深度信念模型

包含很多层的隐变量，可以有效地学习数据的内部特征表示，也可以作为一种有效的非线性降维方法。

玻尔兹曼机和深度信念网络都是生成模型，借助隐变量来描述复杂的数据分布。作为概率图模型，玻尔兹曼机和深度信念网络的共同问题是推断和学习问题。因为这两种模型都比较复杂，并且都包含隐变量，它们的推断和学习一般通过 MCMC 方法来进行近似估计。

玻尔兹曼机（Boltzmann Machine）

每个变量的状态都以一定的概率受到其它变量的影响，可以用概率无向图模型来描述。

1. 每个随机变量是二值的，所有随机变量可以用一个二值的随机向量 $\mathbf{X} \in \{0,1\}^K$ 来表示，其中可观测变量表示为 $\mathbf{V}$，隐变量表示为 $\mathbf{H}$；
2. 所有节点之间是全连接的。每个变量 X_i 的取值依赖于所有其它变量 $\mathbf{X}_{\setminus i}$；
3. 每两个变量之间的相互影响（$X_i \to X_j$ 和 $X_j \to X_i$）是对称的。

变量 $\mathbf{X}$ 的联合概率由玻尔兹曼分布得到，即

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(\frac{-E(\mathbf{x})}{T}\right), \tag{12.1}$$

其中 Z 为配分函数，能量函数 $E(\mathbf{x})$ 的定义为

$$\begin{aligned} E(\mathbf{x}) &\triangleq E(\mathbf{X} = \mathbf{x}) \\ &= -\left(\sum_{i<j} w_{ij}\, x_i\, x_j + \sum_i b_i\, x_i\right), \end{aligned} \tag{12.2}$$

如果两个变量 X_i 和 X_j 的取值都为 1 时，一个正的权重 $w_{ij}>0$ 会使得玻尔兹曼机的能量下降，发生的概率变大；相反，一个负的权重会使得能量上升，发生的概率变小。因此，如果令玻尔兹曼机中的每个变量 X_i 代表一个基本假设，其取值为 1 或 0 分别表示模型接受或拒绝该假设，那么变量之间连接的权重为可正可负的实数，代表了两个假设之间的弱约束关系。一个正的权重表示两个假设可以相互支持。也就是说，如果一个假设被接受，另一个也很可能被接受。相反，一个负的权重表示两个假设不能同时被接受。

一类是搜索问题。当给定变量之间的连接权重，需要找到一组二值向量，使得整个网络的能量最低。另一类是学习问题。当给一组定部分变量的观测值时，计算一组最优的权重。

生成模型

全条件概率

吉布斯采样需要计算每个变量 X_i 的全条件概率 $p(x_i|\mathbf{x}_{\setminus i})$，其中 $\mathbf{x}_{\setminus i}$ 表示除变量 X_i 外其它变量的取值。

定理 12.1－玻尔兹曼机中变量的全条件概率： 对于玻尔兹曼机中的一个变量 X_i，当给定其它变量 $\mathbf{x}_{\setminus i}$ 时，全条件概率 $p(x_i|\mathbf{x}_{\setminus i})$

为

$$p(x_i = 1|\mathbf{x}_{\setminus i}) = \sigma\left(\frac{\sum_j w_{ij}\, x_j + b_i}{T}\right), \tag{12.6}$$

$$p(x_i = 0|\mathbf{x}_{\setminus i}) = 1 - p(x_i = 1|\mathbf{x}_{\setminus i}), \tag{12.7}$$

其中 σ 为logistic sigmoid 函数。

证明. 首先，保持其它变量 $\mathbf{x}_{\setminus i}$ 不变，改变变量 X_i 的状态，从0（关闭）和1（打开）之间的能量差异（Energy Gap）为

$$\Delta E_i(\mathbf{x}_{\setminus i}) = E(x_i = 0, \mathbf{x}_{\setminus i}) - E(x_i = 1, \mathbf{x}_{\setminus i}) \tag{12.8}$$

$$= \sum_j w_{ij}\, x_j + b_i, \tag{12.9}$$

其中 $w_{ii} = 0$。

又根据玻尔兹曼机的定义可得

$$E(\mathbf{x}) = -T \log p(\mathbf{x}) - T \log Z, \tag{12.10}$$

$$\Delta E_i(\mathbf{x}_{\setminus i}) = -T \ln p(x_i = 0, \mathbf{x}_{\setminus i}) - (-T \ln p(x_i = 1, \mathbf{x}_{\setminus i})) \tag{12.11}$$

$$= T \ln \frac{p(x_i = 1, \mathbf{x}_{\setminus i})}{p(x_i = 0, \mathbf{x}_{\setminus i})} \tag{12.12}$$

$$= T \ln \frac{p(x_i = 1|\mathbf{x}_{\setminus i})}{p(x_i = 0|\mathbf{x}_{\setminus i})} \tag{12.13}$$

$$= T \ln \frac{p(x_i = 1, |\mathbf{x}_{\setminus i})}{1 - p(x_i = 1|\mathbf{x}_{\setminus i})}, \tag{12.14}$$

结合公式(12.14)和(12.14)，得到

$$p(x_i = 1|\mathbf{x}_{\setminus i}) = \frac{1}{1 + \exp\left(-\frac{\Delta E_i(\mathbf{x}_{\setminus i})}{T}\right)} \tag{12.15}$$

$$= \sigma\left(\frac{\sum_j w_{ij}\, x_j + b_i}{T}\right). \tag{12.16}$$

吉布斯采样

随机选择一个变量X_i，然后根据其全条件概率 $p(x_i|x_{\setminus i})$来设置其状态，即以$p(x_i = 1|x_{\setminus i})$的概率将变量$X_i$设为 1，否则为 0。在固定温度 T 的情况下，在运行足够时间之后，玻尔兹曼机会达到热平衡。此时，任何全局状态的概率服从玻尔兹曼分布 p(x)，只与系统的能量有关，与初始状态无关。

要使得玻尔兹曼机达到热平衡，其收敛速度和温度 T 相关。当系统温度非常高 $T\to\infty$时，$p(x_i = 1|x_{\setminus i})\to 0.5$，即每个变量状态的改变十分容易，每一种系统状态都是一样的，而从很快可以达到热平衡。当系统温度非常低 $T\to 0$ 时，如果 $\Delta E_i(x_{\setminus i}) > 0$ 则$p(x_i = 1|x_{\setminus i})\to 1$，如果$\Delta E_i(x_{\setminus i}) < 0$ 则$p(x_i = 1|x_{\setminus i})\to 0$。

$$x_i = \begin{cases} 1(\sum_j w_{ij}x_j + b_i \geq 0) \\ 0(otherwise) \end{cases}$$

当 T→0 时，随机性方法变成了确定性方法。这时，玻尔兹曼机退化为一个 Hopfield 网络。

Hopfield 网络是一种确定性的动力系统，而玻尔兹曼机是一种随机性的动力系统。Hopfield 网络的每次的状态更新都会使得系统的能量降低，而玻尔兹曼机则以一定的概率使得系统的能量上升。

能量最小化与模拟退火

让系统刚开始在一个比较高的温度下运行达到热平衡，然后逐渐降低，直到系统在一个比较低的温度下达到热平衡。这样我们就能够得到一个能量全局最小的分布。模拟退火算法所得解依概率收敛到全局最优解。

参数学习

给定一组可观测的向量 $D = \{\hat{v}^{(1)}, \hat{v}^{(2)}, \cdots, \hat{v}^{(N)}\}$ 作为训练集，我们要学习玻尔兹曼机的参数 W 和 b 使得训练集中所有样本的对数似然函数最大。训练集的对数似然函数定义为

$$\mathcal{L}(\mathcal{D}|W, \mathbf{b}) = \frac{1}{N}\sum_{n=1}^{N} \log p(\hat{\mathbf{v}}^{(n)}|W, b) \tag{12.18}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \log \sum_{\mathbf{h}} p(\hat{\mathbf{v}}^{(n)}, \mathbf{h}|W, \mathbf{b}) \tag{12.19}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \log \frac{\sum_{\mathbf{h}} \exp\left(-E(\hat{\mathbf{v}}^{(n)}, \mathbf{h})\right)}{\sum_{\mathbf{v},\mathbf{h}} \exp\left(-E(\mathbf{v}, \mathbf{h})\right)}. \tag{12.20}$$

对数似然函数$L(D|W,b)$中参数θ的偏导数为

$$\frac{\mathcal{L}(\mathcal{D}|W, \mathbf{b})}{\partial\theta} = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial\theta} \log \sum_{\mathbf{h}} p(\hat{\mathbf{v}}^{(n)}, \mathbf{h}|W, \mathbf{b}) \tag{12.21}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial\theta}\left(\log \sum_{\mathbf{h}} \exp\left(-E(\hat{\mathbf{v}}^{(n)}, \mathbf{h})\right) - \log \sum_{\mathbf{v},\mathbf{h}} \exp\left(-E(\mathbf{v}, \mathbf{h})\right)\right) \tag{12.22}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \sum_{\mathbf{h}} \frac{\exp\left(-E(\hat{\mathbf{v}}^{(n)}, \mathbf{h})\right)}{\sum_{\mathbf{h}} \exp\left(-E(\hat{\mathbf{v}}^{(n)}, \mathbf{h})\right)} \left[\frac{\partial E(\hat{\mathbf{v}}^{(n)}, \mathbf{h})}{\partial\theta}\right]$$

$$- \sum_{\mathbf{v},\mathbf{h}} \frac{\exp\left(-E(\mathbf{v}, \mathbf{h})\right)}{\sum_{\mathbf{v},\mathbf{h}} \exp\left(-E(\mathbf{v}, \mathbf{h})\right)} \left[\frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial\theta}\right] \tag{12.23}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \sum_{\mathbf{h}} p(\mathbf{h}|\hat{\mathbf{v}}^{(n)}) \left[\frac{\partial E(\hat{\mathbf{v}}^{(n)}, \mathbf{h})}{\partial\theta}\right] - \sum_{\mathbf{v},\mathbf{h}} p(\mathbf{v}, \mathbf{h}) \left[\frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial\theta}\right] \tag{12.24}$$

$$= \mathbb{E}_{\hat{p}(\mathbf{v})}\mathbb{E}_{p(\mathbf{h}|\mathbf{v})}\left[\frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial\theta}\right] - \mathbb{E}_{p(\mathbf{v},\mathbf{h})}\left[\frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial\theta}\right], \tag{12.25}$$

整个训练集的对数似然函数$L(D|W,b)$对每个权重w_{ij}和偏置b_i的偏导数为

$$\frac{\partial \mathcal{L}(\mathcal{D}|W,\mathbf{b})}{\partial w_{ij}} = \mathbb{E}_{\hat{p}(\mathbf{v})}\mathbb{E}_{p(\mathbf{h}|\mathbf{v})}[x_i x_j] - \mathbb{E}_{p(\mathbf{v},\mathbf{h})}[x_i x_j], \tag{12.26}$$

$$\frac{\partial \mathcal{L}(\mathcal{D}|W,\mathbf{b})}{\partial b_i} = \mathbb{E}_{\hat{p}(\mathbf{v})}\mathbb{E}_{p(\mathbf{h}|\mathbf{v})}[x_i] - \mathbb{E}_{p(\mathbf{v},\mathbf{h})}[x_i], \tag{12.27}$$

为了近似近似这个期望，我们可以固定住可观测变量，只对 h 进行吉布斯采样。当玻尔兹曼机达到热平衡状态时，采样$x_i x_j$的值。在训练集上所有的训练样本上重复此过程，得到$x_i x_j$的近似期望$\langle x_i x_j \rangle_{data}$。第二项为玻尔兹曼机在没有任何限制时，$x_i x_j$的期望。我们可以对所有变量进行吉布斯采样。当玻尔兹曼机达到热平衡状态时，采样$x_i x_j$的值，得到近似期望$\langle x_i x_j \rangle_{model}$。

当采用梯度上升法时，权重w_{ij}可以用下面公式近似地更新

$$w_{ij} = w_{ij} + \alpha(\langle x_i x_j \rangle_{data} - \langle x_i x_j \rangle_{model})$$

虽然我们优化目标是整个网络的能量最低，但是每个权重的更新只依赖于它连接的相关变量的状态。

受限玻尔兹曼机

全连接的玻尔兹曼机在理论上十分有趣，但是由于其复杂性，目前为止并没有被广泛使用。虽然基于采样的方法在很大程度提高了学习效率，但是每更新一次权重，就需要网络重新达到热平衡状态，这个过程依然比较低效，需要很长时间。

受限玻尔兹曼机（Restricted Boltzmann Machine， RBM）是一个二分图结构的无向图模型。受限玻尔兹曼机中的变量也分为隐藏变量和可观测变量。我们分别用可观测层和隐藏层来表示这两组变量。同一层中的节点之间没有连接，而不同层一个层中的节点与另一层中的所有节点连接，这和两层的全连接神经网络的结构相同。

一个受限玻尔兹曼机由m_1个可观测变量和m_2个隐变量组成，其定义如下：

- 可观测的随机向量 $\mathbf{v} = [v_1, \cdots, v_{m_1}]^{\mathrm{T}}$；
- 隐藏的随机向量 $\mathbf{h} = [h_1, \cdots, h_{m_2}]^{\mathrm{T}}$；
- 权重矩阵 $W \in \mathbb{R}^{m_1 \times m_2}$，其中每个元素 w_{ij} 为可观测变量 v_i 和隐变量 h_j 之间边的权重；
- 偏置 $\mathbf{a} \in \mathbb{R}^{m_1}$ 和 $\mathbf{b} \in \mathbb{R}^{m_2}$，其中 a_i 为每个可观测的变量 v_i 的偏置，b_j 为每个隐变量 h_j 的偏置。

受限玻尔兹曼机的能量函数定义为

$$E(v,h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i w_{ij} h_j = -a^T v - b^T h - v^T W h$$

受限玻尔兹曼机的联合概率分布 $p(v,h)$定义为

$$p(v,h) = \frac{1}{Z}\exp\big(-E(v,h)\big) = \frac{1}{Z}\exp(a^T v)\exp(b^T h)\exp(v^T W h)$$

生成模型

受限玻尔兹曼机的联合概率分布$p(v,h)$一般也通过吉布斯采样的方法来近似，生成一组服从$p(v,h)$分布的样本。

全条件概率

吉布斯采样需要计算每个变量V_i和H_j的全条件概率。受限玻尔兹曼机中同层的变量之间没有连接。从无向图的性质可知，在给定可观测变量时，隐变量之间相互条件独立，同样在给定隐变量时，可观测变量之间也相互条件独立。

$$p(v_i|\, v_{\setminus i}, h) = p(v_i|h)$$

$$p(v_j|\, v_{\setminus j}, h) = p(v_j|h)$$

定理 12.2－受限玻尔兹曼机中变量的条件概率： 在受限玻尔兹曼机中，每个可观测变量和隐变量的条件概率为

$$p(v_i = 1|\mathbf{h}) = \sigma\left(a_i + \sum_j w_{ij} h_j\right), \tag{12.35}$$

$$p(h_j = 1|\mathbf{v}) = \sigma\left(b_j + \sum_i w_{ij} v_i\right), \tag{12.36}$$

其中σ为logistic sigmoid函数。

吉布斯采样

- （给定）或随机初始化一个可观测的向量v_0，计算隐变量的概率，并从中采样一个隐向量h_0；
- 基于h_0，计算可观测变量的概率，并从中采样一个个可观测的向量v_1；
- 重复 t 次后，获得$(\mathrm{v}_t, \mathrm{h}_t)$；
- 当 t→∞时，$(\mathrm{v}_t, \mathrm{h}_t)$的采样服从$p(v,h)$分布。

参数学习

受限玻尔兹曼机通过最大化似然函数来找到最优的参数 W, a, b。给定一组训练样本$D = \{\hat{v}^{(1)}, \hat{v}^{(2)}, \cdots, \hat{v}^{(N)}\}$，其对数似然函数为

$$L(D|W,a,b) = \frac{1}{N}\sum_{n=1}^{N} logp(\hat{v}^{(n)}|W,a,b)$$

$$\frac{\partial \mathcal{L}(\mathcal{D}|W,\mathbf{a},\mathbf{b})}{\partial w_{ij}} = \mathbb{E}_{\hat{p}(\mathbf{v})}\mathbb{E}_{p(\mathbf{h}|\mathbf{v})}[v_i h_j] - \mathbb{E}_{p(\mathbf{v},\mathbf{h})}[v_i h_j], \tag{12.53}$$

$$\frac{\partial \mathcal{L}(\mathcal{D}|W,\mathbf{a},\mathbf{b})}{\partial a_i} = \mathbb{E}_{\hat{p}(\mathbf{v})}\mathbb{E}_{p(\mathbf{h}|\mathbf{v})}[v_i] - \mathbb{E}_{p(\mathbf{v},\mathbf{h})}[v_i], \tag{12.54}$$

$$\frac{\partial \mathcal{L}(\mathcal{D}|W,\mathbf{a},\mathbf{b})}{\partial b_j} = \mathbb{E}_{\hat{p}(\mathbf{v})}\mathbb{E}_{p(\mathbf{h}|\mathbf{v})}[h_j] - \mathbb{E}_{p(\mathbf{v},\mathbf{h})}[h_j], \tag{12.55}$$

首先，将可观测向量 v 设为训练样本中的值并固定，然后根据条件概率对隐向量 h 进行采样，这时受限玻尔兹曼机的值记为$\langle\cdot\rangle_{data}$。然后在不固定可观测向量 v，通过吉布斯采样来轮

流更新 v 和 h。当达到热平衡状态时，采集 v 和 h 的值，记为$\langle\cdot\rangle_{model}$。

$$w_{ij} \leftarrow w_{ij} + \alpha\Big(\langle v_i h_j\rangle_{\texttt{data}} - \langle v_i h_j\rangle_{\texttt{model}}\Big), \tag{12.56}$$

$$a_i \leftarrow a_i + \alpha\Big(\langle v_i\rangle_{\texttt{data}} - \langle v_i\rangle_{\texttt{model}}\Big), \tag{12.57}$$

$$b_j \leftarrow b_j + \alpha\Big(\langle h_j\rangle_{\texttt{data}} - \langle h_j\rangle_{\texttt{model}}\Big), \tag{12.58}$$

根据受限玻尔兹曼机的条件独立性，可以对可观测变量和隐变量进行分组轮流采样。这样受限玻尔兹曼机的采样效率会比一般的玻尔兹曼机有很大提高，但一般还是需要通过很多步采样才可以采集到符合真实分布的样本。

对比散度学习算法

受限玻尔兹曼机的特殊结构，可以使用对比散度（Contrastive Divergence）仅需 k 步吉布斯采样。为了提高效率，对比散度算法用一个训练样本作为可观测向量的初始值。然后，交替对可观测向量和隐藏向量进行吉布斯采样，不需要等到收敛，只需要 k 步就足够了。

算法 12.1: 单步对比散度算法

输入: 训练集: $\hat{\mathbf{v}}^{(n)}, n = 1, \cdots, N$;

学习率：α

初始化：$W \leftarrow 0, \mathbf{a} \leftarrow 0, \mathbf{b} \leftarrow 0$;

for $t = 1 \cdots T$ **do**

  for $n = 1 \cdots N$ **do**

    选取一个样本 $\hat{\mathbf{v}}^{(n)}$，用公式 (12.48) 计算 $p(\mathbf{h} = \mathbf{1}|\hat{\mathbf{v}}^{(n)})$，并根据这个分布采集一个隐向量 $\mathbf{h}$;

    计算**正向梯度** $\hat{\mathbf{v}}^{(n)}\mathbf{h}^{\mathrm{T}}$;

    根据 $\mathbf{h}$，用公式 (12.49) 计算 $p(\mathbf{v} = \mathbf{1}|\mathbf{h})$，并根据这个分布采集重构的可见变量 $\mathbf{v}'$;

    根据 $\mathbf{v}'$，重新计算 $p(\mathbf{h} = \mathbf{1}|\mathbf{v}')$ 并采样一个 $\mathbf{h}'$;

    计算**反向梯度** $\mathbf{v}'\mathbf{h}'^{\mathrm{T}}$;

    更新参数:

    $W \leftarrow W + \alpha(\hat{\mathbf{v}}^{(n)}\mathbf{h}^{\mathrm{T}} - \mathbf{v}'\mathbf{h}'^{\mathrm{T}})$;

    $\mathbf{a} \leftarrow \mathbf{a} + \alpha(\hat{\mathbf{v}}^{(n)} - \mathbf{v}')$;

    $\mathbf{b} \leftarrow \mathbf{b} + \alpha(\mathbf{h} - \mathbf{h}')$;

  end

end

输出: $W, \mathbf{a}, \mathbf{b}$

受限玻尔兹曼机的类型

- **“伯努利-伯努利”受限玻尔兹曼机**（BernoulliBernoulli RBM, BB-RBM）就是上面介绍的可观测变量和隐变量都为二值类型的受限玻尔兹曼机。
- **“高斯-伯努利”受限玻尔兹曼机**（GaussianBernoulli RBM, GB-RBM）是假设可观测变量为高斯分布，隐变量为伯努利分布，其能量函数定义为

$$\mathrm{E}(\mathrm{v},\mathrm{h}) = \sum_i \frac{(v_i - u_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_i \sum_j \frac{v_i}{\sigma_i} w_{ij} h_j$$

- **“伯努利-高斯”受限玻尔兹曼机**（BernoulliGaussian RBM, BG-RBM）是假设可观测变量为伯

努利分布，隐变量为高斯分布，其能量函数定义为

$$\mathrm{E}(\mathrm{v},\mathrm{h}) = \sum_i a_i v_i - \sum_j \frac{(h_j - u_j)^2}{2\sigma_j^2} - \sum_j b_j h_j - \sum_i \sum_j v_i w_{ij} \frac{h_i}{\sigma_i}$$

深度信念网络

一种深层的概率有向图模型，其图结构由多层的节点构成。每层节点的内部没有连接，相邻两层的节点之间为全连接。网络的最底层为可观测变量，其它层节点都为隐变量。最顶部的两层间的连接是无向的，其他层之间的连接是有向的。

对一个有 L 层隐变量的深度信念网络，令$v = h^{(0)}$表示最底层（第 0 层）为可观测变量，$h^{(1)}, \cdots, h^{(L)}$表示其余每层的变量。顶部的两层是一个无向图，可以看做是一个受限玻尔兹曼机，用来产生$p(h^{(L-1)})$的先验分布。除了最顶上两层外，每一层变量$h^{(l)}$依赖于其上面一层$h^{(l+1)}$，

$$p\left(h^{(l)} \middle| h^{(l+1)}, \cdots, h^{(L)}\right) = p(h^{(l)} | h^{(l+1)})$$

深度信念网络中所有变量的联合概率可以分解为

$$p\left(v, h^{(1)}, \cdots, h^{(L)}\right) = p\left(v \middle| h^{(1)}\right)\left(\prod_{l=2}^{L-2} p\left(h^{(l)} \middle| h^{(l+1)}\right)\right) p\left(h^{(L-1)}, h^{(L)}\right)$$

$$= \left(\prod_{l=0}^{L-1} p\left(h^{(l)} \middle| h^{(l+1)}\right)\right) p\left(h^{(L-1)}, h^{(L)}\right)$$

其中$p\left(h^{(l)} \middle| h^{(l+1)}\right)$使用 Sigmoid 型条件概率分布：

$$p\left(h^{(l)} \middle| h^{(l+1)}\right) = \sigma(a^{(l)} + W^{(l+1)} h^{(l+1)})$$

生成模型

深度信念网络是一个生成模型，可以用来生成符合特定分布的样本。隐变量用来描述在可观测变量之间的高阶相关性。

在生成样本时，首先在最顶两层进行足够多次的吉布斯采样，生成$h^{(L-1)}$，然后依次计算下一层隐变量的分布。因为在给定上一层变量取值时，下一层的变量是条件独立的，因为可以独立采样。这样，我们可以从第 L-1 层开始，自顶向下进行逐层采样，最终得到可观测层的样本。

参数学习

最直接的训练方式可以通过最大似然方法使得可观测变量的边际分布 p(v)在训练集合上的似然达到最大。但在深度信念网络中，隐变量 h 之间的关系十分复杂，由于“贡献度分配问题”，很难直接学习。

为了有效地训练深度信念网络，我们将每一层的 Sigmoid 信念网络转换为受限玻尔兹曼机。

这样做的好处是隐变量的后验概率是相互独立的，从而可以很容易地进行采样。这样，深度信念网络可以看作是由多个受限玻尔兹曼机从下到上进行堆叠，第 l 层受限玻尔兹曼机的隐层作为第 l+1 层受限玻尔兹曼机的可观测层。进一步地，深度信念网络可以采用逐层训练的方式来快速训练，即从最底层开始，每次只训练一层，直到最后一层。

先通过逐层预训练将模型的参数初始化为较优的值，再通过传统学习方法对参数进行精调。

逐层预训练

将深度信念网络的训练简化为对多个受限玻尔兹曼机的训练。

自下而上依次训练每一层的受限玻尔兹曼机。假设我们已经训练好了前 l-1 层的受限玻尔兹曼机，那么可以计算隐变量自下而上的条件概率

$$p\left(h^{(i)}\middle|h^{(i-1)}\right)=\sigma(b^{(i)}+W^{(i)}h^{(i-1)})$$

算法 12.2: 深度信念网络的逐层训练方法

输入: 训练集: $\hat{\mathbf{v}}^{(n)}, n=1,\cdots,N$;
学习率: α, 深度信念网络层数: L, 第 l 层权重: $W^{(l)}$, 第 l 层偏置 $\mathbf{a}^{(l)}$, 第 l 层偏置 $\mathbf{b}^{(l)}$;

```
for l = 1 ··· L do
    初始化: W^(l) ← 0, a^(l) ← 0, b^(l) ← 0;
    从训练集中采样 ĥ^(0);
    for i = 1 ··· l − 1 do
        根据分布 p(h^(i) | ĥ^(i−1)) 采样 ĥ^(i);
    end
    将 ĥ^(l−1) 作为训练样本，充分训练第 l 层受限玻尔兹曼机 W^(l), a^(l), b^(l);
end
```

输出: $\{W^{(l)},\mathbf{a}^{(l)},\mathbf{b}^{(l)}\}, 1\le l\le L$

逐层预训练可以产生非常好的参数初始值，从而极大地降低了模型的学习难度。

精调

经过预训练之后，再结合具体的任务（监督学习或无监督学习），通过传统的全局学习算法对网络进行精调（fine-tuning），使模型收敛到更好的局部最优点。

- **作为生成模型的精调** 除了顶层的受限玻尔兹曼机，其它层之间的权重被分成向上的认知权重（recognition weights）W′和向下的生成权重（generative weights）W。认知权重用来进行后验概率计算，而生成权重用来进行定义模型。认知权重的初始值$W'^{(l)}=W^{(l)^T}$。

深度信念网络一般采用 contrastive wake-sleep 算法进行精调，其算法过程是：

1）**Wake 阶段：**认知过程，通过外界输入（可观测变量）和向上认知权重，计算每一层隐变量的后验概率并采样。然后，修改下行的生成权重使得下一层的变量的后验概率最大。也就是“如果现实跟我想象的不一样，改变我的权重使得我想象的东西就是这样的”；

2）**Sleep 阶段：**生成过程，通过顶层的采样和向下的生成权重，逐层计算每一层的后验概率并采样。然后，修改向上的认知权重使得上一层变量的后验概率最大。也就是“如果梦中的景象不是我脑中的相应概念，改变我的认知权重使得这种景象在我看来就是这个概念”；

3）交替进行 Wake 和 Sleep 过程，直到收敛。

- **作为深度神经网络的精调**　深度信念网络的一个应用是作为深度神经网络的预训练部分，提供神经网络的初始权重。

在深度信念网络的最顶层再增加一层输出层，然后再使用反向传播算法对这些权重进行调优。特别是在训练数据比较少时，预训练的作用非常大。因为不恰当的初始化权重会显著影响最终模型的性能，而预训练获得的权重在权值空间中比随机权重更接近最优的权重，避免了反向传播算法因随机初始化权值参数而容易陷入局部最优和训练时间长的缺点。这不仅提升了模型的性能，也加快了调优阶段的收敛速度