Machine Learning II: Fractal 3

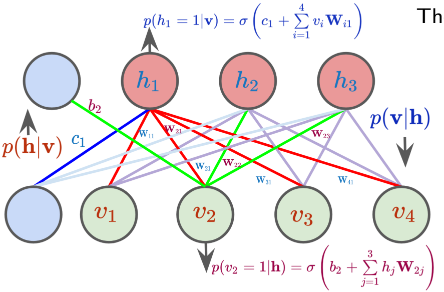
Rajendra Nagar

Assistant Professor
Department of Electircal Engineering
Indian Institute of Technology Jodhpur
http://home.iitj.ac.in/∼rn/

November 14, 2021

Restricted Boltzmann Machine

- Given a set of points $\{\mathbf{v}_i\}_{i=1}^{N}$ model $p(\mathbf{v})$ that is as close as possible to true distribution.
- An undirected graphical model with no intra-layer interactions. The joint distribution between the observed units and the hidden units is defined as: $p(\mathbf{v},\mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z}$.



The energy function is defined as

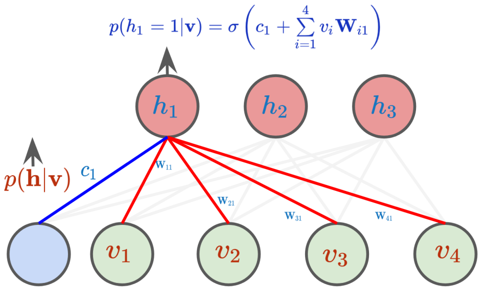
$$
\begin{aligned}
E(\mathbf{v},\mathbf{h}) &= -\mathbf{v}^\top \mathbf{W}\mathbf{h} - \mathbf{b}^\top \mathbf{v} - \mathbf{c}^\top \mathbf{h} \\
p(\mathbf{h}|\mathbf{v}) &= \prod_j p(h_j|\mathbf{v}) \\
p(h_j = 1|\mathbf{v}) &= \sigma\left(c_j + \mathbf{v}^\top \mathbf{w}_j\right) \\
p(v_i = 1|\mathbf{h}) &= \sigma\left(b_i + \mathbf{h}^\top \bar{\mathbf{w}}_i\right)
\end{aligned}
$$

$$
\begin{aligned}
\ell(\boldsymbol{\theta}) &= \log p_{\boldsymbol{\theta}}(\mathbf{v}_i) \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mathbf{w}_{ij}} &= \mathbb{E}_{\mathbf{h}\sim p(\mathbf{h}|\mathbf{v}_i)}[v_i h_j] - \mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p(\mathbf{v},\mathbf{h})}[v_i h_j] \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial b_i} &= \mathbb{E}_{\mathbf{h}\sim p(\mathbf{h}|\mathbf{v}_i)}[v_i] - \mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p(\mathbf{v},\mathbf{h})}[v_i] \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial c_j} &= \mathbb{E}_{\mathbf{h}\sim p(\mathbf{h}|\mathbf{v}_i)}[h_j] - \mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p(\mathbf{v},\mathbf{h})}[h_j]
\end{aligned}
$$

Restricted Boltzmann Machine

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mathbf{w}_{ij}} = \mathbb{E}_{\mathbf{h}\sim p(\mathbf{h}|\mathbf{v})}[v_i h_j] - \mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p(\mathbf{v},\mathbf{h})}[v_i h_j]$$

The first term $\mathbb{E}_{\mathbf{h}\sim p(\mathbf{h}|\mathbf{v})}[v_i h_j]$ is easy to compute.



$$\begin{aligned}
\mathbb{E}_{\mathbf{h}\sim p(\mathbf{h}|\mathbf{v}_i)}[v_i h_j] &= \sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}) v_i h_j \\
&= \sum_{h_j \in \{0,1\}} p(h_j|\mathbf{v}) v_i h_j \\
&= v_i \sigma(\mathbf{v}^\top \mathbf{w}_j + c_j).
\end{aligned}$$

Restricted Boltzmann Machine

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mathbf{w}_{ij}} = \mathbb{E}_{\mathbf{h}\sim p(\mathbf{h}|\mathbf{v})}[v_i h_j] - \mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p(\mathbf{v},\mathbf{h})}[v_i h_j]$$

- Computing the second term $\mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p(\mathbf{v},\mathbf{h})}[v_i h_j]$ is not straightforward.
- We do not have access to the joint distribution $p_{\boldsymbol{\theta}}(\mathbf{v},\mathbf{h})$.
- Therefore, we can not find samples $(\mathbf{v},\mathbf{h})$ from this distribution to find $\mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p(\mathbf{v},\mathbf{h})}[v_i h_j]$.

Gibbs Sampling:

- Consider two random variables X and Y and their joint distribution $p_{X,Y}(x,y)$.
- How do we sample new points if $p_{X,Y}$ is intractable?
- Assume that sampling from their conditional distributions is easy.

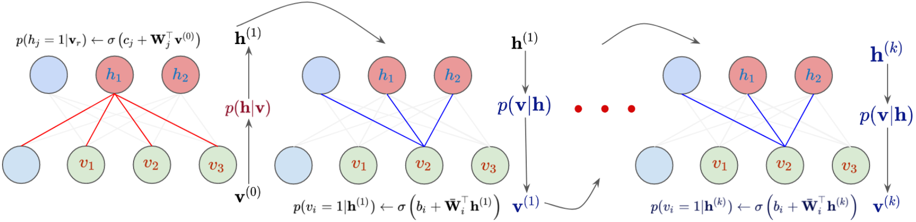
Initialize: (X,Y) as $(x^{(0)}, y^{(0)})$

1: **for** $t \in \{1,2,\ldots,k\}$ **do**
2: Sample $x^{(1)} \sim p_{X|Y=y^{(0)}}$
3: Sample $y^{(1)} \sim p_{Y|X=x^{(1)}}$
4: **end for**

Burn-in for k iterations.
$p_{X,Y}(x^{(k)}, y^{(k)})$ will be close to the true value.

$$
\begin{aligned}
\mathbb{E}_{(\mathbf{v},\mathbf{h})\sim p(\mathbf{v},\mathbf{h})}[v_i h_j] &= \sum_{\mathbf{v}}\sum_{\mathbf{h}} p(\mathbf{v},\mathbf{h}) v_i h_j \\
&= \sum_{\mathbf{v}}\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}) p(\mathbf{v}) v_i h_j \\
&= \sum_{\mathbf{v}} p(\mathbf{v}) \sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}) v_i h_j \\
&= \sum_{\mathbf{v}} p(\mathbf{v}) \sum_{h_j \in \{0,1\}} p(h_j|\mathbf{v}) v_i h_j \\
&= \sum_{h_j \in \{0,1\}} p(h_j|\mathbf{v}^{(k)}) v_i^{(k)} h_j. \text{ (Contrastive Divergence)}
\end{aligned}
$$

- CDs $p(\mathbf{v}|\mathbf{h})$ and $p(\mathbf{h}|\mathbf{v})$ are well defined as $p(h_j = 1|\mathbf{v}) = \sigma\left(c_j + \mathbf{v}^\top \mathbf{w}_j\right)$ and $p(v_i = 1|\mathbf{h}) = \sigma\left(b_i + \mathbf{h}^\top \bar{\mathbf{w}}_i\right)$.
- Use Gibbs sampling.



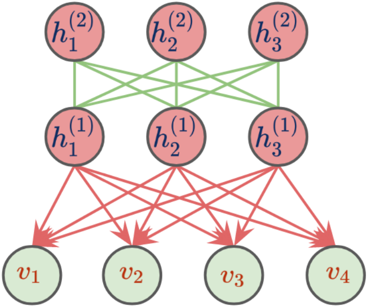
Contrastive Divergence

Input: An RBM and a dataset $\mathcal{S} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$. **Output**: Gradient Values.

Initialize: $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mathbf{w}_{ij}} \leftarrow 0$, $\frac{\partial \ell(\boldsymbol{\theta})}{\partial b_i} \leftarrow 0$, $\frac{\partial \ell(\boldsymbol{\theta})}{\partial c_j} \leftarrow 0$, $\forall i \in \{0, \ldots, m-1\}$, $\forall j \in \{0, \ldots, n-1\}$.

for $\mathbf{v} \in \mathcal{S}$ **do**
    $\mathbf{v}^{(0)} \leftarrow \mathbf{v}$
    for $t \in \{0, 1, \ldots, k-1\}$ **do**
        for $j \in \{0, 1, \ldots, n-1\}$ **do**
            Sample $h_j^{(t)} \sim p(h_j = 1 | \mathbf{v}^{(t)})$
        end for
        for $i \in \{0, 1, \ldots, m-1\}$ **do**
            Sample $v_i^{(t+1)} \sim p(v_i = 1 | \mathbf{h}^{(t)})$
        end for
    end for
    for $i \in \{0, 1, \ldots, m-1\}$ and $j \in \{0, 1, \ldots, n-1\}$ **do**
        $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mathbf{w}_{ij}} \leftarrow \frac{\partial \ell(\boldsymbol{\theta})}{\partial \mathbf{w}_{ij}} + \left(\mathbf{v}_i^{(0)} p(h_j = 1 | \mathbf{v}^{(0)}) - \mathbf{v}_i^{(k)} p(h_j = 1 | \mathbf{v}^{(k)})\right)$
    end for
    for $i \in \{0, 1, \ldots, m-1\}$ **do**
        $\frac{\partial \ell(\boldsymbol{\theta})}{\partial b_i} \leftarrow \frac{\partial \ell(\boldsymbol{\theta})}{\partial b_i} + \left(\mathbf{v}_i^{(0)} - \mathbf{v}_i^{(k)}\right)$
    end for
    for $j \in \{0, 1, \ldots, n-1\}$ **do**
        $\frac{\partial \ell(\boldsymbol{\theta})}{\partial c_j} \leftarrow \frac{\partial \ell(\boldsymbol{\theta})}{\partial c_j} + \left(p(h_j = 1 | \mathbf{v}^{(0)}) - p(h_j = 1 | \mathbf{v}^{(k)})\right)$
    end for
end for

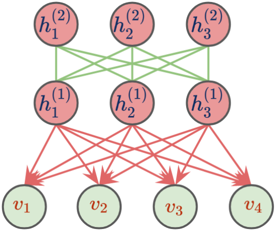
Deep Belief Networks



DBN

- Generative models with several layers of latent variables.
- The latent variables are typically binary.
- While the visible units may be binary or real.
- There are no intra-layer connections.
- The connections between the top two layers are undirected.
- The connections between all other layers are directed, with the arrows pointed toward the layer that is closest to the data.

Deep Belief Networks

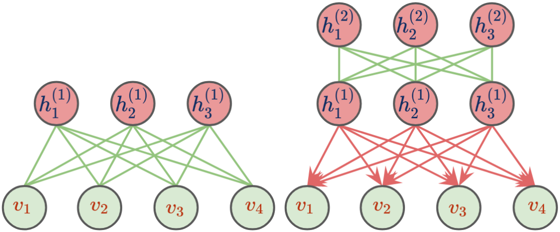


DBN

A DBN with ℓ hidden layers contains ℓ weight matrices: $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(\ell)}$. It also contains $\ell + 1$ bias vectors $\mathbf{b}^{(0)}, \ldots, \mathbf{b}^{(\ell)}$, with $\mathbf{b}^{(0)}$ providing the biases for the visible layer. The probability distribution represented by the DBN is given by:

$$\begin{aligned}
p(\mathbf{h}^{(\ell)}, \mathbf{h}^{(\ell-1)}) &= \frac{1}{Z} e^{(\mathbf{b}^{(\ell)})^\top \mathbf{h}^{(\ell)} + (\mathbf{b}^{(\ell-1)})^\top \mathbf{h}^{(\ell-1)} + (\mathbf{h}^{(\ell-1)})^\top \mathbf{W}^{(\ell)} \mathbf{h}^{(\ell)}}. \\
p(h_i^{(k)} = 1 | \mathbf{h}^{(k+1)}) &= \sigma\left(b_i^{(k)} + (\mathbf{h}^{(k+1)})^\top \mathbf{W}_i^{(k+1)}\right) \ \forall i, \forall k \in \{1, 2, \ldots, \ell - 2\} \\
p(v_i = 1 | \mathbf{h}^{(1)}) &= \sigma\left(b_i^{(0)} + (\mathbf{h}^{1})^\top \mathbf{W}_i^{(1)}\right) \ \forall i.
\end{aligned}$$

Deep Belief Networks



- To train a DBN, we first train an RBM to maximize $\mathbf{E}_{\mathbf{v}\sim p_{\text{data}}}[\log p(v)]$ using the contrastive-divergence algorithm.
- The parameters of the RBM then define the parameters of the first layer of the DBN.
- Next, a second RBM is trained with inputs as the output of the first layer.
- We maximize, $\mathbf{E}_{\mathbf{v}\sim p_{\text{data}}}\left[\mathbf{E}_{\mathbf{h}^{(1)}\sim p^{(1)}(\mathbf{h}^{(1)}|\mathbf{v})}\left[\log p^{(2)}(\mathbf{h}^{(1)})\right]\right]$.
- Here, $p^{(1)}$ is the probability distribution represented by the first RBM.
- $p^{(2)}$ is the probability distribution represented by the second RBM.
- The second RBM is trained to model the distribution defined by sampling the hidden units of the first RBM.
- This procedure can be repeated indefinitely, to add as many layers to the DBN as desired, with each new RBM modeling the samples of the previous one.