Neural Network

1 NEURAL NETWORK

1.1 Back-Propagation

For the training data (x_k, y_k), the output of the neural network is $\mathbf{y}^k = (\hat{y}_1^k, \hat{y}_1^k, \cdots, \hat{y}_l^k)$, where,

$$\hat{y}_l^k = f(\beta_j - \theta_j) \tag{1}$$

The loss of neural network is,

$$E_k = \frac{1}{2}\sum_{i=1}^{l}(\hat{y}_i^k - y_i^k)^2 \tag{2}$$

Using the back-propagation to update the parameters of the network as below.

$$w_{i+1} = w_i + \Delta w \tag{3}$$

Given the error E_k and the learning rate η, Δw is,

$$\Delta w_{hj} = -\eta\frac{\partial E_k}{\partial w_{hj}} \tag{4}$$

Using the chain rule,

$$\begin{aligned}
\frac{\partial E_k}{\partial w_{hj}} &= \frac{\partial E_k}{\partial \hat{y}_j^k}\cdot\frac{\partial \hat{y}_j^k}{\partial \beta_j}\cdot\frac{\partial \beta_j}{\partial w_{hj}} \\
&= \frac{\partial E_k}{\partial \hat{y}_j^k}\cdot\frac{\partial \hat{y}_j^k}{\partial \beta_j}\cdot b_h \\
&= \frac{\partial E_k}{\partial \hat{y}_j^k}\cdot \hat{y}_j^k(1-\hat{y}_j^k)\cdot b_h \\
&= (\hat{y}_i^k - y_i^k)\cdot \hat{y}_j^k(1-\hat{y}_j^k)\cdot b_h \\
\therefore \Delta w_{hj} &= -\eta(\hat{y}_i^k - y_i^k)\cdot \hat{y}_j^k(1-\hat{y}_j^k)\cdot b_h
\end{aligned} \tag{5}$$

Let $g_j = -(\hat{y}_i^k - y_i^k)\cdot\hat{y}_j^k(1-\hat{y}_j^k)$, Therefore $\Delta w_{hj} = \eta g_j b_h$.

$\Delta\theta_j$ is,

$$\begin{aligned}
\Delta\theta_j &= -\eta\frac{\partial E_k}{\partial \hat{y}_j^k}\cdot\frac{\partial \hat{y}_j^k}{\partial \theta_j} \\
&= -\eta\frac{\partial E_k}{\partial \hat{y}_j^k}\cdot \hat{y}_j^k(\hat{y}_j^k - 1) \\
&= -\eta(\hat{y}_i^k - y_i^k)\cdot\hat{y}_j^k(\hat{y}_j^k - 1) \\
&= -\eta g_j
\end{aligned} \tag{6}$$

1.2 Restricted Boltzmann Machine

Restricted Boltzmann Machine is a kind of Gibbs distribution. Using the Harmmersley Clifford Theorem, C_i is the maximal clique, $\psi(x_{ci})$ is the potential function. Therefore,

$$\begin{aligned}
&p(x) = \frac{1}{Z}\prod_{i=0}^{k}\psi(x_{ci}) \\
&Z = \sum_{x_{c1}}\sum_{x_{c2}}\cdots\sum_{x_{cp}}\prod_{i=0}^{k}\psi(x_{ci}), \quad \text{partition function} \\
&\because \quad \psi(x_{ci}) = \exp(-E(x_{ci})) \\
&\therefore \quad p(x) = \frac{1}{Z}\exp(-\sum_{i=0}^{k}E(x_{ci}))
\end{aligned} \tag{7}$$

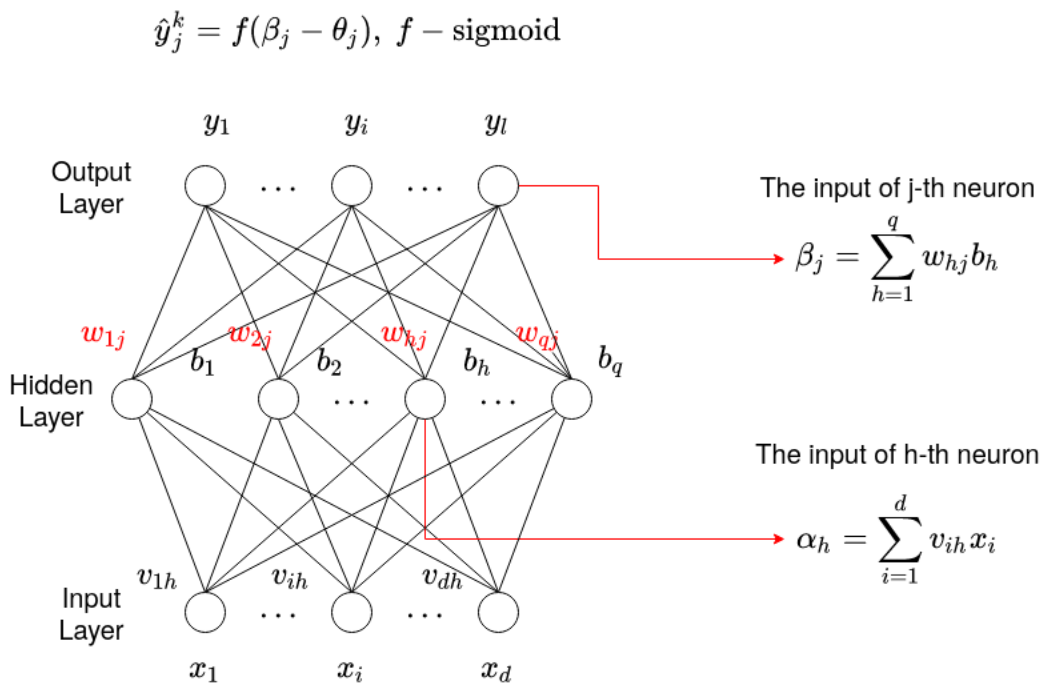

Fig. 1. Neural Network

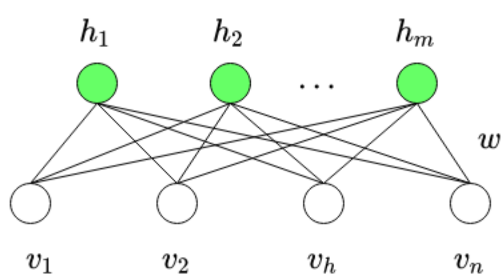

Fig. 2. Restricted Boltzmann Machine

Given the observed variable is v, and the latent variable is h. So the input variable is $x \in \mathbf{R}^p$.

$$x = \begin{cases} h, \\ v \end{cases}, h = \begin{cases} h_1, \\ h_2, \\ \vdots, \\ h_m \end{cases}, v = \begin{cases} v_1, \\ v_2, \\ \vdots, \\ v_n \end{cases}, p = m + n, h \in \{0, 1\} \tag{8}$$

The Restricted Boltzmann Machine is that there are connections between h and v, and there are no connections in themselves, as shown in Fig. 2. Therefore,

$$\begin{aligned} E(x_{ci}) &= E(h, v) = -h^T w v - \alpha^T h - \beta^T v \\ E(h, v) &= -\sum_{i=0}^{m}\sum_{j=0}^{n} h_i^T w_{ij} v_j - \sum_{i=0}^{m} \alpha_i^T h_i - \sum_{j=0}^{n} \beta_j^T v_j \end{aligned} \tag{9}$$

Therefore,

$$\begin{aligned} p(x) &= \frac{1}{Z}\exp(E(x_{ci})) \\ \Rightarrow p(h, v) &= \frac{1}{Z}\exp(E(h, v)) \end{aligned} \tag{10}$$

1.3 RBM Inference

1.3.1 Posterior Probability

The inference of RBM model is to compute the posterior, which are the probabilities of $p(h|v)$ and $p(v|h)$, and the marginal probability $p(v)$. RBM is an undirected graph model and it meets the local Markov. Therefore,

$$p(h_i|v) = p(h_i|h_{-i}, v) \tag{11}$$

where h_{-i} is the set of h without h_i.

$$\begin{aligned}
&p(h|v) = \prod_{i=0}^{m} p(h_i|v) \\
\therefore\ \ &p(h_l = 1|v) = p(h_l = 1|h_{-l}, v) = \frac{p(h_l = 1, h_{-l}, v)}{p(h_{-l}, v)} \\
&= \frac{p(h_l = 1, h_{-l}, v)}{p(h_l = 1, h_{-l}, v) + p(h_l = 0, h_{-l}, v)} \\
\therefore\ \ &p(h_l = 1, h_{-l}, v) = -(\underbrace{\sum_{i=0, i\neq l}^{m}\sum_{j=0}^{n} h_i^T w_{ij} v_j}_{\Delta_1} + \underbrace{h_l \sum_{j=0}^{n} w_{lj} v_j}_{\Delta_2} + \underbrace{\sum_{i=0}^{m} \alpha_i^T v_i}_{\Delta_3} + \underbrace{\sum_{i=0, i\neq l}^{m} \beta_i^T h_i}_{\Delta_4} + \underbrace{\beta_l h_l}_{\Delta_5})
\end{aligned} \tag{12}$$

Let $\Delta_2 + \Delta_5 = h_l(\sum_{j=0}^{n} w_{lj} v_j + \beta_l) = h_l \cdot H_l(v)$ and $\hat{H}_l(h_{-l}, v) = \Delta_1 + \Delta_3 + \Delta_4$.

$$\begin{aligned}
\therefore\ \ &E(h, v) = h_l \cdot H_l(v) + \hat{H}_l(h_{-l}, v) \\
\therefore\ \ &p(h_l = 1, h_{-l}, v) = \frac{1}{Z}\exp(H_l(v) + \hat{H}_l(h_{-l}, v)) \\
&p(h_l = 1, h_{-l}, v) + p(h_l = 0, h_{-l}, v) = \frac{1}{Z}\exp(H_l(v) + \hat{H}_l(h_{-l}, v)) + \frac{1}{Z}\exp(\hat{H}_l(h_{-l}, v))
\end{aligned} \tag{13}$$

Therefore,

$$p(h_l = 1|v) = \frac{1}{1 + \exp(-H_l(v))} = \sigma(H_l(v)) = \sigma(\sum_{j=0}^{n} w_{lj} + \beta_l) \tag{14}$$

Similarly, the posterior $p(v|h)$ can be obtained using the same method.

1.3.2 Marginal Probability

$$\begin{aligned}
p(v) &= \sum_h p(h, v) = \sum_h \frac{1}{Z}\exp(E(h, v)) \\
&= \sum_h \frac{1}{Z}\exp(h^T w v + \alpha^T v + \beta^T h) \\
&= \exp(\alpha^T v) \cdot \sum_{h_1} \cdots \sum_{h_m} \frac{1}{Z}\exp(h^T w v + \beta^T h) \\
&= \exp(\alpha^T v) \cdot\cdot \sum_{h_1} \cdots \sum_{h_m} \frac{1}{Z}\exp(\sum_{i=1}^{m}(h_i^T w_i v + \beta_i h_i)) \\
&= \exp(\alpha^T v) \cdot\cdot \sum_{h_1} \cdots \sum_{h_m} \frac{1}{Z}\exp(h_1^T w_1 v + \beta_1 h_1) \cdots \exp(h_m^T w_m v + \beta_m h_m) \\
&= \exp(\alpha^T v) \cdot \frac{1}{Z} \sum_{h_1}\exp(h_1^T w_1 v + \beta_1 h_1) \cdots \sum_{h_m}\exp(h_m^T w_m v + \beta_m h_m) \\
\because\ \ & h_i \in \{0, 1\} \\
\therefore\ \ &= \exp(\alpha^T v) \cdot \frac{1}{Z}(1 + \exp(w_1 v + \beta_1)) \cdots (1 + \exp(w_m v + \beta_m)) \\
&= \exp(\alpha^T v) \cdot \frac{1}{Z}\log(\exp(1 + \exp(w_1 v + \beta_1))) \cdots \log(\exp(1 + \exp(w_m v + \beta_m))) \\
&= \frac{1}{Z}\exp(\alpha^T v + \sum_{i=0}^{m} \underbrace{\log(1 + \exp(w_i v + \beta_i))}_{\text{softplus}}) \\
&= \frac{1}{Z}\exp(\alpha^T v + \sum_{i=0}^{m} \text{softplus}(w_i v + \beta_i))
\end{aligned} \tag{15}$$

1.3.3 Learning

Given the observed training set $V = \{v_1, v_2, \cdots, v_n\}$ and the latent variable $H = \{h_1, h_2, \cdots, h_m\}$, the parameters is $\theta = \{W, \alpha, \beta\}$. The learning strategy is to get the maximal likelihood function,

$$\begin{aligned} L(\theta) &= \ln\left(\prod_{k=1}^{n} P(v_k)\right) \\ &= \sum_{k=1}^{n} \ln P(v_k) \\ &= \sum_{k=1}^{n} L_k(\theta) \end{aligned} \tag{16}$$

Therefore,

$$\begin{aligned} \frac{\partial L_k(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta}\left[\ln \sum_{h} \exp(-E(h, v_k))\right] - \frac{\partial}{\partial \theta}\left[\ln \sum_{h,v} \exp(-E(h, v))\right] \\ &= -\sum_{h} \frac{\exp(-E(v_k, h)\frac{\partial E(v_k,h)}{\partial \theta})}{\sum\limits_{h}\exp(-E(v_k, h))} + \sum_{v,h} \frac{\exp(-E(v, h)\frac{\partial E(v,h)}{\partial \theta})}{\sum\limits_{v,h}\exp(-E(v, h))} \\ \because \quad \frac{\exp(-E(v_k, h))}{\sum\limits_{h}\exp(-E(v_k, h))} &= \frac{\frac{\exp(-E(v_k,h))}{Z}}{\sum\limits_{h}\frac{\exp(-E(v_k,h))}{Z}} = \frac{P(v_k, h)}{\sum\limits_{h} P(v_k, h)} = p(h|v_k) \\ \because \quad \frac{\exp(-E(v, h))}{\sum\limits_{v,h}\exp(-E(v, h))} &= \frac{\frac{\exp(-E(v,h))}{Z}}{\sum\limits_{v,h}\frac{\exp(-E(v,h))}{Z}} = \frac{P(v, h)}{\sum\limits_{h} P(v, h)} = p(v, h) \\ \therefore \quad \frac{\partial L_k(\theta)}{\partial \theta} &= -\sum_{h} p(h|v_k)\frac{\partial E(v_k, h)}{\partial \theta} + \sum_{v,h} p(v, h)\frac{\partial E(v, h)}{\partial \theta} \\ &= -\sum_{h} p(h|v_k)\frac{\partial E(v_k, h)}{\partial \theta} + \sum_{v} P(v)\sum_{h} p(h|v)\frac{\partial E(v, h)}{\partial \theta} \end{aligned} \tag{17}$$

Therefore,

$$\begin{aligned} \frac{\partial L_k(\theta)}{\partial w_{ij}} &= -\sum_{h} p(h|v_k)\frac{\partial E(v_k, h)}{\partial w_{ij}} + \sum_{v} p(v)\sum_{h} p(h|v)\frac{\partial E(v, h)}{\partial w_{ij}} \\ \therefore \quad \frac{\partial L_k(\theta)}{\partial w_{ij}} &= -\mathbb{E}_{p(h|v)}\left[\frac{\partial E(v, h)}{\partial \theta}\right] + \mathbb{E}_{p(v,h)}\left[\frac{\partial E(v, h)}{\partial \theta}\right] \end{aligned} \tag{18}$$

$$
\begin{aligned}
\sum_h p(h|v)\frac{\partial E(v,h)}{\partial w_{ij}} &= -\sum_h p(h|v)h_i v_j \\
&= -\sum_h \prod_{l=1}^{q} p(h_l|v)h_i v_j \\
&= -\sum_h p(h_i|v) \prod_{l=1,l\neq i}^{q} p(h_l|v)h_i v_j \\
&= -\sum_h p(h_i|v)p(h_1,h_2,\cdots,h_{i-1},h_{i+1},\cdots,h_q)h_i v_j \\
&= -\sum_{h_i} p(h_i|v)h_i v_j \underbrace{\sum_{h_1,\cdots,h_{i-1},h_{i+1},\cdots,h_q} p(h_1,h_2,\cdots,h_{i-1},h_{i+1},\cdots,h_q)}_{=1} \\
&= -\sum_{h_i} p(h_i|v)h_i v_j \\
\because \quad h_1 &\in \{0,1\} \\
\therefore \quad &= -p(h_i=1|v)v_j \\
\therefore \quad \sum_h p(h|v)\frac{\partial E(v,h)}{\partial w_{ij}} &= -p(h_i=1|v)v_j \\
\therefore \quad \sum_h p(h|v_k)\frac{\partial E(v_k,h)}{\partial w_{ij}} &= -p(h_i=1|v_k)v_j \\
\therefore \quad \frac{\partial L_k(\theta)}{\partial w_{ij}} &= p(h_i=1|v_k)v_j - \sum_v p(v)p(h_i=1|v)v_j
\end{aligned}
\tag{19}
$$

Use the similar method to compute the $\sum_h p(h|v)\frac{\partial E(v,h)}{\partial \beta_j}$,

$$
\begin{aligned}
\sum_h p(h|v)\frac{\partial E(v,h)}{\partial \beta_j} &= -\sum_h p(h|v)h_i \\
&= -\sum_h p(h_i|v) \prod_{l=1,l\neq i}^{q} p(h_l|v)h_i \\
&= -\sum_h p(h_i|v)p(h_1,h_2,\cdots,h_{i-1},h_{i+1},\cdots,h_q)h_i \\
&= -\sum_{h_i} p(h_i|v)h_i \underbrace{\sum_{h_1,\cdots,h_{i-1},h_{i+1},\cdots,h_q} p(h_1,h_2,\cdots,h_{i-1},h_{i+1},\cdots,h_q)}_{=1} \\
&= -\sum_{h_i} p(h_i|v)h_i \\
\because \quad h_1 &\in \{0,1\} \\
\therefore \quad &= -p(h_i=1|v)
\end{aligned}
\tag{20}
$$

Final, compute $\sum_h p(h|v)\frac{\partial E(v,h)}{\partial \alpha_i}$,

$$
\begin{aligned}
&\sum_h p(h|v)\frac{\partial E(v,h)}{\partial \alpha_i} \\
&= -\sum_h p(h|v)v_i \\
&= -v_i \sum_h p(h|v) \\
&= -v_i \ \ (\because \sum_h p(h|v) = 1)
\end{aligned}
\tag{21}
$$

Conclusion,

$$\begin{aligned}\frac{\partial \ln p(v)}{\partial w_{ij}} &= -\sum_h p(h|v_k)\frac{\partial E(v_k,h)}{\partial w_{ij}} + \sum_v p(v)\sum_h p(h|v)\frac{\partial E(v,h)}{\partial w_{ij}} \\ &= p(h_i=1|v)v_j - \sum_v p(v)p(h_i=1|v)v_j \\ \frac{\partial \ln p(v)}{\partial \alpha_i} &= -\sum_h p(h|v_k)\frac{\partial E(v_k,h)}{\partial \alpha_i} + \sum_v p(v)\sum_h p(h|v)\frac{\partial E(v,h)}{\partial \alpha_i} \\ &= v_i - \sum_v p(v)v_i,\end{aligned} \tag{22}$$