变分自编码器的充分统计量特征研究

张静雯

2024 年 4 月

中图分类号： TQ028.1

UDC分类号： 540

变分自编码器的充分统计量特征研究

作 者 姓 名 张静雯

学 院 名 称 数学与统计学院

指 导 教 师 孔祥顺教授

答辩委员会主席 杨国孝教授

申 请 学 位 理学硕士

学 科 专 业 应用统计

学 位 授 予 单 位 北京理工大学

论 文 答 辩 日 期 2024 年 4 月

Optimal Distributed Subsampling for Maximum Quasi-Likelihood Estimators With Massive Data

Candidate Name:	Wei Zhang
School or Department:	School of Mathematics and Statistics
Faculty Mentor:	Prof. Xiangshun Kong
Chair, Thesis Committee:	Prof. Guoxiao Yang
Degree Applied:	Master of Science
Major:	Applied Statistics
Degree by:	Beijing Institute of Technology
The Date of Defence:	December，2020

变分自编码器的充分统计量特征研究

北京理工大学

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

作者签名：_____　　签字日期：_____

关于学位论文使用权的说明

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：① 学校有权保管、并向有关部门送交学位论文的原件与复印件；② 学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③ 学校可允许学位论文被查阅或借阅；④ 学校可以学术交流为目的, 复制赠送和交换学位论文；⑤ 学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

作者签名：_____　　导师签名：_____

签字日期：_____　　签字日期：_____

摘要

变分自编码器是一种强大的生成模型，其利用隐变量来表示数据分布的特征，通过优化目标函数来学习隐变量表示。本文提出了一种新的方式，将变分自编码器的隐变量视为输入数据的充分统计量。

根据信息瓶颈理论，我们可以将变分自编码器的目标函数重新解释为最大化输入数据和隐变量之间的互信息。原先的变分自编码器将隐变量视为均值、方差和随机误差的组合，而在充分统计量意义下隐变量可以视为输入数据的确定性函数。通过将变分自编码器的目标函重新解释为信息瓶颈并进行神经网络学习，我们可以将隐变量优化为输入数据的充分统计量。

此外，原先的 VAE 认为隐变量 z 服从高斯分布，但实际数据分布可能存在非线性和非高斯性质。由于高斯分布的连续性，可能无法捕捉到数据分布的复杂性，因此生成的样本可能出现不连续或不真实的现象。现在我们通过将隐变量 z 重新解释为输入变量的充分统计量，不需要对 z 的分布作某种特定的假设，而是认为其服从最广泛的 Gibbs 分布，有效的避免了这一问题。

通过实验验证，我们发现这种方法能够提供更好的样本生成质量和生成能力。此外，我们观察到隐变量与输入数据在神经网络迭代优化过程中的互信息变化规律符合充分统计量特征，这进一步证明了我们的方法的有效性。

综上所述，本文通过将变分自编码器的隐变量视为输入数据的充分统计量，提出了一种基于信息瓶颈的目标函数及其优化方法。实验证明该方法能够提高变分自编码器的性能，并揭示了隐变量和输入数据之间的互信息变化规律。这对于机器学习和深度学习领域的研究具有重要的理论和实践意义。

关键词：变分自编码器；信息瓶颈；充分统计量；互信息

Abstract

Nonuniform subsampling methods are effective to reduce computational burden and maintain estimation efficiency for massive data. Existing methods mostly focus on subsampling with replacement due to its high computational efficiency. If the data volume is so large that nonuniform subsampling probabilities cannot be calculated all at once, then subsampling with replacement is infeasible to implement. This article solves this problem using Poisson subsampling. We first derive optimal Poisson subsampling probabilities in the context of quasi-likelihood estimation under the A- and L-optimality criteria. For a practically implementable algorithm with approximated optimal subsampling probabilities, we establish the consistency and asymptotic normality of the resultant estimators. To deal with the situation that the full data are stored in different blocks or at multiple locations, we develop a distributed subsampling framework, in which statistics are computed simultaneously on smaller partitions of the full data. Asymptotic properties of the resultant aggregated estimator are investigated. We illustrate and evaluate the proposed strategies through numerical experiments on simulated and real datasets.

Key Words: Variational autoencoder; Information bottleneck; Sufficient statistics; Mutual information

主要符号对照表

x_i	输入数据
y_i	输出数据
$\boldsymbol{z}$	潜在变量/中间变量
$\hat{x}$	VAE 的输出
T	充分统计量
$q_\varphi(z\|x)$	VAE 的编码器
$p_\theta(x\|z)$	VAE 的解码器
f_ω	WGAN 的判别器函数
G	WGAN 的生成器
F	WGAN 的判别器

目录

插图

表格

第 1 章　绪论

1.1　本论文研究背景

在当今信息时代，数据的快速增长和复杂性给信息处理和模式识别带来了巨大的挑战。在面对大量、高维的数据时，如何从中提取有用的特征和信息成为了研究人员关注的焦点。为了解决这一问题，自编码器（Autoencoder）作为一种无监督学习方法被广泛应用于数据特征提取和重建。

然而，传统的自编码器在潜在表示的压缩效率和隐变量的表达能力方面面临一定的局限。为了改进这一问题，变分自编码器（Variational Autoencoder, VAE）应运而生。VAE 是一种生成模型，结合了自编码和概率推断方法，以对数据进行有效模型建立和解释。通过引入隐变量，VAE 能够学习数据的潜在分布，并实现数据重建、生成和插值等操作。相比于传统自编码器，它可以引入随机性并在特定分布下实现数据的重构与生成，因此在模型生成能力和特征学习上具有显著优势。

然而，普通的 VAE 还存在一些缺陷，主要表现在以下几个方面：首先，VAE 的潜在变量通常假设服从高斯分布，但实际数据分布可能存在非线性和非高斯性质。由于高斯分布的连续性，可能无法捕捉到数据分布的复杂性，因此生成的样本可能出现不连续或不真实的现象。第二，VAE 的目标是最小化重构误差和潜在变量的 KL 散度，来实现对数据分布的建模。然而，不平衡的权重设置可能导致生成样本过于模糊（KL 散度权重过大）或不够真实（重构误差权重过大）。第三，在训练过程中，VAE 可能会陷入模式崩溃的情况，即生成样本过于集中在某个特定的模式上。这可能是由于 KL 散度项的过度限制导致潜在空间中的多样性被压缩。第四，在反向传播过程中，由于解码器中存在反向采样操作，梯度可能会呈指数级地衰减，并导致模型无法有效地学习和更新。

如何进一步优化 VAE 以提取更有效的信息仍然是一个挑战，该模型在复杂数据和高维度情况下如何有效提取关键信息，是我们希望研究和解决的问题。信息瓶颈理论（Information Bottleneck Theory）是一种新兴的信息理论，试图以一种更严格的数理方法来描述信息处理的本质特性。其灵感源自于通信理论中的信道容量和数据压缩理论中的最优比特率，而这些概念可以形象地理解为信息传递的“瓶颈”。在复杂的信息处理问题中，信息瓶颈理论提供了一种寻找最优解决方案的新视角，通过构造有

效的解码器和编码器，以达到降维和提取关键信息的目标。

1.2 本文研究的内容

基于以上背景和动机，本研究的目标是基于信息瓶颈理论改进变分自编码器，以实现对输入数据的更好特征提取。具体而言，我们将重新解释变分自编码器的目标函数，并将其重新表述为最大化输入数据与隐变量之间的互信息。通过在神经网络学习过程中使用信息瓶颈框架，我们可以将隐变量作为输入数据的充分统计量进行优化，进而提高自编码器的特征提取能力。我们进一步提出一种新的优化策略，其核心思想是对隐变量的定义进行微调，让它不再是均值、标准偏差和随机误差的混合，而是将隐变量视作输入数据的确定性函数，使得他们的关联性得到极大提升。在研究过程中，我们积极借鉴信息瓶颈理论的优势，结合神经网络作为工具，并利用 WGAN 进行优化，引导模型在学习过程中接近目标函数。在理论与实际应用中，我们通过实验来证明，这种优化策略对于新的编码器模型学习是有效和可行的，可以极大地提升模型生成能力、特征学习能力，从而最终达到优化 VAE 的效果。

本论文将按照以下结构进行组织：第一部分是绪论，包括研究背景、研究内容、研究意义以及充分统计量、信息瓶颈和变分自编码器的国内外研究现状；第二部分将系统介绍充分统计量、变分自编码器和信息瓶颈理论的基本原理，并给出充分统计量的等价定义及其证明；第三部分将详细介绍所提出的算法和方法；第四部分将报道实验结果和分析；最后，第五部分将总结全文并对未来的研究方向提出建议。

1.3 研究意义

在训练深度神经网络时，充分统计量提供了足够的信息来准确地估计模型的参数。它在提高模型的性能和泛化能力方面起到关键作用。深度网络通常拥有大量的参数，而训练数据相对较少。在这种情况下，如果统计量不充分，模型可能会受到过拟合的困扰，即在训练数据上表现良好，但在未见过的数据上表现较差。因此，充分的统计量可以提供足够的信息，以克服过拟合问题。同时，充分的统计量有助于减小模型的方差和偏差，提高模型的稳定性和泛化能力。它能够帮助深度网络更好地学习数据的分布和隐含特征，从而提高模型在未知数据上的表现能力。

利用反映充分统计量的指标作为待优化的目标函数，能够更好地约束表示学习过程，使模型能够从数据中学习到更有意义和有效的特征表示。这有助于提高模型的泛

化能力，使其在未见过的数据上表现更好。这一点对于许多应用来说都非常重要，例如在图像识别、自然语言处理等领域。

但在变分自编码器中，充分统计量未被充分研究。将变分自编码器的隐变量 z 看作输入数据 x 的充分统计量，可以帮助我们更好地理解变分编码器的工作机制。这包括它是如何从输入数据中学习并提取特征的，以及它是如何将这些特征整合起来进行编码和解码的。这有助于提高模型的可解释性，使得模型生成的结果更容易被人类理解和解释。

原先的 VAE 将隐变量 z 看作是均值和方差的线性组合，认为其服从高斯分布，但实际数据分布可能存在非线性和非高斯性质。由于高斯分布的连续性，可能无法捕捉到数据分布的复杂性，因此生成的样本可能出现不连续或不真实的现象。现在我们通过将隐变量 z 重新解释为输入变量的充分统计量，不需要对 z 的分布作某种特定的假设，而是认为其服从最广泛的 Gibbs 分布，有效的避免了这一问题。

最后，为了解决从吉布斯后验采样需要时间昂贵的迭代方法，如马尔可夫链蒙特卡罗（MCMC）方法，我们利用非参数变分推理算法，即 SVGD，从吉布斯算法中进行有效的采样，同时避免了求解归一化参数的过程，大大简化了计算。

1.4 国内外研究现状及发展趋势

1.4.1 信息瓶颈国内外研究综述

信息瓶颈理论是由 Naftali Tishby 等人于 1999 年 [1] 提出的一种精确的方法来分析深度神经网络的训练过程和学习动态的信息理论方法，用于对复杂系统中的信息流进行建模和分析。其基本原理是通过最大化目标变量与输入变量之间的互信息，实现对系统主要特征的捕捉和压缩。信息瓶颈已广泛应用于数据挖掘、模式识别、自然语言处理等领域，并取得了一些重要的研究成果。

一方面，Tishby N 等人 [2] 提出了一种基于互信息衡量的聚类方法，在这个聚类方法中，互信息被用来衡量在马尔科夫过程的弛豫过程中，信息的衰减情况。Kraskov A 等人 [3] 进一步优化了基于 k 近邻距离的熵估计方法。并且当数据的准稳定结构被发现，聚类能够最有效地捕获起始点的信息。另一方面，Shwartz-Ziv R 等人 [4] 指出深度神经网络（DNN）的主要目标是优化信息瓶颈（IB）在压缩和预测之间的权衡。这样的过程主要通过信息压缩以提高训练效率，而不是仅仅拟合训练标签。[5] 从

统计学和信息论原理出发，探讨深度神经网络中不变因子与学习表示的信息最小化等价关系。通过分解交叉熵损失并限制过度拟合项，提出规范损失的两种等效方式：Kullbach-Leibler 项与权重信息作为复杂度指标。最终揭示网络学习表示的不变性、独立性与权重信息的关系，可预测欠拟合与过拟合的相变并优化训练。

之后的研究发展展示了 IB 理论的一些实际应用。尤其是，Alemi, Alexander A 等人 [6] 介绍了“深度变分信息瓶颈”（Deep VIB），这种变分方法允许以有效的方式利用神经网络来参数化信息瓶颈模型。其训练模型在泛化性能和对抗攻击的鲁棒性方面优于其他形式的正则化模型。Barber D[7] 等人通过变分近似降低计算嘈杂信道互信息的难度，并将其应用于线性压缩、群体编码和 CDMA 等实际示例。研究发现，该方法可以对编码和解码方案进行优化，并提高计算效率。另外，通过最大化深度神经网络编码器输入和输出之间的互信息进行无监督学习，提出了 Deep InfoMax 方法 [8]。这一方法在无监督学习表示的发展中起到重要作用。Shamir, Ohad and Sabato 等人 [9] 研究了信息瓶颈方法在学习和泛化中的应用。通过优化信息瓶颈框架中的互信息量来平衡学习复杂性和准确性，提出了一种新的学习理论视角。他们还探讨了信息瓶颈方法在无监督学习和监督学习中的应用，并就分类错误和互信息之间的关系进行了讨论。谢盛嘉 [10] 研究了信息熵和信息瓶颈算法在图像聚类中的应用，实验结果表明，提出的方法具有良好的聚类效果。Slonim, N 等人 [11] 介绍了多元 IB 方法的一般原则框架，允许考虑相互关联的多个系统数据分区。

信息瓶颈方法也被应用在低资源环境中 [12]，解决了变分方法在机器翻译中的表现不佳的问题，而基于信息瓶颈理论的混合压缩方案 [13]，则引入到移动或嵌入式设备中部署神经网络的计算和存储限制问题。信息瓶颈理论在文本分类提取方法中也有应用，例如基于概念特征的文本分类提取方法 [14]，利用信息瓶颈法对关键词进行聚类，并将聚类结果映射到知网义原作为分类特征，该方法表现出的鲁棒性强和特征维数低的优势，克服了概念词典中新词无定义以及需要维护更新词典的问题。

然而，信息瓶颈理论并非没有争议。特别是 Saxe, Andrew M. and Bansal 等人 [15] 的研究质疑了深度网络经历初始拟合和后续压缩两个不同阶段的观点。他们发现信息的平面轨迹主要取决于使用的神经非线性函数，如 ReLU 并不会产生压缩阶段。

总结来说，信息瓶颈理论提供了一种强大的工具来理解和改进深度学习，在国内外的研究中已经得到广泛应用，涵盖了文本分类、图像处理、音频处理、推荐系统等多个领域。最重要的是，它提供了一个框架，可以将深度学习的各个部分（如聚类，

降维，无监督学习和监督学习）统一起来。但信息瓶颈理论但仍然有很多问题和挑战需要进一步探索和解决，包括模型解释能力、理论解释等方面的深入研究。

1.4.2 充分统计量国内外研究综述

随着数据量的不断增长，如何通过数据的概括和结构化来高效地提取信息和进行推理变得越来越关键。为解决这一问题，研究人员提出了各种方法，其中充分统计（sufficient statistic）作为一种概括数据并保留关于参数的所有信息的方法，吸引了越来越多的关注。充分统计量是一种用于确定优化类函数的机器学习模型的训练方法。许多研究人员已经开始尝试使用充分统计量来训练深度网络，并发现在监督学习和不确定性量化基准上取得了竞争性的性能。

Ryan G. James 等人 [16] 通过证明最小充分统计量保留了与其对应变量的信息，为充分统计量的概念提供了理论基础。何鹏光 [17] 综述了充分统计量的两种证明方法，并给出了与充分统计量相关的几个结论。

近年来，充分统计量在机器学习领域的研究逐渐深入，尤其是在深度神经网络中。如 Cvitkovic, Milan 等人 [18] 引入了最小可实现充分统计（MASS）学习，通过训练深度网络，显著提高了监督学习和不确定性量化基准的性能。同时，Yanzhi Chen 等人 [19] 提出了构建充分统计量作为学习互信息最大化数据表示的方法，并利用深度神经网络自动生成摘要统计量。在此基础上，Bai Jiang 和 Tung-Yu Wu 等人 [20] 进一步提高了摘要统计量的准确性和计算效率。为解决高维数据集分析的问题，Joyce 和 Marjoram[21] 开发了一种顺序评分方法用于选择合适的摘要统计量，该方法可以应用于无法使用精确似然方程的高维数据集。

在似然函数不可计算或难以处理的情况下，学者们也为充分统计量提出了相应的解决方法。Johann Brehmer 等人 [22] 针对模拟器难以处理密度问题，提出了基于学习的替代密度方案的技术，以提高推理的样本效率和质量。Jeffrey Chan 等人 [23] 研究了一种基于神经网络的交换式模型，实现了摘要统计量自由、无似然函数的推断。Wiqvist S , Mattei P A 等人 [24] 提出了一种新颖的深度神经架构——部分可交换网络 (PENs)。PENs 是一种深度神经网络架构，利用概率对称性进行建模，并在学习概要统计量方面具有竞争力。

在参数估计和推断方法中，充分统计量也发挥了重要作用。Michael Creel[25] 提出了一种交叉验证方法，用于选择用于近似贝叶斯计算和仿真矩方法等相关估计方法的

统计量。另外，Paul Fearnhead 等人 [26] 通过模拟人工数据与观测数据的摘要统计量进行比较，研究了使用近似贝叶斯计算进行复杂随机模型推断的方法。同时，Diggle P J 等人 [27] 发展了一种方法，用于推断无法计算概率分布理论的隐式统计模型。

充分统计量在高斯信道上的应用通过明确充分统计量在高斯信道上的应用，尹灿斌和贾鑫 [28] 将传统的基于贝叶斯准则的传统估计与基于充分统计量的估计方法进行了性能比较。研究结果表明，在高斯信道中，基于充分统计量的估计完全保留了有关 Y 的信息，且性能优于传统方法。

总之，充分统计量已经在多个领域取得了显著的研究成果和广泛应用，如深度神经网络，高斯信道，无似然函数，参数估计和推断方法等。但充分统计量的理论基础仍需进一步研究和探索，包括充分统计量与其对应变量之间的关系和性质。在深度网络中构建充分统计量的方法和技术仍有待改进，特别是对于复杂模型和问题的适应性。如何选择合适的摘要统计量仍然是一个挑战，可能需要继续发展新的评价方法和算法。充分统计量在机器学习中的应用已取得了显著的进展。尽管仍存在一些问题和挑战，但通过进一步的研究和改进，充分统计量有望在更多的领域发挥作用，并为解决相关问题提供有竞争力的解决方案。未来仍需深入研究和探索，以期在更多领域和场景下取得更好的性能和效果。

1.4.3 变分编码器国内外研究综述

近年来，变分自编码器（VAEs）作为一种有效的无监督学习方法，在各种领域中得到了广泛的应用。变分自编码器是一种基于概率模型的深度学习算法，通过最大化 KL 散度来学习潜在变量的分布。在过去的几年里，VAEs 与其他生成模型相结合，取得了许多突破性的成果。然而，VAEs 的训练目标、解码分布以及潜在变量的利用等方面仍存在一些问题，需要进一步改进。

现有的 VAEs 训练目标可能导致摊后推理分布不准确。这可能会影响模型的生成能力和推断质量。当 VAEs 与解码分布结合时，如果解码分布过于灵活，模型可能会忽略潜在变量。这会影响模型的代表性和泛化能力。为了解决上述问题，Zhao S , Song J[29] 等人提出了一种新的训练目标：InfoVAE，它可以显著提高变分后验的质量并且有效利用潜在特征。InfoVAE 通过最大化 KL 散度和互信息来优化目标函数，从而提高了推断分布的准确性。此外，它还可以使用贝叶斯推断进行参数估计，以提高推荐系统的性能。实验证明，InfoVAE 在多个真实数据集上显著优于其他基线算法。

Mescheder L , Nowozin S 等人 [30] 提出了 Adversarial Variational Bayes (AVB) 方法，用于训练具有任意表达力的 Variational Autoencoders (VAEs)，在非参数极限下能够得到精确的最大似然估计和后验分布。

变分自编码器也被应用在不同任务中。Im 等人 [31]VAE 应用于图像生成和去噪任务。通过学习数据的潜在表示，VAE 能够生成高质量的图像样本，并且在图像去噪任务中表现出色。针对图结构数据，Kipf T N , Welling M[32] 研究了一种用于图结构数据的无监督学习模型——变分图自编码器（VGAE）。VGAE 可以学习出可解释的图的潜在表示，并在链接预测任务上取得了有竞争力的结果。这种方法可以应用于社交网络分析、推荐系统等领域中，帮助我们更好地理解图的结构和行为。Liang D , Krishnan R G 等人 [33] 将 VAE 扩展到隐性反馈的协同过滤中，提出了一种具有多项式似然度的生成模型。这种方法可以提高推荐系统的性能，帮助我们更准确地预测用户的兴趣和行为。实验证明，该方法在多个真实数据集上显著优于其他基线算法。Pu Y , Gan Z , Henao R 等人 [32] 开发了一种新的 VAE 来建模图像以及与之相关的标签或标题，并提出了一种半监督的 CNN 学习框架。这种方法可以应用于图像建模和标签提取任务中，帮助我们从图像中提取更有效的特征和标签。针对数据驱动的 RUL 预测问题，林焱辉、李春波 [34] 提出了使用 VAE 生成多维退化数据特征的方法。该方法通过全局优化模型和条件变分自编码器提取特征，并生成相似数据扩充 RUL 预测模型训练集。同时，利用长短时记忆网络作为 RUL 预测模型并更新生成模型的参数以提高模型效果。另一种应用是人脸图像修复。一种基于变分自编码器的修复方法被提出 [35]，该方法通过设计变种网络引入生成对抗网络来解决修复人脸图像不清晰的问题，并对变分自编码器中的隐变量进行约束以实现特征解耦操作。最后通过动态规划获得最佳分割边界，利用泊松图像编辑得到无缝融合的结果。一些研究工作将 VAE 应用于推荐系统任务。例如，DSVAECF 模型 [36] 被提出用于从用户历史行为中分解静态和动态偏好因素。该模型的两个编码器分别使用多层感知机和循环神经网络对用户行为进行建模，得到用户的静态和动态偏好表示。实验结果表明，与基准方法相比，DSVAECF 在推荐性能上具有显著提升。VAEs 的另一个重要扩展是隐变量模型 [37]。这种模型可以扩展到具有离散隐变量的概率模型中，例如通过离散隐变量进行反向传播。它可以帮助模型学习更复杂的潜在表示，提高模型的表达能力和泛化能力。

尽管现有的研究工作已经取得了显著的进展，但仍存在一些问题需要进一步解决。例如，如何设计更灵活且具有解释性的解码分布以更好地捕捉数据的复杂模式；

如何更有效地利用对抗性训练方法以提高 VAE 的性能；以及如何将 VAE 与其他技术相结合以扩展其应用范围等。未来研究方向包括开发更有效的训练目标和方法来优化 VAEs 的性能，以及将 VAEs 与其他生成模型相结合以扩展其应用范围。

第 2 章　理论介绍

2.1　充分统计量理论

充分统计量（Sufficient Statistic）是概率论和数理统计中的重要概念，它在描述随机数据的特征和参数估计中发挥着关键作用。充分统计量的理论不仅在统计推断中具有重要意义，而且在实际问题中也有着广泛的应用。以下是对充分统计量理论的介绍：

2.1.1　关于充分统计量的一些等价定义

Definition 1. 设样本 X 的样本分布族为 $\{f(\theta,x),\theta\in\Theta\}$，$\Theta$ 是参数空间，令 $T=T(X)$ 为一统计量，若在已知 T 的条件下，样本 X 的条件分布于参数 θ 无关，则 $T(X)$ 为 θ 的充分统计量。

$$P(X|T,\theta)=P(X|T) \tag{2.1}$$

即 X 的后验分布与 θ 无关。

Theorem 2. （因子分解定理）若样本联合密度函数可写为

$$f(\boldsymbol{x},\theta)=g(T(\boldsymbol{x}),\theta)h(\boldsymbol{x}) \tag{2.2}$$

其中 $h(\boldsymbol{x})$ 与 θ 无关，则 $T(\boldsymbol{X})$ 为充分统计量。

证明. 以 $\boldsymbol{X}$ 有密度的情形为例，设统计量 $\boldsymbol{T}=(T_1,\ldots,T_k)$，$T_1,\ldots,T_k$ 都是一维的随机变量，k 一般是较小的自然数，并可以找到 $n-k$ 维统计量 $\boldsymbol{Y}=(Y_1,\ldots,Y_{n-k})$ 使得变换

$$\boldsymbol{X}=(X_1,\ldots,X_n)\to(\boldsymbol{T},\boldsymbol{Y})=(T_1,\ldots,T_k,Y_1,\ldots,Y_{n-k})$$

是一个一一对应的变换，且具有一阶连续偏导数。假定 X 的样本空间 χ 和 T 的样本

空间 $\mathcal{T}$ 皆为欧氏空间，即 $X=\mathrm{R_n}, \mathcal{T}=\mathrm{R_k}$。由于变换是一一对应的，所以

$$\begin{gathered}X_i=X_i(\boldsymbol{T}, \boldsymbol{Y})=X_i\left(T_1, \ldots, T_k, Y_1, \ldots, Y_{n-k}\right), i=1,2, \ldots, n \\ T_j=T_j\left(X_1, \ldots, X_n\right), j=1,2, \ldots, k \\ Y_l=Y_l\left(X_1, \ldots, X_n\right), \quad l=1,2, \ldots, n-k\end{gathered}$$

变换的 Jacobi 行列式

$$|J|=\left|\frac{\partial\left(x_1, \ldots, x_n\right)}{\partial(\boldsymbol{t}, \boldsymbol{y})}\right|=w(\boldsymbol{t}, \boldsymbol{y})$$

此处的 $\boldsymbol{t}$ 为 $\boldsymbol{T}$ 的观察值，$\boldsymbol{y}$ 为 $\boldsymbol{Y}$ 的观察值。

充分性的证明：已知因式分解式成立，即

$$f(\boldsymbol{x}, \theta)=g(T(\boldsymbol{x}), \theta) h(\boldsymbol{x})$$

故 $(\boldsymbol{T}, \boldsymbol{Y})=(T_1, \ldots, T_k, Y_1, \ldots, Y_{n-k})$ 的联合密度为

$$\begin{aligned}k_\theta(\boldsymbol{t}, \boldsymbol{y}) & =f(\boldsymbol{x}, \theta)|J|=g(T(\boldsymbol{x}), \theta) h(\boldsymbol{x}) w(\boldsymbol{t}, \boldsymbol{y}) \\ & =g(\boldsymbol{t}, \theta) h\left(x_1(\boldsymbol{t}, \boldsymbol{y}), \ldots, x_n(\boldsymbol{t}, \boldsymbol{y})\right) w(\boldsymbol{t}, \boldsymbol{y}) \\ & =g(\boldsymbol{t}, \theta) \mu(\boldsymbol{t}, \boldsymbol{y})\end{aligned}$$

这里 $\mu(\boldsymbol{t}, \boldsymbol{y})=h\left(x_1(\boldsymbol{t}, \boldsymbol{y}), \ldots, x_n(\boldsymbol{t}, \boldsymbol{y})\right) w(\boldsymbol{t}, \boldsymbol{y})$ 与 θ 无关，而 $\boldsymbol{T}=T(\boldsymbol{X})$ 的边缘密度为

$$V_\theta(t)=\int_{R_{n-k}} k_\theta(\boldsymbol{t}, \boldsymbol{y}) d \boldsymbol{y}=g(\boldsymbol{t}, \theta) \int_{R_{n-k}} \mu(\boldsymbol{t}, \boldsymbol{y}) d \boldsymbol{y}$$

给定 $\boldsymbol{T}=t$ 时，$\boldsymbol{Y}$ 的条件密度为

$$q(\boldsymbol{y} \mid \boldsymbol{t})=\frac{k_\theta(\boldsymbol{t}, \boldsymbol{y})}{V_\theta(t)}=\frac{g(\boldsymbol{t}, \theta) \mu(\boldsymbol{t}, \boldsymbol{y})}{g(\boldsymbol{t}, \theta) \int_{R_{n-k}} \mu(\boldsymbol{t}, \boldsymbol{y}) d \boldsymbol{y}}=\frac{\mu(\boldsymbol{t}, \boldsymbol{y})}{\int_{R_{n-k}} \mu(\boldsymbol{t}, \boldsymbol{y}) d \boldsymbol{y}}$$

与 θ 无关，

$$\begin{gathered}q(y|t)=q(y|t, \theta) \\ q(y, t|t)=q(y|t)\end{gathered}$$

根据结论: 若 $\boldsymbol{T}=T(\boldsymbol{X})$ 为 θ 的充分统计量，$S=\varphi(\boldsymbol{T})$ 是单值可逆函数，则 $S=\varphi(\boldsymbol{T})$ 也是 θ 的充分统计量。可知 $\boldsymbol{T}=T(\boldsymbol{X})$ 是充分统计量。

必要性的证明: 已知 $\boldsymbol{T}=T(\boldsymbol{X})$ 是充分统计量，因此当给定 $T(\boldsymbol{X})=t$ 时，$\boldsymbol{Y}$ 的条件密度 $q(\boldsymbol{y}|\boldsymbol{t})$ 与 θ 无关。$(\boldsymbol{T},\boldsymbol{Y})$ 的联合密度为

$$k_\theta(\boldsymbol{t},\boldsymbol{y})=q(\boldsymbol{y}|\boldsymbol{t})g(\boldsymbol{t},\theta)$$

通过前面的一一变换可知，$(X_1,\ldots,X_n)$ 的联合密度为

$$\begin{aligned}f(\boldsymbol{x},\theta)&=k_\theta(\boldsymbol{t},\boldsymbol{y})\left|\frac{\partial(\boldsymbol{t},\boldsymbol{y})}{\partial(x_1,\ldots,x_n)}\right|\\&=q(\boldsymbol{y}\mid\boldsymbol{t})g(\boldsymbol{t},\theta)\left|\frac{\partial(\boldsymbol{t},\boldsymbol{y})}{\partial(x_1,\ldots,x_n)}\right|\\&=g(\boldsymbol{t},\theta)h(\boldsymbol{x})\end{aligned}$$

将 $\boldsymbol{t}=(t_1(x_1,\ldots,x_n),\ldots,t_k(x_1,\ldots,x_n))$ 和 $\boldsymbol{y}=(y_1(x_1,\ldots,x_n),\ldots,y_{n-k}(x_1,\ldots,x_n))$ 带入到 $q(\boldsymbol{y}|\boldsymbol{t})\left|\frac{\partial(\boldsymbol{ty})}{\partial(x_1,\ldots,x_n)}\right|$ 的表达式中, 可见它是 $x_1,\ldots,x_n$ 的函数, 用 $h(\boldsymbol{x})$ 表示, 即

$$h(\boldsymbol{x})=q(\boldsymbol{y}|\boldsymbol{t})\left|\frac{\partial(\boldsymbol{ty})}{\partial(x_1,\ldots,x_n)}\right|$$

，显然 $h(\boldsymbol{x})$ 与 θ 无关。因此因子分解定理成立，即证。 □

Corollary 3. 若条件概率密度满足

$$P(X|\theta)=P(X|T)P(T|\theta) \tag{2.3}$$

则 T 为充分统计量。

证明. 根据因子分解定理, 样本的联合密度函数可写为

$$f(\boldsymbol{x},\theta)=g(T(\boldsymbol{x}),\theta)h(\boldsymbol{x})$$

其中 $T(X)$ 为充分统计量。则条件密度

$$\begin{aligned}f(\boldsymbol{x}|\theta)&=\frac{f(\boldsymbol{x},\theta)}{P(\theta)}=\frac{g(T(\boldsymbol{x}),\theta)h(\boldsymbol{x})}{P(\theta)}\\&=P(T(\boldsymbol{x})|\theta)P(X|T=t)\end{aligned}$$

即 $P(X|\theta) = P(X|T)P(T|\theta)$ □

Corollary 4. 若

$$P(\theta|T) = P(\theta|X) \tag{2.4}$$

则 T 为 X 关于 θ 的充分统计量。即用样本分布算得的后验分布与用充分统计量算得的后验分布相同。

Corollary 5. 若 T 是 X 关于 θ 的充分统计量，则

$$P(\theta|X,T) = P(\theta|T) \tag{2.5}$$

证明.

$$P(\theta|X,T) = \frac{P(X|\theta,T)P(\theta|T)}{P(X|T)} = \frac{P(X|T)P(\theta|T)}{P(X|T)} = P(\theta|T)$$

□

Corollary 6. 样本的条件概率与未知参数的后验分布独立，即

$$P(X|T)P(\theta|T) = P(X,\theta|T) \tag{2.6}$$

证明.

$$\begin{aligned} P(X,\theta|T) &= P(X|T)P(\theta|X,T) = P(X|T)\frac{P(X|\theta,T)P(\theta|T)}{P(X|T)} \\ &= P(X|\theta,T)P(\theta|T) = P(X|T)P(\theta|T) \end{aligned}$$

□

Corollary 7. 设 $\boldsymbol{T} = T(\boldsymbol{X})$ 为 θ 的充分统计量，$S = \varphi(\boldsymbol{T})$ 是单值可逆函数，则 $S = \varphi(\boldsymbol{T})$ 也是 θ 的充分统计量。

证明. 由于 $S = \varphi(\boldsymbol{T})$ 是单值可逆函数，所以

$$\{\boldsymbol{X} : T(\boldsymbol{X}) = t_0\} = \{\boldsymbol{X} : S = \varphi(\boldsymbol{T}) = s_0\}$$

表示相同的事件，故对任意事件 A，

$$P\left(A \mid \boldsymbol{T} = t_0\right) = P\left(A \mid S = s_0\right)$$

与 θ 无关, 所以 $S=\varphi(\boldsymbol{T})$ 也是 θ 的充分统计量。 □

Corollary 8. 若 T 为 X 关于 θ 的充分统计量，则有

$$P(X,T|\theta)=P(X|\theta) \tag{2.7}$$

Theorem 9. $\theta\sim P(\theta)$，$x\sim P(x\mid\theta)$，s : $X->\mathcal{S}$ 是一个确定性的函数，则 $s=s(x)$ 是 $P(X|\theta)$ 的充分统计量，当且仅当

$$s=\arg\max_{\mathcal{S}:X\to\mathcal{S}} I(\theta;S(x)) \tag{2.8}$$

其中 $\mathcal{S}$ 为确定性映射，$I(\cdot;\cdot)$ 为随机变量之间的互信息。

证明. 数据处理不等式: 若 $X->Y->Z$，则有 $I(X;Y)\geq I(X,Z)$。

证明数据处理不等式: 若 $X->Y->Z$，

$$P(x,z|y)=\frac{P(x,y,z)}{P(y)}=\frac{P(x,y)P(z|y)}{P(y)}=P(x|y)P(z|y)$$

即当 Y 给定时，X 与 Z 是条件独立的。

且

$$I(X;Y,Z)=I(X;Z)+I(X;Y|Z)=I(X;Y)+I(X;Z|Y)$$

由于 $I(X;Z|Y)=0$，而 $I(X;Y|Z)\geq 0$，所以

$$I(X;Y)\geq I(X,Z)$$

下面证明定理: 设 $s(\cdot)$ 是一个充分的统计量, 通过充分统计量的定义，我们知道

$$P(x|\theta)=P(x|s)P(s|\theta)$$

即给定 θ，通过生成统计量 s，我们可以获得观测数据 x，然后我们有了马尔科夫链

$$\theta->s->x$$

另一方面，由于 $x \sim P(x|\theta)$ 和 $\mathcal{S}$ 是一个确定性函数，给定参数 θ，通过生成数据 x，可以反推统计量 s，我们有马尔科夫链

$$\theta -> x -> s$$

通过数据处理不等式，我们有

$$I(\theta; s(x)) \geq I(\theta; x) \text{ (第一条链)}$$
$$I(\theta; x) \geq I(\theta; s(x)) \text{(第一条链)}$$

这说明 $I(\theta; s(x)) = I(\theta; x)$, 即 s 是 $I(\theta; s(x))$ 的最大值。

从另一个方向看, 由于

$$I(\theta; s(x)) = \max_{\mathcal{S}} I(\theta; S(x))$$

我们有

$$I(\theta; s(x)) = I(\theta; x)$$

注意 $\theta -> x -> s$ 是一个马尔科夫链，根据条件独立性定义，θ 和 x 在给定的 s 的条件下是独立的，即

$$P(\theta, x|s) = P(x|s)P(\theta|s)$$

所以 s 为充分统计量。

这一重要命题表明，通过最大化 θ 和 s 之间的互信息

$$\text{ML} : I(\theta; s) = KL[P(\theta, s) \| P(\theta)P(s)]$$

我们可以找到似然函数 $P(x|\theta)$ 的充分统计量 $s = s(x)$。 □

Corollary 10. $Z = f(X)$ 是离散随机变量 Y 的充分统计量, 当且仅当

$$I(Z, Y) = \max_{S'} I\left(S'(X), Y\right) \tag{2.9}$$

一个分布族的充分统计量往往不止一个，一个好的统计量应满足两点：1. 样本中

尽可能包含未知参数的全部信息，信息损失越少越好，即充分性的要求。2. 统计量越简化越好，即满足最小性。因此我们去寻找最小充分统计量。

Definition 11. 若 $T^* = T^*(X)$ 为极小充分统计量，那么对于任意的充分统计量 $T = T(X)$，存在映射 φ，使得

$$T^*(X) = \varphi(T(X)) \tag{2.10}$$

Theorem 12. 设样本联合密度函数为 $f(x;\theta)$，如果 $\frac{f(x;\theta)}{f(y;\theta)}$ 与 θ 无关的充要条件为

$$T^*(x) = T^*(y) \tag{2.11}$$

且 T^* 充分，那么 T^* 一定极小充分。

Corollary 13. 令 $Z = f(X)$ 为 X 关于 Y 的充分统计量，Z 对于一组函数 $\mathcal{F}$ 是最小可达的，$f \in \mathcal{F}$，如果对于任何 Lipschitz 连续、不可逆的函数 g，其中 $g \circ f \in \mathcal{F}$，$g(Z)$ 于 Y 不再是充分的。

这是因为不可逆函数意味着存在一些样本数据的不同组合，它们具有相同的函数值，这就导致了在逆函数操作时，无法得到原始数据的唯一性。具体来说，如果一个充分统计量 T 经过一次不可逆函数的转换得到 T'，而 T' 无法唯一地反推回 T，那么 T' 就不能被定义为最小充分统计量。因为最小充分统计量需要保持原始数据的全部信息，并且通过它可以唯一地重建样本数据。

Definition 14. 通过最大化表示 Z 和目标变量 Y 之间的互信息，同时最小化表示 Z 和输入变量 X 之间的互信息，使得表示 Z 仅保留与最任务相关的那部分内容，称表示 Z 是 X 关于 Y 的充分统计量。

$$R_{IB}(\theta) = I(Z,Y;\theta) - \beta I(Z,X;\theta) \tag{2.12}$$

但是，直接去求解互信息是非常困难的：一是对于高维连续变量，二重积分难以求解；二是数据的分布在大多数情况下是未知的。幸运的是，我们并不关注互信息的具体数值，而是通过施加互信息限制迫使模型学到更有用的信息。因此，可以通过寻找互信息的 boundary 来间接的优化模型。

Theorem 15. 设 X 是连续型随机变量，Y 是离散型随机变量，$\mathcal{F}$ 是具有公共输出空间

的 Lipschitz 连续函数组合 (例如，一个深度网络的不同参数设置)。如果，

$$\begin{aligned} & f \in \arg\min_{S \in \mathcal{F}} C(X, S(X)) \\ & \text{s.t. } I(S(X), Y) = \max I\left(S'(X), Y\right) \end{aligned} \tag{2.13}$$

则 $f(X)$ 是 X 关于 Y 的最小可达充分统计量。

$$C(X, f(X)) := H(f(X)) - E_X\left[\log\left(J_f(X)\right)\right] \tag{2.14}$$

为守恒微分信息 (CDI)。

证明. (CDI 数据处理不等式) 对于具有相同输出空间的 Lipschitz 连续函数 f 和 g，

$$C(X, f(X)) \geq C(X, g(f(X)))$$

当且仅当 g 几乎到处都是可逆的, 等号成立。

首先，$I(S(X), Y) = \max_S I\left(S'(X), Y\right)$ 保证了 $S(X)$ 是充分的。假设这样的 f 不是最小可逆的，根据最小可达充分统计量定义，存在一个不可逆的、Lipschitz 连续的 g，这样的 $g(f(X))$ 是充分的，由 CDI 数据处理不等式，$\exists C(X, g(f(X))) < C(X, f(X))$，这与 f 最小化 $C(X, S(X))$ 相矛盾。 □

2.1.2 充分统计量的意义

充分统计量的主要作用是在不损失信息的前提下对数据进行了压缩表达。具体来说，当我们获得了充分统计量 $T(X)$ 的取值后，对参数 θ 的估计将不会因为样本 X 的具体取值而受到影响。换句话说，充分统计量包含了样本中关于参数 θ 的全部信息，因此能够在做参数估计时提供更高效的分析。

本文中采用最小可达充分统计量，它可以提供更紧凑的信息，因为它是充分统计量的一个函数，能够包含与参数估计相关的所有重要信息。其次，最小充分统计量有助于减少数据处理和计算量，使得估计过程更加高效。最后，使用最小充分统计量可以简化统计推断的过程，因为它能够减少参数估计的不确定性，提高估计的精确度。因此，相比直接使用充分统计量，使用最小充分统计量通常更为实用和有效。

2.1.3 充分统计量的性质

充分统计量满足以下两个重要性质：

a. 费希信息量不增原理：充分统计量包含了样本中所有关于参数 θ 的信息，因此不会因为样本的具体取值而产生信息重复。这就意味着在给定充分统计量的条件下，任何其他的样本信息不会进一步提高对参数 θ 的估计精度。

b. 充分统计量的最小性：在充分统计量的条件下，任何其他的统计量都不具备额外的信息量，因此充分统计量是最小均方误差的估计。

2.1.4 充分统计量的常见构造方法

构造充分统计量的常见方法包括：封闭性，因子分解定理，典型率和贝叶斯估计等。这些方法均通过不同的途径获取了关于参数的充分信息，并构造出了对应的充分统计量。

然而，人为构造充分统计量存在一些不足之处。首先，人为构造充分统计量需要领域专家的先验知识和经验，这在实践中往往很困难。其次，人为构造充分统计量需要进行手工设计，需要耗费大量时间和精力。此外，由于隐式模型的概率密度函数通常是非解析的，人为构造充分统计量的方法在高维空间中往往面临维度灾难的挑战。

相比之下，基于隐式模型的充分统计量可以通过深度神经网络自动学习，而无需事先的领域知识和手工设计。通过利用统计充分性与信息论之间的联系，将学习充分统计量的任务转化为学习信息最大化的数据表示的任务。这种方法不需要估计任何概率密度或概率密度比，可以通过学习数据的紧凑、接近充分的统计量来提高隐式生成模型的性能。因此，考虑基于隐式模型的充分统计量可以克服人为构造充分统计量的不足，并提供一种自动的、高效的方法来构造适用于隐式生成模型的充分统计量。

具体做法是利用深度神经网络来学习数据的信息最大化表示，从而构造充分统计量。通过利用统计充分性和信息论之间的关联，将学习充分统计量的任务形式化为学习数据的信息最大化表示。

2.1.5 充分统计量的应用

充分统计量在统计推断中有着广泛的应用，例如在后验推断、贝叶斯估计、最大似然估计等方面发挥着重要作用。此外，充分统计量也被应用在实际问题中，比如在

工程和科学研究中对参数的估计和假设检验等方面。

总之，充分统计量作为统计学中的重要概念，不仅具有重要的理论意义，而且在实际问题中也有着广泛的应用。它在统计推断中起着关键作用，并且对于理解和处理随机数据具有重要意义。

2.2 信息瓶颈理论

信息瓶颈理论（Information Bottleneck Theory）是由约翰·特基利（John T. Wixted）和纳丁·扎伊茨曼（Naftali Tishby）提出的一种新的信息理论框架。信息瓶颈理论试图解释信息压缩和学习的关系，旨在寻求最少信息损失的情况下逼近最为充足的解码信息，用于解释生物学、认知科学和机器学习等领域的广泛观察现象，并提供理论指导和工程实践。

2.2.1 信息瓶颈理论的基本概念

信息瓶颈理论的核心概念是“信息瓶颈”，指的是在数据传输或学习过程中，输入数据与输出数据之间的信息传递受到了限制或瓶颈。具体来说，当输入数据的维度较高且包含丰富信息时，我们期望通过某种方法将其映射到较低维度的表示，保留输入数据中重要的信息，而丢弃无用的冗余信息。同时，这种映射需要尽可能少地损失重要的信息，以保持对输入数据的表征能力。信息瓶颈理论试图找到这种转换的理论界限以及最优的信息压缩策略。

2.2.2 信息瓶颈理论与深度学习

信息瓶颈理论提出的是，在大量信息中只有部分信息能够得到处理，这与深度学习中的信息传递和处理存在一定联系。深度学习是一种模仿人脑行为的机器学习方法，可以处理大量信息，通过分层次的数据表示和抽象，自动挖掘数据中的复杂模式。

深度学习网络中的隐藏层可以看做不同放缩程度和旋转程度的信息处理单元，它们相互连接构成一个具有层次结构的网络。每一个隐藏层通过对输入的信息进行变换，然后将这种变换的信息传递给下一层。更深层次的隐藏层可以理解为对信息的进一步压缩和抽象，就像信息瓶颈理论描述的那样，选择性取舍和压缩信息。

在深度学习一个重要的网络结构——卷积神经网络（CNN）中，每一层都包含许

多卷积核，这些卷积核作为信息过滤器，与输入数据进行卷积运算，从而提取出输入数据中的某种特定特征。这种运算方式与信息瓶颈理论中的早期处理方式较为相似，它们都是通过一定的方式从大量的信息中提取出重要的部分信息。

在深度学习中，我们有另一个关键概念——特征选择，这是指挑选出对模型预测结果影响大的特征，而忽略其它特征。如果我们将一段时间内收集到的所有信息视为全部信息，那么特征选择就是这个全部信息的信息瓶颈。事实上，特征选择不仅能提高模型的准确率，还能达到降维的效果，使模型更容易理解，减少过拟合的可能性。

另外，深度学习中的压缩网络也具有与信息瓶颈理论类似的意义，也就是自动编码器。在自动编码器中，有一个隐藏层的节点数小于输入和输出层，这时，模型需要学习如何在压缩到尽量少的节点时存储尽量多的信息，这一过程可以看作是信息瓶颈的过程。

总的来说，深度学习与信息瓶颈理论在对信息的处理上，有许多共通之处。深度学习中的一些结构和算法，如 CNN、特征选择、自动编码器等，都在一定程度上实现了信息瓶颈理论的主张，即挑选重要的信息，忽视不重要的信息。这些结构和算法在这一过程中，提取出了对模型预测有重要影响的特征，从而提高了模型的预测能力。

因此，深度学习并不是完全摒弃了信息瓶颈理论，而是在实现中充分利用了信息瓶颈的思想，将其融入到各个层次和过程中，从而使得模型能够高效地从大量信息中挑选出重要的特征，完成各种复杂的预测和分类任务。这一点，对于我们理解深度学习的模型构造，理解信息的处理过程，都有着重要的启示和意义。

2.2.3 深度变分信息瓶颈（VIB）

将一些中间层的内部表示看作是输入源 X 的随机编码 Z, 由参数编码器 $P(Z|X;\theta)$ 定义。其网络图如图 2.1 所示

目标：学习 Z，s.t $I(Z,Y;\theta)$ max 且 $I(X,Z;\theta) \leq I_c, I_c$ 为信息约束。

目标函数：

$$\max R_{IB}(\theta) = I(Z,Y;\theta) - \beta I(Z,X;\theta) \tag{2.15}$$

对于 $I(Z,Y)$):

$$\begin{aligned} I(Z,Y) &= \int \mathrm{dydz}\, p(y,z) \log \frac{p(y,z)}{p(y)p(z)} \\ &= \int \mathrm{dydz}\, p(y,z) \log \frac{p(y|z)}{p(y)} \end{aligned} \tag{2.16}$$

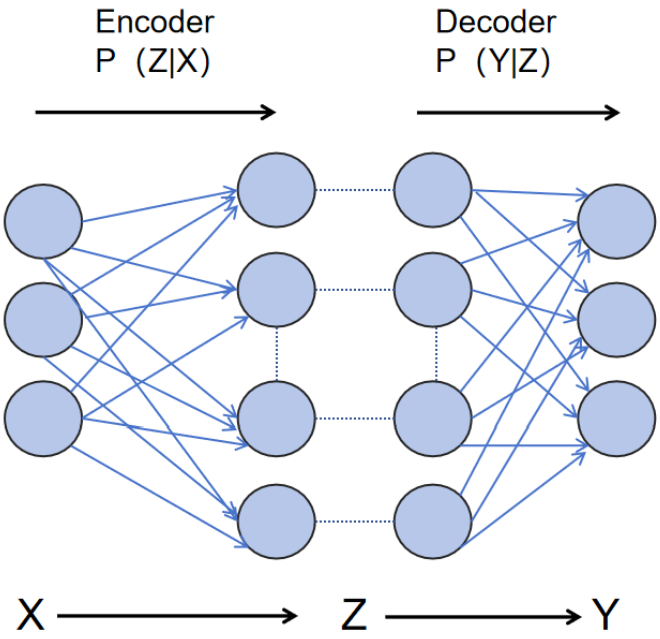

图 2.1　深度网络信息表征过程

其中 $p(y|z)$ 完全由编码器和 Markov 链定义,

$$\begin{aligned} p(y|z) &= \int \mathrm{dx}\, p(x,y|z) = \int \mathrm{dx}\, p(y|x)p(x|z) \\ &= \int \mathrm{dx}\, \frac{p(y|x)p(z|x)p(x)}{p(z)} \end{aligned} \tag{2.17}$$

但是上式难以处理, 所以采用变分近似的思想, 设 $q(y|z)$ 是 $p(y|z)$ 的变分近似, 将其看作是另一个有自己参数集的神经网络。由于

$$\begin{aligned} &\mathrm{KL}[p(Y|Z), q(Y|Z)] \geq 0 \Rightarrow \int p(y|z) \log \tfrac{p(y|z)}{q(y|z)} dy \geq 0 \\ &\Rightarrow \int \mathrm{dy}\, p(y|z) \log(y|z) \geq \int \mathrm{dy}\, p(y|z) \log q(y|z) \end{aligned} \tag{2.18}$$

所以

$$\begin{aligned} &I(Z,Y) = \int \mathrm{dydz}\, p(y,z) \log \tfrac{p(y|z)}{p(y)} \\ &\geq \int \mathrm{dydz}\, p(y,z) \log \tfrac{q(y|z)}{p(y)} \\ &= \int \mathrm{dydz}\, p(y,z) \log q(y|z) - \int \mathrm{dy}\, p(y) \log p(y) \\ &= \int \mathrm{dydz}\, p(y,z) \log q(y|z) + H(Y) \end{aligned} \tag{2.19}$$

其中 $H(Y)$ 与优化过程无关，可以忽略。将

$$p(y,z)=\int \mathrm{dx}\, p(x,y,z)=\int \mathrm{dx}\, p(x)p(y|x)p(z|x) \tag{2.20}$$

带入上式，得

$$\begin{aligned} I(Z,Y) &\geq \int p(y,z)\log q(y|z)\,\mathrm{dydz} \\ &= \int \mathrm{dxdydz}\, p(x)p(y|x)p(z|x)\log q(y|z) \end{aligned} \tag{2.21}$$

这只需要来自数据分布的样本和来自随机编码器的样本, 它需要我们获得 $q(y|z)$ 中可处理的变分近似。对于 $I(Z,X)$：

$$\begin{aligned} I(Z,X) &= \int \mathrm{dzdx}\, p(x,z)\log\frac{p(z|x)}{p(z)} \\ &= \int \mathrm{dzdx} p(x,z)\log p(z|x) - \int \mathrm{dz} p(z)\log p(z) \end{aligned} \tag{2.22}$$

其中 $p(z)=\int \mathrm{dx} p(z|x)p(x)$ 计算困难，设 $\gamma(z)$ 是 $p(z)$ 的变分近似。

$$\begin{aligned} &\mathrm{KL}[p(z),\gamma(z)]\geq 0 \Rightarrow \textstyle\int p(z)\log\frac{p(z)}{\gamma(z)}\,\mathrm{dz}\geq 0 \\ &\Rightarrow \textstyle\int \mathrm{dz} p(z)\log p(z)\geq \int \mathrm{dz}\, p(z)\log\gamma(z) \end{aligned} \tag{2.23}$$

所以，

$$I(Z,X)\leq \int \mathrm{dxdz}\, p(x)p(z|x)\log\frac{p(z|x)}{\gamma(z)} \tag{2.24}$$

结合两项可得新的变分下界为：

$$I(Z,Y)-\beta I(Z,X)\geq \int \mathrm{dxdydz}\, p(x)p(y|x)p(z|x)\log q(y|z)-\beta\int \mathrm{dxdz}\, p(x)p(z|x)\log\frac{p(z|x)}{\gamma(z)}=L \tag{2.25}$$

其经验估计为:

$$\begin{aligned} \mathrm{L} &\approx \frac{1}{N}\sum_{n=1}^{N}\left[\int \mathrm{dz}\, p\left(z|x_n\right)\log q\left(y_n|z\right)-\beta p\left(z|x_n\right)\log\frac{p\left(z|x_n\right)}{\gamma(z)}\right] \\ J_{IB} &= \frac{1}{N}\sum_{n=1}^{N} E_{\varepsilon\sim p(\varepsilon)}\left[-\log q\left(y_n\mid f\left(x_n,\varepsilon\right)\right)+\beta\,\mathrm{KL}\left[p\left(Z|x_n\right),\gamma(Z)\right]\right] \end{aligned} \tag{2.26}$$

2.2.4 信息瓶颈理论的关键问题

在信息瓶颈理论中，有一些重要的问题是需要解决的。首先，如何衡量输入数据和输出数据之间的信息传递量，以及如何量化信息的重要性和损失程度。其次，如何在信息瓶颈的约束下，设计有效的信息压缩方法和学习算法，以获取具有良好表征能力的数据表示。此外，如何将信息瓶颈理论应用于实际问题，并在实际任务中取得良好的性能也是一个重要的挑战。

2.2.5 信息瓶颈理论与机器学习

信息瓶颈理论在机器学习、深度学习、模式识别和数据挖掘等领域具有广泛的应用前景。通过信息瓶颈理论，我们可以更好地理解模型学习的本质，帮助我们设计和优化机器学习模型，提高其对数据的表征能力和泛化性能。同时，信息瓶颈理论也为我们提供了一种新的思路和方法，能够帮助我们解决实际问题中的信息压缩和学习挑战，推动相关领域的发展。综上所述，信息瓶颈理论作为一种重要的信息理论框架，对机器学习和深度学习等领域具有重要意义。通过深入研究信息瓶颈理论，我们有望更好地理解模型学习的本质，并设计出更加高效和可靠的学习算法和模型结构。同时，信息瓶颈理论也将为我们提供一种新的思路和方法，帮助解决实际问题中的信息压缩和学习挑战，推动相关领域的发展，并为我们提供更广阔的研究空间和应用前景。

2.3 变分自编码器理论

变分自编码器（Variational Autoencoder，VAE）是一种基于概率推断的生成模型，可以用于学习数据的潜在表示并进行生成。它结合了自动编码器和变分推断的思想，通过最大化数据的边缘似然来学习数据的分布，并通过抽样技术来生成新的样本。我们将详细介绍变分自编码器的理论原理和工作流程。

2.3.1 自动编码器（Autoencoder）简介

自动编码器是一种无监督学习模型，它由编码器和解码器两部分组成。编码器将输入数据映射到潜在空间的表示，解码器将潜在表示映射回原始数据空间。自动编码器可以通过最小化输入和重构输出之间的误差来学习数据的表示。然而，传统的自动编码器在学习过程中容易受到噪声干扰和过拟合等问题的影响。

2.3.2 变分推断（Variational Inference）基础

变分推断是一种处理概率模型中隐变量后验分布的近似推断方法。它通过最大化近似后验分布和真实后验分布之间的相似度来求解模型中的隐变量分布。变分推断的目标是最大化变分下界（Variational Lower Bound），以近似推断隐变量的后验分布。这一思想被引入到自动编码器中，形成了变分自编码器。

2.3.3 变分自编码器的工作流程

变分自编码器将自动编码器的编码器部分扩展为一个生成模型，以概率分布的形式来学习数据的潜在表示。在变分自编码器中，编码器将输入数据映射到潜在空间的均值向量和方差向量，并通过这两个参数构建隐变量的分布。解码器则从这个隐变量分布中进行采样，生成重构的数据。整个过程可以通过最大化边缘似然来进行训练。

1. 编码器网络（Encoder Network）：编码器网络将输入数据映射为潜在空间中的均值向量 μ 和方差向量 σ，可以表示为潜在空间分布的参数。

2. 重参数化技巧（Reparameterization Trick）：为了实现可微分，可以使用重参数化技巧，从标准正态分布中采样参数，并通过均值和方差进行变换得到潜在表示。这样可以使整个网络成为可训练的。

3. 解码器网络（Decoder Network）：解码器网络从编码器的潜在表示重构原始数据，并最大化重构数据与原始数据之间的相似性。

4. KL 散度损失（KL Divergence Loss）：为了近似真实的后验分布，需要最大化编码器学习的潜在表示的分布与标准正态分布的相似度，这一项也被称为 KL 散度损失。

5. 训练过程：通过最大化重构损失和最小化 KL 散度损失来训练整个网络，可以通过随机梯度下降等优化算法来实现。

2.3.4 变分自编码器的原理

假设有一批数据样本 $X = \{X_1, \ldots, X_n\}$，如果我们能根据 $\{X_1, \ldots, X_n\}$ 得到 X 的分布 $p(X)$，直接根据分布 $p(X)$ 来采样，就可以获得所有可能的 X 了（包括 $\{X_1, \ldots, X_n\}$ 以外的）。但是，这个理想的生成模型很难实现，所以我们考虑将分布

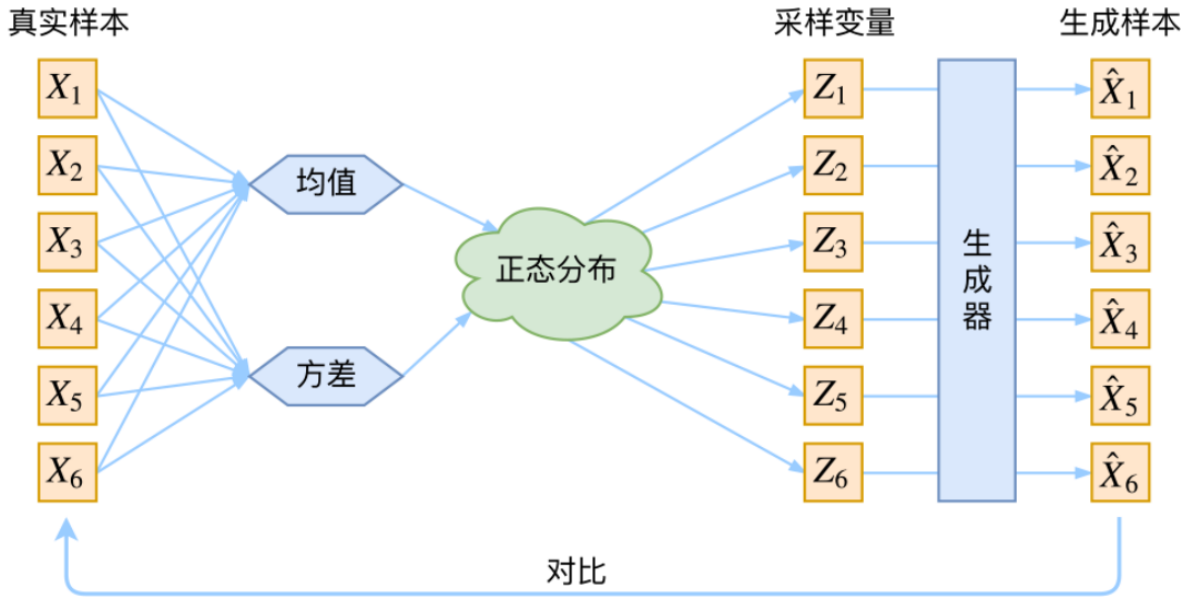

图 2.2　自编码器的实现框架

改为

$$p(X) = \sum_{Z} p(X|Z)p(Z) \tag{2.27}$$

在我们的情景中，不涉及具体的数学操作类型，不论是求和还是积分。我们关注的是条件概率分布 $p(X|Z)$，它描述了一个从潜在变量 Z 生成观测数据 X 的模型。特别地，我们假设潜在变量 Z 遵循标准正态分布，即 $p(Z) = N(0,1)$。这一假设使得我们可以轻松地从这一标准分布中抽取样本，并随后利用条件分布 $p(X|Z)$ 生成对应的 X 值。接下来就是结合自编码器来实现重构，保证有效信息没有丢失，再加上一系列的推导，最后把模型实现。框架的示意图为图 2.2

从这张图来看，我们不能直接确定经过重新采样得到的 Z_k，是否仍与原始的 X_k 相对应，因此，直接最小化 $D\left(\hat{X}_k, X_k\right)^2$（这里 D 代表某种距离函数）并不科学。

因此，在 VAE 模型中，我们并非基于 $p(Z)$ 是正态分布的假设，而是假设条件概率 $p(Z|X)$ 服从正态分布。这意味着对于每一个真实的样本 X_k，我们假设存在一个特定的分布 $p\left(Z|X_k\right)$（即后验分布），且这个分布是独立的、多元的正态分布。之所以强调“特定”，是因为我们需要训练一个生成器 $X = g(Z)$，使得从 $\mathrm{p}\left(\mathrm{Z}|X_k\right)$ 中采样得到的 Z_k 能够还原为 X_k。如果假设 $p(Z)$ 是正态分布，并从其中采样 Z，则难以确定这个 Z 与哪个真实的 X 相对应。然而，当 $p\left(Z|X_k\right)$ 特定于 X_k 时，我们可以合理地期望从这个分布中采样得到的 Z 能够还原为 X_k。

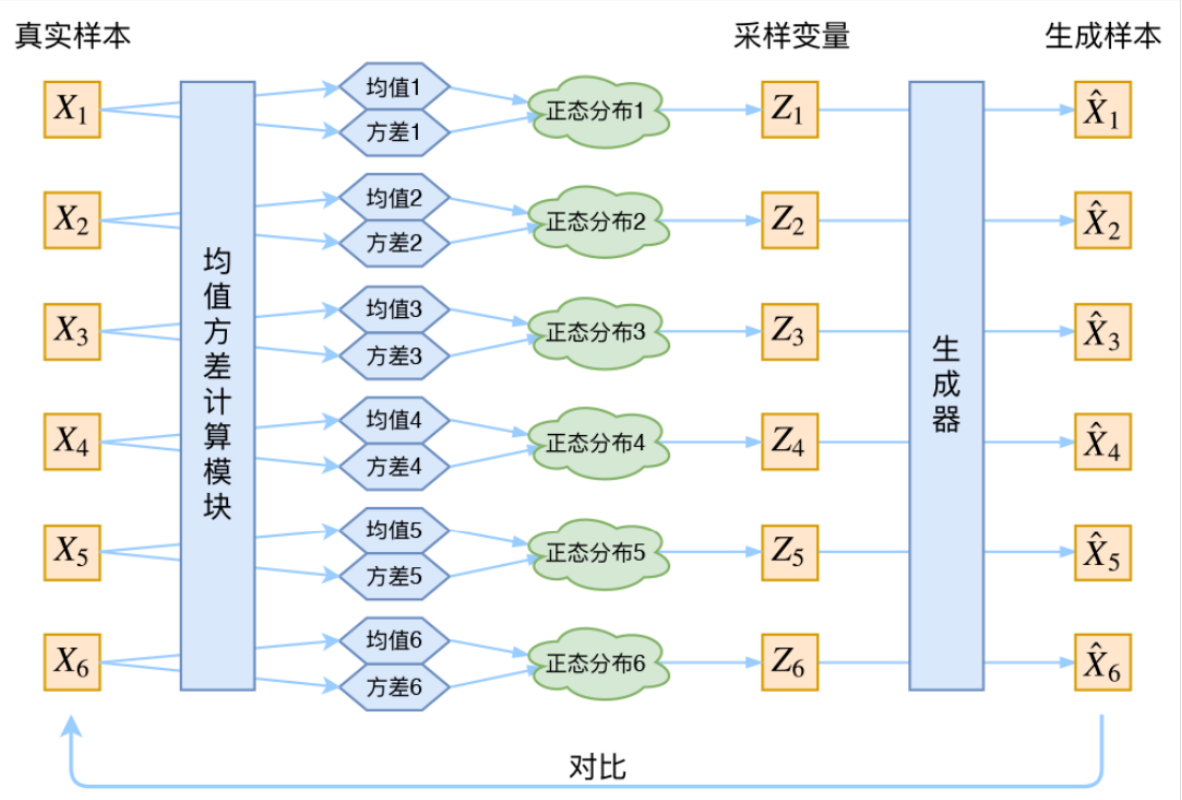

图 2.3　变分自编码器的实现框架

这时每个 X_k 都有对应的的正态分布，我们是用神经网络来拟合这些分布的均值和方差。具体地，我们构建两个神经网络 $\mu_k = f_1(X_k)$ 和 $\log \sigma_k^2 = f_2(X_k)$ 来分别计算均值和方差的对数。选择拟合 $\log \sigma_k^2$，而非直接拟合 σ_k^2，是为了避免对方差进行非负性约束，因为对数方差可以是任意实数。

一旦我们得到了特定于 X_k 的均值和方差，我们就可以确定这个正态分布，并从中采样得到 Z_k。接着，我们将 Z_k 输入到生成器中得到 $\widehat{X}_k = g(Z_k)$，并最小化 $D\left(\widehat{X}_k, X_k\right)^2$。由于 Z_kZ 是从特定于 X_k 的分布中采样得到的，因此生成器应该能够将原始的 X_k 还原回来。于是可以画出 VAE 的示意图如图 2.3

如果所有的条件概率分布 $p(Z|X)$ 都非常接近于标准正态分布 $N(0, I)$，那么从定义上来看，Z 的边缘分布 $p(Z)$ 将会是这些条件分布的加权平均。具体来说，

$$p(Z) = \sum_X p(Z|X)p(X) = \sum_X N(0, I)p(X) = N(0, I)\sum_X p(X) = N(0, I) \tag{2.28}$$

由于每个 $p(Z|X)$ 都接近标准正态分布 $N(0, I)$，这个求和过程实际上是将一系列接近标准正态分布的分布进行加权平均。如果 $p(X)$ 的分布相对均匀，那么最终的 $p(Z)$ 也

会非常接近标准正态分布。

因此，在这种情况下，我们的先验假设——即 $p(Z)$ 是标准正态分布——得到了验证。一旦我们确认了 $p(Z)$ 是标准正态分布，就可以直接从 $N(0, I)$ 中采样来生成图像。这意味着我们不再需要针对每个 X 去拟合一个特定的 $p(Z|X)$，而是可以直接利用这个统一的先验分布来生成数据。

然而，需要指出的是，虽然这种情况在数学上是可能的，但在实际应用中，尤其是在复杂的图像生成任务中，很难保证所有的 $p(Z|X)$ 都严格接近标准正态分布。因此，在实际构建 VAE 模型时，我们通常会显式地建模 $p(Z|X)$，使其能够捕获数据中的复杂结构和变化，从而提高生成图像的质量和多样性。

变分编码器的目标函数为:

$$\min KL(q(z|x)||p(z|x)) \tag{2.29}$$

考虑似然函数 $\log(p(x))$，

$$\begin{aligned} L &= \log(p(x)) = \sum_z q(z|x) \log[p(x)] \\ &= \sum_z q(z|x) \log \frac{p(z,x)}{p(z|x)} \\ &= \sum_z q(z|x) \log \left(\frac{p(z,x)}{q(z|x)} \frac{q(z|x)}{p(z|x)} \right) \\ &= \sum_z q(z|x) \log \left(\frac{p(z,x)}{q(z|x)} \right) + \sum_z q(z|x) \log \left(\frac{q(z|x)}{p(z|x)} \right) \\ &= L^v + D_{KL}(q(z|x)||p(z|x)) \end{aligned} \tag{2.30}$$

由 KL 散度性质可知，$D_{KL}(q(z|x)\|p(z|x)) \geq 0$，所以 $L \geq L^v, L^v$ 为 L 的变分下界，L

为定值，所以最小化 p，q 之间的散度，即最大化 L^v。

$$\begin{aligned} L^v &= \sum_z q(z|x) \log\left(\frac{p(z,x)}{q(z|x)}\right) \\ &= \sum_z q(z|x) \log\left(\frac{p(x|z)p(z)}{q(z|x)}\right) \\ &= \sum_z q(z|x) \log\left(\frac{p(z)}{q(z|x)}\right) + \sum_z q(z|x) \log(p(x|z)) \\ &= -D_{KL}(q(z|x)\|p(z)) + E_{q(z|x)}(\log(p(x|z))) \end{aligned} \tag{2.31}$$

最大化 L^v 即最小化 $q(z|x)$ 和 $p(z)$ 的 KL 散度, 同时最大化上式右边第二项。

由于 $q(z|x)$ 是通过深度网络来实现的，我们预先设定了 z 本身遵循高斯分布的假设。这样做的目的是使编码器的输出尽可能贴近高斯分布。通过这样的设定，我们可以利用高斯分布的特性和性质来简化和优化模型的训练过程，同时也为后续的生成过程提供了便利。通过让编码器的输出服从高斯分布，我们能够更好地捕捉数据的潜在结构，并生成更符合原始数据分布的样本。

$$\begin{aligned} p_\theta(z) &= N(0, I) \\ q_\phi(z|x) &= N\left(z; \mu_z(x,\phi), \sigma_z^2(x,\phi)\right) \end{aligned} \tag{2.32}$$

L^v 的第一项可分解如下:

$$L_1 = \int q_\phi(z|x) \log p(z) dz - \int q_\phi(z|x) \log q_\phi(z|x) dz \tag{2.33}$$

其中，

$$\begin{aligned} \int q_\phi(z|x) \log p(z) dz &= \int N\left(z; \mu, \sigma^2\right) \log N(z; 0, 1) dz \\ &= E_{z\sim N(\mu,\sigma^2)}[\log N(z; 0, 1)] \\ &= E_{z\sim N(\mu,\sigma^2)}\left[\log \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\right] \\ &= -\frac{1}{2}\log 2\pi - \frac{1}{2} E_{z\sim N(\mu,\sigma^2)}\left[z^2\right] \\ &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\left(\mu^2 + \sigma^2\right) \end{aligned} \tag{2.34}$$

$$\begin{aligned}\int q_\phi(z|x)\log q_\phi(z|x)dz &= \int N\left(z;\mu,\sigma^2\right)\log N\left(z;\mu,\sigma^2\right)dz \\ &= E_{z\sim N(\mu,\sigma^2)}\left[\log N\left(z;\mu,\sigma^2\right)\right] \\ &= E_{z\sim N(\mu,\sigma^2)}\left[\log\frac{1}{\sqrt{2\pi}}e^{-\frac{(z-\mu)^2}{2}}\right] \\ &= -\frac{1}{2}\log 2\pi-\frac{1}{2}\log\sigma^2-\frac{1}{2\sigma^2}E_{z\sim N(\mu,\sigma^2)}\left[(z-\mu)^2\right] \\ &= -\frac{1}{2}\log 2\pi-\frac{1}{2}\left(\log\sigma^2+1\right)\end{aligned} \tag{2.35}$$

所以，

$$L_1=\frac{1}{2}\sum_{j=1}^{J}\left[1+\log\left(\sigma_j\right)^2-\left(\mu_j\right)^2-\left(\sigma_j\right)^2\right] \tag{2.36}$$

这里的 J 表示隐变量 Z 的维度数，即隐空间的大小，而 μ_j 和 $(\sigma_j)^2$ 分别对应于一般正态分布中均值向量和方差向量的第 j 个元素。

$$\begin{aligned}q_\phi(z|x) &= N\left(\mu(x,\phi),\sigma^2(x,\phi),I\right) \\ p_\theta(x|z) &= N\left(\mu(x,\phi),\sigma^2(z,\theta),I\right)\end{aligned} \tag{2.37}$$

采用 MC 算法，将 L^v 的第二项 L_2 等价于

$$\begin{gathered}L_2=E_{q(z|x)}(\log(p(x|z)))\approx\frac{1}{L}\sum_{l=1}^{L}\log p\left(x|z^{(l)}\right) \\ z^{(l)}\sim q(z|x)\end{gathered} \tag{2.38}$$

重参数技巧: 从 $N\left(\mu,\sigma^2\right)$ 重采样一个 Z，实际上等价于首先从标准正态分布 $N(0,I)$ 中采样一个 ε，然后通过变换 $Z=\mu+\varepsilon\sigma$ 得到所需的 Z。这种采样方式巧妙地将随机性从梯度下降过程中分离出来，使得采样结果本身参与到模型的训练过程中。这样做的好处是，我们可以避免采样操作对梯度计算的影响，使得整个模型变得可训练。通过这种方式，我们可以有效地学习 μ 和 σ 这两个参数，从而更准确地描述隐变量 Z 的分布，并生成更真实的样本。

2.3.5 变分自编码器的性质和应用

1. 生成样本：通过从编码器的潜在表示中采样并通过解码器生成新的样本，实现对数据的生成。

2. 插值和生成：在潜在空间中进行插值，可以生成具有连续变化的样本。

3. 数据重构：通过编码器和解码器学习数据的潜在表示并实现重构，可以用于数据去噪和特征提取。

4. 卷积变分自编码器（Convolutional Variational Autoencoder）：对图像数据进行特征学习和生成的应用。总的来说，变分自编码器通过结合自动编码器和变分推断的思想，能够学习数据的潜在表示，并利用这些表示进行生成和分析。它在生成模型、降维和特征学习等领域都有着广泛的应用前景。

第 3 章　信息瓶颈下的变分自编码器

在论文的推导部分，我们首先详细阐述了信息瓶颈理论以及其在变分自动编码器（Variational Autoencoder, VAE）中的使用。然后，我们探讨了如何将变分自编码器的目标函数重新进行解释，并以此为基础，我们采取了基于信息瓶颈理论的优化方法。

3.1　变分自编码器（VAE）介绍和目标函数

变分自编码器（VAE）是一种深度生成模型，通过学习数据的分布来实现数据的生成和重构。它由一个编码器网络和一个解码器网络组成，编码器网络将输入数据映射到潜在空间中的概率分布参数，解码器网络则将潜在变量重构为原始输入数据。

对于给定的输入数据 x，VAE 试图学习一个编码器 $q_\varphi(z|x)$ 将输入数据 x 映射到潜在空间 z 中的概率分布上，这个映射是可逆的，同时学习一个解码器 $p_\theta(x|z)$ 将潜在变量 z 映射回重构数据 x 的概率分布上。

在训练过程中，VAE 尝试最小化两个损失函数：重构误差和潜在变量的 KL 散度。重构误差衡量重构样本与原始输入之间的差异，而 KL 散度衡量潜在变量的分布与先验分布之间的差异。通过最小化这两个损失，VAE 可以同时实现数据的压缩表示和潜在空间的建模。由于潜在变量 z 的引入，估计这个概率是困难的，因此需要使用变分推断来优化目标函数。

我们常用的变分编码器的目标函数为：

$$\min KL(q(z|x)||p(z|x)) \tag{3.1}$$

考虑最大似然函数 $L = \log(p(x))$

$$
\begin{aligned}
L = \log(p(x)) &= \int_z q(z|x) \log[p(x)] dz \\
&= \int_z q(z|x) \log \frac{p(z,x)}{p(z|x)} dz \\
&= \int_z q(z|x) \log \left(\frac{p(z,x)}{q(z|x)} \frac{q(z|x)}{p(z|x)} \right) dz \\
&= \int_z q(z|x) \log \left(\frac{p(z,x)}{q(z|x)} \right) dz + \int_z q(z|x) \log \left(\frac{q(z|x)}{p(z|x)} \right) dz \\
&= L^v + D_{KL}(q(z|x) \| p(z|x)) \\
&\geq \int_z q(z|x) \log \left(\frac{p(x|z)p(z)}{q(z|x)} \right) dz \\
&= \int_z q(z|x) \log p(x|z) dz - \int_z q(z|x) \log \left(\frac{q(z|x)}{p(z)} \right) dz \\
&= E_{Z \sim q(z|x)} \log p(x|z) - KL(q(z|x)||p(z))
\end{aligned}
\tag{3.2}
$$

由于 L 为定值, 最小化 p，q 之间的散度，即最大化

$$
L^v = \int_z q(z|x) \log \left(\frac{p(z,x)}{q(z|x)} \right) dz = E_{Z \sim q(z|x)} \log p(x|z) - KL(q(z|x)||p(z))
$$

所以,

$$
\begin{aligned}
E_{X \sim q(x)}[\log p(x)] &> E_{X \sim q(x)} \left[E_{Z \sim q(z|x)} \log p(x|z) \right] - E_{X \sim q(x)}[KL(q(z|x)||p(z))] \\
&= E_{X \sim q(x)} \left[E_{Z \sim q(z|x)} \log p(x|z) \right] - \iint q(z|x)q(x) \log \frac{q(z|x)q(z)}{p(z)q(z|x)} dzdx \\
&- \iint q(z|x)q(x) \log \frac{q(z|x)q(x)}{q(z)q(x)} dzdx \\
&= E_{X \sim q(x)} \left[E_{Z \sim q(z|x)} \log p(x|z) \right] - \iint q(z|x)q(x) \log \frac{q(z)}{p(z)} dzdx \\
&- \iint q(x,z) \log \frac{q(x,z)}{q(z)q(x)} dzdx \\
&= E_{X \sim q(x)} \left[E_{Z \sim q(z|x)} \log p(x|z) \right] - KL(q(z)||p(z)) - I(Z;X)
\end{aligned}
\tag{3.3}
$$

可知, VAE 的变分推断可以分解为: 隐空间和输入的互信息 $I(Z;X)$，加于隐空间的先

验 p(z) 与隐空间的后验 q(z) 的 KL 散度，以及第一项重构误差。其中，

$$\begin{aligned}
&E_{X\sim q(x)}\left[E_{Z\sim q(z|x)}\log p(x|z)\right]-E_{X\sim q(x)}[KL(q(z|x)||p(z))]\\
&\quad=E_{X\sim q(x)}\left[\int q(z|x)\log p(x|z)dz-\int q(z|x)\log\frac{q(z|x)}{p(z)}dz\right]\\
&\quad=E_{X\sim q(x)}\left[\int q(z|x)\log\frac{p(x|z)p(z)}{q(z|x)}dz\right]\\
&\quad=E_{X\sim q(x)}\left[\int q(z|x)\log\frac{p(x,z)}{q(z|x)}dz\right]\\
&\quad=-E_{X\sim q(x)}\left[\int q(z|x)\log\frac{q(z|x)}{p(x,z)}dz\right]
\end{aligned}\tag{3.4}$$

由于 $E_{X\sim q(x)}[\log q(x)]$ 为常数，不影响优化，且 $\int q(z\mid x)dz=1$，所以上式可改写为

$$\begin{aligned}
&E_{X\sim q(x)}\left[E_{Z\sim q(z|x)}\log p(x|z)\right]-E_{X\sim q(x)}[KL(q(z|x)||p(z))]\\
&\quad=-E_{X\sim q(x)}\left[\int q(z|x)\log\frac{q(z|x)}{p(x,z)}dz\right]-E_{X\sim q(x)}[\log q(x)]\\
&\quad=-E_{X\sim q(x)}\left[\int q(z|x)\log\frac{q(z|x)}{p(x,z)}dz\right]-E_{X\sim q(x)}\left[\log q(x)\int q(z|x)dz\right]\\
&\quad=-E_{X\sim q(x)}\left[\int q(z|x)\log\frac{q(x)q(z|x)}{p(x,z)}dz\right]\\
&\quad=-\int q(x)q(z|x)\log\frac{q(x,z)}{p(x,z)}dz\\
&\quad=-KL(q(x,z)||p(x,z))
\end{aligned}\tag{3.5}$$

因此，最大化 L^v，即最小化 $q(x,z)$ 和 $p(x,z)$ 之间的 KL 散度，两种目标函数实际上是等价的，本文采用

$$\min KL(q(x,z)||p(x,z))\tag{3.6}$$

作为目标函数。其中，$q(x,z)=q(x)q(z|x)$ 是编码器网络的联合分布，$p(x,z)=p(x|z)p(z)$ 是解码器网络的联合分布。

3.2 信息瓶颈理论与目标函数重新解释

信息瓶颈理论认为，对于给定的输入数据，模型应该提取出对于预测结果最关键的信息，而忽略掉那些对于预测结果不重要的信息。

当我们将信息瓶颈理论应用到变分自编码器中时，我们需要将变分自编码器的目

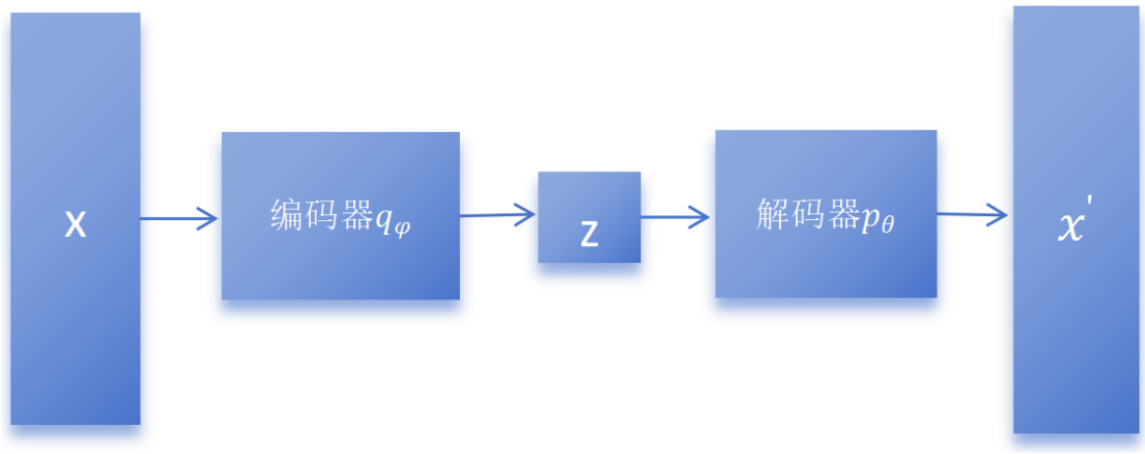

图 3.1　变分自编码器示意图

标函数进行重新解释。根据信息瓶颈理论，我们希望学习到的表示 z 能够尽可能地包含输入数据 x 的重要信息，同时丢弃冗余的信息。因此，我们重新解释 VAE 的目标函数为最小化输入数据 x 和隐变量 z 之间的互信息，同时最大化隐变量 z 和输出数据 $\hat{x}$ 的互信息，然后引入充分统计量。将变分编码器表示成如图 3.1 所示。

由于

$$\begin{aligned} KL(q(x,z)||p(x,z)) &= \iint q(x,z)\log\frac{q(x,z)}{p(x,z)}\mathrm{dxdz} \\ &= \iint q(x,z)\log\frac{q(x,z)}{q(x)q(z)}\mathrm{dxdz} + \iint q(x,z)\log\frac{q(x)q(z)}{p(x,z)}\mathrm{dxdz} \\ &= I(X,Z) - \iint q(x,z)\log\frac{p(x,z)}{q(x)q(z)}\mathrm{dxdz} \end{aligned} \tag{3.7}$$

根据第二章中信息瓶颈理论，对于输入源 X 的随机编码 Z 和输出源 Y 之间的互信息，其下界为：

$$I(Z,Y) \geq \iint p(y,z)\log q(y|z)\mathrm{dydz} \tag{3.8}$$

其中，$q(y|z)$ 是 $p(y|z)$ 的变分近似，$p(y|z)$ 完全由编码器和 Markov 链定义，

$$\begin{aligned} p(y|z) &= \int \mathrm{dx}p(x,y|z) = \int \mathrm{dx}p(y|x)p(x|z) = \int \mathrm{dx}\frac{p(y|x)p(z|x)p(x)}{p(z)} \\ p(y,z) &= \int \mathrm{dx}p(x,y,z) = \int \mathrm{dx}p(x)p(y|x)p(z|x) \end{aligned} \tag{3.9}$$

在变分编码器场合下，输入源 x 的随机编码 z 即为变分编码器的隐变量 z，输出数据 y 即为变分编码器的输出数据 $\hat{x}$，先验分布 $p(x)$ 即为变分自编码器场合中的原

始数据的分布 $q(x)$，后验分布 $p(z|x)$ 即为编码器 $q(z)$，条件分布 $q(y|z)$ 即为解码器 $p(x|z)$。所以，上述不等式在变分编码器场合中为：

$$I(Z,\hat{X}) \geq \iint q(x,z)\log(x|z)\mathrm{dydz} \tag{3.10}$$

所以

$$\begin{aligned}
KL(q(x,z)\|p(x,z)) &= I(X,Z) - \iint q(x,z)\log\frac{p(x,z)}{q(x)q(z)}\mathrm{dxdz} \\
&= I(X,Z) - \iint q(x,z)\log\frac{p(x,z)}{p(z)}\mathrm{dxdz} - \iint q(x,z)\log\frac{p(z)}{q(x)q(z)}\mathrm{dxdz} \\
&= I(X,Z) - \iint q(x,z)\log \mathrm{p(x|z)}\mathrm{dxdz} + \iint q(x,z)\log\frac{q(z)}{p(z)}\mathrm{dxdz} \\
&+ \iint q(x,z)\log q(x)\mathrm{dxdz} \\
&= I(X,Z) - \iint q(x,z)\log \mathrm{p(x|z)}\mathrm{dxdz} + \iint q(x,z)\mathrm{dx}\log\frac{q(z)}{p(z)}\mathrm{dz} \\
&+ \iint q(x,z)dz\log q(x)\mathrm{dx} \\
&= I(X,Z) - \iint q(x,z)\log \mathrm{p(x|z)}\mathrm{dxdz} + \int q(z)\log\frac{q(z)}{p(z)}\mathrm{dz} \\
&+ \int q(x)\log q(x)\mathrm{dx} \\
&= I(X,Z) - \iint q(x,z)\log \mathrm{p(x|z)}\mathrm{dxdz} + KL(q(z)\|p(z)) + \int q(x)\log q(x)\mathrm{dx}
\end{aligned} \tag{3.11}$$

其中 $\int q(x)\log q(x)\mathrm{dx}$ 为常数，与优化过程无关，可省略，$KL(q(z)||p(z)) \geq 0$，所以上式可进行放缩：

$$KL(q(x,z)||p(x,z)) \geq I(X,Z) - I(Z,\hat{X}) \tag{3.12}$$

右边是信息瓶颈的形式，即变分编码器的目标函数的下界是信息瓶颈。因此我们可以认为隐变量 z 是输入数据 x 的充分统计量。

3.3 隐变量优化和神经网络学习

3.3.1 隐变量优化

为了优化变分自编码器的目标函数，我们通过神经网络学习将输入数据 x 映射到潜在变量 z 的概率分布上，并且学习将潜在变量 z 映射回重构数据 x 的概率分布 $q_\varphi(z|x)$ 上，并且学习将潜在变量 z 映射回重构数据 x 的概率分布 $p_\theta(z|x)$ 上。在这个过程中，我们可以利用信息瓶颈的概念来指导潜在变量 z 的学习，使其能够充分地包含输入数据 x 的重要信息，并且过滤掉冗余信息。通过神经网络和信息瓶颈的思想，我们可以将隐变量优化为输入数据的充分统计量，即 z 包含了输入数据 x 的最重要和最显著的特征，通过优化隐变量 z，我们可以使其更好地表示输入数据的统计特征。

3.3.2 MASS 学习

为了实现隐变量优化，我们可以设计神经网络结构，以最大化输入数据和隐变量之间的互信息。通过神经网络的学习，我们能够使隐变量 z 更好地表示输入数据的充分统计量，从而实现对输入数据的高效表示和生成。

由于在机器学习中，存在一个问题: 一个深度网络的输入和输出之间的互信息是无限的，即对于一个连续型随机变量 X 和一个连续、非常数函数 f，互信息 $I(X, f(X))$ 是无穷大的。这使得 $I(X, f(X))$ 不适合在某些学习目标中使用，例如，当 f 是一个标准的深度网络时。在之前的研究中，$I(X, f(X))$ 的无穷大问题通常被两种方法规避: 一种是离散 X 和 $f(X)$，另一种是使用分布 $P(Z|X)$ 的随机变量 Z 作为 X 的表示, 而不是使用 $f(X)$ 。$P(Z|X)$ 通常通过向以 X 为输入的深度网络中添加噪声来实现。

本文参考 $MinimalAchievableSufficientStatisticLearning$[18] 使用守恒微分信息 (CDI）来避免 $I(X, f(X))$ 无穷。

Definition 16. 对于连续型随机变量 $X \in R^d$, 和一个 Lipschitz 连续函数 $f : R^d \to R^r$，守恒微分信息 (CDI) 为

$$C(X, f(X)) := H(f(X)) - E_X \left[\log \left(J_f(X)\right)\right] \tag{3.13}$$

其中 H 为微分熵，$H(Z)=-\int p(z)\log p(z)\mathrm{dz}$，$J_f$ 是 f 的雅可比行列式，

$$J_f(x)=\sqrt{\det\left(\frac{\partial f(x)}{\partial x^T}\left(\frac{\partial f(x)}{\partial x^T}\right)^T\right)}$$

根据定理 15，将其转化为函数 f 上的学习目标，通过将严格约束放宽为具有拉格朗日乘子 $\frac{1}{\beta}$ 的拉格朗日公式：

$$C(X,f(X))-\frac{1}{\beta}I(f(X),Y) \tag{3.14}$$

其中 β 的价值越大，我们的目标就会越鼓励最小化而不是充分性。我们使用恒等式 $I(f(X),Y)=H(Y)-H(Y|f(X))$ 来简化这个公式，这给出了以下优化目标:

$$L_{\text{MASS}}(f):=H(Y|f(X))+\beta H(f(X))-\beta E_X\left[\log\left(J_f(X)\right)\right] \tag{3.15}$$

将最小化这一目标称为 MASS 学习。

在变分自编码器中，隐变量 z 即是充分统计量 $f(X)$，输出数据为 $\hat{x}$。在实践中，我们感兴趣的是使用有限数据集 $\{x_i\}_{i=1}^N$ 的分布 $p(x)$ 采样的 N 个点，利用 MASS 学习训练一个参数为 φ 的编码器 q_φ 和一个参数为 θ 的解码器 p_θ。假设 $z_i=f_\varphi(x_i)$ 是从编码器 $q_\varphi(z|x)$ 中采样出来的，我们最小化了以下对 L_{MASS} 的经验上界:

$$L_{MASS}\leq\hat{L}_{MASS}(\theta,\varphi):=\frac{1}{N}\sum_{i=1}^{N}-\log p_\theta\left(\hat{x}_i|f_\varphi(x_i)\right)-\beta\log p\left(f_\varphi(x_i)\right)-\beta\log_{f_\varphi}(x_i) \tag{3.16}$$

为了实现上述优化，等式中两种分布的参数形式必须被指定。对于解码器 $p_\theta(\hat{x}|f_\varphi(x))$，我们假设它有一个 Gibbs 分布形式，即

$$p_\theta(\hat{x}|f_\varphi(x))=p_\theta(\hat{x}|z)=\frac{1}{C_1}\exp\left\{-\frac{\|x-D_\theta(z)\|^2}{2\sigma^2}\right\} \tag{3.17}$$

其中，$D_\theta(z)$ 表示由解码器生成的数据，常数 C_1 为满足 $\int p_\theta(x|z)dx=1$ 的归一化因子。

对于分布 $p\left(f_{\varphi}(x)\right)$，我们假设它的 Gibbs 分布形式为

$$p\left(f_{\varphi}(x)\right)=p(z)=\exp\{F(z)\}/C_2 \tag{3.18}$$

其中 $F(z)$ 即为 Wasserstein GAN（WGAN）中的判别器。常数 C_2 为满足 $\int p(z)dz=1$ 的归一化因子。

3.3.3 WGAN 理论推导

生成对抗网络（GAN）是一种用于生成模型的深度学习架构，它由生成器（Generator）和判别器（Discriminator）两个神经网络组成，通过博弈的方式来训练生成模型。生成模型的任务是通过输入随机噪声生成和原始数据相似的伪造的样本，判别模型的任务是对伪造样本和真实样本进行分类来学习区分它们（真实样本来源于数据集，伪造样本来源于生成模型）。

GAN 的训练过程是一个博弈的过程，生成器和判别器相互对抗并不断优化，以达到生成接近真实数据的样本的目标。在 GAN 训练中，我们希望判别器 D 能够准确地将真实样本与生成的样本区分开来。对于给定的真实样本 x，我们希望最大化 $D(x)$，表示判别器 D 将真实样本标记为真的概率。对于生成器 G 生成的样本 $G(z)$，我们希望最小化 $\log(1-D(G(z)))$，表示判别器 D 将生成样本标记为假的概率。

在训练过程中，采取交替迭代的策略。在每个训练步骤中，固定一个网络（例如判别器 D），然后更新另一个网络（生成器 G）的参数。这样做的目的是优化生成器 G，使得它能够生成更加真实的样本，从而更容易欺骗判别器 D。通过交替迭代训练，判别器 D 和生成器 G 相互博弈，通过最小化对方的损失来提高自身的性能。最终，生成器 G 学习到了数据的分布，能够生成更加真实的样本。

GAN 的目标函数为：

$$\min_{G}\max_{D}V(G,D)=\mathbb{E}_{\boldsymbol{x}\sim p_r(\boldsymbol{x})}[\log D(\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))] \tag{3.19}$$

GAN 网络整体示意图如图 3.2。

需要注意的是，GAN 的训练是一个非常复杂的过程，需要仔细地平衡判别器 D 和生成器 G 的优化目标。同时，GAN 的训练也可能面临不稳定性和模式坍塌等问题，需要通过合适的技巧和调整来提高训练的效果和稳定性。

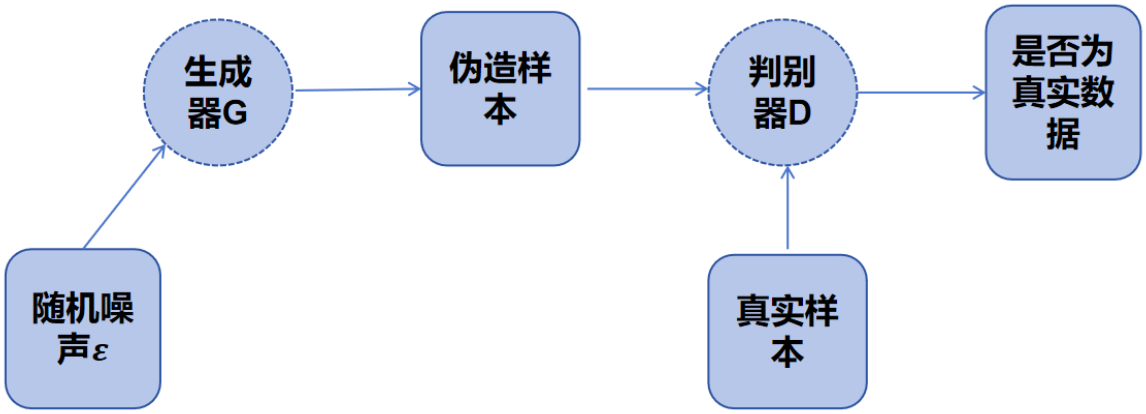

图 3.2　GAN 网络示意图

WGAN 是一种改进的生成对抗网络（GAN），它使用 Wasserstein 距离作为训练目标，旨在解决 GAN 在训练中的一些问题。WGAN 相对于传统 GAN 的一个主要优势是解决了梯度消失问题。传统 GAN 的生成器和判别器之间使用交叉熵作为损失函数，容易出现梯度消失或爆炸的问题。而 WGAN 使用 Wasserstein 距离作为损失函数，避免了梯度消失的情况，使得训练更加稳定。另一个优势是提供了更好的训练稳定性。传统 GAN 的训练过程不稳定，容易陷入模式坍缩或模式崩溃的问题。WGAN 通过最小化 Wasserstein 距离，可以提供更稳定的训练过程，生成器和判别器的学习相对独立，减少互相干扰。此外，WGAN 还能够解决模式塌陷问题，即生成器只能生成少数几种样本的问题。通过 Wasserstein 距离，WGAN 可以更好地捕捉到数据分布的细节，使得生成器能够生成更多样的样本。

所以，WGAN 通过解决梯度消失问题、提供训练稳定性和解决模式塌陷等问题，在生成对抗网络中具有显著的优势。这使得 WGAN 能够生成更高质量、多样性的样本。

WGAN 的生成器 G 和判别器 F 的损失函数分别为：

$$J(F) = E_{x\sim P_g}\left[f_w(G(x))\right] - E_{x\sim P_r}\left[f_w(x)\right] \tag{3.20}$$

$$J(G) = -E_{x\sim P_g}\left[f_w(G(x))\right] \tag{3.21}$$

Theorem 17. 极大似然与熵正则化的 WGAN 等价，且 $f(x,\theta)$ 为极大似然模型的吉布

斯分布指数项

$$\max_p E_q \log p = \max_\theta \min_g [E_q f(x,\theta) - E_g f(x,\theta) - H(g)] = \arg\max_p \min_g [KL(g\|p) - KL(q\|p)] \tag{3.22}$$

证明. WGAN 原始表示为

$$\max_\theta \min_g [E_q f(x,\theta) - E_g f(x,\theta)] \tag{3.23}$$

其中 g 为生成数据概率分布，q 为实际数据分布，p 为模型分布。借鉴 Contrastive Divergence 思想

$$\begin{aligned} KL(g\|p) - KL(q\|p) &= E_g \log g - E_g \log p - E_q \log q + E_q \log p \\ KL(g\|p) - E_q \log q + E_q \log p &= E_q f(x,\theta) - E_g f(x,\theta) + H(q) - H(g) \\ L(p,g) \doteq KL(g\|p) + E_q \log p &= E_q f(x,\theta) - E_g f(x,\theta) - H(g) \\ &\geq E_q \log p \doteq L(p) \end{aligned} \tag{3.24}$$

一方面

$$\begin{aligned} L(p,g) &\geq L(p) \\ \max_p \min_g L(p,g) &\geq \max_p L(p) \end{aligned} \tag{3.25}$$

另一方面

$$\begin{aligned} \min_g L(p,g) &\leq L(p,p) = L(p) \\ \max_p \min_g L(p,g) &\leq \max_p L(p) \end{aligned} \tag{3.26}$$

所以

$$\max_p \min_g L(p,g) = \max_p L(p) \tag{3.27}$$

□

因此，对于我们假设 $p\left(f_\varphi(x)\right)$ 的 Gibbs 分布形式

$$p\left(f_\varphi(x)\right) = p(z) = \exp\{F(z)\}/C_2 \tag{3.28}$$

其中 $F(z)$ 就是 Wasserstein GAN（WGAN）中的判别器。

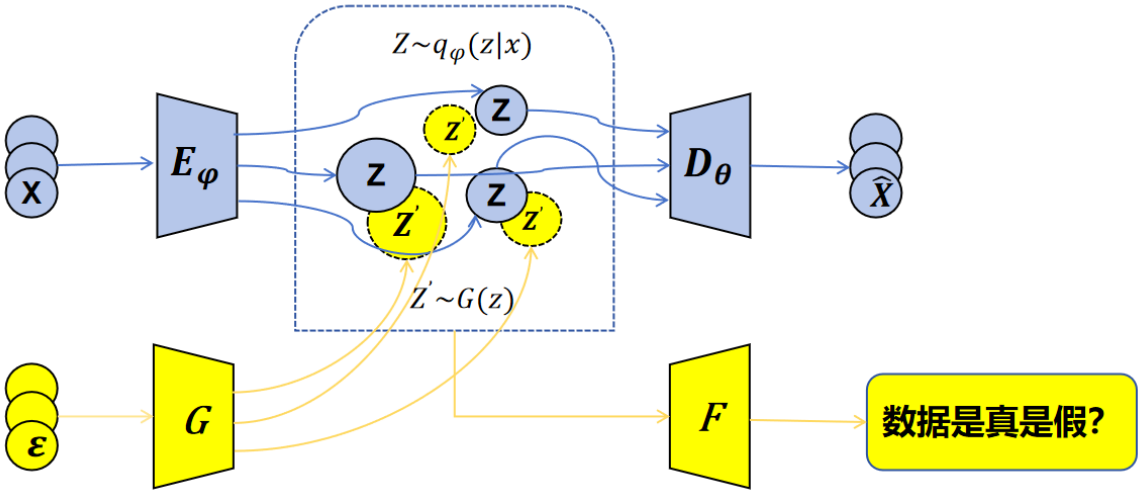

图 3.3　WGAN-VAE

3.3.4　WGAN-VAE

由于 $p(z)$ 的分布形式中 $F(z)$ 的具体形式未知，我们设计 WGAN-VAE 来同时优化判别器 $F(z)$ 和参数 θ、φ。WGAN-VAE 模型如图 3.3。

对于生成器 G，我们给定任意输入噪声 ϵ，生成伪造样本 Z'，真实样本 Z 由分布 $q_\varphi(z|x)$ 采样得到，通过比较伪造样本和真实样本来优化判别器 F。在每一轮训练完判别器 F 后，$p(z)$ 的分布形式已知，则可以根据目标函数 3.16 训练变分自编码器中编码器和解码器的两个参数 θ 和 φ，由此产生新的真实样本 z，依次迭代至收敛。

我们提出了算法 1，利用 WGAN 来训练 VAE 的过程。

3.4　结论

通过将变分自编码器的目标函数重新解释为最大化输入数据和隐变量之间的互信息，并结合信息瓶颈理论进行神经网络学习，我们可以实现隐变量的优化为输入数据的充分统计量，从而提高了对输入数据的特征表示和生成能力，为变分自编码器的优化和应用提供了新思路和方法。

Algorithm 1: WGAN-VAE 的第 t 次迭代训练过程

Input:

一组数据 $x \sim p(x)$;
任意噪声 ϵ;
解码器 $E_{\varphi}^{(t)}$ 和编码器 $D_{\theta}^{(t)}$ 的模型参数;
生成器 $G_{g}^{(t)}$ 和判别器 $F_{\omega}^{(t)}$ 的模型参数

Output:

更新的参数：$E_{\varphi}^{(t+1)}$、$D_{\theta}^{(t+1)}$、$G_{g}^{(t+1)}$、$F_{\omega}^{(t+1)}$;

1: 初始化参数 φ、θ、g、ω，设置学习率 α;
2: 得到解码器和编码器的输出：$z = E_{\varphi}^{(t)}(x)$，$\hat{x} = D_{\theta}^{(t)}(z)$;
3: 根据目标函数 3.16 计算 $\hat{L}_{MASS}(\theta, \varphi)$;
4: 生成器根据输入的随机噪声 ϵ 生成伪造样本 z';
5: 根据损失函数 3.20 更新判别器的参数 ω:
$\omega \leftarrow \omega - \alpha \frac{\partial}{\partial \omega} J(F)$;
6: 更新编码器和解码器的参数 φ、θ：
$\varphi \leftarrow \varphi - \alpha \frac{\partial}{\partial \varphi} \hat{L}_{MASS}(\theta, \varphi)$
$\theta \leftarrow \theta - \alpha \frac{\partial}{\partial \theta} \hat{L}_{MASS}(\theta, \varphi)$;
7: 根据损失函数 3.21 更新生成器的参数 g：
$g \leftarrow g - \alpha \frac{\partial}{\partial g} J(G)$;
8: **return** $E_{\varphi}^{(t+1)}$、$D_{\theta}^{(t+1)}$、$G_{g}^{(t+1)}$、$F_{\omega}^{(t+1)}$ 的参数;

第 4 章　LDC-VAE

在本章中，我们提出了潜在分布一致性 VAE（LDC-VAE）的方法，旨在解决 ELBO 优化中后验潜在分布与先验潜在分布之间的实质性不一致性问题。我们利用信息瓶颈理论，假设解码器分布服从 Gibbs 分布，潜在变量 z 的分布也是 Gibbs 分布形式，以此来估计 z 的后验分布。然而，传统的吉布斯后验方法在近似中没有解析解，并且使用基于迭代采样的 MCMC 的传统近似方法非常耗时。为了解决这个问题，我们采用了 Stein 变分梯度下降（SVGD）方法来近似吉布斯后验。

4.1　Stein 变分梯度下降（SVGD）

SVGD 是一种非参数变分推理算法，它集成了来自 Stein 方法 [38]、核方法 [39] 和变量推理 [40] 的思想。近年来，SVGD 在机器学习中取得了成功。

Proposition 18. 我们假设 $q(x) : X \subset Rn$ 是一个连续可微的、正的概率密度函数，并且 $\Phi(x) : \mathbb{R}^n \to \mathbb{R}^n$ 等于 $[\phi_1(x), \ldots, \phi_d(x)]^T$，这是一个平滑的向量函数。

SVGD 的目标是传输一组初始粒子 $\{x_i\}_{i=1}^n$ 来近似给定的目标后验分布 $p(x)$。粒子集合 $\{x_i\}_{i=1}^n$ 从分布 $q(x)$ 中采样。SVGD 通过利用有效的确定性动力学迭代更新初始粒子集 $\{x_i\}_{i=1}^n : x_i \leftarrow x_i + \epsilon \boldsymbol{\Phi}^*(x_i)$ 来实现近似，其中 ϵ 是一个小步长，$\boldsymbol{\Phi}^*$ 是一个选择最大限度地减少 KL 散度的函数。$\boldsymbol{\Phi}^*$ 是

$$\boldsymbol{\Phi}^* = \arg\max_{\boldsymbol{\Phi} \in \mathcal{B}} \lim_{\epsilon \to 0} \left\{ -\frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathrm{KL}\left(q_{[\epsilon\boldsymbol{\Phi}]}(x) \| p(x)\right) \right\} \tag{4.1}$$

其中，$q_{[\epsilon\Phi]}$ 是更新后利用 $x \leftarrow x + \epsilon\boldsymbol{\Phi}(x)$ 的粒子分布，$x \sim q(x)$；$\mathcal{B}$ 表示再生核希尔伯特空间（RKHS）的单位球：$\mathcal{H}^d := \mathcal{H}_0 \times \mathcal{H}_0 \cdots \mathcal{H}_0$ 和 $\mathcal{B} = \left\{\boldsymbol{\Phi} \in \mathcal{H}^d \mid \|\boldsymbol{\Phi}\|_{\mathcal{H}^d} \leq 1\right\}$。幸运的是，Liu, Qiang 等人 [38] 证明了等式的结果 4.1 可以转化为 $\boldsymbol{\Phi}$ 的线性函数，

$$\begin{gathered} \lim_{\epsilon \to 0} -\frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathrm{KL}\left(q_{[\epsilon\boldsymbol{\Phi}]}(x) \| p(x)\right) = \mathbb{E}_{x \sim q(x)}\left[\operatorname{trace}\left(\mathcal{T}_p^\top \boldsymbol{\Phi}(\boldsymbol{x})\right)\right] \\ \mathcal{T}_p^\top \boldsymbol{\Phi}(x) = \nabla_x \log p(x)^\top \boldsymbol{\Phi}(x) + \nabla_x^\top \boldsymbol{\Phi}(x) \end{gathered} \tag{4.2}$$

其中 $\mathcal{T}_p$ 被称为 Stein 算符。$\mathcal{T}_p$ 和导数 ∇_x 被认为是 $\mathbb{R}^n$ 列向量，因此 $\mathcal{T}_p^\top \boldsymbol{\Phi}(x)$ 和 $\nabla_x^\top \boldsymbol{\Phi}(x)$ 可以看作是内积。例如，$\nabla_x^\top \boldsymbol{\Phi}(x) = \sum_{j=1}^d \nabla_{x^j} \phi_j(x) = \left\langle \nabla_x^\top, \boldsymbol{\Phi} \right\rangle$，其中 x^j 和 ϕ_j 是向量

x 和 $\mathbf{\Phi}$ 的第 j 个变量。由于 $\mathbf{\Phi} \in \mathcal{H}^d$，等式 4.2 等于

$$\mathbf{\Phi}^*_{q,p}(\cdot) = \mathbb{E}_{x\sim q(x)}\left[\mathcal{K}(x,\cdot)\nabla_x \log p(x) + \nabla_x \mathcal{K}(x,\cdot)\right] \tag{4.3}$$

其中，$\mathcal{K}(x,\cdot)$ 是一个与 RKHS $\mathcal{H}^d$ 相关联的正定义核。我们得到了 Stein 变分梯度来近似于 $\{x_i\}_{i=1}^n \sim q(x)$ 到 $p(x)$。

4.2 LDC-VAE

与现有的试图改进 ELBO 的工作不同，我们基于信息瓶颈提出了一种改进变分推理的替代方法，而不需要优化 VAEs 中的 ELBO。我们对真实联合分布 $p_\theta(x,z)$ 和近似联合分布 $q_\varphi(x,z)$ 之间的 KL 散度进行建模，并通过近似后验 $q_\varphi(z|x)$ 来优化它，而不是最大化 ELBO。我们的目的方法是通过假设 $p_\theta(x|z)$ 的分布是 Gibbs 分布来实现的。为了实现有效的训练，我们使用了 Stein 变分梯度下降（SVGD）计算学习到的后验 $q_\varphi(z|x)$ 和 $p_\theta(z|x)$ 之间的 KL 散度梯度，这是受 Stein-VAE [41] 的启发。最后，为了解决从吉布斯后验采样需要时间昂贵的迭代方法，如马尔可夫链蒙特卡罗（MCMC）方法 [17]，我们用 SVGD 训练一个采样器网，从吉布斯后验进行有效采样。

第 5 章　实证分析

5.1　仿真研究

结论

本文推导了拟似然估计的最佳泊松子抽样概率，并且提出分布式最优子抽样算法。我们研究了所提出方法的理论特性，并在模拟数据集和真实数据集上进行了大量的数值实验，以评估它们的实际性能。理论结果和数值结果都表明了该方法在从海量数据中提取有用信息方面的巨大潜力。

参考文献

[1] Tishby N. The information bottleneck method [J], 1999.

[2] Bardera A, Feixas M, Boada I, et al. Registration-Based Segmentation Using the Information Bottleneck Method [J]. Springer-Verlag, 2007.

[3] Kraskov A H, Stogbauer P. Grassberger. Estimating mutual information [J], 2004.

[4] Shwartz-Ziv R, Tishby N. Opening the Black Box of Deep Neural Networks via Information [J], 2017.

[5] Achille A, Soatto S. On the Emergence of Invariance and Disentangling in Deep Representations [J], 2017.

[6] Alemi A A, Fischer I, Dillon J V, et al. Deep Variational Information Bottleneck [J], 2016.

[7] Barber D, Agakov F. Information Maximization in Noisy Channels : A Variational Approach [J], 2022.

[8] Hjelm R D, Fedorov A, Lavoie-Marchildon S, et al. Learning deep representations by mutual information estimation and maximization [J], 2018.

[9] Shamir O, Sabato S, Tishby N. Learning and generalization with the information bottleneck [J]. Theoretical Computer Science, 2010, 411 (29–30): 2696–2711.

[10] 谢盛嘉, 梁竞敏. 信息熵和信息瓶颈算法在图像聚类中的应用 [J]. 计算机工程与应用, 2010, 46 (34): 4.

[11] Slonim N, Friedman N, Tishby N. Multivariate Information Bottleneck [J]. Neural Computation, 2006, 18 (8): 1739.

[12] 于志强, 余正涛, 黄于欣, et al. 基于变分信息瓶颈的半监督神经机器翻译 [J]. 自动化学报, 2022, 48 (7): 12.

[13] 卓越, 姜黎. 一种基于信息瓶颈的神经网络混合压缩方法 [J]. 计算机应用研究, 2021.

[14] 贺一帆, 江铭虎. 网络文本分类中基于信息瓶颈的特征提取 [J]. 清华大学学报：自然科学版, 2010 (1): 5.

[15] Saxe A M, Bansal Y, Dapello J, et al. On the information bottleneck theory of deep learning [C]. In ICLR 2018, 2018.

[16] Crutchfield J P, Mahoney J R, James R G. Trimming the Independent Fat: Sufficient Statistics, Mutual Information, and Predictability from Effective Channel States. 2017.

[17] 何鹏光. 充分统计量的证明及其相关结论 [J]. 阜阳师范学院学报：自然科学版, 2006, 23 (3): 3.

[18] Cvitkovic M, Koliander G. Minimal Achievable Sufficient Statistic Learning [J], 2019.

[19] Chen Y, Zhang D, Gutmann M, et al. Neural Approximate Sufficient Statistics for Implicit Models [J], 2020.

[20] Jiang B, Wu T Y, Zheng C, et al. Learning Summary Statistic for Approximate Bayesian Computation via Deep Neural Network [J]. Statistics, 2015.

[21] Joyce P, Marjoram P. Approximately Sufficient Statistics and Bayesian Computation [J]. Statistical Applications in Genetics and Molecular Biology, 2008, 7 (1): Article26.

[22] Johann B, Gilles L, Juan P, et al. Mining gold from implicit models to improve likelihood-free inference. [J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117: 5242–5249.

[23] Chan J, Perrone V, Spence J P, et al. A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks [J]. Advances in neural information processing systems, 2018, 8603-8614.

[24] Wiqvist S, Mattei P A, Picchini U, et al. Partially Exchangeable Networks and Architectures for Learning Summary Statistics in Approximate Bayesian Computation [J], 2019.

[25] Creel M, Kristensen D. On selection of statistics for approximate Bayesian computing (or the method of simulated moments) [J]. Computational Statistics and Data Analysis, 2016, 100: 99–114.

[26] Fearnhead P, Prangle D. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2012, 74 (3).

[27] Gratton D R J. Monte Carlo Methods of Inference for Implicit Statistical Models [J]. Journal of the Royal Statistical Society. Series B (Methodological), 1984, 46 (2): 193–227.

[28] 尹灿斌, 贾鑫. 充分统计量在数字信号处理中的应用研究 [J]. 宇航计测技术, 2006, 26 (3): 6.

[29] Zhao S, Song J, Ermon S. InfoVAE: Information Maximizing Variational Autoencoders [J], 2017.

[30] Mescheder L, Nowozin S, Geiger A. Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks [J], 2017.

[31] Kingma D P, Welling M. Auto-Encoding Variational Bayes [J]. arXiv.org, 2014.

[32] Kipf T N, Welling M. Variational Graph Auto-Encoders [J], 2016.

[33] Liang D, Krishnan R G, Hoffman M D, et al. Variational Autoencoders for Collaborative Filtering. 2018.

[34] 林焱辉, 李春波. 基于变分自编码器的多维退化数据生成方法 [J]. 北京航空航天大学学报, 2023, 49 (10): 2617–2627.

[35] 张雪菲, 程乐超, 白升利, et al. 基于变分自编码器的人脸图像修复 [J]. 计算机辅助设计与图形学学报, 2020, 32: 401–409.

[36] 伍美霖, 黄佳进, 秦进. 用于协同过滤的序列解耦变分自编码器 [J]. 计算机科学, 2022, 49 (12): 7.

[37] Rolfe J T. Discrete Variational Autoencoders [J], 2016.

[38] Liu Q, Wang D. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm [J], 2016.

[39] Liu Q, Lee J D, Jordan M I. A Kernelized Stein Discrepancy for Goodness-of-fit Tests [J]. JMLR.org, 2016.

[40] Liu Q, Wang D. Stein Variational Gradient Descent as Moment Matching [J], 2018.

[41] Pu Y, Gan Z, Henao R, et al. VAE Learning via Stein Variational Gradient Descent [J], 2017.

[42] Wang H, Ma Y. Optimal subsampling for quantile regression in big data [J]. Biometrika, 2020+.

附录 A　L-optimality and A-optimality

考虑线性分位数模型：对于给定协变量 $\boldsymbol{x}_i$，响应变量 Y_i 的第 τ 分位数表示为：

$$q_\tau(Y_i|\boldsymbol{x}_i) = \boldsymbol{\beta}^T x_i$$

设 $\epsilon_i = y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i$，且 $f_{\epsilon|X}(\epsilon, \boldsymbol{x}_i)$ 是 ϵ_i 的概率密度函数。则定理A.1和定理A.2给出 L-optimality 和 A-optimality，两定理来自于 Wang H, Ma Y.[42]。

定理 A.1 (L-optimality). *if the sampling probabilities* p_i，$i = 1, \ldots, N$, *are chosen as*

$$p_i^{\text{Lopt}} = \frac{|\tau - I(\varepsilon_i < 0)| \, \|x_i\|}{\sum_{j=1}^{N} |\tau - I(\varepsilon_j < 0)| \, \|x_j\|}, i = 1, 2, \ldots, N \tag{A–1}$$

then the total asymptotic MSE *of* $D_N\widetilde{\beta}$, $\text{tr}(V_p)/n$, *attains its minimum.*

定理 A.2 (A-optimality). *If the sampling probabilities* $p_i, i = 1, \ldots, N$ *are chosen as*

$$p_i^{Aopt} = \frac{|\tau - I(\varepsilon_i < 0)| \, \left\|D_N^{-1} x_i\right\|}{\sum_{j=1}^{N} |\tau - I(\varepsilon_j < 0)| \, \left\|D_N^{-1} x_j\right\|}, i = 1, 2, \ldots, N \tag{A–2}$$

then the total asymptotic MSE of $\widetilde{\beta}$, $\text{tr}\left(D_N^{-1} V_p D_N^{-1}\right)/n$, *attains its minimum.*

其中，n 是子抽样样本集样本个数，

$$V_p = \sum_{i=1}^{N} \frac{\{\tau - I(\epsilon_i < 0)\}^2 \boldsymbol{x}_i \boldsymbol{x}_i^T}{N^2 p_i},$$

D_N 定义为:

$$D_N = \frac{1}{N} \sum_{i=1}^{N} f_{\epsilon|X}(0, \boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^T$$

致谢

本论文的工作是在孔祥顺老师的指导下完成的，感谢孔老师！

作者简介

本人张静雯，就读于北京理工大学数学与统计学院，师从王岩华，专业是应用统计。