Boltzmann Machine

Chen Gong

08 April 2020

目录

1 Introduction **1**

2 基于极大似然的梯度上升 **2**

2.1 似然导数求解 2

2.2 似然梯度下降法汇总 3

2.3 小结 3

3 基于 MCMC 的似然梯度下降 **4**

3.1 MCMC 似然梯度求解总述 4

3.2 条件概率推导 5

3.3 小结 7

4 变分推断法求解 **7**

4.1 平均场理论求解 7

5 总结 **10**

玻尔兹曼机（Boltzmann Machine）在“受限玻尔兹曼机”那一章就有了简单的描述。在那一章我们就较为详细的分析过了，由于 Boltzmann machine 中的依赖关系过于复杂，它的 Learning 和 Inference 问题基本是 intractable。所以，为了简化而提出了受限玻尔兹曼机（Restricted Boltzmann Machine）。但是，为什么又重新谈谈这个似乎不太好的模型呢？主要原因是 Boltzmann Machine 是深度信念网络（DBN），前馈神经网络等网络结构的基础，大名鼎鼎的变分推断（Variational Inference）也是 Hinton 为求解 Boltzmann machine 而提出的。

1 Introduction

Boltzmann machine 节点之间为任意连接，节点可以分为可观测变量 v 和不可观测变量 h。每个节点都符合 $\{0,1\}$ 的伯努利分布。Boltzmann machine 模型的概率图示意图如下所示：

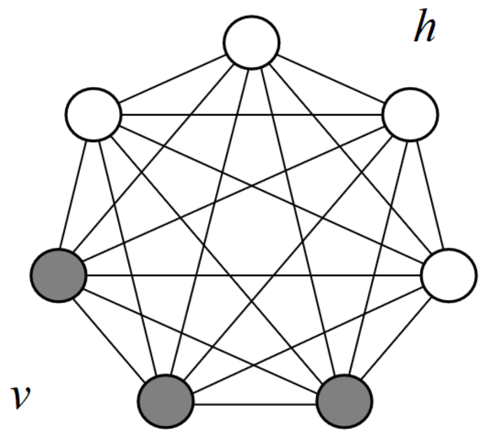

图 1: Boltzmann machine 模型的概率图

其中，$v_{D\times 1}\in\{0,1\}^D$，$h_{P\times 1}\in\{0,1\}^P$。根据“受限玻尔兹曼”那节的知识，可以得出，概率图的联合概率分布为：

$$\begin{cases} P(v,h)=\frac{1}{2}\exp\{-\mathrm{E}(v,h)\} \\ \mathrm{E}(v,h)=-\left(v^\top\cdot W\cdot h+\frac{1}{2}v^\top\cdot Lv+\frac{1}{2}h^\top\cdot J\cdot h\right) \end{cases} \tag{1}$$

其中，$L=[L_{ij}]_{D\times D}$，$J=[J_{ij}]_{P\times P}$，$W=[w_{ij}]_{D\times P}$。我相信坚持学到这里的同学们，对于机器学习的数学推导变换一定有了较好的基础。**实际上矩阵的相乘就是用简单的方式来表示连加，在涉及到求导运算时，矩阵相乘来代替连加符号，可以简化推导过程**。比如，$v^\top\cdot W\cdot h=\sum_{i=1}^D\sum_{j=1}^P v_iw_{ij}h_j$。而 $\frac{1}{2}v^\top\cdot Lv$ 前面为什么要乘上 1/2 呢？实际上打开就知道了，$v^\top\cdot Lv=\sum_{i=1}^D\sum_{j=1}^D v_iw_{ij}v_j$，那么很显然 $v_iw_{ij}v_j=v_jw_{ji}v_i$。所以，所有的值都被加了两次，而我们的目的只要求 v 集合中的任意两个点的乘积，只需要加一次即可，当然需要乘上 $\frac{1}{2}$。

而这个 $\frac{1}{2}$ 又乘与不乘都没有关系，因为 $\frac{1}{2}$ 可以藏在 L 里面，在 Learning 的过程中，自动的缩小 $\frac{1}{2}$ 就可以了。在此问题中，要学习的参数集合为 $\theta=\{W,L,J\}$。

2 基于极大似然的梯度上升

既然是基于极大似然的梯度上升，显然离不开两个部分，极大似然函数和梯度。总所周知，**极大似然估计的主要思路是，使极大似然函数最大时的参数**。首先明确一下，要求的参数为 $\theta=\{W,L,J\}$。样本集合 v，$|v|=D$。那么，似然函数为：

$$\sum_v P(v)=\sum_v\sum_h P(v,h)$$

那么对数似然函数为（实际上 $\frac{1}{D}$ 加不加对于求解没有什么关系，为了严谨起见还是加上）：

$$\frac{1}{D}\sum_v \log P(v)$$

2.1 似然导数求解

那么，下一步就是对对数似然函数求导，即为：

$$\frac{\partial}{\partial\theta}\frac{1}{D}\sum_v \log p(v)=\frac{1}{D}\sum_v\frac{\partial \log p(v)}{\partial\theta} \tag{2}$$

在“直面配分函数”那章的公式（27），我们已经详细的推导了 Boltzmann Distribution 的 log 似然梯度，

$$\frac{1}{D}\frac{\partial}{\partial\theta}\log P(v)=\frac{1}{D}\left(\sum_h\sum_v P(h,v)\frac{\partial}{\partial\theta}\mathrm{E}(h,v)-\sum_h P(h|v)\frac{\partial}{\partial\theta}\mathrm{E}(h,v)\right) \tag{3}$$

我们主要研究的是对 w 的求导，对其他两个参数矩阵的求导都一样，而且比 w 要更简单一点，这里主要是对 w 求导。小编狠下心来，系统的看了一下矩阵求导，迟早都要学的，建议大家也可以系统的看看，挺有帮助的。那么对 w 参数矩阵的求导如下所示：

$$\begin{aligned}\frac{\partial \log p(v)}{\partial W}&=\sum_v\sum_h p(v,h)\cdot-\left(vh^\top\right)-\sum_h p(h|v)\cdot-\left(vh^\top\right)\\&=\sum_i p(h|v)\cdot vh^\top-\sum_\top\sum_h p(v,h)\cdot vh^\top\end{aligned} \tag{4}$$

其中，$\mathrm{E}(v,h)=-\left(v^\top Wh+\frac{1}{2}v^\top Lv+\frac{1}{2}h^\top Jh\right)$。注意一下，这里的 v 和 h 矩阵的大小分别为，$D\times 1$ 和 $P\times 1$。$v^\top Wh$ 是一个一维的，那么对 $W_{D\times P}$ 求导，得到的也必然是一个 $D\times P$ 的矩阵。那么，很简单可以得到：

$$\frac{1}{D}\sum_v\frac{\partial \log P(v)}{\partial W}=\frac{1}{D}\sum_v\sum_h p(h|v)\cdot vh^T-\frac{1}{D}\sum_v\sum_v\sum_h P(v,h)\cdot vh^T \tag{5}$$

看到其中的 $\frac{1}{D}\sum_v\sum_v\sum_h P(v,h)\cdot vh^T$，对 v 和 h 求完和以后，显然 $\sum_v\sum_h P(v,h)\cdot vh^T$ 是一个常数 C。所以，$frac1D\sum_v\sum_v\sum_h P(v,h)\cdot vh^T=\frac{1}{D}\sum_v C=\frac{1}{D}D\cdot C=\sum_v\sum_h P(v,h)\cdot vh^T$。所以，公式（5）可以改写为：

$$\frac{1}{D}\sum_v\frac{\partial \log P(v)}{\partial W}=\frac{1}{D}\sum_v\sum_h P(h|v)\cdot vh^T-\sum_v\sum_h P(v,h)\cdot vh^T \tag{6}$$

而公式（6）可以被简写为：

$$\frac{1}{D}\sum_v\frac{\partial \log P(v)}{\partial W}=\mathbb{E}_{P_{\text{data}}}\left[vh^\top\right]-\mathbb{E}_{P_{\text{model}}}\left[vh^\top\right] \tag{7}$$

其中，

$$
\begin{aligned}
P_{\text{data}} &= P_{\text{data}}(v) \cdot P_{\text{model}}\,(h|v) \\
P_{\text{model}}\, &= P_{\text{model}}\,(v, h)
\end{aligned}
\tag{8}
$$

为什么这样表达呢? 实际上老师说的很模糊，我谈谈自己的理解。在 $\sum_v \sum_h P(v,h)$ 中，$P(v,h)$ 是生成模型，本身就是我们建立的模型，所以被称为 P_{model}。而在 $\sum_v \sum_h P(h|v)$ 首先从经验分布 $P(v)$ 从采样得到 v，然后利用模型分布来求解 $P(h|v)$，所以 $P_{\text{data}} = P_{\text{data}}(v) \cdot P_{\text{model}}\,(h|v)$。采样出 $P_{\text{model}}\,(h|v)$ 和 $P_{\text{model}}\,(v)$ 就可以求解出 $P_{\text{model}}\,(h,v)$ 了。按照同样的方法可以求得对 $\{L, J\}$ 的导数。

2.2 似然梯度下降法汇总

- Boltzmann Machines 中的节点可以分为可观测变量集合 v 和不可观测变量集合 h。每个节点属于 0/1 分布，$v_{D\times 1} \in \{0,1\}^D$，$h_{P\times 1} \in \{0,1\}^P$。
- 参数集合为：$\theta = \{W, L, J\}$。参数矩阵的大小为：$L = [L_{ij}]_{D\times D}$，$J = [J_{ij}]_{P\times P}$，$W = [w_{ij}]_{D\times P}$。
- Boltzmann Distribution 的模型表示为：

$$
\left\{
\begin{array}{l}
P(v,h) = \frac{1}{2}\exp\{-\mathrm{E}(v,h)\} \\
\mathrm{E}(v,h) = -\left(v^\top \cdot W \cdot h + \frac{1}{2} v^\top \cdot L v + \frac{1}{2} h^\top \cdot J \cdot h\right)
\end{array}
\right.
\tag{9}
$$

- 求解参数用到极大似然估计，Log-Likelihood Function 为：

$$
\frac{1}{D} \sum_v \log P(v) \tag{10}
$$

- 通过计算可以得到每个参数矩阵的似然梯度为：

$$
\left\{
\begin{array}{l}
\Delta W = \alpha \left(\mathbb{E}_{p_{data}}\left[v h^\top\right] - \mathbb{E}_{p_{\text{model}}}\left[v h^\top\right]\right) \\
\Delta L = \alpha \left(\mathbb{E}_{p_{data}}\left[v v^\top\right] - \mathbb{E}_{p_{\text{model}}}\left[v v^\top\right]\right) \\
\Delta J = \alpha \left(\mathbb{E}_{p_{data}}\left[h h^\top\right] - \mathbb{E}_{p_{\text{model}}}\left[h h^\top\right]\right)
\end{array}
\right.
\tag{11}
$$

其中：

$$
\left\{
\begin{array}{l}
P_{\text{data}} = P_{\text{data}}(v) \cdot P_{\text{model}}\,(h|v) \\
P_{\text{model}}\, = P_{\text{model}}\,(v, h)
\end{array}
\right.
\tag{12}
$$

2.3 小结

通过上述的求解发现，梯度的统计量只和 v, h 相关，只不过分布不一样而已。RBM 也是一种特殊的 Boltzmann Machines，RBM 的求解比较的简单。在“直面配分函数”那一章中可以看到，RBM 在化简完毕后，$P_{\text{data}} = P_{\text{data}}(v)$ 不需要考虑 $P_{\text{model}}\,(h|v)$，这样计算起来就非常简单，梯度在理论上很干净。在前馈神经网络中 Gradient 需要使用链式求导法则，计算起来非常的复杂。而这里就不一样，只要解决了后验 $P_{\text{model}}\,(h|v)$ 就可以了。那么，下一个重点就是如何从后验 $P_{\text{model}}\,(h|v)$ 中进行采样。

3 基于 MCMC 的似然梯度下降

3.1 MCMC 似然梯度求解总述

在第二小节中，我们已经讲到了，使用梯度上升法来使 log 似然函数达到最大，从而求解对应的最优参数。参数更新公式为：

$$\theta^{(t+1)} = \theta^{(t)} + \triangle\theta \tag{13}$$

其中，$\triangle\theta = \{\triangle W, \triangle L, \triangle J\}$。以 $\triangle W$ 为例，$\triangle W$ 是一个矩阵 $\triangle W = [\triangle w_{ij}]$。其中，

$$\triangle w_{ij} = \alpha \left[\underbrace{\mathbb{E}_{P_{\text{data}}}[v_i h_j]}_{\text{Postive phase}} - \underbrace{\mathbb{E}_{P_{\text{model}}}[v_i h_j]}_{\text{Negative phase}} \right] \tag{14}$$

这个 Postive 和 Negative phase 的说法，我们在“直面配分函数”那章有详细的描述。那么，**现在的难点就是 $v_i h_j$ 从何而来。**

回忆一下，在 RBM 中，$P(h|v)$ 是可以直接求出来的。

$$P(h|v) = \prod_{l=1}^{m} P\left(h_l|v\right) = \left(\sigma\left(\sum_{j=1}^{n} w_{lj} v_i + \beta_l\right)\right)^k \left(1 - \sigma\left(\sum_{j=1}^{n} w_{lj} v_i + \beta_l\right)\right)^{m-k} \tag{15}$$

而 P_{data} 直接从样本中进行采样就可以了，而 $P_{\text{model}}(v, h)$ 为：

$$\begin{aligned} P(h, v) h_i v_j &= \sum_h \sum_v P(v) P(h|v) h_i v_j \\ &= \sum_v P(v) \sum_h P(h|v) h_i v_j \\ &= \frac{1}{Z} \exp\left(\alpha^T v + \sum_{i=1}^{m} \log\left(1 + \exp\left(w_i v + \beta_i\right)\right)\right) \sigma\left(\sum_{j=1}^{n} w_{ij} v_i + \beta_i\right) v_j \end{aligned} \tag{16}$$

这个分布过于复杂，当时采用的是基于对于散度的 Gibbs 采样来解决。而在 Boltzmann Machines 中，Postive phase 和 Negative phase 都是 Intractable。所以，Hinton 提出了用 MCMC 来对 $P(h|v)$ 进行采样。

这里再明确一下逻辑，在求解 $\triangle W$ 中，主要是解决三个部分，$P_{\textbf{data}}(v), P_{\textbf{model}}\ (h|v), P_{\textbf{model}}\ (v, h)$，其中 $P_{\textbf{model}}\ (v, h) = P_{\textbf{model}}\ (h|v) \cdot P_{\textbf{model}}\ (v)$。所以，而 $P_{\textbf{data}}(v)$ 和 $P_{\textbf{model}}\ (v)$ 相对比较简单，所以难点在于 $P_{\textbf{model}}\ (h|v)$ 的求解。而在 RBM 中 $P_{\textbf{model}}\ (h|v)$ 比较容易求解，而 $P_{\textbf{model}}\ (v, h)$ 过于复杂，所以要采用 MCMC 来解决。而在 Boltzmann Machines 中，由于关系过于复杂，没有办法分解，甚至最大团分解都没有用，因为最大团就是自己，那么连 $P_{\textbf{model}}\ (h|v)$ 都求不出来，那么 Postive phase 和 Negative phase 都是 Intractable。

很幸运的是，通过推导，可以得到：

$$\begin{array}{l} P\left(v_i = 1|h, v_{-i}\right) = \sigma\left(\sum_{j=1}^{P} w_{ij} h_j + \sum_{k=1\setminus i}^{D} L_{ik} v_k\right) \\ P\left(h_j = 1|v, h_{-j}\right) = \sigma\left(\sum_{i=1}^{D} w_{ij} v_i + \sum_{m=1\setminus j}^{p} J_{jm} h_n\right) \end{array} \tag{17}$$

解释一下，这两个公式是什么意思。公式表达的是，在已知一个节点以外的所有的点的条件下，这个节点的条件概率是可求的。其中 $1 \setminus i$ 表达的意思是 $1 \sim D$ 但不包括 i 的所有节点。

为什么说很幸运呢？因为真实的后验是求不出来的，但是 MCMC 提供了一种一维一维的采样的方法（Gibbs 采样法）。而每一个维的概率分布可以求出来，那么 Gibbs 采样就可以很愉快的被使用了。而且，这个结论同时也可以在 RBM 中使用，下面我们来举个例子，假设有一个 RBM，如下图所示：

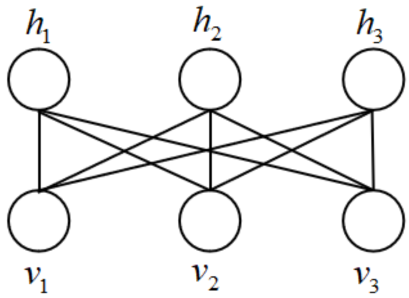

图 2: RBM 概率图模型

由于在已知 v 的情况下，h 中的节点都是相互独立的，所以:

$$P(h|v)=\prod_{j=1}^{3} p\left(h_j|v\right) \tag{18}$$

同理可得:

$$P\left(h_j=1|v\right)=P\left(h_j=1|v,h_{-j}\right)=\sigma\left(\sum_{i=1}^{D} w_{ij}v_i+0\right) \tag{19}$$

为什么 $\sum_{m=1\backslash j}^{p} J_{jm}h_n=0$ 呢？因为，h 节点内部都是相互独立的，没有边，所有都是 0。实际上在 RBM 那一章，后验是花了较大的功夫去求的。而使用公式（17）给出的结论，我们可以较为简单的写出。可以看到，由于 RBM 的特殊性质，h 集合之间相互独立，分解起来非常简单。在 BM 就没有这么好了，虽然每一维可以求出来，由于无法分解，求解起来根本就不可能。

3.2 条件概率推导

在 3.1 节中，给出了两个条件概率分布:

$$\begin{aligned} P\left(v_i=1|h,v_{-i}\right)=\sigma\left(\sum_{j=1}^{P} w_{ij}h_j+\sum_{k=1\backslash i}^{D} L_{ik}v_k\right) \\ P\left(h_j=1|v,h_{-j}\right)=\sigma\left(\sum_{i=1}^{D} w_{ij}v_i+\sum_{m=1\backslash j}^{p} J_{jm}h_n\right) \end{aligned} \tag{20}$$

这一节，就来详细的推导一下:

$$\begin{aligned} P\left(v_i|h,v_{-i}\right)=\frac{P(v,h)}{P(h,v_{-i})} &= \frac{\frac{1}{Z}\exp\{-\mathbb{E}(v,h)\}}{\sum_{v_i}\frac{1}{Z}\exp(-\mathbb{E}(v,h)\}}=\frac{\exp\left\{v^\top Wh+\frac{1}{2}v^\top Lv+\frac{1}{2}h^\top Jh\right\}}{\sum_{v_i}\exp\left\{v^\top Wh+\frac{1}{2}v^\top Lv+\frac{1}{2}h^\top Jh\right\}} \\ &= \frac{\exp\left\{v^\top Wh+\frac{1}{2}v^\top Lv\right\}\cdot\exp\left\{\frac{1}{2}h^\top Jh\right\}}{\exp\left\{\frac{1}{2}h^\top Jh\right\}\cdot\sum_{v_i}\exp\left\{v^\top Wh+\frac{1}{2}v^\top Lv\right\}} \end{aligned} \tag{21}$$

由于 $\exp\left\{\frac{1}{2}h^\top Jh\right\}$ 和 v 没有关系，所以被单独提出来准备约掉。那么有:

$$P\left(v_i|h,v_{-i}\right)=\frac{\exp\left\{v^\top Wh+\frac{1}{2}v^\top Lv\right\}}{\sum_{v_i}\exp\left\{v^\top Wh+\frac{1}{2}v^\top Lv\right\}} \tag{22}$$

令 $v_i=1$ 和分母部分没有关系，因为 $\sum_{v_i}$ 之后，是和 v_i 无关的部分了。所以，

$$P\left(v_i=1|h,v_{-i}\right)=\frac{\exp\left\{v^\top Wh+\frac{1}{2}v^\top Lv\right\}|_{v_i=1}}{\exp\left\{v^\top Wh+\frac{1}{2}v^\top Lv\right\}|_{v_i=1}+\exp\left\{v^\top Wh+\frac{1}{2}v^\top Lv\right\}|_{v_i=0}} \tag{23}$$

为了简化公式，我们将公式简写为：

$$P\left(v_i=1|h,v_{-i}\right)=\frac{\triangle|_{v_i=1}}{\triangle|_{v_i=0}+\triangle|_{v_i=1}} \tag{24}$$

下一步很自然的想到，将包含 v_i 的项，从公式中分离，然后赋予相应的值。

$$\begin{aligned}\Delta v_i=&\exp\left\{v^\top wh+\frac{1}{2}v^\top Lv\right\}=\exp\left\{\sum_{\hat{i}=1}^{D}\sum_{j=1}^{P}v_{\hat{i}}w_{\hat{i}j}h_j+\frac{1}{2}\sum_{\hat{i}=1}^{D}\sum_{k=1}^{D}v_{\hat{i}}L_{\hat{i}k}v_k\right\}\\=&\exp\left\{\sum_{\hat{i}=1\backslash i}^{D}\sum_{j=1}^{P}v_{\hat{i}}w_{\hat{i}j}h_j+\sum_{j=1}^{P}v_iw_{ij}h_j\right.\\&\left.+\frac{1}{2}\left(\sum_{\hat{i}=1\backslash i}^{D}\sum_{k=1}^{D}v_{\hat{i}}L_{\hat{i}k}v_k+v_iL_{ii}v_i+\sum_{\hat{i}=1\backslash i}^{D}v_{\hat{i}}L_{\hat{i}i}v_i+\sum_{\hat{k}=1\backslash i}^{D}v_iL_{ik}v_k\right)\right\}\end{aligned} \tag{25}$$

又因为 $L_{ii}=0$，且 L 矩阵是对称的，所以 $\sum_{\hat{i}=1\backslash i}^{D}v_{\hat{i}}L_{\hat{i}i}v_i=\sum_{\hat{k}=1\backslash i}^{D}v_iL_{ik}v_k$。所以，

$$\begin{aligned}\Delta v_i=&\exp\left\{\sum_{\hat{i}=1\backslash i}^{D}\sum_{j=1}^{P}v_{\hat{i}}w_{\hat{i}j}h_j+\sum_{j=1}^{P}v_iw_{ij}h_j+\frac{1}{2}\left(\sum_{\hat{i}=1\backslash i}^{D}\sum_{k=1}^{D}v_{\hat{i}}L_{\hat{i}k}v_k+2\sum_{\hat{k}=1\backslash i}^{D}v_iL_{ik}v_k\right)\right\}\\=&\exp\left\{\sum_{\hat{i}=1\backslash i}^{D}\sum_{j=1}^{P}v_{\hat{i}}w_{\hat{i}j}h_j+\sum_{j=1}^{P}v_iw_{ij}h_j+\frac{1}{2}\sum_{\hat{i}=1\backslash i}^{D}\sum_{k=1}^{D}v_{\hat{i}}L_{\hat{i}k}v_k+\sum_{\hat{k}=1\backslash i}^{D}v_iL_{ik}v_k\right\}\end{aligned} \tag{26}$$

其中，$\frac{1}{2}\sum_{\hat{i}=1}^{D}\sum_{k=1}^{D}v_{\hat{i}}L_{\hat{i}k}v_k$ 是按这样的方式进行分解：

$$\begin{cases}\hat{i}\neq i,k\neq i & (D-1)(D-1)\\\hat{i}=i,k=i & 1\\\hat{i}=i,k\neq i & (D-1)\\\hat{i}\neq i,k=i & (D-1)\end{cases} \tag{27}$$

而 $(D-1)(D-1)+(D-1)+(D-1)+1=D^2$。

那么，使用公式（26）的推导结果，可以得到：

$$\Delta v_{i=0}=\exp\left\{\sum_{\hat{i}=1\backslash i}^{D}\sum_{j=1}^{P}v_{\hat{i}}w_{\hat{i}j}h_j+\frac{1}{2}\sum_{\hat{i}=1\backslash i}^{D}\sum_{k=1}^{D}v_{\hat{i}}L_{\hat{i}k}v_k\right\}=\exp\left\{A+B\right\} \tag{28}$$

其中，$A=\sum_{\hat{i}=1\backslash i}^{D}\sum_{j=1}^{P}v_{\hat{i}}w_{\hat{i}j}h_j,B=\frac{1}{2}\sum_{\hat{i}=1\backslash i}^{D}\sum_{k=1}^{D}v_{\hat{i}}L_{\hat{i}k}v_k$。同理可得：

$$\Delta v_{i=1}=\exp\left\{A+B+\sum_{j=1}^{P}w_{ij}h_j+\sum_{\hat{k}=1\backslash i}^{D}L_{ik}v_k\right\} \tag{29}$$

所以，将公式（28）和（29）的结果代入到公式（24）中可得：

$$
\begin{aligned}
P\left(v_i=1|h,v_{-i}\right)=&\frac{\triangle|_{v_i=1}}{\triangle|_{v_i=0}+\triangle|_{v_i=1}}\\
=&\frac{\exp\left\{A+B+\sum_{j=1}^{P}w_{ij}h_j+\sum_{\hat{k}=1\backslash i}^{D}L_{ik}v_k\right\}}{\exp\left\{A+B+\sum_{j=1}^{P}w_{ij}h_j+\sum_{\hat{k}=1\backslash i}^{D}L_{ik}v_k\right\}+\exp\left\{A+B\right\}}\\
=&\frac{\exp\left\{\sum_{j=1}^{P}w_{ij}h_j+\sum_{\hat{k}=1\backslash i}^{D}L_{ik}v_k\right\}}{\exp\left\{\sum_{j=1}^{P}w_{ij}h_j+\sum_{\hat{k}=1\backslash i}^{D}L_{ik}v_k\right\}+1}\\
=&\sigma\left(\sum_{j=1}^{P}w_{ij}h_j+\sum_{\hat{k}=1\backslash i}^{D}L_{ik}v_k\right)
\end{aligned}
\tag{30}
$$

而 $P\left(h_j=1|v,h_{-j}\right)=\sigma\left(\sum_{i=1}^{D}w_{ij}v_i+\sum_{m=1\backslash j}^{p}J_{jm}h_n\right)$ 的计算采用的也是同样的思路。

3.3 小结

本小节主要讲述了基于 MCMC 的似然梯度下降法，不同于 RBM，在 BM 中后验分布 $P(h|v)$ 过于复杂，所以采用 MCMC 采样的思路来求解。幸运的是，$P(h|v)$ 的条件概率是可求的，所以，可以用 Gibbs 采样。然后，给出了条件概率的详细推导。

4 变分推断法求解

我们采用的是梯度上升法，那么在每一次求解梯度的过程中，都要采样得到 $vh^\top$。在采样的过程中，主要是对 $P_{\text{model}}(h|v)$ 进行采样，使用 MCMC 采样的劣势大家都很清楚，无法求解大规模问题。如何求解大规模问题一直是难点，直到 90 年代初，Hinton 提出了变分推断法（Variational Inference）来求 $P_{\text{model}}(h|v)$。

4.1 平均场理论求解

这部分的基础思想，在“近似推断”那一章有非常详细的描述。大体上说就是通过优化下界 ELBO，来达到求解的效果，有兴趣的同学请回顾“近似推断”。公式近似推断中的公式（5）可得：

$$
\begin{aligned}
\mathcal{L}=&\text{ELBO}=\log P_\theta(v)-\text{KL}(Q_\phi||P_\theta)\\
=&\sum_h Q_\phi(h|v)\log P_\theta(v,h)+H(Q_\phi)
\end{aligned}
\tag{31}
$$

根据平均场理论（假设分布可以分解成几个部分之积），假定 $Q_\phi(h|v)=\prod_{j=1}^{P}Q_\phi(h_j|v)$，令 $Q_\phi(h_j=1|v)=\phi_j$，ϕ 就可以认为是 $\{\}$。那么推导过程如下所示：

$$
\begin{aligned}
\hat{\phi}_j=&\arg\max_{\phi_j}\mathcal{L}=\arg\max_{\phi_j}\sum_h Q_\phi(h|v)\log P_\theta(v,h)+H(Q_\phi)\\
=&\arg\max_{\phi_j}\sum_h Q_\phi(h|v)\left[-\log Z+v^\top Wh+\frac{1}{2}v^\top Lv+\frac{1}{2}h^\top Jh\right]+H(Q_\phi)
\end{aligned}
\tag{32}
$$

$$=\arg\max_{\phi_j}\sum_h Q_\phi(h|v)\left[-\log Z+\frac{1}{2}v^\top Lv\right]+\arg\max_{\phi_j}\sum_h Q_\phi(h|v)\left[v^\top Wh+\frac{1}{2}h^\top Jh\right]+H(Q_\phi) \quad (33)$$

其中,ϕ_j 是和 h 相关的参数,$\left[-\log Z+\frac{1}{2}v^\top Lv+\right]$ 与 ϕ 没有关系,那么 $\sum_h Q_\phi(h|v)\left[-\log Z+\frac{1}{2}v^\top Lv+\right]$ 可以写成 $\left[-\log Z+\frac{1}{2}v^\top Lv\right]\sum_h Q_\phi(h|v)$。很显然,$\sum_h Q_\phi(h|v)=1$,所以,$\arg\max_{\phi_j}\sum_h Q_\phi(h|v)\left[-\log Z+\frac{1}{2}v^\top Lv\right]$ 和 ϕ 没有关系，可以直接约掉。化简之后,

$$\begin{aligned}\hat{\phi}_j&=\arg\max_{\phi_j}\sum_h Q_\phi(h|v)\left[v^\top Wh+\frac{1}{2}h^\top Jh\right]+H(Q_\phi)\\&=\arg\max_{\phi_j}\sum_h Q_\phi(h|v)v^\top Wh+\frac{1}{2}\sum_h Q_\phi(h|v)h^\top Jh+H(Q_\phi)\\&=\arg\max_{\phi_j}\ ①\ +\ ②\ +\ ③\end{aligned} \quad (34)$$

那么，下一步工作就是将 h_j 分离出来。

$$\begin{aligned}①\ &=\sum_h Q_\phi(h|v)\cdot\sum_{i=1}^D\sum_{j=1}^P v_iw_{ij}h_j\\&=\sum_h\prod_{j=1}^P Q_\phi\left(h_{\hat{j}}|v\right)\cdot\sum_{i=1}^D\sum_{j=1}^P v_iw_{ij}h_j\end{aligned} \quad (35)$$

$\sum_{i=1}^D\sum_{j=1}^P v_iw_{ij}h_j$ 中一共有 $D\times P$ 项，这里太复杂了，我们先挑一项来分析一下。

$$\begin{aligned}①\ &=\sum_h\prod_{j=1}^P Q_\phi\left(h_{\hat{j}}|v\right)\cdot v_1w_{12}h_2\\&=\sum_{h_2}Q_\phi\left(h_2|v\right)\cdot v_1w_{12}h_2\sum_{h\backslash h_2}\prod_{\hat{j}=1\backslash 2}^P Q_\phi\left(h_{\hat{j}}|v\right)\end{aligned} \quad (36)$$

这里将 $\sum_{h\backslash h_2}\prod_{\hat{j}=1\backslash 2}^P Q_\phi\left(h_{\hat{j}}|v\right)$ 提出了分析一下,

$$\sum_{h\backslash h_2}\prod_{\hat{j}=1\backslash 2}^P Q_\phi\left(h_{\hat{j}}|v\right)=\sum_{h_1}Q_\phi\left(h_1|v\right)\sum_{h_3}Q_\phi\left(h_3|v\right)\sum_{h_4}Q_\phi\left(h_4|v\right)\cdots \quad (37)$$

显然，$\sum_{h_1}Q_\phi\left(h_1|v\right)=\sum_{h_3}Q_\phi\left(h_3|v\right)=\sum_{h_4}Q_\phi\left(h_4|v\right)=\cdots=1$。所以，$\sum_{h\backslash h_2}\prod_{\hat{j}=1\backslash 2}^P Q_\phi\left(h_{\hat{j}}|v\right)=1$。那么,

$$\begin{aligned}\sum_{h_2}Q_\phi\left(h_2|v\right)\cdot v_1w_{12}h_2\sum_{h\backslash h_2}\prod_{\hat{j}=1\backslash 2}^P Q_\phi\left(h_{\hat{j}}|v\right)=&\sum_{h_2}Q_\phi\left(h_2|v\right)\cdot v_1w_{12}h_2\\=&Q_\phi\left(h_2=1|v\right)\cdot v_1w_{12}\times 1+Q_\phi\left(h_2=0|v\right)\cdot v_1w_{12}\times 0\\=&Q_\phi\left(h_2=1|v\right)\cdot v_1w_{12}=\phi_2v_1w_{12}\end{aligned} \quad (38)$$

那么，依次类推，可以得出:

$$①\ =\sum_{i=1}^D\sum_j^P\phi_jv_iw_{ij} \quad (39)$$

而 ② 的做法相对复杂一些，基本思想和 ① 的分解，基本一致，也是要想办法将 h_j 分解出来。那么，目标为将其中和 h_j 相关的项分解出来：

$$\sum_{\hat{j}=1}^{P}\sum_{m=1\backslash j}^{P} h_{\hat{j}} J_{\hat{j}m} h_m \tag{40}$$

大体求解思路是可以分成如下四个部分：

$$\begin{cases} \hat{j} \neq j, m \neq j \\ \hat{j} = j, m = j \\ \hat{j} = j, m \neq j \\ \hat{j} \neq j, m = j \end{cases} \tag{41}$$

其中，$\hat{j} = j, m = j$ 的情况下 $J_{jj} = 0$，直接省略掉。$\hat{j} = j, m \neq j$ 和 $\hat{j} \neq j, m = j$ 是对称的，相加起来可以抵掉 $\frac{1}{2}$ 这个系数，而 $\hat{j} \neq j, m \neq j$ 的情况与 h_j 无关。所以：

$$② = \sum_{j=1}^{P}\sum_{m=1\backslash j}^{P} \phi_j \phi_m J_{jm} \tag{42}$$

最后一项 ③ 的化简为：

$$\begin{aligned} ③ &= \sum_{j=1}^{P}\left[\phi_j \log\frac{1}{\phi_j} + (1-\phi_j)\log\frac{1}{(1-\phi_j)}\right] \\ &= -\sum_{j=1}^{P}\left[\phi_j \log\phi_j + (1-\phi_j)\log(1-\phi_j)\right] \end{aligned} \tag{43}$$

我们想得到使 ELBO 最大时对应的 ϕ_j，那么就对 ϕ_j 求偏导，可以得到：

$$\begin{cases} \dfrac{\partial ①}{\partial \phi_j} = \displaystyle\sum_{i=1}^{D} v_i w_{ij} \\ \dfrac{\partial ②}{\partial \phi_j} = \displaystyle\sum_{m=1\backslash j}^{P} \phi_m J_{jm} \\ \dfrac{\partial ③}{\partial \phi_j} = -\log\dfrac{\phi_j}{1-\phi_j} \end{cases} \tag{44}$$

合并起来即为：

$$\frac{\partial[① + ② + ③]}{\partial \phi_j} = 0$$

解得：

$$\phi_j = \sigma\left(\sum_{i=1}^{D} v_i w_{ij} + \sum_{m=1\backslash j}^{P} \phi_m J_{jm}\right) \tag{45}$$

观察一下 ϕ_j 的结果，里面有一个项为 $\sum_{m=1\backslash j}^{P}\phi_m$。所以，利用公式（45）求解最终结果的方法依然比较坎坷。

首先，$\{\phi_j\}_{j=1}^P$ 都赋予一个初始值。然后依次计算 $\phi_1, \phi_2, \cdots, \phi_P$，得到的结果为第一次迭代 $\{\phi^{(1)}\}$。不断的重复这个过程，直到最后收敛为止，收敛时得到的结果 $\{\hat{\phi}_j\}_{j=1}^P$ 就是最终的答案。实际上就是求解不动点方程——公式（45），采用的是坐标上升法求解。利用不动点方程的求解结果，可以得到 Q_ϕ：

$$\{\hat{\phi}_j\}_{j=1}^P \Longrightarrow Q_\phi \tag{46}$$

而 $Q_\phi(h|v) \approx P_{\text{model}}(h|v)$。那么，公式（12）中 P_{data} 的计算基本解决了。那么，就不需要再进行采样了。而对于 P_{model} 还是用 MCMC，实际上 $P_{\text{model}}(h|v)$，采样 $P_{\text{model}}(h,v)$ 难度就小了很多了。理论上，我们给出了一个实际可行的方法。但是，每一步正向用 VI，负向用 Gibbs，计算复杂度还是较大的。而有很多改进的方法，比如之前讲的用基于对比散度的 Gibbs 采样，还有后来的概率对比散度，Deep Boltzmann Machines 等。

5 总结

理一下这章的逻辑思路。首先，我们描述了什么是玻尔兹曼机（Boltzmann Machines），描述了其模型表示。下一个问题，就是如何利用观测数据集来求解参数，我们介绍了基于极大似然的梯度上升，经过推导得出了似然梯度的方向。但是，似然梯度中涉及到对 P_{model} 和 P_{data} 的采样。那么难点就转移到了，如何从 P_{model} 和 P_{data} 中进行采样。通过分析，得到玻尔兹曼机求解主要的难点就是 $P_{\text{model}}(h|v)$ 很难求解。

我们和受限玻尔兹曼机的采样进行了对比，受限玻尔兹曼机中的后验 $P_{\text{model}}(h|v)$ 可以直接计算，而玻尔兹曼机中不行。所以，为了求解后验分布，介绍了 MCMC 中的 Gibbs 采样的思想。Gibbs 采样是一维一维的采样，那么需要满足单个节点的条件概率分布可以求出。幸运的是，Boltzmann Machines 中可以求出。下一步则进行了单个节点条件概率的详细推导。

MCMC 虽然提供了一个理论上的可行方法。可惜，无法解决大规模求解的问题。所以，介绍了 Hinton 提出的变分推断（Variational Inference），用一个简单分布 Q_ϕ 来近似 $P_{\text{model}}(h|v)$。通过推导，我们得到了 ϕ 的不动点方程，使用坐标上升法即可得到 ϕ 的参数表达式。从而成功的求解 $P_{\text{model}}(h|v)$。