C_1}\|\sum_d (\mu_d - \mu_d^*)\mathbb{E}[\mathbf{h}_d]\|_2^2 - \sum_d [\epsilon(\mu_d + \mu_d^*) - y_d(\mu_d - \mu_d^*)] \\ &\text{s.t.}: \mu_d, \mu_d^* \in [0, C_2], \forall d, \end{aligned} \tag{4-14}$$

这个优化问题可以使用标准的QP solver或者已有的成熟算法（如SVM-light[125]）求解，从而计算对偶变量μ_d 和μ_d^*。

可以发现，在基于最大间隔的隐层空间回归模型中，$\mathbf{W}$和$\mathbf{U}$的梯度中包含关于拉格朗日乘子μ_d和μ_d^*的额外项，即如公式（4-12）中的最后一项。这一项对于学习隐层空间表示起到正则化作用。即在考虑代价损失函数情况下的模型预测值$\mathbf{V}^\top\mathbb{E}[\mathbf{h}_d]$与真实值$y_d$ 的绝对值差别大于ϵ时，拉格朗日乘子μ_d或μ_d^*（由于KKT条件的制约，两者最多有一项非零）将为非零，从而这一项将会使得模型朝着更加有益于发现判别式的隐层空间表示（即更有利于回归）的方向改进。如作者所见，这个“偏差项”将使最大间隔的隐层空间回归模型更利于发现具有强区分性的隐层空间表示。

值得说明的是，为了更加深入地研究最大间隔学习准则的有效性并与之前相关工作（如DWH，TWH）比较，与上一章分类模型相似，在实验部分将主要研究模型中输入变量间相互独立时的最大间隔Harmonium回归模型（Max Margin Harmonium，简称MMH）。上述方法可以用来方便地进行参数估计。为了推理q_0和q_1, $\mathbf{x}$，$\mathbf{z}$和$\mathbf{h}$的分布都完全地满足因子化，可很容易地计算MMH回归模型中的次梯度。

4.3　有结构输入的隐层空间马尔可夫网络

本文已在第3章和本章第4.2.3节详细研究了最大间隔隐层空间马尔可夫网络分类和回归模型，以及其特例：当输入变量间满足条件独立时的最大间隔Harmonium模型。本节研究一个具体的有结构输入数据的隐层空间马尔可夫网络模型。特别地，本节选用Hotel Review数据为例，对有段落依赖关系的评论文本

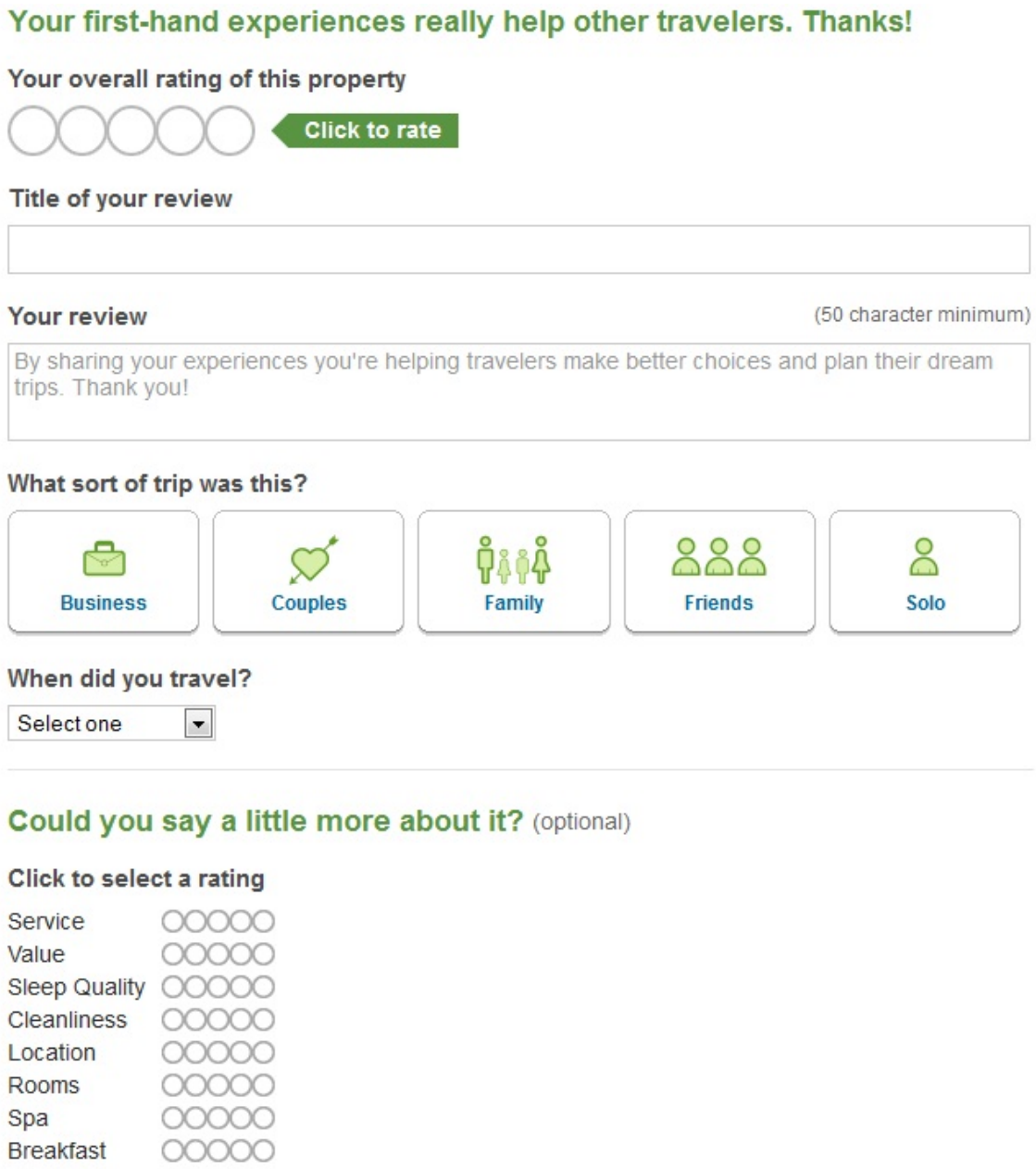


图 4.3　TripAdvisor宾馆评论网站的用户评论输入界面。

数据建模。如图4.3所示，在TripAdvisor ① 中有若干种已经定义好的评论点（如房间、地理位置、整洁度、服务质量、早餐等），这些评论点可引领用户按此撰写评论信息。由于这些评论点的有序性，用户提交的评论内容通常也有着相同的顺序。虽然其他的可能依赖关系（如句子、句子间的顺序关系）也存在，在此只考虑段落之间的依赖信息，然后设计一种有结构的隐层空间马尔可夫网络：

如图4.4所示，对于某一文档样本，用一个$P \times N$的观测矩阵$\mathbf{X}$表示，其中P为

① www.tripadvisor.com

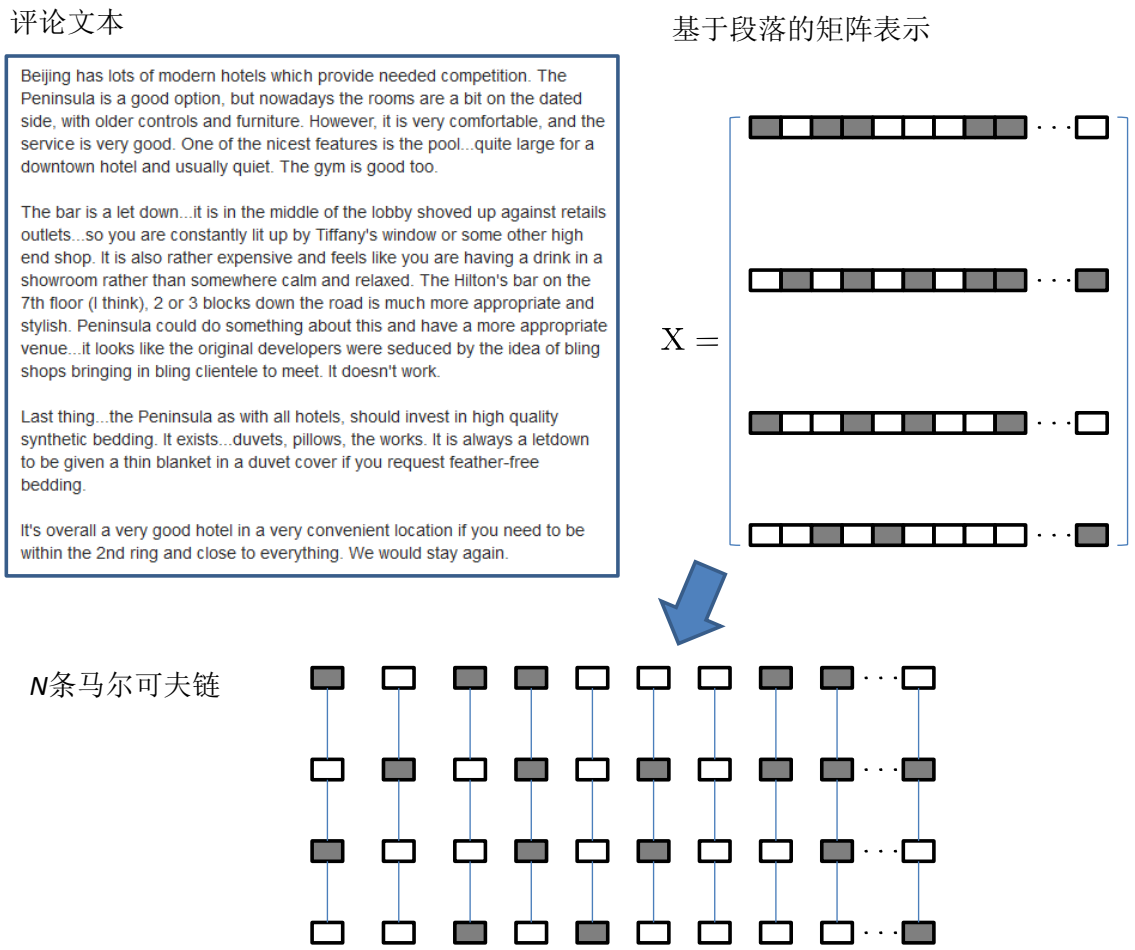


图 4.4　基于段落的矩阵表示，图中左侧文本由4段组成，每段分别对应右侧矩阵**X**的一行。灰色的格子表示对应的单词出现在某个段落里，白色的格子表示对应的单词没有出现在相应的段落里。

该文档里面段落的数目，N为词典大小。每一行$\mathbf{x}_p$为一个向量，$x_{pi}=1$表示第i个词在第p个段落中出现，否则不出现。每一列$\mathbf{x}_{\cdot i}$表示第i个词在所有段落中出现的模式。为了考虑各段落间的顺序信息，这里对每一个$\mathbf{x}_{\cdot i}$定义一阶马尔可夫链，并假设不同的$\mathbf{x}_{\cdot i}$满足条件独立性。即条件分布

$$p(\mathbf{x}|\mathbf{h})=\prod_{i=1}^{N}p(\mathbf{x}_{\cdot i}|\mathbf{h}),$$

其中每一个$p(\mathbf{x}_{\cdot i}|\mathbf{h})$都为一个具有链式结构的条件随机场（Conditional Random Fields，简称CRF）[115]。该模型事实上为一个N模态的隐层空间马尔可夫网络，其中第i个模态有P个变量$\{X_{pi}\}_{p=1}^{P}$，这些变量通过一个线性链连接，如图4.4所示。

由第3章3.2节的定义（如公式（3-1）和公式（3-2）），这里需要为每一模态输入指定边的集合（Edge Sets），并定义其特征函数。令E表示某一模态对应

的广义边（Generalized Edges）的集合，其元素包括该模态[①]对应的链式结构上的独立顶点（Singleton Vertices）以及顶点与顶点之间的边（pairwise edges）。即$E = \{(1), \cdots, (P), (1,2), \cdots, (P-1, P)\}$，其中$(p)$表示一个退化的边（Degenerate Edge）（即顶点p）。于是，特征函数ϕ包括：定义在单一变量（Single Variable）上的*零阶*特征函数g，以及四个定义在两个变量上的*一阶*特征函数 ϕ_0, ϕ_1, ϕ_2和ϕ_3。这里，零阶特征函数定义为

$$g(x_{pi}) = x_{pi},$$

四个一阶特征函数定义为

$$\forall j = 0, \cdots, 3: \ \phi_j(x_{pi}, x_{p+1,i}) = \begin{cases} 1, & \text{若 } 2x_{pi} + x_{p+1,i} = j \\ 0, & \text{否则} \end{cases}$$

其中α 和β_j $(j = 0, \cdots 3)$分别表示与特征函数g和ϕ对应的权值。为了使模型更加丰富，假设不同模态变量有不同的特征权值，但是在每一模态中，所有的边都共享同一组权值。为了简化记号，定义

$$g(\mathbf{x}_{\cdot i}) \stackrel{\text{def}}{=} \sum_{p=1}^{P} g(x_{pi})$$

为g在序列$\mathbf{x}_{\cdot i}$上的累积特征函数（Accumulated Function），

$$\phi_j(\mathbf{x}_{\cdot i}) \stackrel{\text{def}}{=} \sum_{p=1}^{P-1} \phi_j(x_{pi}, x_{p+1,i})$$

为 ϕ_j在序列$\mathbf{x}_{\cdot i}$上的累积特征函数（Accumulated Functions）。下面，定义输入变量$\mathbf{X}$和隐变量$\mathbf{H}$之间的关联项（Interaction Terms）为

$$\sum_{i=1}^{N} g(\mathbf{x}_{\cdot i}) \mathbf{U}^i \mathbf{h} + \sum_j \phi_j(\mathbf{x}_{\cdot i}) \mathbf{W}_j^i \mathbf{h},$$

其中$\mathbf{W}_j^i$和$\mathbf{U}^i$为K维实数值向量。此外，还需在指数族联合分布里引入一个真实变量$\mathbf{H}$的二次能量函数。综上定义，可以得到模型的联合概率分布

$$p(\mathbf{x}, \mathbf{h}) \propto \exp\Big\{ \sum_i g(\mathbf{x}_{\cdot i})(\alpha^i + \mathbf{U}^i \mathbf{h}) + \sum_{ij} \phi_j(\mathbf{x}_{\cdot i})(\beta_j^i + \mathbf{W}_j^i \mathbf{h}) - \frac{1}{2} \mathbf{h}^\top \mathbf{h} \Big\},$$

以及条件概率分布

$$p(\mathbf{x}_{\cdot i} | \mathbf{h}) \propto \exp\Big\{ g(\mathbf{x}_{\cdot i})(\alpha^i + \mathbf{U}^i \mathbf{h}) + \sum_j \phi_j(\mathbf{x}_{\cdot i})(\beta_j^i + \mathbf{W}_j^i \mathbf{h}) \Big\}$$

① 根据定义，所有模态都有相同的边的集合。

$$p(h_k|\mathbf{X}) = \mathcal{N}\Big(h_k|\sum_{i=1}^{N}(g(\mathbf{x}_{.i})\mathbf{U}_k^i + \sum_j \phi_j(\mathbf{x}_{.i})\mathbf{W}_{jk}^i), 1\Big)$$

于是可根据第3.3节和第4.2.3节中所述，进行参数学习和推理。由于每一模态都满足一阶马尔可夫链，本章运用前向-后向消息传递方法（Forward-Backward Message Passing Scheme）进行模型推理，与文献[115]中方法相同，在此不再赘述。

4.4　时间复杂度

本小节讨论有监督的隐层空间马尔可夫网络回归模型及有结构的隐层空间马尔可夫网络的时间复杂度。

由于本章提出的最大间隔隐层空间马尔可夫网络回归模型只定义了一个局部（*partial*）（即只定义在输入数据$(\mathbf{x}, \mathbf{z})$上）的似然函数，所以不需迭代，一步即可高效地推理得到判别式的隐空间表示，和第3章介绍的分类模型相同，其计算复杂度关于输入特征的维度成线性关系。相反，基于MLE的有监督隐层空间回归模型，它定义一个关于输入变量$(\mathbf{x}, \mathbf{z})$和响应变量y的完全（full）的似然函数$p(\mathbf{x}, \mathbf{z}, y)$，因此在测试过程中，$Y$是不可观测的，这时需要迭代下面两步

（1） 推理得到Y的后验分布；

（2） 推理得到隐层空间表示。

因此，基于MLE的有监督隐层空间回归模型的测试时间是用最大间隔方法的一个恒定倍的时间。但总体上说，这些无向图模型的推理效率与有向图相比更加高效，本章将在4.5.4节实验部分详细比较。

对于具有一阶马尔可夫链式结构输入的隐层空间马尔可夫网络，每一个文档的消息传递（Message Passing）方法的时间复杂度是$O(N \times P \times S^2)$，其中S为每一个变量X_{pi}的可能值。因为X_{pi}为二值的，于是有$S^2 = 4$（一个非常小的常量）。另外，段落P的数量平均来看也非常小，例如在本文用到的Hotel Review数据集中，平均的段落数为9。因此，总体时间复杂度与特征维数N（即已知字典的大小）成线性关系。

4.5　实验结果与分析

下面，本章通过在两种不同Hotel Review真实数据集上的实验，分别定性与定量地评估上述回归模型及有结构化输入的隐层空间模型的性能，包括隐层空间表

示、预测性能、以及时间效率等。首先，深入研究最大间隔Harmonium回归模型，并与流行的用于回归分析的隐层空间模型进行比较。然后，在4.5.3.2节，展示考虑具有链式结构输入的隐层空间马尔可夫网络在解决分类问题中的性能。

4.5.1　数据集与特征

本实验使用两种Hotel review数据集①，分别用于评价回归问题和考虑有结构化输入的分类问题。

（1）第一种Hotel Review数据集[117]包含5000个随机从TripAdvisor② 网站上下载收集到的宾馆评论（Hotel Reviews）文本。每一个评论文档都由两种特征，即12000维词袋BOW特征和14维描述单词属性的特征[117]组成，以及一个全局评价分数和五个局部评价分数。其中，14维描述单词属性的特征包括：单词的词性标注，这里考虑形容词、名词、动词和副词；单词的褒、贬或者中性等，对于单词的褒贬特性，我们使用一些种子单词（褒义词如good, excellent等；贬义词如bad, painful等），通过在WordNet③ 上查找同义词和反义词获得。全局评价分数从1到5排序，将其归一化为(0, 1)的实数值。五种局部评价指标分别为：价格、房间、地理位置、整洁度、服务。在本实验中，只关注评论文档的全局评价分数的预测任务，随机地将整个数据集平分为训练集和测试集。需要说明的是，在下述所有实验中，离散的文本BOW特征被转化为0/1二值特征（即当一个词出现时，特征值为1，否则为0），并假设其服从伯努利分布。

（2）第二个Hotel Review数据集同样是从TripAdvisor网站下载收集到的，该数据集中每一条评论都包含段落号及每段落的12000维BOW特征，所有文档的平均段落数是9。每一种不同评价分数都对应着600条评论。将所有文档按照五种全局评价分数的不同分为1到5类。本文实验中此数据集用于有结构输入的分类问题，随机选择一半用于训练，另一半用于测试。不使用第一种Hotel Review数据集的原因是第一个数据集中的很多样本只有一个段落。

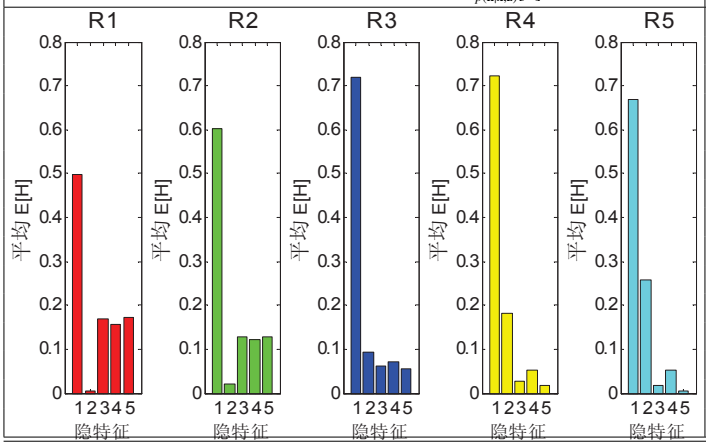
4.5.2　判别性隐层空间表示

本小节展示隐层空间马尔可夫回归模型在第一种Hotel Review文本数据集上自动学习得到的隐层空间表示的判别性。实验中将5中评价分数按照从小到大的顺序依次命名为 $R1, R2, \cdots R5$。表4.1所示为5维隐特征的MMH、TWH和DWH回

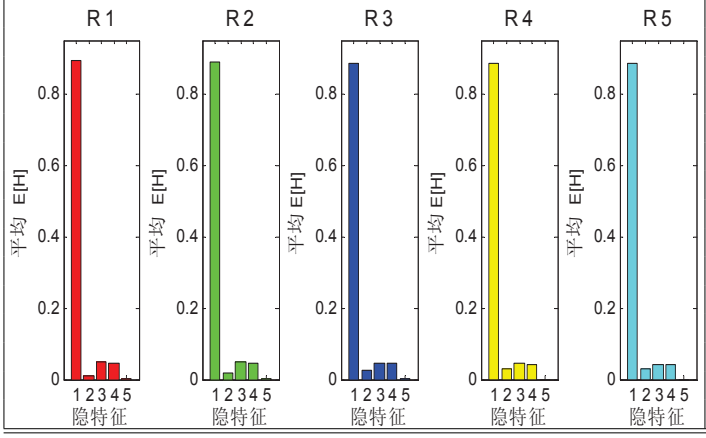
① http://www.cs.cmu.edu/ junzhu/data.htm。
② http://www.tripadvisor.com
③ http://wordnet.princeton.edu/

表 4.1　5维隐特征的MMH，TWH和DWH回归模型得到的隐层空间表示。

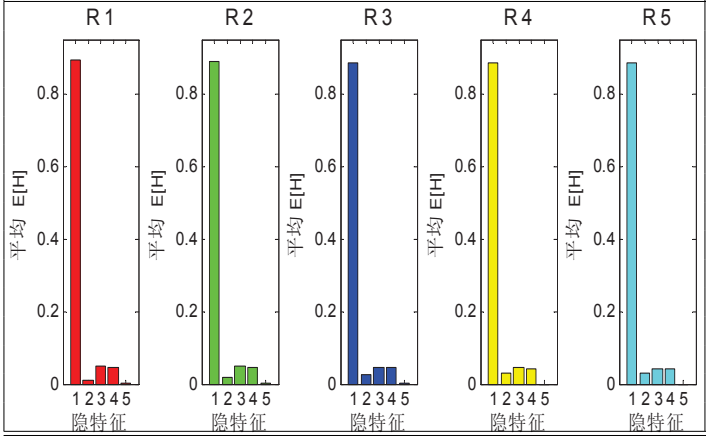
Max Margin Harmonium (Avg-KL: 3.568)

5种不同评价分数对应的样本的平均$\mathbb{E}_{p(\mathbf{h}\|\mathbf{x},\mathbf{z})}[\mathbf{h}]$	F1	F2	F3	F4	F5
	room	*great*	small	worst	worst
	hotel	*loved*	worst	small	dirty
	n't	arrived	dirty	dirty	small
	time	*enjoyed*	shower	shower	shower
	stay	*fantastic*	broken	broken	broken
	day	bit	smell	smell	smell
	night	*wonderful*	paying	paying	paying
	good	*lovely*	bathroom	bathroom	poor
	staff	pool	poor	poor	toilet
	pool	trip	toilet	toilet	refund
	back	beach	staying	refund	manager
	rooms	*fun*	refund	staying	bathroom
	food	*happy*	breakfast	walls	walls
	area	pools	hotel	hotel	carpet
	nice	*perfect*	walls	carpet	paid

Tri-Wing Harmonium (Avg-KL: 0.045)

5种不同评价分数对应的样本的平均$\mathbb{E}_{p(\mathbf{h}\|\mathbf{x},\mathbf{z})}[\mathbf{h}]$	F1	F2	F3	F4	F5
	room	beach	bathroom	resort	experience
	hotel	food	parking	beach	breakfast
	n't	pool	tv	ocean	stay
	time	*great*	area	trip	service
	stay	bar	coffee	vacation	*beautiful*
	day	resort	kitchen	desk	dinner
	night	restaurants	bed	check	guests
	good	drinks	street	time	made
	staff	restaurant	floor	front	*wonderful*
	pool	view	*large*	*great*	feel
	back	lunch	small	called	trip
	rooms	*good*	tub	call	house
	food	sea	*comfortable*	property	visit
	area	*beautiful*	location	*beautiful*	special
	nice	walk	internet	people	*comfortable*

Dual-Wing Harmonium (Avg-KL: 0.038)

5种不同评价分数对应的样本的平均$\mathbb{E}_{p(\mathbf{h}\|\mathbf{x},\mathbf{z})}[\mathbf{h}]$	F1	F2	F3	F4	F5
	room	beach	food	breakfast	belize
	hotel	food	told	reception	brett
	n't	pool	asked	bathroom	cam
	time	resort	holiday	bed	canapes
	stay	*great*	reception	shower	canoeing
	day	restaurants	day	holiday	caracol
	night	bar	bar	coffee	hosts
	good	drinks	staff	evening	nadege
	staff	restaurant	back	small	underway
	pool	lunch	manager	*clean*	wineries
	back	sea	people	bar	adopted
	rooms	*beautiful*	evening	hotel	amanda
	food	entertainment	entertainment	*good*	aurora
	area	pools	arrived	main	begun
	nice	view	hotel	tea	boasted

归模型得到的隐层空间表示。表中左侧为五种不同评价分数R_x对应样本的5维隐特征$\mathbb{E}[H_k]$值的归一化平均分布，表中右侧为每一维隐特征F_k对应的出现频率最高的15个词。

可以发现，基于最大间隔准则的MMH得到的隐层空间表示比基于最大化似然估计的DWH和TWH的结果更具有区分性。例如，MMH的隐特征F2的期望值对表示评价分数值由低到高（由$R1$到$R5$）的文本呈现逐渐增长的趋势，更具体地说，特征F2表示评价分数更高（$R5$和$R4$）的文本概率较大，但表示评价分数较低（如$R1$和$R2$）的文本概率较小（甚至降为零）。相应的，表4.1右侧部分F2对应着很多褒义情感的高频词（如‘great’，‘fantastic’，‘wonderful’，‘perfect’等），所以更适于表示评价分数值高的样本。相反，隐特征F3，F4和F5的期望值对评价分数值由低$R1$到高$R5$的样本呈现平滑下降的趋势。对应到表的右侧，隐特征F3，F4和F5对应着很多贬义情感的高频词（如‘worst’，‘dirty’，‘poor’等），所以更适于表示评价分数值较低的样本。此外，可以发现特征F1的期望值比其他特征都高很多，这是因为评论文本中包含大量关于宾馆信息的中性词（如‘room’，‘hotel’，‘food’，‘area’等），因此，这种结果也是合理的。相比之下，基于最大似然估计的DWH和TWH回归模型得到的隐特征期望值，对具有不同评价分数的样本并没有明显的区分性。最后，作者计算所有类别间的$\mathbb{E}[\mathbf{H}]$平均KL 散度值，用来定量地衡量隐层空间表示的区分性程度，其中MMH、DWH、TWH回归模型得到的平均KL散度值分别是3.568，0.038，0.045，这与定性分析的结果吻合，即MMH回归模型得到的隐层空间表示更具有区分性。

4.5.3　预测性能

下面比较MMH回归模型与文献[3,35,117]中其他算法在Hotel Review文本评价分数预测问题，以及有段落依赖关系的Hotel Review文本分类问题的预测性能。

4.5.3.1　文本回归结果

与文献[117]相似，本实验将预测Hotel Review数据集上的评价分数看作一个回归分析问题。实验中将MMH回归模型与无监督的DWH、有监督的TWH、sCTRF（即有监督条件主题随机场模型supervised Conditional Topic Random Fields，简称sCTRF）[117]以及MedLDA[40]回归模型比较。其中，多模态MMH、TWH、DWH、sCTRF模型的输入为12000维的BOW特征及14维的上下文相关（Contextual）特征，而单模态MedLDA模型输入仅为12000维的BOW特征。对于无监督DWH模型，需要在隐层空间表示基础上建立一个线性SVR回归预测模型。

其他的许多基准算法（如有监督LDA）的结果不如sCTRF[117]，在此不再一一列出，具体结果可参见论文[117]。考虑每一种文本的两种类别输入：其中输入**X**表示BOW（bag-of-word）特征，而另一种输入**Z**表示14维contextual特征[117]。实验中使用预测性R2值作为回归分析性能的评价标准，R2值的计算公式为

$$predictive\ R2 = 1 - \frac{\sum_d (y_d^* - y_d)^2}{\sum_d (y_d - \bar{y})^2} \tag{4-15}$$

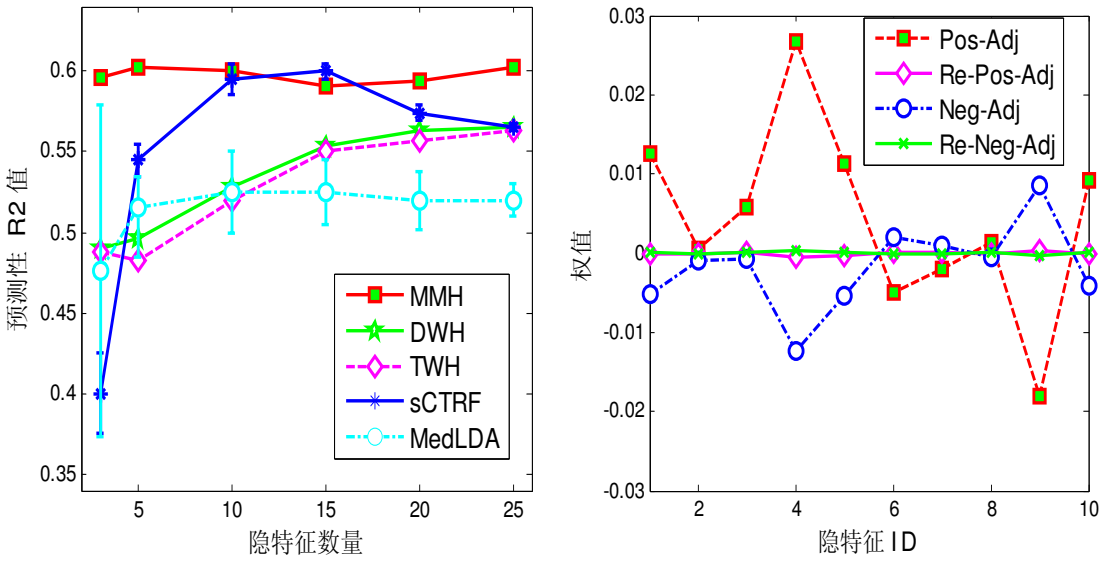
其中$\bar{y}$是真实标注的平均值。

图4.5（a）所示为各回归模型的预测性R2值。可以发现，当隐特征维数K变化时，基于最大间隔方法的MMH基本都得到优于其他模型的性能。相比之下，基于最大似然估计的有监督TWH与无监督的DWH模型并没有表现出优越的性能。此结果表明最大间隔的学习方法对于提高模型的回归预测性能具有显著作用。另外，多模态的MMH与单模态的MedLDA模型（仅使用12000维BOW特征）相比较结果更好的原因是，MMH可以同时学习多模态的特征，这也证明了考虑具有互补信息的多模态输入比仅考虑单模态特征更有助于提高模型预测性能。事实上，与第一种特征相比，第二种输入特征更适于预测。图4.5（b）中展示了使用10维隐特征的MMH回归模型学习得到的第二个模态输入中4种特征对应的权重，这四种特征分别是：（1）‘Pos-Adj’表示褒义的形容词（Positive Adjective），如‘good’，‘perfect’等；（2）‘Re-Pos-Adj’表示前面有否定词（Denying Word）的褒义形容词；（3）‘Neg-Adj’表示贬义的形容词（Negative Adjective），如‘worst’，‘dirty’等；（4）‘Re-Neg-Adj’表示前面有否定词的贬义形容词。可以看出，无论是‘Pos-Adj’还是‘Neg-Adj’特征都可以发现对于评价分数更具区分性的隐特征（如F4和F9）。MMH的最好结果与sCTRF相比不相上下，而sCTRF模型是一种有向图，因为存在第1章介绍的V-结构，其后验推理更加复杂。在第4.5.4节中读者将会看到，基于无向图的MMH回归模型与基于有向图的sCTRF相比，训练和测试时更加高效。

4.5.3.2　结构化有段落依赖关系的文本分类结果

图4.6为具有一阶马尔可夫链式结构输入的隐层空间模型在第二个TripAdvisor数据集上的分类结果（回归结果也可容易地得到）。实验中比较隐特征维度变化时，结构化的有监督MMH模型（记为structMMH）、结构化的无监督DWH模型（记为structDWH）与不考虑结构输入的最优模型（即MMH模型）对第二种Hotel Review数据的分类结果。

从图中结果观察发现，当隐特征维度变化时，基于最大间隔的有结



(a) 预测性R2值　　(b) 10维隐特征MMH的特征权值

图 4.5　回归预测问题的结果比较。

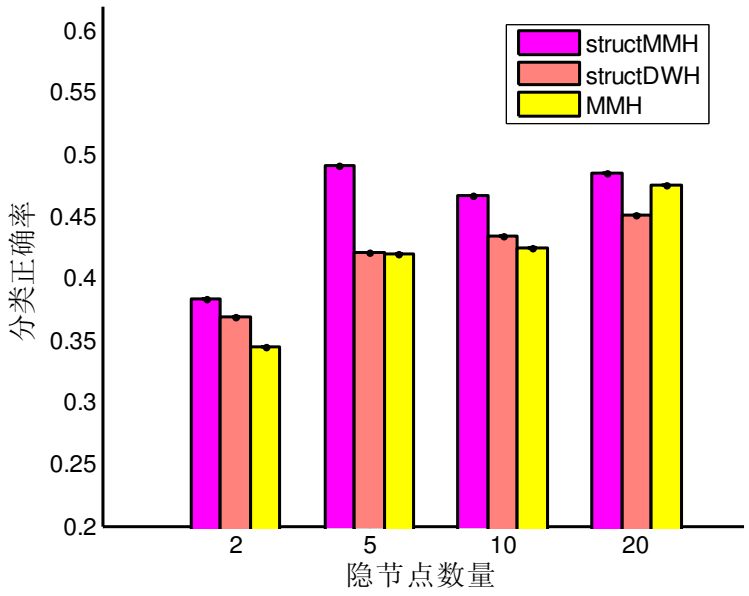
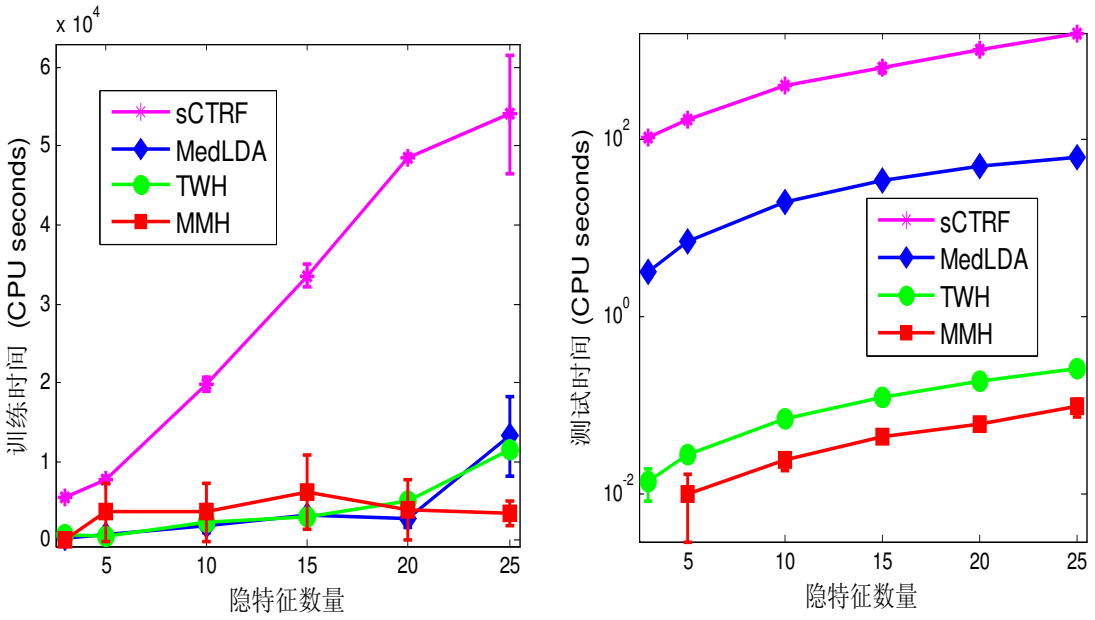


图 4.6　结构化有监督MMH、结构化无监督DWH及无结构有监督MMH模型在TripAdvisor数据集上的分类正确率。

构structMMH模型始终比无结构的最大间隔MMH模型以及无监督的有结构structDWH模型的预测性能更好，这个观测结果表明：（1）考虑段落间的依赖关系更有助于提高模型的预测性能；（2）考虑有监督信息可以显著提高隐空间表示的判别能力。



(a) 训练时间　　(b) 测试时间

图 4.7　在第一种Hotel Review数据集上各种回归预测模型的训练和测试时间比较[117]，其中纵轴坐标经过对数变换。

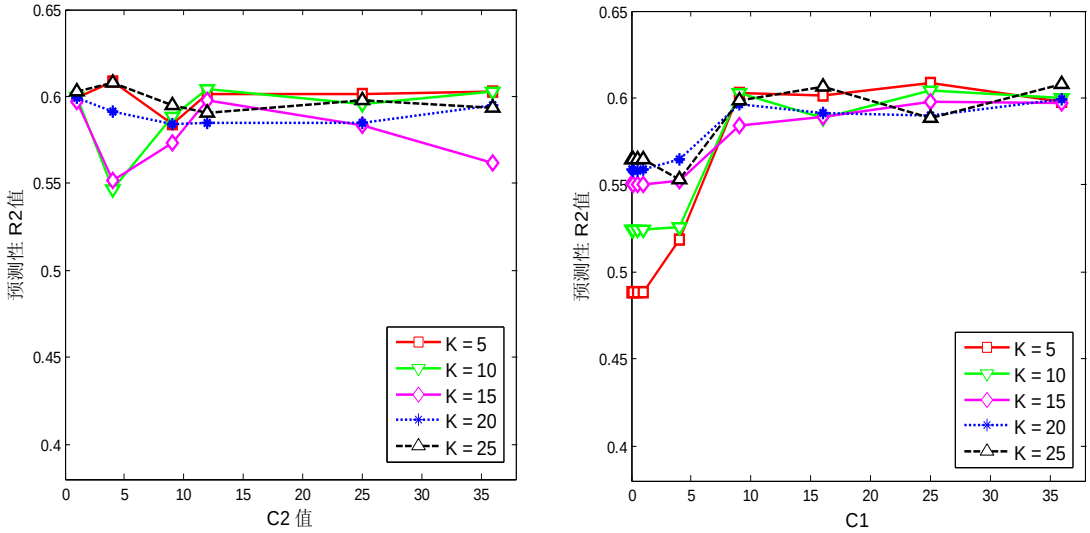
4.5.4　运行时间分析

图4.7展示不同回归分析模型，包括基于无向图的MMH回归模型与TWH回归模型以及基于有向图的回归模型（包括MedLDA和sCTRF[117]）在第一种Hotel Review 数据集回归问题上的时间效率。从图中测试结果可以看出：

（1）　与需要复杂的迭代步骤推理得到隐变量表示的有向图模型MedLDA相比，基于无向图的MMH和TWH模型在训练和测试中时间效率都更加高效；

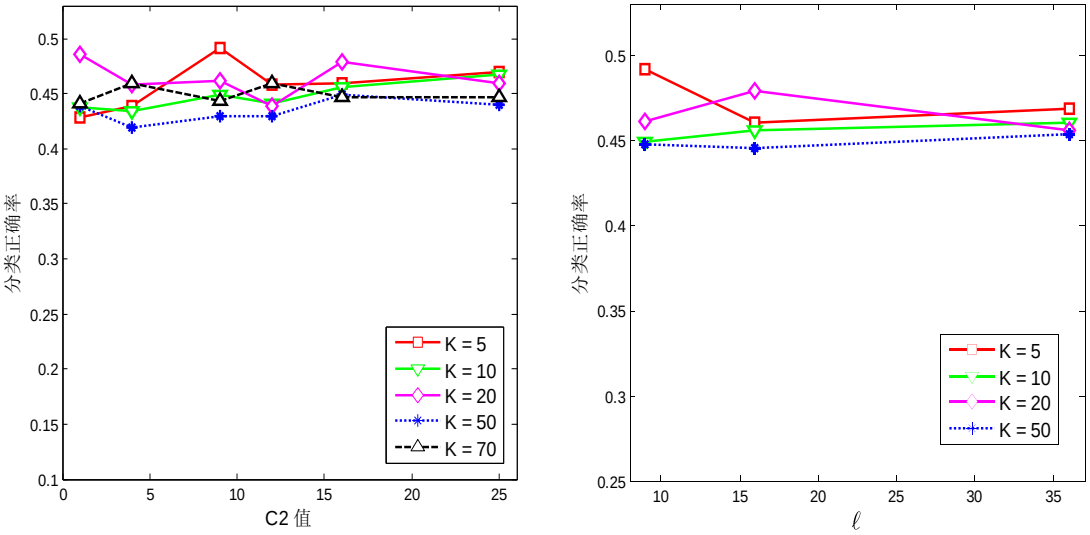
（2）　同样基于无向图的TWH比MMH在测试时间效率上要慢很多倍，其中原因已在第4.4节中说明，主要是因为TWH需要同时推理响应变量的不确定性；

（3）　最后，基于有向图的有结构sCTRF模型比同样基于有向图的无结构MedLDA模型的运行时间慢大约10倍，同时比基于无向图的MMH模型慢10000倍。sCTRF之所以速度慢的原因是sCTRF对每个文档中的每句话都用一个马尔可夫链建模。因此，需花费大量时间进行消息传递（读者可参考文献[117]中详细的介绍）。在训练过程中，MMH与TWH和MedLDA的时间大致相差不多，而仍然比sCTRF（在训练过程中推理仍然复杂）高效。



(a) 参数C_2敏感度分析　　(b) 参数C_1敏感度分析

图 4.8　MMH回归模型的参数敏感度分析。



(a) 参数C_2敏感度分析　　(b) 参数ℓ敏感度分析

图 4.9　有结构输入的MMH分类模型的参数敏感度分析。

4.5.5　参数敏感度分析

最后，分析本章提出的参数化MMH回归模型对于参数C_2、C_1的敏感度。实验中使用grid搜索来选择参数，图4.8(a)、和图4.8(b)展示的是MMH回归模型在Hotel Review数据集上的预测性R2值对于参数C_2、C_1的变化曲线[①]。可以观察发现，C_2和 C_1的变化不会影响模型在两种数据集上的预测性能。当正则化参

① 本实验使用代价函数$\ell_d^{\Delta}(y) = \ell\mathbb{I}(y_d = y)$，其中$\mathbb{I}$为指示函数，$\ell$为大于0的参数。

数$C_2 < 10$时，模型的预测性能有一些振荡，但当$C_2 \geq 10$时，预测性能趋近于稳定。对于参数C_1，当$C_1 \geq 10$时，模型的预测性能趋于稳定。总地来说，敏感度分析中最优的预测结果与图4.5(a)中的最优结果一致。

图4.9(a)和图4.9(b)展示的是有链式结构的MMH分类模型在有段落依赖关系的Hotel Review数据集上的预测性R2值对于参数C_2、ℓ的变化曲线。类似地，可以观察发现，模型在两种数据集上的预测性能对于参数C_2和ℓ的变化并不敏感。对于正则化参数C_2，当$C_2 = 10$时，模型的预测性能得到最优值；对于参数ℓ，当$\ell = 9$时，模型的预测性能最优。总地来说，敏感度分析中最优的预测结果与图4.6中的最优结果一致。

4.6　本章小结

本章提出基于最大间隔的有监督隐层空间回归模型，以及对结构化有段落依赖关系的文本数据建模的隐层空间分类模型。通过在两种Hotel Review真实数据集上的预测结果可以发现，基于最大间隔的方法有助于提高多种隐层空间模型的回归分析性能。此外，考虑输入数据的结构信息也能够进一步提高模型的预测性能。

第 5 章　非参数化马尔可夫网络隐层空间模型

本文第3，4章以典型的多模态数据分析为例，分别介绍了基于最大间隔的参数化隐层空间分类、回归以及有结构的多模态数据分类模型，具体研究了基于无向图的隐层空间统计模型的模型表示和参数学习问题。然而，这些参数化模型存在一个公认难题：即如何确定模型的复杂度，这里具体指隐层空间节点的数目。因为很难事先确定隐层空间模型的隐节点数量，通常需要进行代价很高的模型选择。本章研究基于无向图隐空间模型的非参数化贝叶斯推理方法，该方法可自动确定基于无向图的隐层空间模型的复杂度。

具体地说，本章系统研究非参数化马尔可夫网络隐空间模型。拟在无向图马尔可夫网络中引入非参数贝叶斯方法，以第3，4章提出的多模态隐层空间马尔可夫（以及其特例Harmonium）模型为例，提出无限维的指数族Harmonium（Infinite Exponential Family Harmonium，简称iEFH）模型。在基础理论方面，本章将基于有向图贝叶斯网络的正则化贝叶斯推理理论推广至链图及带有参数的经验贝叶斯推理模型中，在此基本框架下，本章引入最大间隔后验约束，提出最大间隔无限维指数族Harmonium模型，该模型不仅可以避开代价很高的模型选择问题，同时可以学习判别式的隐层空间表示、提高模型的预测性能等。

5.1　研究动机与内容

学习隐变量模型（Latent Variable Models）的一个最常见问题就是如何确定隐变量的数量。最常用的解决办法是使用模型选择（Model Selection），比如交叉验证（Cross-Validation）或者似然比测试（Likelihood Ratio Test[126]）等方法，往往代价很大。近年来，非参数化贝叶斯隐变量模型在统计及机器学习等领域得到了广泛的关注及长足的发展。其中一个重要原因是其“非参数”的优良属性，该性质可让统计模型有效“避开” 代价较高的模型选择过程。例如，通过引入合适的先验分布，在隐类别模型（或混合模型）中自动确定类别（或混合成份）的数量[47,58]；在隐特征模型中自动确定隐特征空间的维度[57,93]等。其中非参数化隐特征模型也是本章的主要研究对象。最常用的非参数化先验分布包括本文第2章曾介绍过的狄利克雷过程（Dirichlet Process，简称DP）先验[47]、印度自助餐过程（Indian Buffet Process，简称IBP）先验[93]、以及定义在函数空间上的高斯过程（Gaussian Process，简称GP）先验[127]等。

但是，如本文第1章1.2.3节所述，标准的非参数化贝叶斯模型一般局限于对观测数据做严格但和实际问题不相符的假设。例如，大部分已有非参数化贝叶斯方法假设观测值为同质（Homogeneous）或可互换（Exchangeable）的。随着智能处理复杂数据的需求不断增加，很多最新的关于非参数化贝叶斯的研究工作试图放宽这些约束，以更好地适应复杂数据处理的要求。若干成功的例子包括：（1）为了处理异质（Heterogenous）的观测数据，文献[51]提出属性依赖的狄利克雷（Predictor-Dependent）随机过程；（2）为了松弛可互换假设，最近的研究工作提出将多种关联结构（Correlation Structures）引入到非参数化随机过程中，其中包括层次结构（Hierarchical Structures）[52,128]，时域/空域依赖（Temporal or Spatial Dependencies）结构[53,129]，以及随机排序依赖（Stochastic Ordering Dependencies）结构[54,55]等。但所有这些方法都是单纯通过设计一些具有特殊结构的非参数化先验分布，通过结合似然模型 ① 间接地影响模型的后验分布。但事实上，后验概率分布才是研究者想要得到的最终对象，它包含描述问题本质的隐含结构信息。因此，有必要研究更加直接地控制后验分布性质的理论与方法。事实上，一个更加直接地影响后验分布的方法是引入后验正则化因子，即在后验分布上加入正则化项，这也正是本章接下来的研究内容。使用后验正则化因子的另一个原因是：在很多情况下，在正则化的框架下引入领域知识更加自然和简单，例如直接在后验分布（而非先验分布）中引入最大间隔约束[40,56] 或者流形约束[59]等。本章将会详细介绍最大间隔的后验约束方法以提高隐空间表示的判别性。

5.1.1　正则化贝叶斯推理

后验正则化，即在隐变量的后验分布中直接引入约束或通过信息投影的方式实现，已被广泛用于从部分可观测数据中学习有限维参数化对数线性模型，代表工作包括广义期望[130]，后验正则化准则[131]，以及交替投影[132]模型等，所有这些模型都使用最大似然估计方法，通过最大化似然函数学习单一的模型参数集合。在学习模型参数的后验分布方面，相关的工作包括从设定的一些度量知识中学习有限维对数线性模型[133]，基于最大间隔准则的最大熵判别式学习[56]以及最大熵判别式话题模型[40]等。但这些方法局限于有限维的参数化模型。据本文作者所知，目前为止，将后验正则化用于非参数贝叶斯隐变量模型的研究仍是一个空白。为了填补上述空白，本文作者参与合作提出了一种通用的满足合适后验

① 非参数化贝叶斯方法主要关注如何设计和使用不同特性的随机过程先验分布。对于似然函数，大部分工作都是默认使用最基本的模型，例如，对于实值的数据使用高斯似然函数，而对于离散输入数据，使用多项式或泊松分布等似然函数。在本章及本文第6章中，也假设似然函数是给定的，不做过多讨论。

约束的非参数贝叶斯推理理论框架[57]，并且特别着重研究以下两种非参数化隐特征模型的具体实例，即无限维隐变量支持向量机（Infinite Latent Support Vector Machines，简称iLSVM）和多任务无限维隐特征支持向量机（Multi-Task Infinite Latent Support Vector Machines，简称MT-iLSVM），两者分别适于多类别分类和多任务学习[134,135]（更多细节请参见附录B。非参数贝叶斯方法和最大间隔学习方法具有各自独特的优点，它们是机器学习两大重要的子领域，并且在过去的二十年里分别取得了显著的成就和发展。但是，长久以来，它们一致被当作两个互相孤立的研究方向。以上述两个统计模型iLSVM和MT-iLSVM以及本章将要介绍的非参数化无向图隐空间模型为例，这些研究成果积极推进了这两大子领域的有机融合。

在技术层面上，对于基于最大似然估计的方法，虽然直觉上可以很自然地在最大似然估计的目标函数中引入一个关于隐变量后验概率分布的正则化项及后验约束，但是，对于贝叶斯方法，却不是显而易见的。其中一个重要原因是经典的贝叶斯推理方法依赖于贝叶斯准则，它并没有一个目标函数可以扩展到包含合适的正则化项。直接在贝叶斯推理过程中考虑后验约束将是非常困难的。首先，对于硬约束，虽然可以使用诸如拒绝采样的方法进行推理，这种方法的效率通常很低，尤其是在高维采样空间中；其次，对于带有软约束的问题，将很难直接采用采样的方法进行近似推理。解决上述问题的突破口是1988年统计学家Zellner教授提出的贝叶斯对偶理论[62]，他将贝叶斯推理的过程等价地描述为一个求最优解的优化问题。在此基础上，可以系统地引入正则化后验约束进行正则化贝叶斯推理（Regularized Bayesian Inference），其中带有一个惩罚项来衡量约束（包括硬约束和软约束）是否被满足。

5.1.2　马尔可夫网络隐空间模型

与基于有向贝叶斯网络的隐层空间模型相似，无向图隐层空间模型中的一个公认难题是：如何自动确定模型的复杂度（对于无向图隐特征模型，例如第3，4章所研究的EFH和MMH模型，这里具体指模型的隐节点数量）。最常用的解决办法是使用模型选择（Model Seletion），如交叉验证、似然比测试等。但是如前所述，这种方法的缺点是通常代价很高。幸运的是，非参数化贝叶斯方法的发展与应用对于避免模型选择带来了新的契机。然而，如上小节所述，绝大多数非参数隐变量模型都是在有向图贝叶斯网络框架下提出的[53,64,65]。至今为止，几乎没有任何工作使用非参数贝叶斯方法解决无向图隐空间马尔可夫网络中的模型选择问题。然而，如本文第3，4章所述，基于无向图的隐层空间马尔可夫网络（如指数

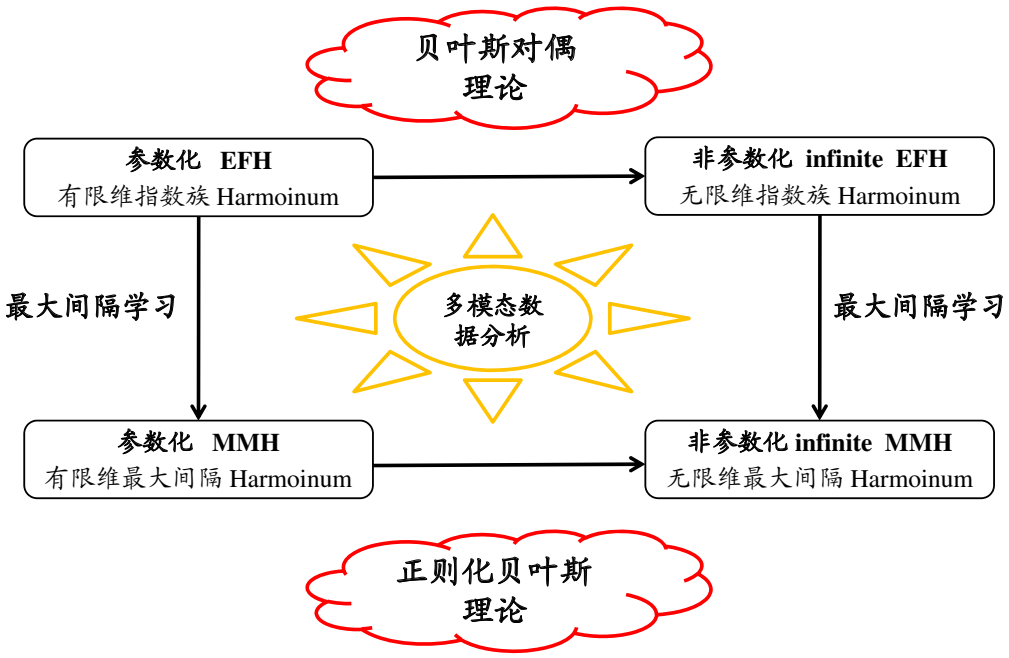


图 5.1　参数化无向图隐空间模型（EFH、MMH）与非参数化无向图隐空间模型（iEFH、iMMH）之间的关系。

族Harmonium模型[14]）是一类重要的隐特征模型，并且与有向贝叶斯网络相比，具有很多互补的优势（例如该模型建立在弱条件独立性假设基础上，所以推理效率高），EFH模型及其各种扩展已被成功用于图像分类、图像检索和图像标注等多种任务中[3,35,36]。

为了推广非参数化贝叶斯方法在无向隐变量模型中的使用，本章提出无限维指数族Harmonium模型（Infinite Exponential Family Harmoniums，简称iEFH）。假设模型中有无限维隐特征，每个特征与一个二值化的指示变量相关联，然后在这些二值变量（即二值矩阵）上假设一个稀疏的印度自助餐过程（IBP）先验[49]，用来选择一个有限维的隐特征子集。由此得到的模型为一个链图模型[66]。根据Zellner[62]教授提出的贝叶斯对偶理论，本章将其扩展到链图和经验贝叶斯推理中。同时本章将正则化贝叶斯推理基本框架推广至链图模型及有经验参数的模型，并且在此理论框架下，通过引入最大间隔约束来正则化隐变量后验分布的性质，将iEFH扩展到有监督无限维最大间隔Harmonium模型（Infinite Max-margin Harmonium，简称iMMH），发现判别式的隐层空间表示，同时进行分类预测。最后，在大量的真实数据集上的实验结果证明，本章提出的方法与其他流行方法相比具有更加优异的预测性能。

参数化无向图隐空间模型（EFH、MMH）与非参数化无向图隐空间模型

（iEFH、iMMH）之间的关系请见图5.1所示。其中多模态数据分析是一个典型的应用场景，是本文各章节在应用方面的一条主线。

5.2　正则化贝叶斯推理及其在无向隐空间模型上的推广

下面首先介绍统计学家Zellner教授在1988年提出的贝叶斯对偶理论[62]，然后基于此理论，介绍带有后验约束的正则化贝叶斯推理的通用理论框架以及其在基于无向马尔可夫网络的隐空间模型中的推广和具体计算模型。

5.2.1　贝叶斯对偶理论

令$\mathbb{M}$表示包含所有随机变量的空间，其后验分布可从经验数据中推理得出；$\mathbf{X}$表示可观测数据；$\pi(\mathcal{M})$表示模型$\mathcal{M} \in \mathbb{M}$的先验分布；$p(\mathbf{x}|\mathcal{M})$表示模型$\mathcal{M} \in \mathbb{M}$的似然函数。当已知一系列观测数据$\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$时，由贝叶斯理论，模型的后验分布为

$$p(\mathcal{M}|\mathcal{D}) = \frac{\pi(\mathcal{M})p(\mathcal{D}|\mathcal{M})}{p(\mathcal{D})} \tag{5-1}$$

其中$p(\mathcal{D})$为观测数据的边缘似然函数。根据最优信息处理准则（Optimal Information Processing Principle），Zellner教授在1988年首先提出[62]：由贝叶斯理论推理得到的模型后验分布和下面的优化问题的最优解是相同的

$$\begin{aligned} &\min_{p(\mathcal{M})} \quad \mathrm{KL}(p(\mathcal{M})||\pi(\mathcal{M})) - \int \log p(\mathcal{D}|\mathcal{M})p(\mathcal{M})d\mathcal{M} \\ &\text{s.t.}: p(\mathcal{M}) \in \mathcal{P}_{\text{prob}}, \end{aligned} \tag{5-2}$$

其中$\mathrm{KL}(p(\mathcal{M})||\pi(\mathcal{M}))$表示后验概率$p(\mathcal{M})$与先验概率$\pi(\mathcal{M})$的KL散度，而$\mathcal{P}_{\text{prob}}$表示一个有效的概率分布空间，在不引起歧义的情况下，本文将忽略该概率分布空间的维度。

5.2.2　有后验约束的正则化贝叶斯推理

上述将贝叶斯推理表示为在对偶空间的优化问题具有重要的意义。例如，统计及物理学家E.T. Jaynes教授在文献[62]中评价上述工作道："贝叶斯对偶理论作为贝叶斯推理的新颖解释，可以使贝叶斯方法变得更有吸引力并得到广泛传播，同时可能激发贝叶斯一般理论的新发展"。在这个贝叶斯对偶理论基础上，本章主要研究如何通过引入正则化的后验约束来提出应用范围更广的正则化贝叶斯推理理论。具体地说，标准贝叶斯推理中的约束（即$p(\mathcal{M}) \in \mathcal{P}_{\text{prob}}$）是对概率分布最基本

的约束，它并不能充分表示研究者通常希望得到的后验分布$p(\mathcal{M})$的性质。为了扩展贝叶斯推理的自由度以及更方便地控制后验分布的性质，本文作者合作提出了正则化贝叶斯推理（ Regularized Bayesian Inference）理论[57]，并将它定义为求解下面有约束的优化问题

$$\begin{aligned}&\min_{p(\mathcal{M}),\boldsymbol{\xi}} \mathrm{KL}(p(\mathcal{M})\|\pi(\mathcal{M})) - \sum_{d=1}^{D}\int \log p(\mathbf{x}_d|\mathcal{M})p(\mathcal{M})d\mathcal{M} + U(\boldsymbol{\xi}) \\ &\text{s.t.}: p(\mathcal{M}) \in \mathcal{P}_{\text{post}}(\boldsymbol{\xi}),\end{aligned} \tag{5-3}$$

其中$\mathcal{P}_{\text{post}}(\boldsymbol{\xi})$表示满足一系列约束的概率分布子空间。辅助参数$\boldsymbol{\xi}$表示非负的松弛变量。为了让该问题具有较好的性质，这里假设$U(\boldsymbol{\xi})$是一个凸函数，下面将会看到，$U(\boldsymbol{\xi})$通常与一个预测准则的代价损失函数（如铰链损失函数）相关联。当后验分布的约束为线性约束时，上述问题为一个凸优化的问题。

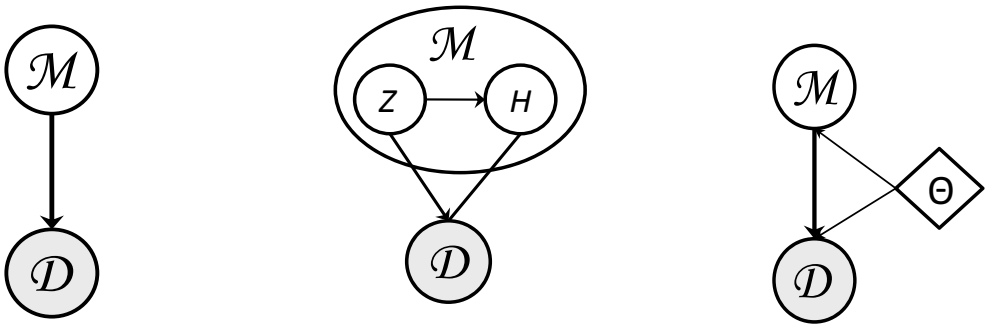
基于上述定义，可以使用基于凸优化理论的迭代过程来求解一般的正则化贝叶斯推理问题。一个常用的方法为拉格朗日法。这里，引入拉格朗日乘子$\boldsymbol{\omega}$。于是迭代求解过程为：（1）固定$\boldsymbol{\omega}$和$\boldsymbol{\xi}$，迭代求解$p(\mathcal{M})$；交替地，（2）固定$p(\mathcal{M})$，求解 $\boldsymbol{\omega}$和$\boldsymbol{\xi}$。对于第一步，可以使用采样或者变分推理的方法[136]近似求解；在某些特定条件下，例如在后验期望中使用约束[130]时，第二步可使用凸优化技术高效地求解。

5.2.3　链图上的贝叶斯对偶理论

上述介绍的贝叶斯对偶理论[62]以及有后验约束的正则化贝叶斯推理理论，它们的基本过程直观上可用图5.2(a)[①] 所示的贝叶斯网络描述，其中$\mathcal{M}$表示包含了所有随机变量的模型，$\mathcal{D}$表示观测数据。模型的联合分布$p(\mathcal{M},\mathcal{D})$可写成先验分布$\boldsymbol{\pi}(\mathcal{M})$与似然函数$p(\mathcal{D}|\mathcal{M})$的乘积形式，即$p(\mathcal{M},\mathcal{D}) = \boldsymbol{\pi}(\mathcal{M})p(\mathcal{D}|\mathcal{M})$。为了将非参数化贝叶斯推理用于解决无向图隐特征模型的模型复杂度问题，本章将贝叶斯对偶理论推广至基于无向图的隐变量模型（如Harmoniums模型），以及模型中除了后验推理还需要估计未知参数Θ的情况。后者也被称为经验贝叶斯方法，该方法经常被研究者采用。

链图模型： 图5.2(b) 所示为一个典型的链图模型，其中模型$\mathcal{M}$包含两个随机变量子集。其中一个子集$\mathbf{H}$通过无向边与观测变量$\mathcal{D}$相关联，而另一个子集$\mathbf{Z}$与$\mathbf{H}$和观测变量$\mathcal{D}$通过有向边相关联。根据链图模型的马尔可夫属性[66]，可

① $\mathcal{M}$的结构可为任意的，如有向图，无向图，或混合链图。



(a) 贝叶斯产生式模型　　(b) 链图　　(c) 经验贝叶斯方法

图 5.2　贝叶斯对偶理论在三种模型上的图示。

以得出它的联合分布有因子乘积的形式

$$p(\mathcal{M},\mathcal{D}) = p(Z)p(H,\mathcal{D}|Z), \tag{5-4}$$

其中，$p(H,\mathcal{D}|Z)$ 为一个马尔可夫随机场。

显然，$p(\mathcal{M},\mathcal{D})$并不能严格地用先验和似然函数写为$\boldsymbol{\pi}(\mathcal{M})p(\mathcal{D}|\mathcal{M})$形式。但仍然可以将贝叶斯对偶理论推广至无向图隐特征模型。其最关键思想来源于如下事实：问题 (5-2)的贝叶斯对偶理论中的目标函数记为$\mathcal{L}_B(p(\mathcal{M}))$实际上为一个KL散度

$$\begin{aligned}\mathcal{L}_B(p(\mathcal{M})) &\overset{\text{def}}{=} \mathrm{KL}(p(\mathcal{M})||\boldsymbol{\pi}(\mathcal{M})) - \int \log p(\mathcal{D}|\mathcal{M})p(\mathcal{M})d\mathcal{M} \\ &= \mathrm{KL}(p(\mathcal{M})||p(\mathcal{M},\mathcal{D})),\end{aligned} \tag{5-5}$$

其中$p(\mathcal{M},\mathcal{D})$ 为联合分布。因此，只要定义了联合分布$p(\mathcal{M},\mathcal{D})$，都可以将贝叶斯推理过程表示为等价的优化问题，为了定义联合分布$p(\mathcal{M},\mathcal{D})$，这里可以选择有向贝叶斯网络或者本章要讨论的无向马尔可夫网络。本章将要介绍的无限维指数族Harmonium模型以及贝叶斯马尔可夫网络[72]都是链图模型的特例。

经验贝叶斯推理： 图5.2(c) 所示为有未知参数Θ的经验贝叶斯模型的图结构。对于该模型，可以联合使用最大似然估计（MLE）以及后验推理对此模型求解

$$\begin{aligned}\min_{\Theta,p(\mathcal{M}|\Theta)}\ & \mathrm{KL}(p(\mathcal{M}|\Theta)||\pi(\mathcal{M})) - \mathbb{E}_{p(\mathcal{M}|\Theta)}[\log p(\mathcal{D}|\mathcal{M},\Theta)] \\ \text{s.t.}:\ & p(\mathcal{M}|\Theta) \in \mathcal{P}_{\text{prob}}(\Theta),\end{aligned} \tag{5-6}$$

命题（5.1）显示，以上问题可以通过一个EM过程求解。

命题 5.1： 问题 (5-6)中的后验分布$p(\mathcal{M}|\Theta)$ 的最优解为贝叶斯定理中Θ为任意值时

推理得到的后验分布; 而最优的Θ^*通过最大似然估计得到结果

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \log p(\mathcal{D}|\Theta).$$

证明 (摘要) 由于KL散度的凸性以及期望的线性，问题 (5-6)是一个凸优化的问题，通过引入拉格朗日乘子，可以推导出上述命题中的结论。 □

5.2.4　链图中的正则化贝叶斯推理

同理，上述介绍的正则化贝叶斯推理（Regularized Bayesian Inference，简称RegBayes）方法也是面向有向图贝叶斯网络的。基于上述链图中的贝叶斯推理框架，本章将RegBayes推广至无向图隐变量模型中。具体地说，基于链图上的贝叶斯对偶理论，含有未知参数的链图正则化贝叶斯推理可写成

$$\begin{aligned} \min_{\Theta, p(\mathcal{M}|\Theta), \boldsymbol{\xi}} \ & \mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta)) + U(\boldsymbol{\xi}) \\ \text{s.t.}: \ & p(\mathcal{M}|\Theta) \in \mathcal{P}_{\text{post}}(\Theta, \boldsymbol{\xi}), \end{aligned} \tag{5-7}$$

其中$\mathcal{P}_{\text{post}}(\Theta, \boldsymbol{\xi})$ 为满足一系列约束的概率分布子空间，而$\mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta))$是问题 (5-6)的目标函数。辅助参数$\boldsymbol{\xi}$通常是非负的松弛变量。同样地，为了让优化问题具有良好的性质，这里假设$U(\boldsymbol{\xi})$ 是一个凸函数，通常与一个预测准则的替代损失函数函数（如铰链损失函数）相关联，将在下文中用到。

以上结论也同样适用于本文第3，4章提出的参数化多模态隐层空间马尔可夫网络分类和回归模型，因此，**本文的所有研究内容都可以统一到正则化贝叶斯推理的基本框架下。这也是本文研究内容系统化的一个重要体现**。

5.3　无限维指数族Harmonium模型

基于上述在链图模型上的贝叶斯推理理论框架，本章以无限维指数族Harmonium（Infinite Exponential Family Harmonium，简称iEFH）模型为典型特例，提出将非参数贝叶斯方法运用到无向概率图模型的基本方法与理论。

5.3.1　有限维Beta-Bernoulli Harmonium模型

为了方便理解，本节首先介绍一种包含二值隐变量的有限维指数族Harmonium模型，它可以从大量但有限个候选隐特征中选择用于描述数据的隐层特征集合。如图5.3所示，为有限维Beta-Bernoulli Harmonium模型的图结构，图中蓝色部分是Beta-Bernoulli先验，红色部分是基本的EFH模型，其中每个

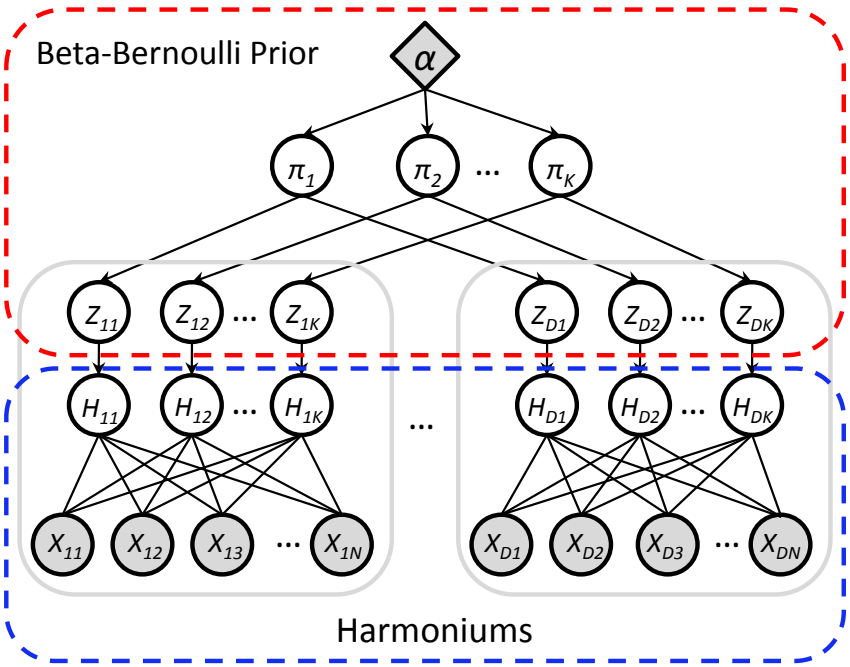


图 5.3　基于Beta-Bernoulli先验的有限维指数族Harmonium模型。

样本对应一个EFH，隐变量$\mathbf{H}_d$表示样本d在隐特征空间的表示。读者将会看到，这个有限模型可以容易地扩展至一个无限维模型。令K表示隐特征数，对每一个样本d，引入二值变量$\mathbf{Z}_d$，每一个Z_{dk}与隐特征H_{dk}相关联。于是，对每一数据d 定义一个实际**有效的**隐层特征为

$$\widetilde{\mathbf{h}}_d = \mathbf{z}_d \circ \mathbf{h}_d$$

其中$\circ$算子表示对两个向量的所有对应元素做乘积。当二值元素$z_{dk} = 1$，数据d包含特征k，否则不包含，h_{dk}为对应特征的值。

为了完善此贝叶斯模型，假设二值变量Z_{dk}服从伯努利分布

$$z_{dk} \sim \text{Bernoulli}(\pi_k), \tag{5-8}$$

而均值参数π_k服从Beta分布

$$\pi_k \sim \text{Beta}(\frac{\alpha}{K}\ 1), \tag{5-9}$$

其中α为超参数。这种Beta-Bernoulli先验有一个很好的属性：矩阵$\mathbf{Z}$中非零元素的期望数量以$N\alpha$为上界[49]。

基于以上定义，带有Beta-Bernoulli先验的有限维EFH模型的联合分布为

$$p(\boldsymbol{\pi}, \{\mathbf{z}_d, \mathbf{h}_d, \mathbf{x}_d\}) = \prod_{k=1}^{K} p(\pi_k) \prod_{d=1}^{D} \Big(p(\mathbf{z}_d|\boldsymbol{\pi}) p(\mathbf{x}_d, \mathbf{h}_d|\mathbf{z}_d) \Big),$$

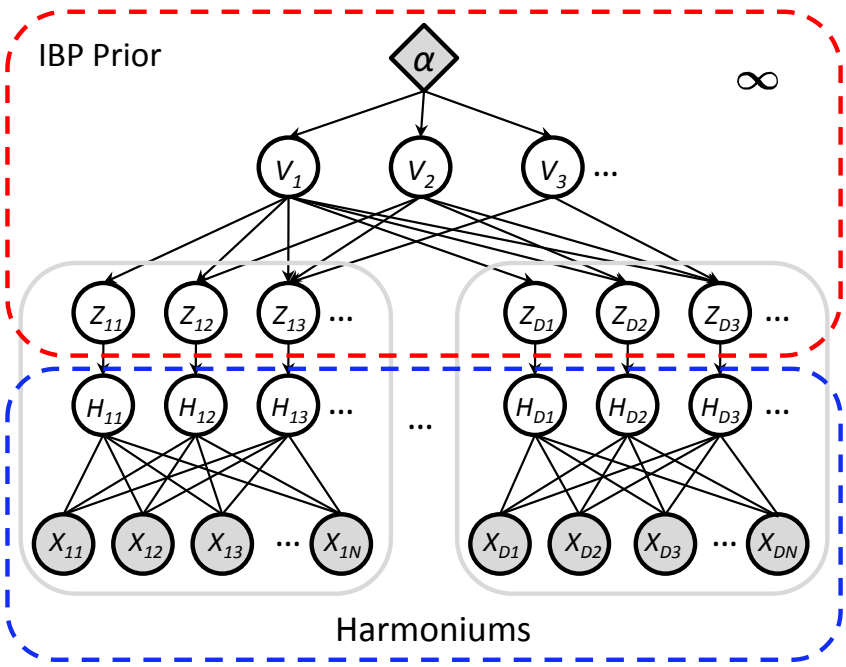


图 5.4　基于IBP stick-breaking 表示先验的无限维指数族Harmonium模型。

其中$p(\mathbf{x}_d, \mathbf{h}_d|\mathbf{z}_d)$与公式 (2-5)形式相同，唯一的区别是公式中的$\mathbf{h}_d$被替换为$\widetilde{\mathbf{h}}_d$，但与$\mathbf{h}_d$相关的平方项除外。

值得一提的是，无论是Beta-Bernoulli模型，还是下面即将介绍的无限维指数族Harmonium模型，都可退化成只包含二值隐变量的情况，只需要将$\mathbf{H}$的取值固定为恒值1。因此，这些方法都包含了受限波尔兹曼机。本章介绍的方法同样可以用于自动确定受限波尔兹曼机模型的隐特征的数量。

5.3.2　无限维指数族Harmonium模型

下面，在有限维Beta-Bernoulli 指数族Harmonium模型中令$K \to \infty$，即可将其推广至无限维指数族Harmonium（简称iEFH）。根据文献[49]中所述，当使用lof-等价类矩阵时，有限维Beta-Bernoulli先验可推广至无限维，此时得到的$\mathbf{Z}$的边缘分布即为大家熟知的印度自助餐过程（IBP）先验分布。图5.4所示为iEFH模型的图结构。这里使用IBP的Stick-Breaking表示，引入中间变量$\mathbf{V} = \{V_1, V_2, \cdots\}$。第2章已经介绍过，为方便读者，这里再次简单描述一下。$\mathbf{Z}$的IBP先验可以表示为一个产生式模型，即$\forall d \geq 1$, $z_{dk} \sim \text{Bernoulli}(\pi_k)$，而$\boldsymbol{\pi} = \{\pi_1, \pi_2, \cdots\}$从一个Stick-breaking过程[96]中产生出来，即$\pi_1 = v_1$, $\pi_k = v_k \pi_{k-1} = \prod_{i=1}^{k} v_i$，其中$v_i \sim \text{Beta}(\alpha, 1)$。在iEFH模型中，每一个样本都有无限维隐特征，即隐特征矩阵$\mathbf{Z}$有无数列。

如图5.4所示，iEFH模型可被理解为由下面两部分组成：（1）用于表示观测值

的“似然模型”；（2）从无限维候选特征中选出有限维子集的“ 选择模型”。根据链图模型[66]的马尔可夫属性，iEFH的联合分布表示为

$$p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\}) = p(\mathbf{v}) \prod_d (p(\mathbf{z}_d|\mathbf{v}) p(\mathbf{x}_d, \mathbf{h}_d|\mathbf{z}_d)), \tag{5-10}$$

其中$p(\mathbf{x}_d, \mathbf{h}_d|\mathbf{z}_d)$为用二值变量$\mathbf{z}_d$选择有效隐特征的指数族Harmonium模型，而$p(\mathbf{x}_d, \mathbf{h}_d|\mathbf{z}_d)$与公式 (2-5)有相同的形式，将公式 (2-5)的指数项中除$\mathbf{h}_d$的平方项以外的包含$\mathbf{h}_d$的项都用$\widetilde{\mathbf{h}}_d$代替。注意：公式 (5-10)中忽略了变量$\boldsymbol{\pi}$，这是因为它是$\mathbf{v}$的确定性函数。

根据上述定义，模型的后验推理以及估计参数Θ的过程可以表述为求解优化问题 (5-6)，其中$\mathcal{M} = \{\mathbf{V}, \mathbf{H}, \mathbf{Z}\}$以及 $\Theta = \{\alpha, \boldsymbol{\beta}, \mathbf{W}, \sigma_{d1}, \sigma_{d2}\}$。由公式 (5-5)的结果，模型的目标函数为

$$\mathcal{L}_B(p(\mathcal{M}|\Theta), \Theta) = \mathrm{KL}(p(\mathcal{M}|\Theta)||p(\mathcal{M}, \mathcal{D}|\Theta)).$$

下面介绍无限维iEFH模型的优化方法，对于有限维Beta-Bernoulli模型，后验推理及参数估计可以同理得到。

5.3.2.1　模型优化

由于所有的输入数据（每个样本对应一个EFH）使用同一个先验分布$p(\mathbf{Z})$，它们是相关联在一起的（如图5.4所示）。因此，iEFH模型中的训练和推理过程比标准的EFH模型更加复杂。与第3，4章相同，这里使用 Contrastive Divergence的变分推理方法。具体地说，用下面$\mathcal{L}(\Theta, q_0, q_1)$近似$\mathcal{L}_B(p(\mathcal{M}|\Theta), \Theta)$

$$\begin{aligned}\mathcal{L}(\Theta, q_0, q_1) \stackrel{\text{def}}{=} & \mathrm{KL}(q_0(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})||p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})) \\ & - \mathrm{KL}(q_1(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})||p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})),\end{aligned} \tag{5-11}$$

这里，变分分布q_0中的$\mathbf{x}$是可观测的，在推理过程中保持不变。只需推理$q_0(\mathbf{h})$；而q_1中所有变量都是未知的。为了简化计算，进一步限制q（表示q_0 或q_1）满足截断结构化均值场假设

$$q(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\}) = \prod_{k=1}^{T} q(v_k) \prod_{d=1}^{D} (q(\mathbf{x}_d) q(\mathbf{h}_d) q(\mathbf{z}_d)), \tag{5-12}$$

其中$q(v_k) = \mathrm{Beta}(v_k; \gamma_{k1}, \gamma_{k2})$，$T$为截断上界。基于这些近似，可以高效地评估目标函数，用如下所示的迭代方法进行模型推理。

推理: 对于变量$\mathbf{H}$, $\mathbf{X}$和$\mathbf{Z}$，其均值场迭代公式为$q(\mathbf{x}_d) = \prod_{n=1}^{N} q(x_{dn})$, $q(\mathbf{h}_d) = \prod_{k=1}^{T} q(h_{dk})$ 以及$q(\mathbf{z}_d) = \prod_{k=1}^{T} q(z_{dk})$，其中

$$
\begin{aligned}
q(x_{dn}) &= \mathcal{N}(x_{dn}; \mathbb{E}_q[x_{dn}], \sigma_{d1}^2) \\
q(h_{dk}) &= \mathcal{N}(h_{dk}; \mathbb{E}_q[h_{dk}], \sigma_{d2}^2) \\
q(z_{dk}) &= \text{Bernoulli}(z_{dk}; \nu_{dk}),
\end{aligned}
$$

这里，$\mathbb{E}_q[x_{dn}] = \sigma_{d1}^2(\alpha_n + \mathbf{W}_{n.}(\boldsymbol{\nu}_d \circ \mathbb{E}_q[\mathbf{h}_d]))$; $\mathbb{E}_q[h_{dk}] = \sigma_{d2}^2(\nu_{dk}\beta_k + \nu_{dk}\mathbb{E}_q[\mathbf{x}_d]^\top \mathbf{W}_{.k})$；其中$\mathbf{W}_{n.}$ ($\mathbf{W}_{.k}$) 分别表示$\mathbf{W}$矩阵中的第n行（第k列）。

ν的均值场迭代公式为

$$
\nu_{dk} = \frac{1}{1 + \exp\{\tau_1^k - \tau_2^k - \mathbb{E}_q[h_{dk}](\beta_k + \mathbb{E}_q[\mathbf{x}_d]^\top \mathbf{W}_{.k})\}},
$$

其中$\tau_1^k = \mathbb{E}_q[\log(1 - \prod_{j=1}^{k} v_j)]$, $\tau_2^k = \sum_{j=1}^{k} \mathbb{E}_q[\log v_j]$。这里使用digamma函数$\psi$即$\mathbb{E}_q[\log v_k] = \psi(\gamma_{k1}) - \psi(\gamma_{k1} + \gamma_{k2})$计算$\tau_2^k$。对于$\tau_1^k$，当$k > 1$，可以使用多项式上界[136]来近似$\tau_1^k$，文献[136]中已证明，这是一种高效的近似方法。虽然比用泰勒展开式得到的近似弱一些。当$k = 1$时，有$\mathbb{E}_q[\log(1 - \prod_{j=1}^{k} v_j)] = \psi(\gamma_{12}) - \psi(\gamma_{11} + \gamma_{12})$。$\boldsymbol{\gamma}$ 的迭代公式与文献[136]中相同。

参数估计: 使用Contrastive Divergence近似推理的方法得到q_0和q_1之后，使用梯度下降（如拟牛顿方法[121]），通过优化目标函数$\mathcal{L}(\Theta, q_0, q_1)$来学习模型参数。令$\Delta\mathbb{E}[\cdot] \stackrel{\text{def}}{=} \mathbb{E}_{q_1}[\cdot] - \mathbb{E}_{q_0}[\cdot]$，参数$(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$的梯度公式为

$$
\begin{aligned}
\nabla_{\alpha_n}\mathcal{L} &= \sum_d \Delta\mathbb{E}[x_{dn}], \\
\nabla_{\beta_k}\mathcal{L} &= \sum_d \Delta\mathbb{E}[\widetilde{h}_{dk}], \\
\nabla_{\mathbf{W}_{nk}}\mathcal{L} &= \sum_d \Delta\mathbb{E}[x_{dn}\widetilde{h}_{dk}]
\end{aligned} \tag{5-13}
$$

对于取值为正数的变分参数σ_{d1}^2和σ_{d2}^2，为了避开正数的约束，可以在对数空间计算梯度。令$t_{d1} = \log\sigma_{d1}^2$，$t_{d2} = \log\sigma_{d2}^2$。于是，梯度公式分别为

$$
\begin{aligned}
\nabla_{t_{d1}}\mathcal{L} &= \frac{1}{2\sigma_{d1}^2}\Delta\mathbb{E}[\mathbf{x}_d^\top \mathbf{x}_d], \\
\nabla_{t_{d2}}\mathcal{L} &= \frac{1}{2\sigma_{d2}^2}\Delta\mathbb{E}[\mathbf{h}_d^\top \mathbf{h}_d],
\end{aligned}
$$

对于输入数据$\mathbf{x}$，有$\mathbb{E}_{q_0}[x_{dn}] = x_{dn}$。

5.4　无限维最大间隔Harmonium模型

无限维iEFH模型是一种无监督的非参数化隐特征模型。如第3，4章所述，在很多实际任务中，研究者期望学习更适于具体任务（如预测问题）的隐层空间表示。例如，当使用学习得到的隐特征进行文本、图像分类时，通常希望这些隐特征对于不同的类别尽可能的具有判别性（或可区分性）。但是，无监督的iEFH模型由于没有考虑有监督信息，并不能提供一个区分多种类别的明确的机制，所以它推理得到的隐特征表示对于分类任务来说通常并不是最理想（最优）的。理论上说，许多方法都可用于考虑有监督信息。例如，可以通过定义已知输入变量$\mathbf{X}$时，响应变量Y的条件概率分布，进而采用最大化条件似然估计的方法进行参数学习[35]。但是，如文献[36]中以及本文在第3，4章所述，这种最大似然估计方法会导致模型在预测性能和发现判别式的隐特征表示方面得到不理想的结果。

借鉴本文第3，4章的研究成果，本小节介绍无限维最大间隔Harmonium（Infinite Max-margin Harmonium，简称iMMH）模型，此模型将iEFH模型扩展到可以考虑有监督信息，来发现判别式的隐层空间表示。于是问题的难点即为如何将最大间隔预测准则引入非参数化的无向图隐层空间模型，因为非参数化贝叶斯方法和最大间隔学习一直以来被认为是机器学习研究中互相分离的子领域。事实上，本章接下来将要介绍的iMMH模型可理解为是一种将这两种方法用于无向图隐特征模型中的最新成功尝试。在论文[57]中，本文作者已合作提出将最大间隔学习与非参数化贝叶斯方法在正则化贝叶斯推理框架下于有向图贝叶斯隐变量模型中巧妙结合。而本章研究的iMMH是正则化贝叶斯推理理论方法在无向图隐变量模型中一个重要的推广。

5.4.1　用于分类任务的无限维最大间隔Harmonium模型

基于前文所述的链图中的正则化贝叶斯推理框架，下面着重介绍用于多类别分类任务的无限维最大间隔Harmonium（iMMH）模型，其中响应变量Y取值为一个有限集合$\mathcal{Y} = \{1, 2, \cdots, L\}$。二值分类问题可以通过相似的推理过程得到结果。

为了建立分类器，下面使用有效的隐特征$\widetilde{\mathbf{h}}_d = \mathbf{z}_d \circ \mathbf{h}_d$作为数据$d$的隐特征表示[①]。当$\mathbf{H}$和$\mathbf{Z}$已知时，定义线性判别函数为

$$F(y, \mathbf{h}, \mathbf{z}; \mathbf{x}, \Phi) \stackrel{\text{def}}{=} \Phi_y^{\top} \widetilde{\mathbf{h}} = \Phi^{\top} \mathbf{f}(y, \widetilde{\mathbf{h}}), \tag{5-14}$$

其中$\mathbf{f}(y, \widetilde{\mathbf{h}})$是由$L$个子向量拼接而成的向量，其中第$y$个子向量是$\widetilde{\mathbf{h}}$，其他的子向量

① 观测特征可被连接到$\widetilde{\mathbf{h}}_d$之后。

都是0；Φ 是一个由L个子向量Φ_y拼接而成的权值向量。为了去除隐含特征的不确定性，和前面章节同样使用线性的期望算子，定义有效判别式函数为

$$\begin{aligned} F(y;\mathbf{x}) &\stackrel{\text{def}}{=} \mathbb{E}_{p(\mathbf{H},\mathbf{Z},\Phi)}[F(y,\mathbf{h},\mathbf{z};\mathbf{x},\Phi)] \\ &= \mathbb{E}_{p(\mathbf{H},\mathbf{Z},\Phi)}[\Phi^\top \mathbf{f}(y,\widetilde{\mathbf{h}})]. \end{aligned} \tag{5-15}$$

线性期望算子的优点本文已经在前面章节阐述。根据上述定义，模型的预测准则可以很自然地写为

$$y^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{y\in\mathcal{Y}} F(y;\mathbf{x}). \tag{5-16}$$

令$\mathcal{D} = \{(\mathbf{x}_d, y_d)\}_{d=1}^D$ 表示一个有标签的训练数据集。使用上述定义，可以在RegBayes框架下定义iMMH模型，即求解问题 (5-7)，令$U(\boldsymbol{\xi}) = C_1 \sum_d \xi_d$ 以及

$$\mathcal{P}_{\text{post}}(\Theta,\boldsymbol{\xi}) = \left\{ p(\mathbf{H},\mathbf{Z},\Phi|\Theta) \left| \begin{array}{c} \forall d : F_d^\Delta(y) \geq \ell_d^\Delta(y) - \xi_d, \forall y \\ \xi_d \geq 0 \end{array} \right. \right\},$$

其中

$$F_d^\Delta(y) = F(y_d;\mathbf{x}_d) - F(y;\mathbf{x}_d)$$

为真实标签y_d与任意类别标签y之间的期望间隔，而$\ell_d^\Delta(y)$ 为预测类别标签为y时的代价函数。可以看出，在公式 (5-7)的有约束子空间中，最小化$U(\boldsymbol{\xi}) = C_1 \sum_d \xi_d$ 与最小化预测准则 (5-16)的铰链损失函数

$$\mathcal{R}_h(p(\mathbf{H},\mathbf{Z},\Phi|\Theta),\Theta) = C_1 \sum_d \max_y \{\ell_d^\Delta(y) - \Delta F(y;\mathbf{x}_d)\}$$

是相等价的。

为了完善模型，下面需要定义模型的联合概率分布$p(\Phi,\mathbf{v},\{\mathbf{h}_d,\mathbf{z}_d\})$。为方便起见，假设$\Phi$与其他的变量独立，即

$$p(\Phi,\mathbf{v},\{\mathbf{h}_d,\mathbf{z}_d\}) = p_0(\Phi)p(\mathbf{v},\{\mathbf{h}_d,\mathbf{z}_d\}),$$

其中$p(\mathbf{v},\{\mathbf{h}_d,\mathbf{z}_d\})$与公式 (5-10)中相同。最常用的$p_0(\Phi)$选择是一个零均值各方向方差相同的正态分布$\mathcal{N}(0,\sigma_0^2 I)$，当然也可使用如拉普拉斯（Laplace）分布[137] 等其他先验分布。

5.4.2　用于回归任务的无限维最大间隔Harmonium模型

同理，基于前文所述的链图中的正则化贝叶斯推理框架，下面介绍用于回归

分析任务的无限维最大间隔Harmonium（iMMH）模型，其中响应变量Y的取值为实数集合或者其特定子集，如在一个特定范围的实数集合等。

为了建立基于最大间隔准则的回归分析模型，这里采用和第4章类似的思想。同理，下面使用有效的隐特征$\widetilde{\mathbf{h}}_d = \mathbf{z}_d \circ \mathbf{h}_d$作为数据$d$的隐特征表示[①]。当$\mathbf{H}$和$\mathbf{Z}$已知时，定义线性回归分析模型为

$$F(\mathbf{h}, \mathbf{z}; \mathbf{x}, \Phi) \stackrel{\text{def}}{=} \Phi^{\top} \widetilde{\mathbf{h}}, \tag{5-17}$$

其中Φ是权值向量。为了去除隐含特征的不确定性，和前面章节同样使用线性的期望算子，定义回归分析的预测函数为

$$y^* \stackrel{\text{def}}{=} \mathbb{E}_{p(\mathbf{H},\mathbf{Z},\Phi)}[F(\mathbf{h}, \mathbf{z}; \mathbf{x}, \Phi)], \tag{5-18}$$

其中，线性期望算子的优点本文已经在前面章节阐述。

令$\mathcal{D} = \{(\mathbf{x}_d, y_d)\}_{d=1}^{D}$表示一个有标签的训练数据集。使用上述定义以及结合第4章介绍的基于最大间隔准则的ϵ-不敏感损失函数，可以在RegBayes框架下定义iMMH回归分析模型，即求解问题 (5-7)，令$U(\boldsymbol{\xi}) = C_1 \sum_d (\xi_d + \xi_d^*)$ 以及

$$\mathcal{P}_{\text{post}}(\Theta, \boldsymbol{\xi}) = \left\{ p(\mathbf{H}, \mathbf{Z}, \Phi | \Theta) \left| \begin{array}{l} \forall d : y_d - \mathbb{E}_{p(\mathbf{H},\mathbf{Z},\Phi)}[F(\mathbf{h}_d, \mathbf{z}_d; \mathbf{x}_d, \Phi)] \leq \epsilon - \xi_d, \forall y \\ \quad -y_d + \mathbb{E}_{p(\mathbf{H},\mathbf{Z},\Phi)}[F(\mathbf{h}_d, \mathbf{z}_d; \mathbf{x}_d, \Phi)] \leq \epsilon - \xi_d^*, \forall y \\ \quad \xi_d \geq 0 \end{array} \right. \right\}.$$

可以看出，在公式 (5-7)的有约束子空间中，最小化$U(\boldsymbol{\xi}) = C_1 \sum_d (\xi_d + \xi_d^*)$与最小化预测准则 (5-16)的铰链损失函数

$$\mathcal{R}_{\epsilon}(p(\mathbf{H}, \mathbf{Z}, \Phi | \Theta), \Theta) = C_1 \sum_d \max(0, |y_d - \mathbb{E}_{p(\mathbf{H},\mathbf{Z},\Phi)}[F(\mathbf{h}_d, \mathbf{z}_d; \mathbf{x}_d, \Phi)]|)$$

是相等价的。

同理，为了完善模型，需要定义模型的联合概率分布$p(\Phi, \mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\})$，这里可以采用和iMMH分类模型中相同的定义方式。

5.4.2.1　模型优化

下面以iMMH分类模型为例，介绍模型学习与推理的算法，对于iMMH回归分析模型，类似的推理算法可以结合第4章内容推导出来，这里不再赘述。具体地说，与iEFH模型相似，这里可以使用近似推理方法实现模型的优化。首先，使用

① 观测特征可被连接到$\widetilde{\mathbf{h}}_d$之后。

结构化均值场假设，即

$$p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\}, \Phi|\Theta) = p(\Phi)p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\}|\Theta).$$

于是，目标函数可以写为

$$\mathcal{L}'_B(\Theta, p(\mathcal{M}|\Theta)) = \mathrm{KL}(p(\Phi)||p_0(\Phi)) + \mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta)).$$

目标函数中的第二项仍然很难计算，需要继续使用近似推理方法（如Contrastive Divergence）近似模型目标函数 $\mathcal{L}'_B$

$$\mathcal{L}'_B(\Theta, q(\mathcal{M}|\Theta)) \approx \mathrm{KL}(p(\Phi)||p_0(\Phi)) + \mathcal{L}(\Theta, q_0, q_1), \tag{5-19}$$

其中$\mathcal{L}(\Theta, q_0, q_1)$是iEFH模型中的Contrastive Divergence近似目标函数，已在公式（5-11）中介绍过。

于是，这里采用一种交替的迭代方法推理得到(q_0, q_1)，并估计$(\Theta, p(\Phi))$。由于推理与iEFH模型相同，这里省略推理步骤。模型的参数估计包括在满足后验约束的条件下，最小化近似目标函数，即使用坐标下降方法。

对于$p(\Phi)$，求解下面的子问题

$$\begin{aligned} \min_{p(\Phi),\boldsymbol{\xi}} \quad & \mathrm{KL}(p(\Phi)||p_0(\Phi)) + C_1 \sum_d \xi_d \\ \forall d, y, \text{ s.t.}: \quad & \mathbb{E}_{p(\Phi)}[\Phi]^\top \mathbf{f}_d^\Delta(y, \mathbb{E}[\widetilde{\mathbf{h}}_d]) \geq \ell_d^\Delta(y) - \xi_d, \end{aligned}$$

其中$\mathbf{f}_d^\Delta(y, \mathbb{E}[\widetilde{\mathbf{h}}_d]) = \mathbf{f}(y_d, \mathbb{E}[\widetilde{\mathbf{h}}_d]) - \mathbf{f}(y, \mathbb{E}[\widetilde{\mathbf{h}}_d])$。这里使用各方向方差相同的高斯先验$p_0(\Phi) = \mathcal{N}(0, \sigma_0^2 I)$，最优解为$p(\Phi) = \mathcal{N}(\boldsymbol{\mu}, \sigma_0^2 I)$，其中均值$\boldsymbol{\mu}$可以通过求解一个SVM原问题得到

$$\begin{aligned} \min_{\boldsymbol{\mu},\boldsymbol{\xi}} \quad & \frac{1}{2\sigma_0^2}\|\boldsymbol{\mu}\|_2^2 + C_1 \sum_d \xi_d \\ \forall d, y, \text{ s.t.}: \quad & \boldsymbol{\mu}^\top \mathbf{f}_d^\Delta(y, \mathbb{E}[\widetilde{\mathbf{h}}_d]) \geq \ell_d^\Delta(y) - \xi_d. \end{aligned}$$

对于Θ，可以使用等价的包含$\mathcal{R}_h$的无约束目标函数，根据次梯度下降方法和点最大函数（Pointwise Maximum Function）属性，得到次梯度公式

$$\begin{aligned} \partial_{\beta_k} f &= \nabla_{\beta_k} \mathcal{L} - C_1 \sum_d (\boldsymbol{\mu}_{y_d} - \boldsymbol{\mu}_{\bar{y}_d}) \zeta_{dk} \\ \partial_{\mathbf{W}_{nk}} f &= \nabla_{\mathbf{W}_{nk}} \mathcal{L} - C_1 \sum_d (\boldsymbol{\mu}_{y_d} - \boldsymbol{\mu}_{\bar{y}_d}) \zeta_{dk} x_{dn} \end{aligned}$$

其中$\bar{y}_d = \mathrm{argmax}_y(\ell_d^\Delta(y) + \boldsymbol{\mu}^\top \mathbf{f}(y, \mathbb{E}[\widetilde{\mathbf{h}}_d]))$ 为考虑代价损失情况下的预测类别；而当$\mathbf{H}$

是常量时（如在二值iMMH中），$\zeta_{dk} = \nu_{dk}(1-\nu_{dk})$；否则，$\zeta_{dk} = \nu_{dk}(1-\nu_{dk})\mathbb{E}[h_{dk}] + \nu_{dk}^2$。

值得一提的是，与无监督模型中的梯度公式（5-13）相比较可以发现，上面的次梯度公式中有额外的一项（即公式中的最后一项）。这一项的作用是：当估计的类别标签$\bar{y}_d$与真实的类别标签y_d不同时，这个偏差项将使模型朝着更易于得到判别式隐层空间表示的方向改进。

5.5　实验结果与分析

下面展示模型在多种数据集（包括TRECVID 2003，13-animal Flickr 及 20 Newsgroups数据集）分类问题上的实验结果。

5.5.1　文本分类

下面首先展示iEFH和iMMH模型在20 Newsgroups数据集上的分类结果。20 Newsgroups数据集① 被广泛地用于文本分类问题，共包含大约20000个文本，分属20种不同类别。为了与判别式受限波尔兹曼机（Discriminative Restricted Boltzmann Machine，简称DRBM）[33]模型比较，本实验使用与文献[33] 相同的数据处理方式，即选取出现频率最高的5000个词作为词典，并将整个数据集平均分为训练集和测试集。实验中将本章提出的无限维无监督iEFH模型、无限维有监督iMMH模型，分别与第3章研究的有限维无监督EFH和有限维有监督MMH模型[36]比较，后两种模型需要“模型选择”确定隐变量数量。训练基于最大间隔的iMMH和MMH模型时，本章使用多类别分类器SVM（康奈尔大学开发的工具包② ）学习模型参数Φ。对于无监督的iEFH和EFH模型，在学习隐层空间表示基础上，训练一个多类别的支持向量机分类器实现分类任务。

5.5.1.1　预测性能

表5.1所示为支持向量机（SVM），神经元网络（Neural Networks，表中简称NNet），产生式受限波尔兹曼机（GRBM）（通过最大化输入变量$\mathbf{X}$和响应变量Y的联合似然函数来学习模型参数），RBM+NNet（使用无监督RBM来初始化含有一层隐变量的神经网络）以及文献[33]中的判别式受限波尔兹曼机

① http://people.csail.mit.edu/jrennie/20Newsgroups/
② http://svmlight.joachims.org/svm multiclass.html

表 5.1　在20 Newsgroups 数据集上的分类错误率。

模型	分类错误率
SVM	0.328 ± 0.0000
NNet	0.282 ± 0.0000
GRBM (K=1000)	**0.249** ± 0.0000
DRBM (K=50)	0.276 ± 0.0000
RBM+NNet	0.268 ± 0.0000
EFH+SVM (K=50)	0.375 ± 0.0000
EFH+SVM (K=180)	0.304 ± 0.0000
MMH (K=50)	0.257 ± 0.0000
MMH (K=180)	**0.251** ± 0.0000
iEFH+SVM	0.283 ± 0.0022
iMMH	**0.252** ± 0.0029

（DRBM）（通过最大化在给定输入$\mathbf{X}$时，响应变量Y的条件似然来学习模型参数）模型的文本分类结果。可以得到下面的观察结果：

（1） 当隐层变量数相同时，基于最大间隔的MMH比基于最大似然估计的DRBM即GRBM方法分类结果好；

（2） 与使用更多隐层节点数（如表中的“1000”）的产生式GRBM模型相比，即使MMH的隐层节点数更少（如表中所示“$K = 50$”），判别式的MMH模型仍然很有竞争力，可以得到与使用1000个隐节点的GRBM不相上下的分类结果；

（3） 通过使用非参数贝叶斯方法，iEFH和iMMH模型可以避开模型选择问题，同时预测性能没有受到很大影响。对于iEFH，无限维iEFH模型的分类性能甚至比有限EFH模型预测效果更好；

（4） 通过引入有监督信息，有监督模型（如MMH和iMMH模型）与无监督方法（EFH和iEFH模型）相比，学习得到的隐空间表示更加具有可分性，也更能提高模型的预测性能。

5.5.1.2　判别性隐层空间表示

为了更直观地展示隐层特征的语义表示，表5.2 所示为iMMH模型得到的隐层特征对应的高频词列表。对每一维隐特征，将所有文档数据按照该特征期望值排序，然后选出在前15个文档中出现频率最高的词显示在表5.2中。总体来看，自

表 5.2　20 Newsgroups数据分类问题中不同隐层特征对应的高频词。

所属类别	iMMH模型得到的隐层特征
religion.	F3: god, people, jesus, life, christian, christ, christians, hell
christian	F73: sandvik, christians, god, people, law, christian
alt.	F83: god, atheism, quadra, mac, problem, strong, belief, atheists
atheism	F92: god, atheism, exist, belief, atheists, people, strong, islam
comp.	F35: work, graphics, windows, information, cylinder
graphics	F97: graphics, windows, linux, image, file, gif, find, program, ftp
sports.	F79: ca, team, hockey, game, play, year, montreal, cup, playoffs
hockey	F85: team, gm, murray, win, good, hockey
sports.	F55: year, runs, team, pitching, games, game, baseball, season
baseball	F81: team, apr, game, games, baseball, series, people, win
misc.	F53: sale, computer, price, drive, things, forsale, pc, misc
forsale	F71: mark, sale, optilink, file, case, email, price, shipping
sci.	F40: orbit, moon, sun, lunar, thing, years, made, space, earth
space	F51: nasa, henry, orbit, comet, baalke, moon, kelvin, space, earth

动学习得到的隐特征对应的词有明显的语义含义。例如，特征F3和特征F37对应的高频词中包括“god”, “jesus”, “christian”等，它们与类别“religion.christian”有很强的相关性，而特征F40和F51对应的高频词包括“orbit”, “moon”, “space”，它们与类别“sci.space”更相关。对于所有其他特征，实验中基本都可以将其语义信息对应到相应的类别中。

5.5.2　图像分类

下面分析iEFH和iMMH模型在两个真实图像数据集（包括TRECVID 2003数据集[35]和13-class animal Flickr图像数据集[36]）上的实验结果。本文在第3章中曾使用两数据集的多模态特征用于多模态分析，本章实验中，同样使用多模态特征用于数据分析。使用两种关于iEFH和iMMH模型的不同配置：第一种配置是，考虑隐层特征为二值的情况（即固定所有连续变量$\mathbf{H}$为恒值1），此时模型记为iEFH和iMMH。第二种配置是，隐层特征不是二值而是实数值，即需要自动推理得到$\mathbf{H}$值，此时模型记为iEFH^r 和 iMMH^r。另外，本章展示不同模态特征的实验结果，即在两个数据集上只使用一种特征（记为单模态特征），以及同时使用两种特征（记为多模态特征）的实验结果。对于单模态模型，实验中使用TRECVID数据集中的文本特征，以及Flickr数据集中的颜色特征，因为这些特征相对而言具有更强的区分性。

表 5.3　不同模型在Trecvid 2003及Flickr image数据集上的分类正确率和F1-score值。

	Trecvid 2003 (Text)		13-animal Flickr (Color)	
模型	正确率	F1值	正确率	F1值
EFH+SVM	0.6320 ± 0.0056	0.4917 ± 0.0062	0.5207 ± 0.0038	0.4959 ± 0.0056
iEFH+SVM	0.5919 ± 0.0105	0.4216 ± 0.0195	0.5170 ± 0.0095	0.4893 ± 0.0117
iEFH^r+SVM	**0.6355** ± 0.0044	**0.5109** ± 0.0143	**0.5320** ± 0.0135	**0.5058** ± 0.0119
MMH	0.6396 ± 0.0035	0.5091 ± 0.0113	0.5342 ± 0.0034	0.5121 ± 0.0031
iMMH	**0.6443** ± 0.0130	0.5170 ± 0.0304	**0.5460** ± 0.0040	**0.5205** ± 0.0061
iMMH^r	0.6394 ± 0.0073	**0.5255** ± 0.0214	0.5306 ± 0.0057	0.5144 ± 0.0046
	Trecvid 2003 (Text+Color)		13-animal Flickr (SIFT+Color)	
模型	正确率	F1值	正确率	F1值
EFH+SVM	0.6295 ± 0.0040	0.5050 ± 0.0173	**0.5341** ± 0.0075	0.5048 ± 0.0085
iEFH+SVM	0.6190 ± 0.0041	0.4721 ± 0.0152	0.5257 ± 0.0056	0.5037 ± 0.0080
iEFH^r+SVM	**0.6465** ± 0.0039	**0.5243** ± 0.0148	0.5294 ± 0.0284	**0.5076** ± 0.0234
MMH	0.6532 ± 0.0065	0.5456 ± 0.0270	0.5438 ± 0.0018	0.5203 ± 0.0044
iMMH	0.6540 ± 0.0089	0.5428 ± 0.0335	**0.5617** ± 0.0040	**0.5351** ± 0.0060
iMMH^r	**0.6702** ± 0.0042	**0.5648** ± 0.0134	0.5430 ± 0.0071	0.5301 ± 0.0072

需要说明的是，iEFH和iMMH两种模型都可以很容易地扩展至考虑两种特征或者多种特征的多模态iEFH和多模态iMMH，需要变化的地方是定义多模态的EFH联合分布，隐变量 **Z**的先验分布不变。与第3章中的DWH模型相似，在此不再赘述。对于实值特征，使用高斯似然模型；对于离散的文本/SIFT特征，使用伯努利似然模型表示输入数据。

5.5.2.1　预测性能

实验中将非参数化的iEFH、iMMH与非参数化的EFH及MMH进行比较。对于无监督的EFH和iEFH模型，分别建立独立的多类别SVM分类器[88]用于预测。表5.3 所示为各模型在两个图像数据集上的分类结果，在此使用分类正确率和F1-score用于评价模型预测性能。可观察得出：

（**1**）单模态和多模态这两种配置中，无限维iEFH与有限维EFH模型的分类结果相差不多（前者甚至更好），而有监督的 iMMH模型也同样与有限维MMH模型的分类结果相差不多（前者甚至更好）；

（**2**）总体上说，有监督模型（iMMH和MMH）比无监督模型（iEFH 和 EFH）的预测结果更好，特别是使用多模态特征时性能提升更显著;

（**3**）使用多模态特征有助于提高模型的预测性能。例如，从TRECVID数据

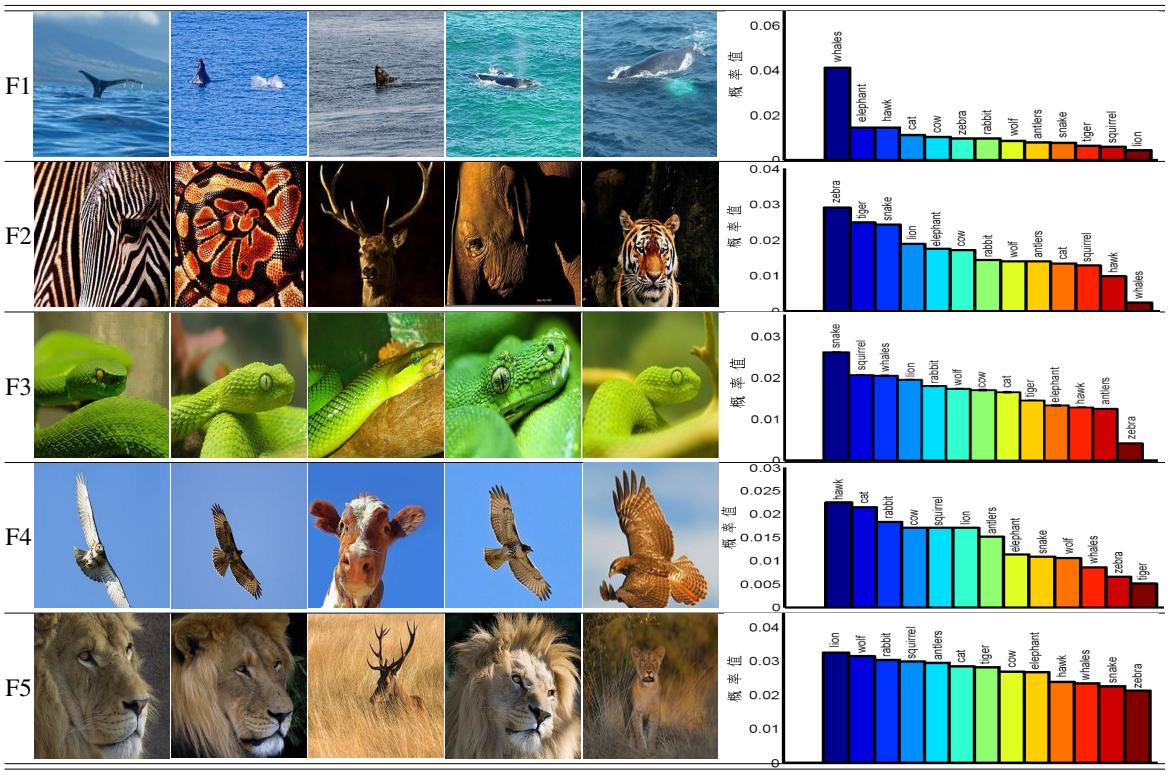


图 5.5　iMMH在Flickr数据集上的每个特征的图像示例。

集上的实验结果可以发现，同时使用文本和图像两种模态特征的iMMH模型比只用文本特征的单模态iMMH结果更好。原因是视觉特征与文本特征可以互补，与使用单模态的文本特征相比，有助于提高预测性能；

（**4**）比较隐变量为实数值的（iMMH^r 和 iEFH^{r}模型）与离散二值化的（iMMH 和iEFH）模型的预测性能。可以发现，两者并没有明显优劣之分。在TRECVID数据集上，使用实数值隐节点的模型可以得到更好的预测性能。而在Flickr数据集上，使用离散二值隐节点的模型能够得到更好的结果。但是从计算复杂度的角度上看，使用离散二值的隐节点，可以节省更多的机器运行时间。

5.5.2.2　判别性隐层空间表示

图5.5所示为iMMH模型用于Flickr数据集分类问题得到的隐特征对应的示例图片。对于每一维隐特征，将所有图像按照该维隐层特征的期望值由高至低排序，然后取前5幅图像作为该特征的代表性图像。为了展示隐特征的区分能力，图5.5中画出每一维特征在所有13类图像上的平均期望分布图。可以发现，自动学习得到的隐特征具有很强的语义含义。举例说明，特征F1更倾向于表示类别“whales”，特征F3更倾向于表示类别“snake”。总的来说，这些特征对于多种类别的样本都具有很强的区分能力。例如，特征F1适于区分“whales”和其他动物，而特征F4更

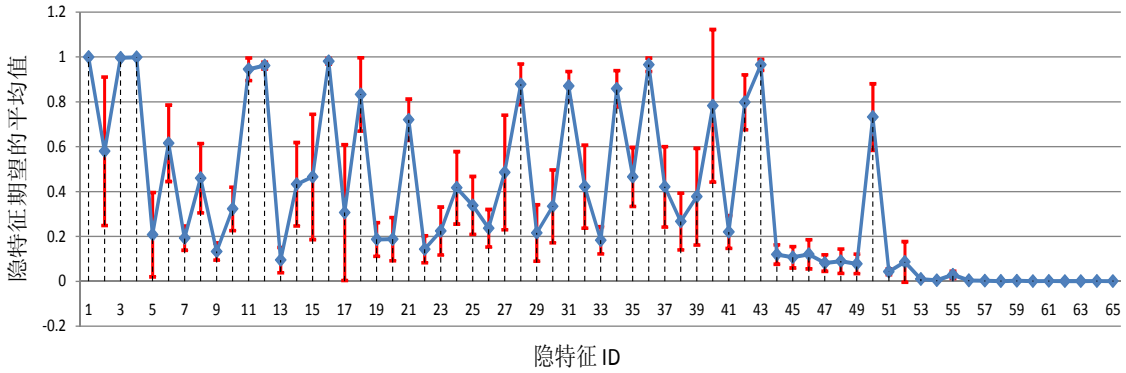


图 5.6　iMMH在Trecvid图像数据集上得到的隐特征的期望在所有5类样本上平均及方差。

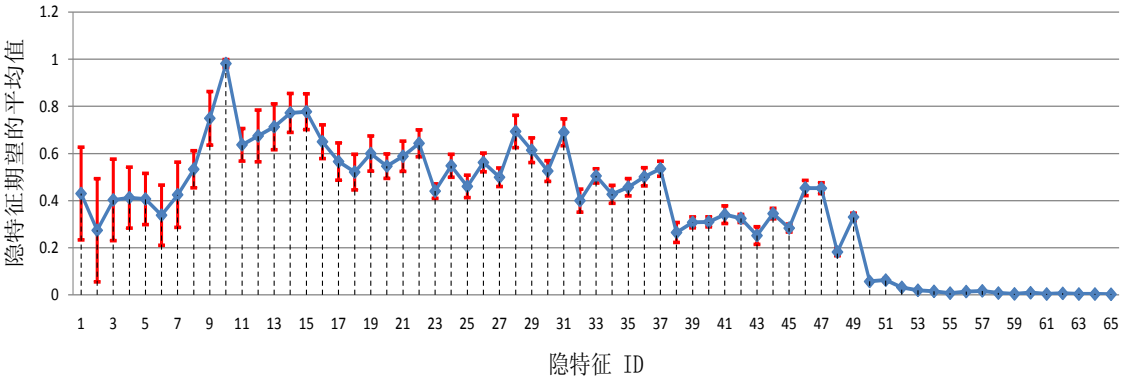
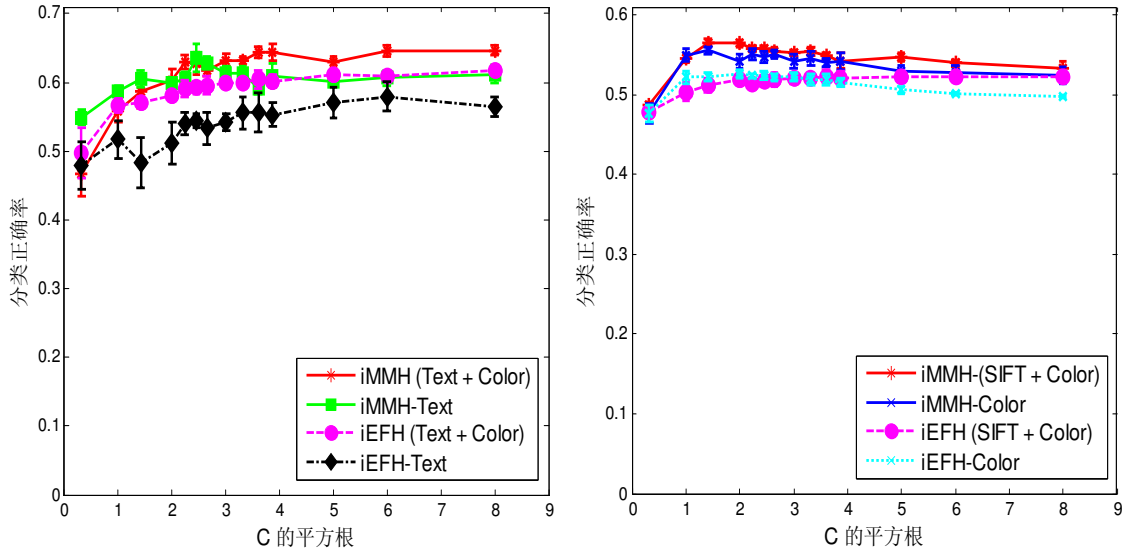


图 5.7　iMMH在Flickr图像数据集上得到的隐特征的期望在所有13类样本上平均及方差。

适于区分类别组合{“hawk”，“tiger”，“zebra”}以及类别“whales”等。相反，有些特征并没有强的区分性。例如，虽然“lion”在特征F5右侧的概率分布中概率值最高，但它与其他动物如“wolf”，“tiger” 和“zebra”等类别的概率值的区别很小，说明特征F5对于所有样本的区分能力与其他特征相比稍弱。为了提供隐层空间表示的整体分布结果，图5.6所示为iMMH模型学习得到的前65维隐特征期望值在所有TRECVID数据集样本上的平均值及方差，由于65维之后的隐特征值都趋近于零，所以在图中忽略。每一维特征k对应的方差一定程度上表示特征k对于所有类别样本的区分能力（即方差越大，区分性越强）。类似的，图5.7所示为iMMH模型学习得到的前65维隐特征期望值在Flickr数据集所有样本上的分布。读者可以观察发现相似的结果。这进一步说明非参数化iMMH模型可以发现判别式的隐层空间表示。

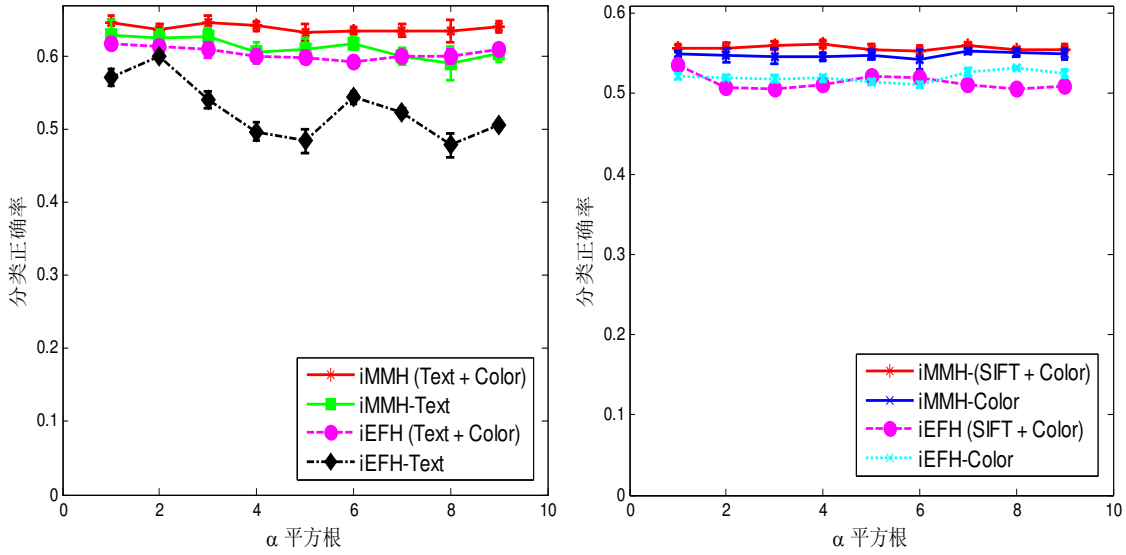
5.5.3　参数敏感度分析

下面分析对于不同的输入特征，本章提出的非参数化iMMH和iEFH模型对于



(a) TRECVID数据集　　(b) Flickr数据集

图 5.8　模型在两个数据集上关于参数C的敏感度分析。

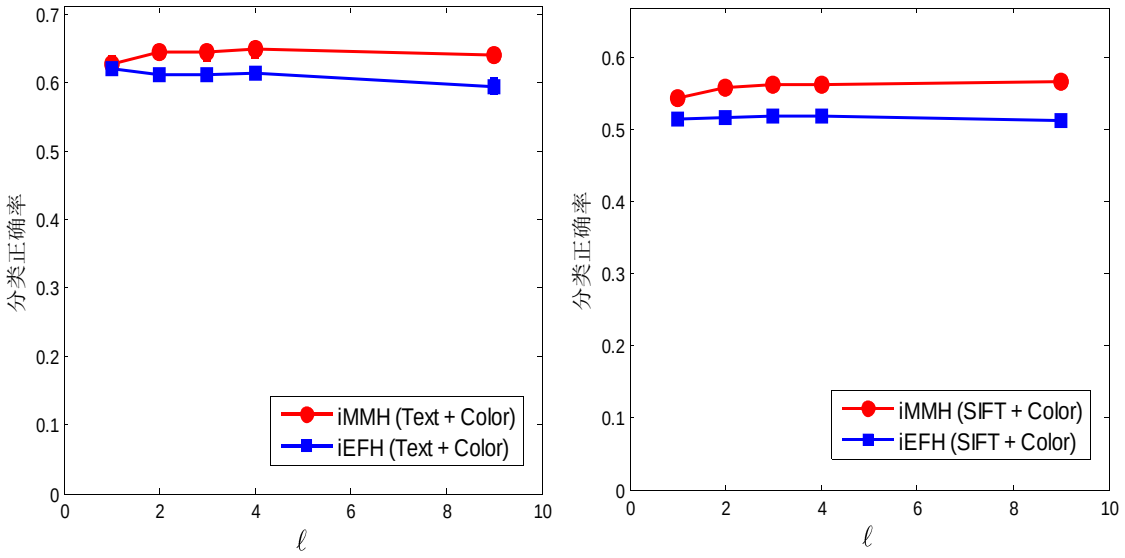


(a) TRECVID数据集　　(b) Flickr数据集

图 5.9　模型在两个数据集上关于参数α的敏感度分析。

参数C、α、ℓ的敏感度分析。实验中使用grid搜索来选择参数，图5.8、图5.9 和图5.10展示的是iMMH以及iEFH模型在TRECVID和Flickr数据集上的分类正确率对于参数C、α、ℓ的变化曲线[①]。可以观察发现，α和 ℓ的变化不会影响模型在两种数据集上的预测性能。当正则化参数$C < 20$时，模型的预测性能有一些振荡，但

① 本实验使用代价函数$\ell_d^{\Delta}(y) = \ell\mathbb{I}(y_d = y)$，其中$\mathbb{I}$为指示函数，$\ell$为大于0的参数。



(a) TRECVID数据集　　(b) Flickr数据集

图 5.10　iMMH和iEFH模型在两个数据集上关于参数ℓ的敏感度分析。

当$C \geq 20$时，预测性能趋近于稳定。总的来说，敏感度分析中最优的预测结果与表5.3的最优结果是一致的。

5.6　本章小结

本章研究在无向图隐层空间马尔可夫网络模型中引入非参数贝叶斯方法，可自动地从经验数据中确定模型复杂度（这里指隐特征的维度）。具体地说，本章提出无限维指数族Harmonium模型（iEFH）。基于正则化贝叶斯推理在带有经验参数的链图上的推广，本章将iEFH模型推广至有监督的无限维最大间隔Harmonium（iMMH）模型，该模型不但可以自动确定隐特征维数，还可用于实现预测任务。iEFH和iMMH模型是迄今为止，将“非参数贝叶斯方法”与“最大间隔学习”在“无向图隐变量模型”中结合的最新尝试。实验结果表明，与其他流行算法相比，iEFH和iMMH模型不仅可以避开复杂的模型选择问题，同时可以得到与参数化模型相媲美（甚至更优）的预测性能。

第 6 章　总结与展望

6.1　本文总结

互联网及数字化技术的飞速发展为用户提供了大规模几乎“免费”的有监督信息。为了有效利用这些大规模的有监督信息，解决有监督隐层空间模型研究中的若干基础性关键问题，本文深入系统地研究了判别式有监督隐空间概率模型，分别从模型的表示、判别式学习与推理方法、以及模型复杂度三个方面，系统建立判别式隐层空间学习的若干关键理论与方法：包括基于无向图马尔可夫网络的多模态数据建模与表示、基于最大间隔准则的有监督判别式学习、正则化贝叶斯及非参数化贝叶斯推理理论与方法等。具体地说，本文的研究内容分为两大部分：第一部分，当模型的复杂度固定时，研究参数化的隐层空间模型的表示和基于最大间隔准则的判别式学习方法；第二部分，当模型的复杂度可随着观测数据的增加而增长时，研究非参数化的隐层空间模型的表示和基于最大间隔准则的判别式后验推理理论方法。本文各章节的关系如图6.1所示。具体来说，本文各部分的主要研究内容可概括为：

（1）参数化隐层空间分类模型

在模型表示方面，为了考虑多模态特征，本文提出一种有监督的无向图统计学习方法，即多模态隐层空间马尔可夫网络。与有向的贝叶斯网络不同，多模态隐层空间马尔可夫网络模型建立在弱条件独立假设基础上：即当一系列隐层变量

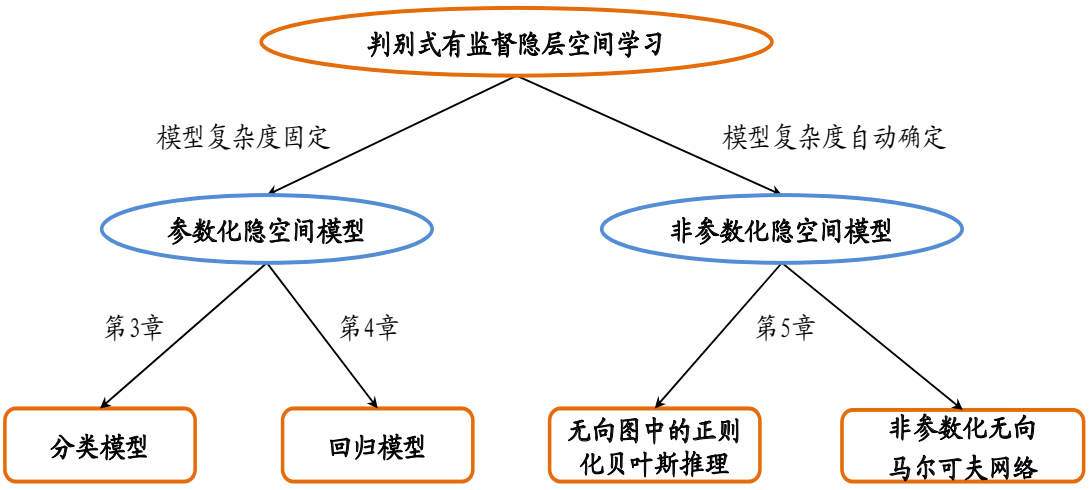


图 6.1　本文各章节关系。

已知时，模型多模态输入变量与响应变量之间满足条件独立性，因此可以高效地进行推理。首先，该模型可以有效利用有监督信息，学习判别式的隐层空间表示。其次，鉴于概率图模型可以处理缺失数据的优点，模型还可以实现多模态输入变量级别的预测（例如在图像标注任务中，已知图像的视觉信息，预测图像的文本标注）。再次，为了克服传统最大似然估计方法存在很难计算的归一化因子的缺点，本文提出使用基于期望的线性算子，在无向图隐层空间概率模型中引入确定性的最大间隔学习方法，联合最小化负对数似然与最小化训练数据的预测损失函数方法进行参数学习，显著提高了模型的预测性能，以及隐层空间表示的判别性。最后，本文提出的最大间隔隐层空间马尔可夫网络不仅可以用于分类问题，还可以用于图像检索与图像标注中。

（2）参数化隐层空间回归模型

在模型表示方面，本文进一步提出用于预测连续响应变量的隐层空间马尔可夫网络回归模型，以及有依赖结构（如输入数据满足一阶马尔可夫链式结构的依赖关系）的输入数据的隐层空间模型预测问题。为了克服最大似然估计方法存在的很难计算的归一化因子的缺点，本文在无向图回归模型中采用判别式最大间隔学习方法，通过对隐变量的期望定义损失函数，最小化负对数似然和ϵ-不敏感损失函数对模型进行参数学习。实验结果证明本文提出的参数化最大间隔隐层空间回归模型不仅可以发现判别式的隐层空间表示，同时可以高效地进行回归分析。最后，考虑输入数据中的结构化信息，可以进一步提高模型的预测性能。

虽然上述工作研究了参数化有监督隐层空间模型的表示和判别式学习等问题，并且显著提高了在多种应用问题上的性能，这些参数化模型和方法都存在一个公认难题：即如何确定模型的复杂度，这里具体指隐节点的数目。因为很难事先确定隐层空间模型的隐节点数量，通常需要进行代价很高的模型选择。为了解决上述问题，本文接下来系统研究了可自动地确定模型复杂度的非参数贝叶斯推理理论与方法。

（3）非参数隐层空间无向马尔可夫网络

绝大多数非参数隐变量模型都是在有向图贝叶斯网络框架下提出的[53,64,65]。至今为止，几乎没有任何工作使用非参数贝叶斯方法解决无向图隐空间马尔可夫网络中的“模型选择”问题。其中一个主要原因为：基于贝叶斯推理的无向图马尔可夫网络是一个链图，它具有和贝叶斯网络不同的马尔可夫性质。于是，本文将Zellner教授提出的经典贝叶斯对偶理论推广至链图及带有经验参数的贝叶斯推理，并在此框架下，系统研究非参数化无向马尔可夫网络隐特征模型的表示及判别式推理问题。以多模态隐层空间马尔可夫（以及其特例Harmonium）模型为例，

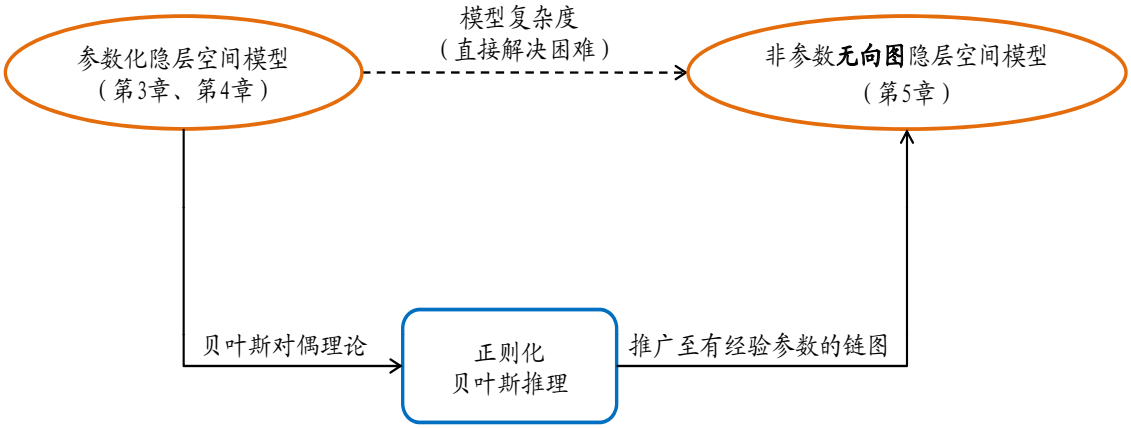


图 6.2　从参数化至非参数化判别式隐层空间模型的研究思路。

提出无限维的指数族Harmonium模型。值得说明的是，为了研究非参数化有监督隐层空间模型，本文作者合作提出带有后验约束的“正则化贝叶斯推理”理论框架[57]。在该框架下，成功地将近20年来互相独立的机器学习子领域非参数贝叶斯推理与最大间隔判别式学习有机地融合在一起。不仅可以避开隐层空间模型的模型选择问题，还可以提高模型的预测性能。但是，对于本文主要研究的对于——基于无向图的隐空间模型，如何有效地利用非参数化贝叶斯方法的长处自动确定其模型复杂度仍然是一个开放的问题。于是，本文又将该正则化贝叶斯推理理论推广至链图及带有经验参数的贝叶斯推理，可以很自然地引入最大间隔后验约束，提出非参数的最大间隔指数族Harmonium模型。该模型不仅可以避开代价较高的模型选择，同时可以学习判别式的隐层空间表示，提高模型的预测性能。

综上所述，本文从参数化判别式隐层空间马尔可夫网络至非参数化判别式隐层空间马尔可夫网络的研究过程，如图6.2中所示。图中的虚线表示在参数化隐层马尔可夫网络基础上，直接研究无向图的非参数化隐层空间马尔可夫网络相对比较困难，所以本文借鉴合作文章[57]中提出的正则化贝叶斯推理理论框架，成功将该基础理论方法推广至带有经验参数的链图中，进而研究基于无向图的判别式非参数化隐层空间模型。

6.2　未来工作展望

本文系统研究了参数化及非参数化的判别式有监督隐层空间模型中的表示、学习和复杂度等基础性关键问题。虽然取得了一些不错的理论和应用结果，判别式隐层空间学习问题中还存在诸多挑战有待研究者们进一步探索。具体地说，可概括为：

（1）无向图隐层空间马尔可夫网络中隐变量期望的可解释性。

如文献[14]中提出，与有监督的有向图模型比较，无向图隐层空间模型有一个稍显不足的地方是：隐层空间表示不能保证非负性，所以直观解释更加复杂一些。虽然本文第3，4，5章对隐空间期望值的变换保持了隐层空间表示的区分性能，但仍然有必要研究一种更具突破性的方法（如在模型优化过程中引入参数权值的非负性约束）来提高隐空间表示的可解释性，同时，研究稀疏的隐空间表示也是增强可解释性的有效手段之一，目前已有的可借鉴的工作包括压缩感知（Compressive Sensing）[138]、稀疏编码（Sparse Coding）[139]、非负性稀疏编码[140]、非负性矩阵分解[141]、Lasso[142,143]、稀疏话题模型[144]等各种变种。本文作者在这方面已经有初步研究成果，如学习稀疏隐层话题表示[145]。

（2）带有其他正则化后验约束的非参数化贝叶斯推理有向图/无向图模型

正则化贝叶斯推理是一种通用的在非参数贝叶斯推理过程中考虑后验正则化的框架。未来的工作中可以研究其他后验正则化约束，例如定义在流形结构[59]上的后验约束，以及探索在其他有趣的非参数贝叶斯模型[52,53]中引入后验正则化约束。此外，本文提出的iEFH和iMMH模型是将“非参数贝叶斯方法”与“最大间隔学习”在“无向图隐变量模型”中结合的一个最新尝试。后续工作可以进一步研究将非参数贝叶斯推理用于更加复杂的无向图模型（例如深度波尔兹曼机（Deep Boltzmann Machines）[146]以及条件随机场（Conditional Random Fields）[147]）的结构。

（3）在大规模数据中的应用处理。

基于本文提出的无向图隐层空间马尔可夫网络模型推理效率高的特点，有希望将其运用于大规模的图像标注以及分类[148]问题。有必要研究更加高效的推理算法（例如多核或者大机群工作模式下的并行推理方法等），将判别式的隐层空间学习用于处理大规模的数据挖掘[149]和计算机视觉[150]等诸多热点问题的研究中。

参考文献

[1] Blei D, Ng A, Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, (3):993–1022.

[2] Hoff P D, Raftery A E, Handcock M S. Latent Space Approaches to Social Network Analysis. Journal of the American Statistical Association, 2005, 97(460):1090–1098.

[3] Xing E P, Yan R, Hauptmann A G. Mining Associated Text and Images with Dual-Wing Harmoniums. Proceedings of Conference on Uncertainty in Artificial Intelligence, 2005.

[4] Rabiner L R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 1989, (77(2)):257–286.

[5] Clark A G. Inference of haplotypes from pcr-amplified samples of diploid populations. Molecular Biology and Evolution, 1990, 7(2):111–122.

[6] Elidan G, Lotner N, Friedman N, et al. Discovering Hidden Variables: A Structure-Based Approach. Proceedings of Advances in Neural Information Processing Systems, 2000.

[7] Lao N, Zhu J, Liu L, et al. Efficient Relational Learning with Hidden Variable Detection. Proceedings of Advances in Neural Information Processing Systems, 2010.

[8] Tenenbaum J B, Kemp C, Griffiths T L, et al. How to Grow a Mind: Statistics, Structure, and Abstraction. Science, 2011, 331(6022):1279–1285.

[9] Parikh A P, Song L, Xing E P. A Spectral Algorithm for Latent Tree Graphical Models. Proceedings of International Conference on Machine Learning, 2011.

[10] Welling M, Hinton G E. A new learning algorithm for mean field Boltzmann machines. Proceedings of International Conference on Artificial Neural Networks, 2001.

[11] Ahmed A, Smola A J. Latent Variable Models on the Internet. WWW Tutorial, 2011..

[12] Lee H, Ekanadham C, Ng A Y. Sparse Deep Belief Net Model for Visual Area V2. Proceedings of Advances in Neural Information Processing Systems 20, 2008. 873–880.

[13] Hofmann T. Probabilistic Latent Semantic Indexing. Proceedings of SIGIR International Conference on Information Retrieval, 1999.

[14] Welling M, Rosen-Zvi M, Hinton G. Exponential Family Harmoniums with an Application to Information Retrieval. Proceedings of Advances in Neural Information Processing Systems, 2004.

[15] Cao L, Fei-Fei L. Spatially coherent latent topic model for concurrent object segmentation and classification. Proceedings of IEEE International Conference on Computer Vision, 2005.

[16] Li L J, Socher R, Fei-Fei L. Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[17] Airoldi E M, Blei D M, Fienberg S E, et al. Mixed Membership Stochastic Blockmodels. Journal of Machine Learning Research, 2008, 9:1981–2014.

[18] Blei D, Jordan M. Modeling Annotated Data. Proceedings of SIGIR International Conference on Information Retrieval, 2003.

[19] Jolliffe I T. Principal Component Analysis: Second Edition. New York, NY: Springer, 2002.

[20] Chen S. Subpattern-based Principal Component Analysis. Pattern Recognition, 2004, 37:1081–1083.

[21] Hyv*ä*rinen A, Karhunen J, Oja E. Independent Component Analysis. John Wiley & Sons, 2001.

[22] Hotelling H. Relations Between Two Sets of Variates. Biometrika, 1936, 28(3/4):321–377.

[23] Diethe T, Hardoon D R, Shawe-Taylor J. Multiview Fisher Discriminant Analysis. Proceedings of NIPS Workshop on Learning from Multiple Sources, 2008.

[24] Bishop C M. Pattern Recognition and Machine Learning. Springer, 2006.

[25] Tipping M E, Bishop C M. Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society, Series B, 1999, (61):611–622.

[26] Roy N, Gordon G. Exponential Family PCA for Belief Compression in POMDPs. Proceedings of Advances in Neural Information Processing Systems, 2003.

[27] Freund Y, Haussler D. Unsupervised learning of distributions of binary vectors using 2-layer networks. Proceedings of Advances in Neural Information Processing Systems, 1992.

[28] Hinton G E. Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation, 2002, 14(8):1771–1800.

[29] Blei D, McAuliffe J D. Supervised Topic Models. Proceedings of Advances in Neural Information Processing Systems, 2007.

[30] Lacoste-Jullien S, Sha F, Jordan M I. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification. Proceedings of Advances in Neural Information Processing Systems, 2008.

[31] Wang C, Blei D, Fei-Fei L. Simultaneous Image Classification and Annotation. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2009.

[32] Titov I, McDonald R. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. Proceedings of Annual Meeting of Association of Computational Linguistics, 2009.

[33] Larochelle H, Bengio Y. Classification using Discriminative Restricted Boltzmann Machines. Proceedings of International Conference on Machine Learning, 2008.

[34] Yang J, Liu Y, Xing E P, et al. Harmonium Models for Semantic Video Representation and Classification. Proceedings of SIAM Conference on Data Mining, 2007.

[35] Yang J, Liu Y, Xing E P, et al. Harmoniums Models for Semantic Video Representation and Classification. Proceedings of SIAM International Conference on Data Mining, 2007.

[36] Chen N, Zhu J, Xing E P. Predictive Subspace Learning for Multi-view Data: A Large Margin Approach. Proceedings of Advances in Neural Information Processing Systems, 2010.

[37] Vapnik V N. Statistical Learning Theory. New York: John Wiley & Sons, Inc., 1998.

[38] Smola A J, Sch*ö*lkopf B. A Tutorial on Support Vector Regression. Statistics and Computing, 2003, 14:199–222.

[39] 张学工. 统计学习理论的本质. 北京: 清华大学出版社, 2000.

[40] Zhu J, Ahmed A, Xing E P. MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification. Proceedings of International Conference on Machine Learning, 2009.

[41] Zhu J, Li L J, Li F F, et al. Large Margin Learning of Upstream Scene Understanding Models. Proceedings of Advances in Neural Information Processing Systems, 2010.

[42] Blei D M, Griffiths T L, Jordan M, et al. Hierarchical Topic Models and the Nested Chinese Restaurant Process. Proceedings of Advances in Neural Information Processing Systems, 2003.

[43] Aldous D J. Exchangeability and Related Topics. Berlin, Germany: Springer, 1985.

[44] Velez F D. The Indian Buffet Process: Scalable Inference and Extensions. Cambridge University Press, 2009.

[45] Ferguson T. A Bayesian Analysis of Some Nonparametric Problems. Annals of Statistics, 1973, (1):209–230.

[46] Rasmussen C E, Williams C K I. Gaussian Processes for Machine Learning. Cambridge, UK: MIT Press, 2006.

[47] Antoniak C. Mixture of Dirichlet Process with Applications to Bayesian Nonparametric Problems. Annals of Statistics, 1974, (273):1152–1174.

[48] Shahbaba B, Neal R. Nonlinear Models Using Dirichlet Process Mixtures. Journal of Machine Learning Research, 2009, 10:1829–1850.

[49] Griffiths T L, Ghahramani Z. Infinite Latent Feature Models and the Indian Buffet Process. Proceedings of UCL technical report, 2005.

[50] Rasmussen C, Ghahramani Z. Infinite Mixtures of Gaussian Process Experts. Proceedings of Advances in Neural Information Processing Systems, 2001.

[51] MacEachern S. Dependent Nonparametric Process. Proceedings of the Section on Bayesian Statistical Science of ASA, 1999.

[52] Teh Y, Jordan M, Beal M, et al. Hierarchical Dirichlet Process. Journal of the American Statistical Association, 2006, 101(476):1566–1581.

[53] Beal M J, Ghahramani Z, Rasmussen C E. The Infinite Hidden Markov Model. Proceedings of Advances in Neural Information Processing Systems, 2002.

[54] Hoff D. Bayesian Methods for Partial Stochastic Orderings. Biometrika, 2003, 90:303–317.

[55] Dunson D, Peddada S. Bayesian Nonparametric Inferences on Stochastic Ordering. ISDS Discussion Paper, 2007, 2.

[56] Jaakkola T, Meila M, Jebara T. Maximum Entropy Discrimination. Proceedings of Advances in Neural Information Processing Systems, 1999.

[57] Zhu J, Chen N, Xing E P. Infinite Latent SVM for Classification and Multi-task Learning. Proceedings of Advances in Neural Information Processing Systems, 2011.

[58] Zhu J, Chen N, Xing E P. Infinite SVM: a Dirichlet Process Mixture of Large-margin Kernel Machines. Proceedings of International Conference on Machine Learning, 2011.

[59] Huh S, Fienberg S. Discriminative Topic Modeling based on Manifold Learning. Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2010.

[60] Chen N, Zhu J, Sun F, et al. Large-margin Predictive Latent Subspace Learning for Multi-view Data Analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence (in press). Preprint: http://www.computer.org/portal/web/csdl/doi/10.1109/TPAMI.2012.64., 2012..

[61] Chen N, Zhu J. MMH: Max Margin Harmoniums. Proceedings of ICML workshop on Topic Models, 2010.

[62] Zellner A. Optimal Information Processing and Bayes' Theorem. American Statistician, 1988, 42:278–280.

[63] Chen N, Zhu J, Sun F. Infinite Exponential Family Harmoniums. Proceedings of NIPS workshop on Bayesian Nonparametrics: Hope or Hype?, 2011.

[64] Blei D, Jordan M I. Variational inference for Dirichlet process mixtures. Bayesian Analysis, 2006, 1:121–144.

[65] Adams R P, Wallach H, Ghahramani Z. Learning the structure of deep sparse graphical models. Proceedings of International Conference on Artificial Intelligence and Statistics, 2010.

[66] Frydenberg M. The Chain Graph Markov Property. Scandinavian Journal of Statistics, 1990, 17:333–353.

[67] Dempster A, Laird N, Rubin D. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B, 1977, (39):1–38.

[68] Jordan M I, Ghahramani Z, Jaakkola T, et al. An introduction to variational methods for graphical models. Cambridge, MA: M. I. Jordan (Ed.), Learning in Graphical Models, Cambridge: MIT Press, 1999.

[69] Expectation – maximization Algorithm. http://en.wikipedia.org/wiki/EM_algorithm.

[70] Xing E P, Jordan M I, Russell S. A Generalized Mean Field Algorithm for Variational Inference in Exponential Families. Proceedings of Conference on Uncertainty in Artificial Intelligence, 2003.

[71] Jordan M I. Learning in Graphical Models: Foundations of Neural Computation. Cambridge: MIT Press, 1998.

[72] Murray I, Ghahramani Z. Bayesian Learning in Undirected Graphical Models: Approximate MCMC algorithms. Proceedings of Conference on Uncertainty in Artificial Intelligence, 2004.

[73] Zhu J, Nie Z, Zhang B, et al. Dynamic Hierarchical Markov Random Fields and their Application to Web Data Extraction. Proceedings of International Conference on Machine Learning, 2007.

[74] Carreira-Perpinan M A, Hinton G E. On Contrastive Divergence Learning. Proceedings of Artificial Intelligence and Statistics, 2005.

[75] Yuille A. The Convergence of Contrastive Divergences. Proceedings of Advances in Neural Information Processing Systems, 2004.

[76] Chen H, Murray A F. Continuous restricted Boltzmann machine with an implementable training algorithm. Proceedings of IEE Proceedings: Vision, Image and Signal Processing, volume 150, 2003. 153–158.

[77] Teh Y W, Welling M, Osindero S, et al. Energy-based Models for Sparse Overcomplete Representations. Journal of Machine Learning Research, 2003, 4:1235–1260.

[78] He X, Zemel R S, Carreira-Perpinan M A. Multiscale conditional random fields for image labeling. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2004.

[79] Quattoni A, Collins M, Darrell T. Conditional Random Fields for Object Recognition. Proceedings of Advances in Neural Information Processing Systems, 2004.

[80] Verbeek J, Triggs B. Scene Segmentation with Conditional Random Fields Learned from Partially Labeled Images. Proceedings of Advances in Neural Information Processing Systems, 2007.

[81] Zhu J, Nie Z, Zhang B, et al. Dynamic Hierarchical Markov Random Fields for Integrated Web Data Extraction. Journal of Machine Learning Research, 2008, (9):1583–1614.

[82] Petrov S, Klein D. Discriminative Log-Linear Grammars with Latent Variables. Proceedings of Advances in Neural Information Processing Systems, 2007.

[83] 朱军. 最大熵判别式马尔可夫网络：理论与应用[D]. 中国北京: 清华大学, 6月, 2009.

[84] Sha F, Saul L K. Large Margin Hidden Markov Models for Automatic Speech Recognition. Proceedings of Advances in Neural Information Processing Systems, 2007.

[85] Yu C, Joachims T. Learning Structural SVMs with Latent Variables. Proceedings of International Conference on Machine Learning, 2009.

[86] Jebara T. Discriminative, Generative and Imitative Learning[D]. USA: Massachusetts Institute of Technology, February, 2002.

[87] Zhu J, Xing E P, Zhang B. Partially Observed Maximum Entropy Discrimination Markov Networks. Proceedings of Advances in Neural Information Processing Systems, 2008.

[88] Crammer K, Singer Y. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. Journal of Machine Learning Research, 2001, (2):265–292.

[89] Joachims T, Finley T, Yu C N. Cutting Plane Training of Structural SVMs. Machine Learning Journal, 2009..

[90] Sethuraman J. A Constructive Definition of Dirichlet Priors. Statistica Sinica, 1994, (4):639–650.

[91] Blackwell D, MacQueen J. Ferguson Distributions via Polya Urn Scheme. Annals of Statistics, 1973, (1):353–355.

[92] Hannah L, Blei D, Powell W. Dirichlet Process Mixtures of Generalized Linear Models. Technical Report, arXiv:0909.5194v2, 2010..

[93] Griffiths T, Ghahramani Z. Infinite Latent Feature Models and the Indian Buffet Process. Proceedings of Advances in Neural Information Processing Systems, 2006.

[94] Miller K, Griffiths T, Jordan M. Nonparametric Latent Feature Models for Link Prediction. Proceedings of Advances in Neural Information Processing Systems, 2009.

[95] Rai P, III H D. Infinite Predictor Subspace Models for Multitask Learning. Proceedings of International Conference on Artificial Intelligence and Statistics, 2010.

[96] Teh Y, Gorur D, Ghahramani Z. Stick-breaking Construction of the Indian Buffet Process. Proceedings of International Conference on Artificial Intelligence and Statistics, 2007.

[97] Blum A, Mitchell T. Combining Labeled and Unlabeled Data with Co-Trainnig. Proceedings of Annual Conference on Learning Theory, 1998.

[98] Xing E P, Yan R, Hauptmann A G. Mining Associated Text and Images with Dual-Wing Harmoniums. Proceedings of Conference on Uncertainty in Artificial Intelligence, 2005.

[99] Ferrari V, Tuytelaars T, Gool L V. Integrating Multiple Model Views for Object Recognition. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2004.

[100] Christoudias C M, Urtasun R, Darrell T. Multi-View Learning in the Presence of View Disagreement. Proceedings of Conference on Uncertainty in Artificial Intelligence, 2008.

[101] Thomas A, Ferrari V, Leibe B, et al. Towards Multi-View Object Class Detection. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[102] Torralba A, Murphy K P, Freeman W T. Sharing visual features for multiclass and multiview object detection. Proceedings of IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007.

[103] Culp M, Michailidis G, Johnson K. On multi-view learning with additive models. Annals of Applied Statistics, 2009, 3(1):292–318.

[104] Burges C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 1998, 2(2):121–167.

[105] Freund Y. An adaptive version of the boost by majority algorithm. Machine Learning, 2001, 43(3):293–318.

[106] Brefeld U, Scheffer T. Co-EM Support Vector Learning. Proceedings of International Conference on Machine Learning, 2004.

[107] Ando K, Zhang T. Two-View Feature Generation Model for Semi-supervised Learning. Proceedings of International Conference on Machine Learning, 2007.

[108] Foster D, Kakade S, Zhang T. Multi-view Dimensionality Reduction via Canonical Correlation Analysis. Technical report, Technical Report TR-2008-4, TTI-Chicago, 2008.

[109] Kakade S M, Foster D P. Multi-view Regression via Conanical Correlation Analysis. Proceedings of Annual Conference on Learning Theory, 2007.

[110] Chaudhuri K, Kakade S M, Livescu K, et al. Multi-View Clustering via Canonical Correlation Analysis. Proceedings of International Conference on Machine Learning, 2009.

[111] Ganchev K, Graca J V, Blitzer J, et al. Multi-View Learning over Structured and Non-Identical Outputs. Proceedings of Conference on Uncertainty in Artificial Intelligence, 2008.

[112] Akaho S. A Kernel Method for Canonical Correlation Analysis. Proceedings of International Meeting on Psychometric Society, 2001.

[113] Salakhutdinov R, Hinton G E. Replicated Softmax: an Undirected Topic Model. Proceedings of Advances in Neural Information Processing Systems, 2009.

[114] Larochelle H, Bengio Y. Classification using Discriminative Restricted Boltzmann Machines. Proceedings of International Conference on Machine Learning, 2008.

[115] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of International Conference on Machine Learning, 2001.

[116] Zhang J, Ghahramani Z, Yang Y. Flexible latent variable models for multi-task learning. Machine Learning, 2008, 73(3):221–242.

[117] Zhu J, Xing E P. Conditional Topic Random Fields. Proceedings of International Conference on Machine Learning, 2010.

[118] Vapnik V. The Nature of Statistical Learning Theory. Springer, 1999.

[119] Xing E P, Jordan M I, Russell S. A generalized mean field algorithm for variational inference in exponential families. Proceedings of Conference on Uncertainty in Artificial Intelligence, 2003.

[120] Wainwright M J, Jordan M I. Graphical Models, Exponential Families, and Variational Inference. Foundations and Trends in Machine Learning, 2008, 1(1–2):1–305.

[121] Liu D C, Nocedal J. On the limited memory BFGS method for large scale optimization. Mathematical Programming, 1989, (45):503–528.

[122] Chua T S, Tang J, Hong R, et al. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. Proceedings of International Conference on Image and Video Retrieval, 2009.

[123] Lowe D G. Object Recognition from Local Scale-invariant Features. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1999.

[124] Maaten L, Hinton G E. Visualizing Data using t-SNE. Journal of Machine Learning Research, 2008, 9:2579–2605.

[125] Joachims T. Making large-Scale SVM Learning Practical. Advances in kernel methods–support vector learning, B. Schölkopf, C. Burges and A. Smola (ed.), MIT-Press, 1999..

[126] Liu X, Shao Y. Asymptotics for Likelihood Ratio Tests Under Loss of Identifiability. Annals of Statistics, 2003, 31(3):807–832.

[127] Rasmussen C, Ghahramani Z. Infinite Mixtures of Gaussian Process Experts. Proceedings of Advances in Neural Information Processing Systems, 2002.

[128] Blei D M, Griffiths T L, Jordan M I. The nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. Journal of the ACM, 2010, 57(2):1–30.

[129] Gershman S J, Frazier P I, Blei D M. Distance Dependent Infinite Latent Feature Models. ArXiv:1110.5454v1, 2011..

[130] Mann G, McCallum A. Generalized Expectation Criteria for Semi-Supervised Learning with Weakly Labeled Data. Journal of Machine Learning Research, 2010, (11):955–984.

[131] Ganchev K, Graca J, Gillenwater J, et al. Posterior Regularization for Structured Latent Variable Models. Journal of Machine Learning Research, 2010, (11):2001–2094.

[132] Bellare K, Druck G, McCallum A. Alternating Projections for Learning with Expectation Constraints. Proceedings of Conference on Uncertainty in Artificial Intelligence, 2009.

[133] Liang P, Jordan M, Klein D. Learning from Measurements in Exponential Families. Proceedings of International Conference on Machine Learning, 2009.

[134] Argyriou A, Evgeniou T, Pontil M. Convex Multi-Task Feature Learning. Proceedings of Advances in Neural Information Processing Systems, 2007.

[135] Bakker B, Heskes T. Task Clustering and Gating for Bayesian Multitask Learning. Journal of Machine Learning Research, 2003, (4):83–99.

[136] Doshi-Velez F, Miller K T, Gael J V, et al. Variational Inference for the Indian Buffet Process. Proceedings of International Conference on Artificial Intelligence and Statistics, 2009.

[137] Zhu J, Xing E P, Zhang B. Laplace Maximum Margin Markov Networks. Proceedings of International Conference on Machine Learning, 2008.

[138] Donoho D L. Compressed Sensing. IEEE Trans. on Information Theory, 2006, 52(4):1289–1306.

[139] Olshausen B A, Field D J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 1996, 381(6583):607–609.

[140] Hoyer P O. Non-negative sparse coding. 2002. 557–565.

[141] Lee D, Seung H. Learning the Parts of Objects by Non-negative Matrix Factorization. Nature, 1999, 401:788 – 791.

[142] Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of Royal Statistics Society, 1996, B(58):267–288.

[143] Kim S, Xing E P. Tree-guided group lasso for multi-task regression with structured sparsity. Proceedings of International Conference on Machine Learning, 2010.

[144] Zhu J, Xing E P. Sparse Topical Coding. Proceedings of Conference on Uncertainty in Artificial Intelligence, 2011.

[145] Zhu J, Lao N, Chen N, et al. Conditional Topical Coding: an Efficient Topic Model Conditioned on Rich Features. Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2011.

[146] Salakhutdinov R, Larochelle H. Efficient Learning of Deep Boltzmann Machines. Proceedings of International Conference on Artificial Intelligence and Statistics, 2011.

[147] Bousmalis K, Morency L, Zafeiriou S, et al. A Discriminative Nonparametric Bayesian Model: Infinite Hidden Conditional Random Fields. Proceedings of NIPS Workshop on Bayesian Nonparametric Methods: Hope or Hype?, 2011.

[148] Weston J, Bengio S, Usunier N. Large Scale Image Annotation: Learning to Rank with Joint Word-Image Embeddings. Proceedings of European Conference on Machine Learning, 2010.

[149] Wallach H M. Topic Modeling: Beyond Bag-of-Words. Proceedings of International Conference on Machine Learning, 2006.

[150] Gökalp D, Aksoy S. Scene Classification Using Bag-of-Regions Representations. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[151] Khan M E, Marlin B, Bouchard G, et al. Variational bounds for mixed-data factor analysis. Proceedings of Advances in Neural Information Processing Systems, 2010.

[152] Joachims T. Transductive Inference for Text Classification using Support Vector Machines. Proceedings of International Conference on Machine Learning, 1999.

[153] Jebara T. Multitask Sparsity via Maximum Entropy Discrimination. Journal of Machine Learning Research, 2011, (12):75–110.

[154] Ando R, Zhang T. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. Journal of Machine Learning Research, 2005, (6):1817–1853.

致 谢

能进入清华大学攻读博士学位，师从孙富春教授，是我今生最荣幸之事。孙老师治学严谨，博学多才，为人耿直。对他的感激之情无以言表，只有在今后的学习中继续努力，以不辜负他的殷切期望。

感谢清华大学计算机系、智能技术与系统国家重点实验室这几年来对我的全面培养！感谢张钹、朱晓燕、孙茂松、马少平、朱纪洪、邓志东、朱军、刘华平、李洪波、汉堡大学张建伟等老师在学习研究过程中给予我的教诲和诸多帮助。感谢汪洪桥、丁林阁、蒋琪夏等师兄弟对我的支持与帮助！

感谢美国卡内基梅隆大学Eric Xing副教授！在美国为期一年的访问交流期间给予我很大自由和空间，让我有各种机会与世界最成功的计算机科学家们一起学习和工作！

感谢我的父母家人！二十七年来我的每一步成长和进步都离不开他们的悉心关爱和全力支持！

最后，谨向在百忙中抽出宝贵时间评审本文的专家、学者致以最由衷的谢意！

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

附录 A 最大间隔Harmonium模型

本附录研究多模态隐层空间马尔可夫网络的特例最大间隔Harmonium模型（Max-margin Harmonium，简称MMH）的参数学习问题。此方法与本文第3.3.1.1节中介绍的学习过程相似，只需做微小改动即可进行参数学习。对于其他多模态隐层空间马尔可夫模型的特例，可以使用相似的学习算法实现。

根据第3.2.1节中的局部条件分布定义，可以直接写出MMH模型的边缘数据似然$p(\mathbf{x},\mathbf{z})$以及模型的联合分布$p(\mathbf{x},\mathbf{z},\mathbf{h})$。其中

$$p(\mathbf{x},\mathbf{z}) \propto \exp\left\{\alpha^{\top}\mathbf{x} + \beta^{\top}\mathbf{z} - \frac{1}{2}\sum_j \frac{z_j^2}{\sigma_j^2} + \frac{1}{2}\sum_k (\mathbf{x}^{\top}\mathbf{W}_{\cdot k} + \mathbf{z}^{\top}\mathbf{U}_{\cdot k})^2\right\}.$$

为了处理难以计算的正则化因子，可以采用Contrastive Divergence方法，引入两种变分分布q_0和q_1。在此使用简单的结构化均值场假设

$$q(\mathbf{x},\mathbf{z},\mathbf{h}) = \prod_i q(x_i)\prod_j q(z_j)\prod_k q(h_k). \tag{A-1}$$

对于q的后验推理，可以得到完全分解的因子化迭代规则

$$\begin{aligned} q(\mathbf{x}) &= \prod_i q(x_i) = \prod_i p(x_i|\mathbb{E}_{q(\mathbf{h})}[\mathbf{h}]) \\ q(\mathbf{z}) &= \prod_j q(z_j) = \prod_j p(z_j|\mathbb{E}_{q(\mathbf{h})}[\mathbf{h}]) \\ q(\mathbf{h}) &= \prod_k q(h_k) = \prod_k p(h_k|\mathbb{E}_{q(\mathbf{x})}[\mathbf{x}],\mathbb{E}_{q(\mathbf{z})}[\mathbf{z}]). \end{aligned}$$

其中对于q_0，(x_i,z_j)固定在它们的观测值，只需计算$q(h_k)$。对于q_1，从q_0为开始，经过几步迭代过程即可得到一个较好的q_1。在推理得到q_0，q_1之后，使用本文3.3.1.2节中介绍的的迭代过程进行参数估计。固定Θ，估计参数$\mathbf{V}$的步骤实际上是学习一个多类别支持向量机的过程

$$\min_{\mathbf{V}} \ \frac{1}{2}C_1\|\mathbf{V}\|_2^2 + C_2\sum_d \max_y [\Delta\ell_d(y) - \mathbf{V}^{\top}\mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d,\mathbf{z}_d)}[\Delta\mathbf{f}_d(y)]].$$

注意到在这个例子中，隐层空间表示（即$\mathbf{H}$的期望值）可被简单的写为$\mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[\mathbf{h}] = \Upsilon$，当输入数据$\mathbf{x}$和$\mathbf{z}$是完全可观测时，$\Upsilon_k = \mathbf{x}^{\top}\mathbf{W}_{\cdot k} + \mathbf{z}^{\top}\mathbf{U}_{\cdot k}$，$\forall 1 \le k \le K$。若$\mathbf{x}$或$\mathbf{z}$存在部分缺失数据时，上式中相应的成份将被替换成他们的期望

值。因此，如本文第3.3.3节所述，在测试过程中预测问题（如分类或检索）可以很容易地实现。

为了进行参数估计，本文采用次梯度下降的方法，其中模型参数Θ的次梯度计算公式为

$$\begin{aligned}
\partial\alpha_i &= -\mathbb{E}_{q_0}[x_i] + \mathbb{E}_{q_1}[x_i], \\
\partial\beta_j &= -\mathbb{E}_{q_0}[z_j] + \mathbb{E}_{q_1}[z_j], \\
\partial(\sigma_j^{-1}) &= -\mathbb{E}_{q_0}[z_j^2\sigma_j^{-1}] + \mathbb{E}_{q_1}[z_j^2\sigma_j^{-1}], \\
\partial\mathbf{W}_{ik} &= -\mathbb{E}_{q_0}[x_i h'_k] + \mathbb{E}_{q_1}[x_i h'_k] - C_2 \sum_d (\mathbf{V}_{y_d k} - \mathbf{V}_{\bar{y}_d k})\mathbb{E}_{q_0}[x_i] \\
\partial\mathbf{U}_{jk} &= -\mathbb{E}_{q_0}[z_j h'_k] + \mathbb{E}_{q_1}[z_j h'_k] - C_2 \sum_d (\mathbf{V}_{y_d k} - \mathbf{V}_{\bar{y}_d k})\mathbb{E}_{q_0}[z_j],
\end{aligned}$$

其中$h'_k = \mathbf{x}^\top \mathbf{W}_{\cdot k} + \mathbf{z}^\top \mathbf{U}_{\cdot k}$，

$$\bar{y}_d = \arg\max_y [\Delta\ell_d(y) + \mathbf{V}^\top \mathbb{E}_{q_0}[\mathbf{f}(y, \mathbf{h}_d)] \tag{A-3}$$

为考虑代价损失函数情况下的预测结果。根据q_0的定义，期望$\mathbb{E}_{q_0}[x_i]$ 和$\mathbb{E}_{q_0}[z_j]$ 事实上分别是x_i和z_j的经验期望。有了上述次梯度，本文采用拟牛顿法自动确定梯度下降的步长，迭代若干次直至收敛或迭代次数大于某个阈值。

附录 B 基于有向贝叶斯网络的无限维隐空间模型

本附录简要介绍基于有向贝叶斯网络的无限维隐空间模型，包括两种正则化贝叶斯推理的具体实例，来进一步阐述上述正则化贝叶斯推理的理论框架。具体地说，本附录介绍包含无限维隐特征的最大间隔分类器，可用于学习单任务分类中的隐空间表示，或多任务学习中多种任务共享的隐含投影变换矩阵。

首先介绍单任务的分类模型。模型的基本设计为：已知D个数据样本，模型将每一样本$\mathbf{x}_d \in \mathcal{X} \subset \mathbb{R}^N$映射为一个隐含的二值[①]特征向量$\mathbf{z}_d$，这里用$\mathbf{Z}$表示所有数据样本的隐特征矩阵，其中$\mathbf{z}_d$对应于$\mathbf{Z}$的第$d$行，$\mathbf{Z}$中每一列表示一个隐含特征在不同样本上的具体表现值，其中$Z_{dk}=1$表示特征k在样本d中出现，否则，不出现。对于有限维隐特征模型，通常假设$\mathbf{Z}$具有有限个列（即特征）。如上讨论所述，有限维隐特征模型需要进行模型选择来确定隐空间的维度。为了自动确定隐空间的维度，这里使用非参数贝叶斯的方法，令$\mathbf{z}$为无穷维。为了确保有效隐特征维度的期望值有限，本章在特征矩阵$\mathbf{Z}$上定义广泛使用的非参数印度自助餐随机过程先验。

印度自助餐随机过程（Indian Buffet Process，简称IBP）在文献[93]中被提出。至今为止，IBP已被成功运用于诸多机器学习相关任务（例如社会网络链接分析[94]，多任务学习[95]等）中。第2章已经具体介绍了IBP的基本内容，包括IBP的 Stick-breaking表示。为了方便读者，这里再简单描述一下。令$\pi_k \in (0,1)$表示二值矩阵$\mathbf{Z}$中对应于第k列的参数。当π_k已知，每一个z_{dk}在第k列独立地从Bernoulli(π_k)中采样得到。参数$\boldsymbol{\pi}$从一个Stick-breaking 过程中产生，可表示为

$$\pi_1 = \nu_1, \text{ and } \pi_k = \nu_k \pi_{k-1} = \prod_{i=1}^{k} \nu_i, \tag{B-1}$$

其中$\nu_i \sim \text{Beta}(\alpha, 1)$。这个过程可以产生一个递减的概率序列$\pi_k$。具体地说，当观测到一个有限的数据集，特征$k$出现的概率随着$k$的增长成指数级下降。

B.1 无限维隐特征支持向量机

在此考虑广义的多类别分类问题，其中每一训练样本都有一个类别编号y，其中$y \in \mathcal{Y} \overset{\text{def}}{=} \{1, \cdots, L\}$。对于二分类问题和回归问题，可用相似步骤，从而定义相

① 实数值特征可以如文献[93]中很容易地考虑进来。

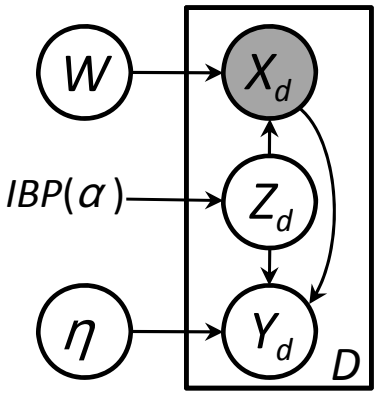


图 B.1 iLSVM 模型的图结构。

应的无限维隐特征支持向量机。具体地说，当隐特征$\mathbf{z}$已知，定义线性隐判别函数为

$$f(y,\mathbf{x},\mathbf{z};\boldsymbol{\eta}) \stackrel{\text{def}}{=} \boldsymbol{\eta}^{\top}\mathbf{g}(y,\mathbf{x},\mathbf{z}), \tag{B-2}$$

其中$\mathbf{g}(y,\mathbf{x},\mathbf{z})$为一个向量，由$L$个子向量拼接而成，其中第$y$个子向量是$\mathbf{z}$[①]，其他都为0。为了进行贝叶斯推理，本章需要保持隐特征$\mathbf{Z}$的完整概率分布。同时，为了对观测数据$\mathbf{x}$进行预测，需要去除$\mathbf{Z}$的不确定性。这里，定义有效判决函数为一个期望值[②]（即考虑所有$\mathbf{Z}$的可能值的带权平均值）。为了使模型为完全贝叶斯，这里将$\boldsymbol{\eta}$也定义为随机变量，然后从输入观测数据中推理后验分布$p(\mathbf{Z},\boldsymbol{\eta})$。具体地说，有效判别函数$f:\mathcal{X}\times\mathcal{Y}\mapsto\mathbb{R}$定义为

$$f(y,\mathbf{x};p(\mathbf{Z},\boldsymbol{\eta})) \stackrel{\text{def}}{=} \mathbb{E}_{p(\mathbf{Z},\boldsymbol{\eta})}[f(y,\mathbf{x},\mathbf{z};\boldsymbol{\eta})] = \mathbb{E}_{p(\mathbf{Z},\boldsymbol{\eta})}[\boldsymbol{\eta}^{\top}\mathbf{g}(y,\mathbf{x},\mathbf{z})]. \tag{B-3}$$

其中$p(\mathbf{Z},\boldsymbol{\eta})$是需要推理得到的后验分布。值得注意的是，尽管假设隐特征的维度可以为无穷，IBP随机过程先验可以确保当输入有限数量的可观测数据时，非零的隐特征数以概率“1”为有限的。另外，为了计算方便，通常会对可能特征的数量设置一个有限的截断上界T，基本要求是T足够大，读者可以参考B.3节和附录C了解详细内容。如文献[136]中所介绍，边缘分布的ℓ_1-距离截断误差随着T的增长呈指数级下降，因此，这种截断方法的近似精度是有理论保证的。

在上面定义的基础上，可以很容易地使用最大间隔学习准则定义问题 (5-3)中

① 在这个定义里，可以很方便地考虑输入特征$\mathbf{X}$或者$\mathbf{X}$的某些统计特性，例如可以在子向量中简单地将$\mathbf{x}$和$\mathbf{z}$拼接起来，定义一个由$\mathbf{x}$和$\mathbf{z}$共同决定的分类决策边界。

② 虽然也可以选用其他的定义（如求模等）。但相比之下，期望是概率分布的线性函数，计算简单；同时，期望比求模更加鲁棒[151]；而且它已被成功用于文献[40,58]中。

的最大间隔后验约束$\mathcal{P}_{\text{post}}(\boldsymbol{\xi})$为

$$\mathcal{P}^{c}_{\text{post}}(\boldsymbol{\xi}) \overset{\text{def}}{=} \left\{ p(\mathbf{Z},\boldsymbol{\eta}) \left| \begin{array}{c} \forall d \in \mathcal{I}_{\text{tr}} : f(y_d,\mathbf{x}_d;p(\mathbf{Z},\boldsymbol{\eta})) - f(y,\mathbf{x}_d;p(\mathbf{Z},\boldsymbol{\eta})) \geq \ell(y,y_d) - \xi_d, \forall y \\ \xi_d \geq 0 \end{array} \right. \right\} \tag{B-4}$$

以及惩罚函数$U^c(\boldsymbol{\xi}) \overset{\text{def}}{=} C\sum_{d\in\mathcal{I}_{\text{tr}}}\xi_n^p$，其中$p \geq 1$。如果$p = 1$，最小化$U^c(\boldsymbol{\xi})$等价于最小化预测准则 (B-7)中的铰链损失函数（或ℓ_1-loss）$\mathcal{R}^c_h$，其中

$$\mathcal{R}^c_h = C\sum_{d\in\mathcal{I}_{\text{tr}}} \max_y (f(y,\mathbf{x}_d;p(\mathbf{Z},\boldsymbol{\eta})) + \ell(y,y_d) - f(y_d,\mathbf{x}_d;p(\mathbf{Z},\boldsymbol{\eta}))) \tag{B-5}$$

若$p = 2$，此时的损失函数为ℓ_2-损失函数。为了方便讨论，本文在此考虑更常用的铰链损失函数。具体地说，本章使用非负的代价函数$\ell(y,y_d)$（例如0/1代价函数）用于度量将真实标签为y_d的数据$\mathbf{x}_d$预测为y时的代价值。$\mathcal{I}_{\text{tr}}$表示训练数据的索引集合。

为了更加鲁棒地估计隐矩阵$\mathbf{Z}$，需要合理数量的观测数据。因此，定义一个似然模型将$\mathbf{Z}$与观测数据$\mathbf{x}$相关联，以提供尽量多的观测数据。这里考虑实值的输入特征，相应地，定义一个关于$\mathbf{X}$的线性高斯似然模型

$$p(\mathbf{x}_d|\mathbf{z}_d,\mathbf{W},\sigma^2_{d0}) = \mathcal{N}(\mathbf{x}_d|\mathbf{W}\mathbf{z}_d^{\top},\sigma^2_{d0}I), \tag{B-6}$$

其中$\mathbf{W}$表示随机变量矩阵，I表示有适当维度的单位矩阵。同时，假设$\mathbf{W}$服从高斯先验分布，即$\pi(\mathbf{W}) = \prod_d \mathcal{N}(\mathbf{w}_d|0,\sigma_0^2 I)$，各变量间的关系请见图B.1所示iLSVM模型的图结构。参数σ_0^2和σ^2_{d0}可预先指定，或从观测数据中估计得到（具体内容请参考附录C.2）。

测试: 为了预测测试数据的类别标签，这里将训练和测试数据放在一起进行正则化贝叶斯推理。区别在于，训练样本有真实类别标签，因此，每个训练数据都对应一组如上定义的最大间隔约束；而对于测试样本，由于它们的真实类别标签是未知的，所以他们不对应任何最大间隔约束进行推理。在推理之后，使用如下预测准则实现类别标签的预测

$$y^* \overset{\text{def}}{=} \arg\max_y f(y,\mathbf{x};p(\mathbf{Z},\boldsymbol{\eta})). \tag{B-7}$$

上述方法之所以能够利用训练数据的信息，对测试数据进行合理的类别预测的原因在于，所有数据都使用相同的$\boldsymbol{\eta}$以及IBP先验。当然也可以将问题表示为一个直推式（Transductive）推理问题，但是需要在测试数据中引入后验约束[152]。因此，这样得到的问题通常更难求解。

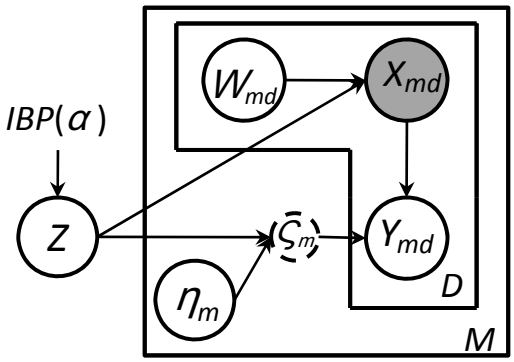
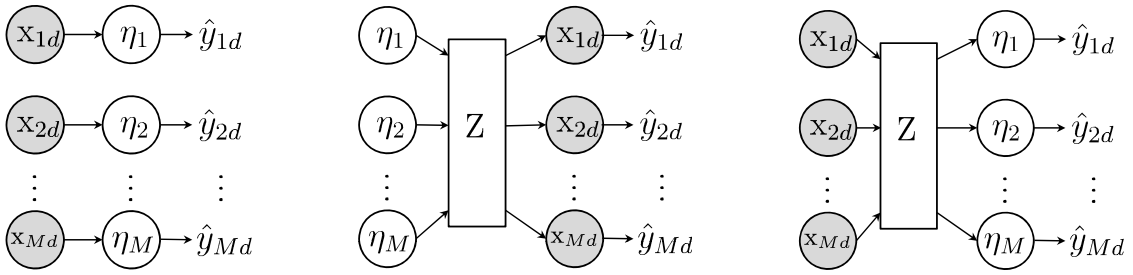


图 B.2 MT-iLSVM 的图结构。图中虚线节点（即 ς_m）用来展示任务之间的关系。为了表示方便，在此省略**W**和$\boldsymbol{\eta}$的先验。



(a) 多个独立单任务学习　(b) MT-iLSVM第一种理解　(c) MT-iLSVM第二种理解

图 B.3 （a）多个独立的单任务学习，其中每一个任务m (表示为模型$\boldsymbol{\eta}_m$) 互相独立地处理输入数据**x**;（b）多任务MT-iLSVM中多个任务关联方式的第一种理解，其中所有的M个任务的模型都需要经过一个共同的转换（即矩阵**Z**）才能处理输入数据**x**;（c）多任务MT-iLSVM中多个任务关联方式的第二种理解，其中输入数据**x**通过多任务共享的隐投影矩阵投影至不同的隐层空间表示。

B.2 多任务无限维隐特征支持向量机

如上所述，分类问题一般表示为一个单任务学习问题。与之不同，多任务学习的目的是通过共享多种任务的统计属性提高所有（或部分）相关任务的性能。对于多任务学习的研究，文献[153]介绍了很多不同的方法。特别地，学习一个所有相关的任务共享的隐空间表示，被认为是一种可以捕获任务之间关系的有效方法[95,134,154]。下面，本章提出多任务无限维隐特征支持向量机（MT-iLSVM）模型，该模型可用于学习多个任务共享的二值投影变换矩阵**Z**，从而捕获多任务的相关性。与iLSVM模型相似，为了自动确定投影矩阵**Z**的维度，本文允许**Z**有无限列，并且在矩阵**Z**中引入IBP先验。

假设有M个相关任务。令$\mathcal{D}_m = \{(\mathbf{x}_{md}, y_{md})\}_{d\in\mathcal{I}_{\text{tr}}^m}$表示第$m$个任务的训练样本集合。在此考虑二分类问题，$\mathcal{Y}_m = \{+1, -1\}$。结合iLSVM的思想，下面方法可以很容易地扩展到多类别分类或者回归问题。为了实现在多个任务上的学习和预测，图B.3(a)展示了一种粗略的不考虑多个任务之间关联的简单方法，即学习多个独立的单任务。为了使多个任务关联在一起并且能共享有用的统计特性，本文在MT-iLSVM模型中引入一个隐投影矩阵$\mathbf{Z}$。当隐矩阵$\mathbf{Z}$已知，定义第m个任务的线性隐判别函数为

$$f_m(\mathbf{x}_{md}, \mathbf{Z}; \boldsymbol{\eta}_m) \overset{\text{def}}{=} (\mathbf{Z}\boldsymbol{\eta}_m)^\top \mathbf{x}_{md} = \boldsymbol{\eta}_m^\top(\mathbf{Z}^\top \mathbf{x}_{md}). \tag{B-8}$$

上述定义可以从两个角度理解M个任务是如何相关联的，分别对应公式（B-8）中的第二和第三项。

（1） 如图B.3(b)所示，若令$\varsigma_m = \mathbf{Z}\boldsymbol{\eta}_m$，那么$\varsigma_m$可以理解为任务$m$的实际模型参数，因此所有不同任务中的模型$\varsigma_m$都通过共享相同的隐矩阵$\mathbf{Z}$而关联在一起；

（2） 如图B.3(c)所示，每一个任务m都有各自的模型参数$\boldsymbol{\eta}_m$，但所有任务都共享相同的隐含投影矩阵$\mathbf{Z}$来抽取有效特征$\mathbf{Z}^\top\mathbf{x}_{md}$（即输入特征$\mathbf{x}_{md}$的投影）。

根据上述定义和解释，此方法可以看作是多任务学习框架ASO（Alternating Structure Optimization）[154]的非参数化贝叶斯扩展，传统的ASO方法学习一个事先指定维数的未知投影变换矩阵来实现多个任务之间的关联。另外，与文献[153]中学习一个事先设定维度的二值向量来选择输入数据$\mathbf{x}$的特征或核不同，这里用非参数贝叶斯的方法来学习一个无限维投影矩阵$\mathbf{Z}$。

如iLSVM模型中所述，本章研究完全贝叶斯的推理方法（即$\boldsymbol{\eta}_m$也是随机变量），然后对每一种任务m都定义一个基于期望的有效判决函数

$$f_m(\mathbf{x}_{md}; p(\mathbf{Z}, \boldsymbol{\eta})) \overset{\text{def}}{=} \mathbb{E}_{p(\mathbf{Z},\boldsymbol{\eta})}[f_m(\mathbf{x}_{md}, \mathbf{Z}; \boldsymbol{\eta}_m)] = \mathbb{E}_{p(\mathbf{Z},\boldsymbol{\eta})}[\mathbf{Z}\boldsymbol{\eta}_m]^\top \mathbf{x}_{md}. \tag{B-9}$$

于是，任务m的预测准则可以很自然地写成$y_{md}^* \overset{\text{def}}{=} \text{sign} f_m(\mathbf{x}_{md})$。相似的，在正则化贝叶斯推理框架下，定义如下面所示的最大间隔后验约束，并定义$U^{MT}(\boldsymbol{\xi}) \overset{\text{def}}{=} C\sum_{m,d\in\mathcal{I}_{\text{tr}}^m}\xi_{md}$，

$$\mathcal{P}_{\text{post}}^{MT}(\boldsymbol{\xi}) \overset{\text{def}}{=} \left\{ p(\mathbf{Z}, \boldsymbol{\eta}) \left| \begin{array}{c} \forall m,\ \forall d \in \mathcal{I}_{\text{tr}}^m :\ y_{md}\mathbb{E}_{p(\mathbf{Z},\boldsymbol{\eta})}[\mathbf{Z}\boldsymbol{\eta}_m]^\top \mathbf{x}_{md} \geq 1 - \xi_{md} \\ \xi_{md} \geq 0 \end{array} \right. \right\}. \tag{B-10}$$

与iLSVM模型相似，最小化$U^{MT}(\boldsymbol{\xi})$等价于最小化多个二分类预测准则的铰链损失函数$\mathcal{R}_h^{MT}$，其中

$$\mathcal{R}_h^{MT} = C \sum_{m,d \in \mathcal{I}_{\text{tr}}^m} \max(0, 1 - y_{md} \mathbb{E}_{p(\mathbf{Z},\boldsymbol{\eta})}[\mathbf{Z}\boldsymbol{\eta}_m]^\top \mathbf{x}_{md}) \tag{B-11}$$

此外，为了使用更多数据来鲁棒地估计隐矩阵$\mathbf{Z}$，对于实值的输入数据，定义高斯似然函数模型

$$p(\mathbf{x}_{md}|\mathbf{w}_{md}, \mathbf{Z}, \lambda_{md}^2) = \mathcal{N}(\mathbf{x}_{md}|\mathbf{Z}\mathbf{w}_{md}, \lambda_{md}^2 I), \tag{B-12}$$

其中$\mathbf{w}_{md}$为一个向量，它的先验分布为$\pi(\mathbf{W}) = \prod_{md} \mathcal{N}(\mathbf{w}_{md}|0, \sigma_{m0}^2 I)$。各变量间的关系读者请参考图B.2所示的MT-iLSVM模型的图结构。

测试过程中，可使用与iLSVM模型相同的策略在训练和测试数据上做贝叶斯推理。区别是训练数据满足最大间隔约束，而测试数据不满足。相似的，超参数σ_{m0}^2和λ_{md}^2可事先指定，或者从数据中估计出来（详情请参考本文附录C.1）。

B.3 基于截断均值场约束的推理方法

下面简要讨论在MT-iLSVM模型中如何求解有最大间隔后验约束的正则化贝叶斯推理[①]问题。为了简化问题，这里使用IBP的 Stick-breaking表示，引入辅助变量$\boldsymbol{\nu}$，推理后验分布$p(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta})$。另外，引入截断均值场约束，即

$$p(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta}) = p(\boldsymbol{\eta}) \prod_{k=1}^{T} \Big(p(\nu_k|\boldsymbol{\gamma}_k) \prod_{n=1}^{N} p(z_{nk}|\psi_{nk}) \Big) \prod_{md} p(\mathbf{w}_{md}|\Phi_{md}, \sigma_{md}^2 I), \tag{B-13}$$

其中T为截断上界；$p(\mathbf{w}_{md}|\Phi_{md}, \sigma_{md}^2 I) = \mathcal{N}(\mathbf{w}_{md}|\Phi_{md}, \sigma_{md}^2 I)$；$p(z_{nk}|\psi_{nk}) = \text{Bernoulli}(\psi_{nk})$；$p(\nu_k|\boldsymbol{\gamma}_k) = \text{Beta}(\gamma_{k1}, \gamma_{k2})$。

首先使用拉格朗日方法，将有约束的优化问题转化为一个寻找驻点的问题。对于每一个最大间隔约束都引入拉格朗日乘子$\boldsymbol{\omega}$，对于$\boldsymbol{\xi}$的非负性约束引入拉格朗日乘子$\mathbf{u}$。令$L(p, \boldsymbol{\xi}, \boldsymbol{\omega}, \mathbf{u})$为拉格朗日函数。于是推理问题可以通过迭代地求解下面两步来求解（具体推导请参考附录C.1）

推理 $p(\boldsymbol{\nu})$，$p(\mathbf{W})$，及$p(\mathbf{Z})$： 对于$p(\mathbf{W})$，由于先验分布和似然函数都是正态分布，可以很容易地推导Φ_{md}和σ_{md}^2的迭代公式[24]。对于$p(\boldsymbol{\nu})$，因为后验约束不直接依赖于它，于是可以得到与文献[136]相同的迭代公式（具体推导请参考附录C.1）。下面着重介绍$p(\mathbf{Z})$的推理过程，以及最大间隔约束如何在推理隐矩阵$\mathbf{Z}$的过程中

① iLSVM模型的正则化贝叶斯推理过程相似

起到正则化作用的原理。由于最大间隔约束是$p(\mathbf{Z})$的线性函数，可以得到均值场迭代公式为

$$\psi_{nk} = \frac{1}{1 + e^{-\vartheta_{nk}}} \tag{B-14}$$

其中

$$\begin{aligned}\vartheta_{nk} = \sum_{j=1}^{k} \mathbb{E}_p[\log v_j] - \mathcal{L}_k^{\gamma} - \sum_{md} \frac{1}{2\lambda_{md}^2}\Big((K\sigma_{md}^2 + (\phi_{md}^k)^2) \\ -2x_{md}^n \phi_{md}^k + 2\sum_{j\neq k} \phi_{md}^j \phi_{md}^k \psi_{nj}\Big) + \sum_{m,d\in\mathcal{I}_{\mathrm{tr}}^m} y_{md}\mathbb{E}_p[\eta_{mk}]x_{md}^n,\end{aligned} \tag{B-15}$$

这里$\mathcal{L}_k^{\gamma}$为$\mathbb{E}_p[\log(1 - \prod_{j=1}^{k} v_j)]$的下界（具体推导请参考附录C.1）。$\vartheta_{nk}$的最后一项是由于使用公式 (B-10)中定义的最大间隔后验约束的结果。

推理$p(\boldsymbol{\eta})$，并求解$\boldsymbol{\omega}$和$\boldsymbol{\xi}$： 只考虑后验分布$p(\boldsymbol{\eta})$优化L，可以得到$p(\boldsymbol{\eta}) = \prod_m p(\boldsymbol{\eta}_m)$，且

$$p(\boldsymbol{\eta}_m) \propto \pi(\boldsymbol{\eta}_m) \exp\{\boldsymbol{\eta}_m^{\top}\boldsymbol{\mu}_m\},$$

其中$\boldsymbol{\mu}_m = \sum_{d\in\mathcal{I}_{\mathrm{tr}}^m} y_{md}\omega_{md}(\boldsymbol{\psi}^{\top}\mathbf{x}_{md})$。这里，假设先验分布$\pi(\boldsymbol{\eta}_m)$服从标准正态分布。于是，可以得到后验分布$p(\boldsymbol{\eta}_m) = \mathcal{N}(\boldsymbol{\eta}_m|\boldsymbol{\mu}_m, I)$。将$p(\boldsymbol{\eta})$的解代入$L$，即可得到$M$个独立的对偶问题

$$\begin{aligned}&\max_{\boldsymbol{\omega}_m} \ -\frac{1}{2}\boldsymbol{\mu}_m^{\top}\boldsymbol{\mu}_m + \sum_{d\in\mathcal{I}_{\mathrm{tr}}^m} \omega_{md} \\ &\text{s.t..}: \ 0 \leq \omega_{md} \leq 1, \forall d \in \mathcal{I}_{\mathrm{tr}}^m\end{aligned} \tag{B-16}$$

此对偶问题（或其原问题）可用一个二分类SVM分类器的学习工具包（如SVM-light）[①] 很高效地求解。

① http://svmlight.joachims.org/svm_multiclass.html

附录 C 无限维隐特征支持向量机的推理算法

本附录介绍无限维隐特征支持向量机分类模型及多任务学习模型的推理算法。

B.1 多任务无限维隐特征支持向量机模型的推理算法

下面首先推导多任务无限维隐特征支持向量机（MT-iLSVM）的推理算法，其基本框架如Algorithm 3中所示。

具体地说，MT-iLSVM模型中$\mathcal{M}$包含所有隐变量$(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta})$。令

$$L_{md}(p) \stackrel{\text{def}}{=} \mathbb{E}_p[\log p(\mathbf{x}_{md}|\mathbf{Z}, \mathbf{w}_{md}, \lambda_{md}^2)] \tag{C-1}$$

表示问题（5-3）中的期望数据似然。于是，根据截断均值场假设 公式(B-13)，可以得到

$$L_{md}(p) = -\frac{\mathbf{x}_{md}^\top \mathbf{x}_{md} - 2\mathbf{x}_{md}^\top \mathbb{E}_p[\mathbf{Z}\mathbf{w}_{md}] + \mathbb{E}_p[\mathbf{w}_{md}^\top \mathbf{U}\mathbf{w}_{md}]}{2\lambda_{md}^2} - \frac{D\log(2\pi\lambda_{md}^2)}{2},$$

其中$\mathbf{x}_{md}^\top \mathbb{E}_p[\mathbf{Z}\mathbf{w}_{md}] = \sum_k \mathbf{x}_{md}^\top \boldsymbol{\psi}_{.k}$；$\boldsymbol{\psi}_{.k} \stackrel{\text{def}}{=} (\psi_{1k} \cdots \psi_{Nk})^\top$为$\boldsymbol{\psi} = \mathbb{E}[\mathbf{Z}]$的第$k$列，且

$$\mathbb{E}_p[\mathbf{w}_{md}^\top \mathbf{U}\mathbf{w}_{md}] = 2\sum_{j<k} \phi_{md}^j \phi_{md}^k \mathbf{U}_{jk} + \sum_k \mathbf{U}_{kk}(K\sigma_{md}^2 + \Phi_{md}^\top \Phi_{md}).$$

而$\mathbf{U} \stackrel{\text{def}}{=} \mathbb{E}[\mathbf{Z}^\top \mathbf{Z}]$是一个$K \times K$的参数矩阵，其中每一个元素是

$$\mathbf{U}_{ij} = \begin{cases} \sum_n \psi_{ni}, & \text{若 } i = j \\ \sum_n \psi_{ni}\psi_{nj}, & \text{否则} \end{cases}$$

对于问题（5-3）中的KL散度一项，可以写为

$$\begin{aligned} \mathrm{KL}(p(\mathcal{M})||\pi(\mathcal{M})) = \mathrm{KL}(p(\boldsymbol{\nu})||\pi(\boldsymbol{\nu})) + \mathrm{KL}(p(\mathbf{W})||\pi(\mathbf{W})) \\ + \mathrm{KL}(p(\mathbf{Z})||\pi(\mathbf{Z})) + \mathrm{KL}(p(\boldsymbol{\eta})||\pi(\boldsymbol{\eta})), \end{aligned}$$

其中每一项分别为

$$\begin{aligned} \mathrm{KL}(p(\boldsymbol{\nu})||\pi(\boldsymbol{\nu})) = \sum_{k=1}^K \Big(& (\gamma_{k1} - \alpha)(\psi(\gamma_{k1}) - \psi(\gamma_{k1} + \gamma_{k2})) + (\gamma_{k2} - 1)(\psi(\gamma_{k2}) - \psi(\gamma_{k1} + \gamma_{k2})), \\ & - \log \frac{\Gamma(\gamma_{k1})\Gamma(\gamma_{k2})}{\Gamma(\gamma_{k1} + \gamma_{k2})} \Big) - K\log\alpha, \end{aligned}$$

$$\begin{aligned}\mathrm{KL}(p(\mathbf{Z})\|\pi(\mathbf{Z})) = \sum_{nk}\Big(&-\psi_{nk}\sum_{j=1}^{k}\mathbb{E}_p[\log \nu_j] - (1-\psi_{nk})\mathbb{E}_p[\log(1-\prod_{j=1}^{k}\nu_j)],\\ &+\psi_{nk}\log\psi_{nk} + (1-\psi_{nk})\log(1-\psi_{nk})\Big)\\ \mathrm{KL}(p(\mathbf{W})\|\pi(\mathbf{W})) = \sum_{md}\Big(&\frac{K\sigma_{md}^2 + \Phi_{md}^\top\Phi_{md}}{2\sigma_{m0}^2} - \frac{K(1+\log\frac{\sigma_{md}^2}{\sigma_{m0}^2})}{2}\Big).\end{aligned}$$

公式中$\psi(\cdot)$表示一个digamma函数，$\mathbb{E}_p[\log \nu_j] = \psi(\gamma_{j1}) - \psi(\gamma_{j1}+\gamma_{j2})$。对于$\mathrm{KL}(p(\boldsymbol{\eta})\|\pi(\boldsymbol{\eta}))$，下面的推导过程可以直接处理，这里不需要明确写出其计算表达式。同时，根据上述均值场假设，有效判别函数可以写成

$$f_m(\mathbf{x}_{md}; p(\mathbf{Z},\boldsymbol{\eta})) = \boldsymbol{\eta}_m^\top\boldsymbol{\psi}^\top\mathbf{x}_{md} = \sum_{k=1}^{K}\mathbb{E}_p[\eta_{mk}]\boldsymbol{\psi}_{\cdot k}^\top\mathbf{x}_{md}. \tag{C-2}$$

在上述各项中，除了$\mathbb{E}_p[\log(1-\prod_{j=1}^{k}\nu_j)]$项以外，所有其他项都可很容易地计算。这里采用多元变分下界（Multivariate Variational Lower Bound）[136]来近似$\mathbb{E}_p[\log(1-\prod_{j=1}^{k}\nu_j)]$

$$\begin{aligned}\mathbb{E}_p[\log(1-\prod_{j=1}^{k}\nu_j)] \geq &\sum_{m=1}^{k}q_{km}\psi(\gamma_{m2}) + \sum_{m=1}^{k-1}(\sum_{n=m+1}^{k}q_{kn})\psi(\gamma_{m1})\\ &-\sum_{m=1}^{k}(\sum_{n=m}^{k}q_{kn})\psi(\gamma_{m1}+\gamma_{m2}) + \mathcal{H}(q_{k\cdot}),\end{aligned}$$

其中变分参数为$q_{k\cdot} = (q_{k1}\cdots q_{kk})^\top$属于$k$维单纯形（$k$-Simplex），而$\mathcal{H}(q_{k\cdot})$表示$q_{k\cdot}$的熵。通过设置最优的$q_{k\cdot}$值可以得到最紧的下界

$$q_{km} = \frac{1}{Z_k}\exp\Big(\psi(\gamma_{m2}) + \sum_{n=1}^{m-1}\psi(\gamma_{n1}) - \sum_{n=1}^{m}\psi(\gamma_{n1}+\gamma_{n2})\Big), \tag{C-3}$$

其中Z_k表示归一化因子，可以保证$q_{k\cdot}$是概率分布。记最紧的下界为$\mathcal{L}_k^\nu$。这里用下界$\mathcal{L}_k^\nu$代替$\mathbb{E}_p[\log(1-\prod_{j=1}^{k}\nu_j)]$一项，于是得到$\mathrm{KL}(p(\mathcal{M})\|\pi(\mathcal{M}))$的上界$\mathcal{L}(p)$。

有了上面各项及上界$\mathcal{L}(p)$，就可以使用拉格朗日方法，对每一间隔约束引入拉格朗日乘子$\boldsymbol{\omega}$，对非负约束$\boldsymbol{\xi}$引入拉格朗日乘子$\mathbf{u}$。于是得到拉格朗日函数为

$$L(p,\boldsymbol{\xi},\boldsymbol{\omega},\mathbf{u}) = \mathcal{L}(p) - \sum_{md}L_{md}(p) - \sum_{m,d\in\mathcal{I}_{\mathrm{tr}}^m}\omega_{md}\Big(y_{md}(\mathbb{E}_p[\boldsymbol{\eta}_m]^\top\boldsymbol{\psi}^\top\mathbf{x}_{md}) - 1 + \xi_{md}\Big) - \mathbf{u}^\top\boldsymbol{\xi}.$$

推理过程即为迭代地求解下面几步

推理$p(\boldsymbol{\nu})$，$p(\mathbf{Z})$和$p(\mathbf{W})$: 对于$p(\mathbf{W})$，由于先验$\pi(\mathbf{W})$和$p(\mathbf{W})$都为高斯分布，可

Algorithm 3 MT-iLSVM模型的推理算法

输入: 数据$\mathcal{D}=\{(\mathbf{x}_{md},y_{md})\}_{m,d\in\mathcal{I}_{\text{tr}}^m}\cup\{\mathbf{x}_{md}\}_{m,d\in\mathcal{I}_{\text{tst}}^m}$，常量$\alpha$和$C$

输出: 分布$p(\boldsymbol{\nu})$, $p(\mathbf{Z})$, $p(\mathbf{W})$, $p(\boldsymbol{\eta})$ 以及超参数σ_{m0}^2 和λ_{md}^2

初始化$\gamma_{k1}=\alpha$，$\gamma_{k2}=1$，$\psi_{nk}=0.5+\epsilon$，其中$\epsilon\sim\mathcal{N}(0,0.001)$，$\Phi_{md}=0$，$\sigma_{md}^2=\sigma_{m0}^2=1,\boldsymbol{\mu}_m=0$，$\lambda_{md}^2$从$\mathcal{D}$中计算得到。

repeat

  repeat

    $\forall 1\le k\le K$，通过公式(C-5)更新参数$(\gamma_{k1},\gamma_{k2})$;

    $\forall m,\forall d,\forall 1\le k\le K$，通过公式(C-4)更新参数$\phi_{md}^k$和$\sigma_{md}^2$;

    $\forall 1\le n\le N,\forall 1\le k\le K$，通过公式(C-6)更新$\psi_{nk}$;

  until L的相对变化小于τ (如$1e^{-3}$) 或者迭代次数为T (如10)

  for $m=1$ **to** M **do**

    使用二分类SVM分类器求解对偶问题 (C-7);

  end for

  通过公式 (C-8)更新超参数σ_{m0}^2，通过公式 (C-9)更新λ_{md}^2；(*可选*)

until L的相对变化小于τ' (如$1e^{-4}$) 或者迭代次数T'（如20）。

以很容易地推导其更新准则

$$\begin{aligned}\phi_{md}^k&=\frac{\sum_n x_{md}^n\psi_{nk}-\sum_{j\neq k}\phi_{md}^j\mathbf{U}_{kj}}{\lambda_{md}^2}\Big(\frac{1}{\sigma_{m0}^2}+\frac{\sum_n\psi_{nk}}{\lambda_{md}^2}\Big)^{-1}\\ \sigma_{md}^2&=\Big(\frac{1}{\sigma_{m0}^2}+\frac{1}{K}\sum_k\frac{\mathbf{U}_{kk}}{\lambda_{md}^2}\Big)^{-1}\end{aligned}\tag{C-4}$$

对于$p(\boldsymbol{\nu})$，可以得到与文献[136]中相似的更新准则

$$\begin{aligned}\gamma_{k1}&=\alpha+\sum_{m=k}^{K}\sum_{n=1}^{N}\psi_{nm}+\sum_{m=k+1}^{K}(D-\sum_{n=1}^{N}\psi_{nm})(\sum_{i=k+1}^{m}q_{mi})\\ \gamma_{k2}&=1+\sum_{m=k}^{K}(D-\sum_{n=1}^{N}\psi_{nm})q_{mk}.\end{aligned}\tag{C-5}$$

对于$p(\mathbf{Z})$，其均值场更新公式为

$$\psi_{nk}=\frac{1}{1+e^{-\vartheta_{nk}}},\tag{C-6}$$

其中

$$\vartheta_{nk}=\sum_{j=1}^{k}\mathbb{E}_p[\log v_j]-\mathcal{L}_k^{\nu}-\sum_{md}\frac{1}{2\lambda_{md}^2}\Big((K\sigma_{md}^2+(\phi_{md}^k)^2)$$

$$-2x_{md}^{n}\phi_{md}^{k}+2\sum_{j\neq k}\phi_{md}^{j}\phi_{md}^{k}\psi_{nj}\Big)+\sum_{m,d\in\mathcal{I}_{\mathrm{tr}}^{m}}y_{md}\mathbb{E}_{p}[\eta_{mk}]x_{md}^{n}$$

推理$p(\boldsymbol{\eta})$ 并求解$\boldsymbol{\omega}$和$\boldsymbol{\xi}$: 只考虑$q(\boldsymbol{\eta})$优化拉格朗日函数L，可以得到

$$p(\boldsymbol{\eta})\propto\pi(\boldsymbol{\eta})\exp\Big\{\sum_{m,d\in\mathcal{I}_{\mathrm{tr}}^{m}}y_{md}\omega_{md}\boldsymbol{\eta}_{m}^{\top}\boldsymbol{\psi}^{\top}\mathbf{x}_{md}\Big\}=\prod_{m=1}^{M}\pi(\boldsymbol{\eta}_{m})\exp\Big\{\boldsymbol{\eta}_{m}^{\top}(\sum_{d\in\mathcal{I}_{\mathrm{tr}}^{m}}y_{md}\omega_{md}\boldsymbol{\psi}^{\top}\mathbf{x}_{md})\Big\}.$$

因此，即使没有假设$p(\boldsymbol{\eta})$为因子化的，仍可推导出因子化形式$p(\boldsymbol{\eta})=\prod_{m}p(\boldsymbol{\eta}_{m})$，其中

$$p(\eta_{m})\propto\pi(\boldsymbol{\eta}_{m})\exp\Big\{\boldsymbol{\eta}_{m}^{\top}(\sum_{d\in\mathcal{I}_{\mathrm{tr}}^{m}}y_{md}\omega_{md}\boldsymbol{\psi}^{\top}\mathbf{x}_{md})\Big\}.$$

这里假设$\pi(\boldsymbol{\eta}_{m})$服从标准正态分布。于是，可以得到$p(\boldsymbol{\eta}_{m})=\mathcal{N}(\boldsymbol{\eta}_{m}|\boldsymbol{\mu}_{m},I)$，其中

$$\boldsymbol{\mu}_{m}=\sum_{d\in\mathcal{I}_{\mathrm{tr}}^{m}}y_{md}\omega_{md}\boldsymbol{\psi}^{\top}\mathbf{x}_{md}.$$

将$p(\boldsymbol{\eta})$的解带入拉格朗日函数L，可以得到M个独立的对偶问题

$$\begin{aligned}&\max_{\boldsymbol{\omega}_{m}}\ -\frac{1}{2}\boldsymbol{\mu}_{m}^{\top}\boldsymbol{\mu}_{m}+\sum_{d\in\mathcal{I}_{\mathrm{tr}}^{m}}\omega_{md}\\&\forall d\in\mathcal{I}_{\mathrm{tr}}^{m},\ \text{s.t..}:\ 0\leq\omega_{md}\leq 1,\end{aligned}\tag{C-7}$$

其中该原问题（及其对偶问题）可使用二分类SVM工具（如SVM-Light）高效地求解。

最后，对于超参数σ_{0}^{2} 和 λ_{md}^{2}，可以选择事先确定或从数据中估计得到。这里采用经验估计方法，根据如上所述的解析解得到

$$\sigma_{m0}^{2}=\frac{\sum_{d=1}^{N_{m}}(K\sigma_{md}^{2}+\Phi_{md}^{\top}\Phi_{md})}{KN_{m}}\tag{C-8}$$

$$\lambda_{md}^{2}=\frac{\mathbf{x}_{md}^{\top}\mathbf{x}_{md}-2\mathbf{x}_{md}^{\top}\mathbb{E}_{p}[\mathbf{Z}\mathbf{w}_{md}]+\mathbb{E}_{p}[\mathbf{w}_{md}^{\top}\mathbf{U}\mathbf{w}_{md}]}{N}.\tag{C-9}$$

B.2 无限维隐特征支持向量机模型的推理算法

本小节介绍基于IBP先验的Stick Breaking表示的iLSVM模型的推理算法，其基本框架如Algorithm 4。

与MT-iLSVM模型的推理相似，这里仍然采用截断均值场假设

$$p(\boldsymbol{\nu},\mathbf{W},\mathbf{Z},\boldsymbol{\eta})=p(\boldsymbol{\eta})p(\mathbf{W}|\Phi,\Sigma)\prod_{d}\Big(\prod_{k=1}^{K}p(z_{dk}|\psi_{dk})\Big)\prod_{k=1}^{K}p(\nu_{k}|\boldsymbol{\gamma}_{k}),\tag{C-10}$$

其中K为截断上界；各因子分布为

$$\begin{aligned} p(\mathbf{W}|\Phi,\Sigma) &= \prod_k \mathcal{N}(\mathbf{W}_{\cdot k}|\Phi_{\cdot k},\sigma_k^2 I) \\ p(z_{dk}|\phi_{dk}) &= \text{Bernoulli}(\phi_{dk}) \\ p(v_k|\boldsymbol{\gamma}_k) &= \text{Beta}(\gamma_{k1},\gamma_{k2}). \end{aligned} \tag{C-11}$$

于是可以使用拉格朗日方法来求解这个有约束问题，对于每一最大间隔约束都引入拉格朗日乘子$\boldsymbol{\omega}$，对非负性约束$\boldsymbol{\xi}$引入拉格朗日乘子 $\mathbf{u}$。相似的，令$L_d(p) \stackrel{\text{def}}{=} \mathbb{E}_p[\log p(\mathbf{x}_d|\mathbf{z}_d,\mathbf{W})]$，可以得到

$$L_d(p) = -\frac{\mathbf{x}_d^\top\mathbf{x}_d - 2\mathbf{x}_d^\top\Phi\mathbb{E}_p[\mathbf{z}_d]^\top + \mathbb{E}_p[\mathbf{z}_d\mathbf{A}\mathbf{z}_d^\top]}{2\sigma_{d0}^2} - \frac{N\log(2\pi\sigma_{d0}^2)}{2}, \tag{C-12}$$

其中$\mathbf{A} \stackrel{\text{def}}{=} \mathbb{E}_p[\mathbf{W}^\top\mathbf{W}]$表示$K\times K$矩阵；$\mathbf{x}_d^\top\Phi\mathbb{E}_p[\mathbf{z}_d]^\top = 2\sum_k \psi_{dk}(\mathbf{x}_d^\top\Phi_{\cdot k})$；而且

$$\mathbb{E}_p[\mathbf{z}_d\mathbf{A}\mathbf{z}_d^\top] = 2\sum_{j<k}\psi_{dj}\psi_{dk}\mathbf{A}_{jk} + \sum_k \psi_{dk}(N\sigma_k^2 + \mathbf{A}_{kk}).$$

有效判别函数为

$$f(y,\mathbf{x}_d) = \sum_k \mathbb{E}_p[\eta_y^k]\psi_{dk} \tag{C-13}$$

为了满足计算的可行性，这里使用下界$\mathcal{L}_k^v$来近似$\mathbb{E}_p[\log(1-\prod_{j=1}^k v_j)]$项。基于这个下界，可以得到一个KL散度项的上界，将拉格朗日函数记为$L(p,\boldsymbol{\xi},\boldsymbol{\omega},\mathbf{u})$。于是，推理过程即为交替地求解下面几步

推理$p(\boldsymbol{v})$, $p(\mathbf{Z})$ 和$p(\mathbf{W})$: 其中$p(\mathbf{W})$有如下更新准则

$$\begin{aligned} \Phi_{\cdot k} &= \sum_d \frac{\psi_{dk}}{\sigma_{d0}^2}\Big(\mathbf{x}_d - \sum_{j\neq k}\psi_{dj}\Phi_{\cdot j}\Big)\Big(1+\sum_d\frac{\psi_{dk}}{\sigma_{d0}^2}\Big)^{-1} \\ \sigma_k^2 &= \Big(1+\sum_d\frac{\psi_{dk}}{\sigma_{d0}^2}\Big)^{-1}. \end{aligned} \tag{C-14}$$

对于$p(\boldsymbol{v})$，可以得到与文献[136]中相似的更新准则，即

$$\begin{aligned} \gamma_{k1} &= \alpha + \sum_{m=k}^K\sum_{d=1}^D\psi_{dm} + \sum_{m=k+1}^K(D - \sum_{d=1}^D\psi_{dm})(\sum_{i=k+1}^m q_{mi}) \\ \gamma_{k2} &= 1 + \sum_{m=k}^K(N - \sum_{d=1}^D\psi_{dm})q_{mk}, \end{aligned} \tag{C-15}$$

其中$q_{\cdot k}$与公式(C-3)中相同。对于$p(\mathbf{Z})$，ψ的均值场更新公式为

$$\psi_{dk} = \frac{1}{1+e^{-\vartheta_{dk}}}, \tag{C-16}$$

Algorithm 4 iLSVM推理算法

输入: 数据$\mathcal{D}=\{(\mathbf{x}_n,y_n)\}_{n\in\mathcal{I}_{\text{tr}}}\cup\{\mathbf{x}_n\}_{n\in\mathcal{I}_{\text{tst}}}$, 常数$\alpha$和$C$

输出: 分布$p(\boldsymbol{\nu})$, $p(\mathbf{Z})$, $p(\mathbf{W})$, $p(\boldsymbol{\eta})$和超参数σ_0^2以及σ_{n0}^2

初始化$\gamma_{k1}=\alpha,\gamma_{k2}=1,\psi_{nk}=0.5+\epsilon$，其中$\epsilon\sim\mathcal{N}(0,0.001),\Phi_{\cdot k}=0,\sigma_k^2=\sigma_0^2=1,\boldsymbol{\mu}=0,\sigma_{n0}^2$ 都通过$\mathcal{D}$计算。

repeat

  repeat

    $\forall 1\le k\le K$，通过公式 (C-15)更新 $(\gamma_{k1},\gamma_{k2})$;

    $\forall 1\le k\le K$，通过公式 (C-14)更新 $\Phi_{\cdot k}$以及σ_k^2;

    $\forall n\in\mathcal{I}_{\text{tr}},\forall 1\le k\le K$，通过公式 (C-16)更新 ψ_{nk};

    $\forall n\in\mathcal{I}_{\text{tst}},\forall 1\le k\le K$，通过公式 (C-16)更新 ψ_{nk}，但是 ϑ_{nk} 中不包含最后一项;

  until L 的相对变化量小于τ（如 $1e^{-3}$）或者迭代次数为T （如 10）

  使用Multi-class SVM分类器，求解问题 (C-17)(或其对偶问题);

  通过公式(C-18)更新超参数σ_0^2，以及公式(C-18)更新σ_{n0}^2； (*可选*)

until L的相对变化量小于τ' (如 $1e^{-4}$) 或者迭代次数为T' （如 20）。

其中

$$\begin{aligned}\vartheta_{dk}=&\sum_{j=1}^{k}\mathbb{E}_p[\log v_j]-\mathcal{L}_k^{\gamma}(p)-\frac{1}{2\sigma_{d0}^2}(N\sigma_k^2+\Phi_{\cdot k}^{\top}\Phi_{\cdot k})\\&+\frac{1}{\sigma_{d0}^2}\Phi_{\cdot k}^{\top}\Big(\mathbf{x}_d-\sum_{j\neq k}\psi_{dj}\Phi_{\cdot j}\Big)+\sum_{y}\omega_d^y\mathbb{E}_p[\eta_{y_d}^k-\eta_y^k].\end{aligned}$$

对于测试数据，由于没有最大间隔约束，所以ϑ_{dk}中不包含最后一项。

推理$p(\boldsymbol{\eta})$并求解$(\boldsymbol{\xi},\boldsymbol{\omega},\mathbf{u})$: 只考虑$q(\boldsymbol{\eta})$优化$L$，可以得到

$$p(\boldsymbol{\eta})\propto\pi(\boldsymbol{\eta})\exp\Big\{\boldsymbol{\eta}^{\top}\big(\sum_{d\in\mathcal{I}_{\text{tr}}}\sum_{y}\omega_d^y\mathbb{E}_p[\mathbf{g}(y_d,\mathbf{x}_d,\mathbf{z}_d)-\mathbf{g}(y,\mathbf{x}_d,\mathbf{z}_d)]\big)\Big\}.$$

由于先验分布$\pi(\boldsymbol{\eta})$服从标准正态分布，因此后验分布$q(\boldsymbol{\eta})$同样服从正态分布，其均值为

$$\boldsymbol{\mu}=\sum_{d\in\mathcal{I}_{\text{tr}}}\sum_{y}\omega_d^y\mathbb{E}_p[\mathbf{g}(y_d,\mathbf{x}_d,\mathbf{z}_d)-\mathbf{g}(y,\mathbf{x}_d,\mathbf{z}_d)],$$

协方差矩阵为单位矩阵。将$p(\boldsymbol{\eta})$的解带入拉格朗日函数，可以得到对偶问题

$$\max_{\boldsymbol{\omega}}\ -\frac{1}{2}\boldsymbol{\mu}^{\top}\boldsymbol{\mu}+\sum_{d\in\mathcal{I}_{\text{tr}}}\sum_{y}\omega_d^y$$

$$\forall d \in \mathcal{I}_{\text{tr}},\ \text{s.t.}:\ 0 \leq \sum_{y} \omega_d^y \leq C \tag{C-17}$$

该问题（以及其原问题）都可使用多类别SVM分类器高效地求解。

与MT-iLSVM算法相似，超参数σ_0^2和σ_{d0}^2可以事先设定或从数据中估计得到。对于iLSVM，参数的经验估计可以简单地通过如上所示的解析解得到

$$\begin{aligned}\sigma_0^2 &= \frac{\sum_{k=1}^{K} N\sigma_k^2 + \Phi_{.k}^{\top}\Phi_k)}{KN} \\ \sigma_{d0}^2 &= \frac{\mathbf{x}_d^{\top}\mathbf{x}_d - 2\mathbf{x}_d^{\top}\Phi\mathbb{E}_p[\mathbf{z}_d]^{\top} + \mathbb{E}_p[\mathbf{z}_d\mathbf{A}\mathbf{z}_d^{\top}]}{N}.\end{aligned} \tag{C-18}$$

个人简历、在学期间发表的学术论文与研究成果

个人简历

1985 年 6 月 19 日出生，籍贯湖北省天门市，成长于山东省烟台市。

2003 年 9 月考入西北工业大学计算机学院计算机科学与技术专业，2007 年 7 月本科毕业并获得工学学士学位（专业第1名）。

2007 年 9 月免试进入清华大学计算机科学与技术系攻读博士学位至今。

2010 年 1 月至 2011 年 2 月被教育部选派至美国卡内基梅隆大学计算机学院机器学习系联合培养，指导老师为Eric P. Xing教授。

所获奖励

2011 年获清华大学“斯伦贝谢”特等奖学金；

2011 年获智能技术与系统国家重点实验室学术优秀奖；

2010至2011年机器学习国际顶级会议NIPS（2010，2011），ICML（2011），WIML（2010，2011），MLSS（新加坡）Student Travel Award；

2009 年获清华大学综合优秀一等奖学金；

2009 年京港博士生论坛最佳口头报告奖；

2009 年清华大学计算机系三堡博士生论坛“最佳论文奖”；

2008 年获“中航CASC”三等奖学金；

2007 年陕西省优秀毕业生。

已发表及在审学术论文

[1] **Ning Chen**, Jun Zhu, Fuchun Sun, Eric P. Xing. Large Margin Predictive Latent Subspace Learning for Multi-view Data Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*（**TPAMI**），（SCI收录，2011年影响因子: 5.03）（**Regular Paper**）（已录用）

Preprint: http://www.computer.org/portal/web/csdl/doi/10.1109/TPAMI.2012.64.

[2] **Ning Chen**, Fuchun Sun, Linge Ding, Hongqiao Wang. An adaptive PNN-DS Approach to Classification using Multi-sensor Information Fusion. *Neural Computing and Applications*（NCA），2009. （SCI收录，检索号：454EP，2009年影响因子: 0.8）

[3] **Ning Chen**, Jun Zhu, Eric P. Xing. Predictive Subspace Learning for Multi-view Data: A Large Margin Approach. *Advances In Neural Information Processing Systems* (NIPS), 2010.（录稿率：24%）

[4] **Ning Chen**, Jun Zhu, Fuchun Sun. Infinite Exponential Family Harmoniums. *In NIPS workshop on Bayesian Nonparametrics: Hope or Hype?*, 2011.

[5] **Ning Chen**, Jun Zhu. MMH: Max Margin Harmoniums. *In ICML workshop on Topic Models.*, 2010.

[6] **Ning Chen**, Jun Zhu, Fuchun Sun. Infinite EFH: An Infinite Undirected Latent Variable Model. *Conference on Uncertainty in Artificial Intelligence*, (UAI) 2012. （在审中）

[7] Jun Zhu, **Ning Chen**, Eric P. Xing. Infinite Latent SVM for Classification and Multi-task Learning. *Advances In Neural Information Processing Systems*, (NIPS) 2011.（录稿率：24%）

[8] Jun Zhu, **Ning Chen**, Eric P. Xing. Infinite SVM: A Dirichlet Process Mixture of Large Margin Kernel Machines. *In Proc. of International Conference on Machine Learning*, (ICML) 2011. （EI收录，检索号：20114014406063）（录稿率：23%）

[9] Jun Zhu, Ni Lao, **Ning Chen**, Eric P. Xing. Conditional Topical Coding: an Efficient Topic Model Conditioned on Rich Features. *In Proc. of ACM SIGKDD*, 2011. (*Research Full Paper*) （EI收录，检索号：20113714332269）（录稿率：11%）

[10] Pei Deli，Fuchun Sun，Hongqiao Wang，**Ning Chen**. Model Based Bridge Recognition in High Resolution SAR Image. *Proceedings of 2009 International Symposium on Multispectral Image proceeding and pattern recognition* , 2009. （EI收录，检索号：20095112550183）（录稿率：29%）

[11] 汪洪桥, 孙富春, 蔡艳宁, **陈宁**, 丁林阁. 多核学习方法. *自动化学报*，第36卷8期，2010。（EI收录，检索号：20103613222049）

[12] Hong-Qiao Wang, Fuchun Sun, Yanning Cai, Linge Ding, **Ning Chen**. An Unbiased

LSSVM Model for Classification and Regression. Soft Computing，14(2), 171-180, 2010.（SCI收录，检索号: 495BH；EI收录，检索号: 20093912343286）

参与的科研项目及取得的科研成果

[1] 国家973项目：基于视觉认知的多模态信息融合与交互，项目号：2007CB311003，二级课题学生排名第一。本文提出的多模态隐层空间概率图模型等工作作为该项目“创新性研究成果”之一得到专家的认可。该973项目被评为**“优秀”**。时间：2007年-2011年。

[2] 中德清华-汉堡CINACS联合培养项目。时间：2007年9月-2011年9月。

[3] 自然科学基金重点项目：基于视听觉感知与认知的脑基交互方法与关键技术研究，项目号：90820304。时间：2010年5月至今。

[4] 清华大学自主科研项目：认知系统的多模态信息处理理论与方法，项目号：20111081111。时间：2011年10月至今。