ECE 544NA Fall 2016 Assignment 4

Jifu Zhao Date: 10/30/2016

I. Pencil-and-Paper

Suppose (V, H) denote the visible and hidden random variable which takes values ($v \in \{0,1\}^m$, $h \in \{0,1\}^n$). And the joint probability is $p(v, h; \theta) = \frac{1}{Z} e^{-E(v,h;\theta)}$, where E is the energy function:

$$\begin{aligned} E(v, h; \theta) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} h_i v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i \\ &= -(\mathbf{v^T W h} + \mathbf{v^T b} + \mathbf{h^T c}) \end{aligned} \tag{1}$$

where $Z = \sum_v \sum_h e^{-E(v,h,\theta)}$ and $\theta = (W, b, c)$.

1. Find $p(v|h, \theta)$ and $\mathbb{E}(v|h, \theta)$.

From the structure of RBM, it means that the hidden variables are independent given the state of the visible variables and the visible variables are independent given the state of the hidden variables. In this way,

$$\begin{aligned} p(v|h, \theta) &= \frac{p(v, h|\theta)}{p(h|\theta)} \\ &= \frac{\prod_i^n \prod_j^m p(v_j, h_i|\theta)}{\sum_v \prod_i^n \prod_j^m p(v_j, h_i|\theta)} \end{aligned} \tag{2}$$

So, we also have:

$$\mathbb{E}(v|h, \theta) = sigmoid(W^T h + b) \tag{3}$$

(The prof will be shown later)

For $p(v_j|h)$,

$$p(v_j|h) = \frac{p(v_j, h)}{p(h)} = \frac{p(v_j, h)}{p(v_j = 0, h) + p(v_j = 1, h)} \tag{4}$$

Since that $p(v_j = 0, h) = \frac{1}{Z} e^{-E(v_j=0,h;\theta)}$ and $p(v_j = 1, h) = \frac{1}{Z} e^{-E(v_j=1,h;\theta)}$. From equation (1), we can think $E(v, h; \theta) = v_j \alpha_j + \beta$, where β doesn't contain v_j.

In this way, we have:

$$\begin{aligned} p(v_j|h) &= \frac{p(v_j, h)}{p(v_j = 0, h) + p(v_j = 1, h)} \\ &= \frac{exp(-v_j \alpha_j - \beta)}{exp(-\beta) + exp(-\alpha_j - \beta)} \\ &= \frac{exp(-v_j \alpha_j)}{1 + exp(-\alpha_j)} \end{aligned} \tag{5}$$

So, we have

$$p(v_j = 1|h) = \frac{exp(-\alpha_j)}{1 + exp(-\alpha_j)} = \frac{1}{1 + exp(\alpha_j)} \tag{6}$$

and

$$p(v_j = 0|h) = \frac{1}{1 + exp(-\alpha_j)} = 1 - \frac{1}{1 + exp(\alpha_j)} \tag{7}$$

From equation (1), we can see that $\alpha_j = -\sum_{i=1}^{n} w_{ij}h_i - b_j$, so we can see that:

$$p(v_j = 1|h) = sigmoid(-\alpha_j) = sigmoid(\textstyle\sum_{i=1}^{n} w_{ij}h_i + b_j) \tag{8}$$

and

$$p(v_j = 0|h) = 1 - sigmoid(-\alpha_j) = 1 - sigmoid(\textstyle\sum_{i=1}^{n} w_{ij}h_i + b_j) \tag{9}$$

From equation (8) and (9), we can see that:

$$\mathbb{E}(v_j|h, \theta) = 1 \cdot p(v_j = 1|h, \theta) + 0 \cdot p(v_j = 0|h, \theta) = sigmoid(\sum_{i=1}^{n} w_{ij}h_i + b_j) \tag{10}$$

In the vector format:

$$\mathbb{E}(v|h, \theta) = sigmoid(W^T h + b) \tag{11}$$

2. Find $p(h|v, \theta)$ and $\mathbb{E}(h|v, \theta)$.

Similar to the procedures in 1, since that the hidden variables are independent given the state of the visible variables. In this way,

$$\begin{aligned} p(h|v, \theta) &= \frac{p(v, h|\theta)}{p(v|\theta)} \\ &= \frac{\prod_i^n \prod_j^m p(v_j, h_i|\theta)}{\sum_h \prod_i^n \prod_j^m p(v_j, h_i|\theta)} \end{aligned} \tag{12}$$

So, we also have:

$$\mathbb{E}(h|v, \theta) = sigmoid(Wv + c) \tag{13}$$

For $p(h_i|v)$,

$$p(h_i|v) = \frac{p(h_i, v)}{p(v)} = \frac{p(h_i, v)}{p(h_i = 0, v) + p(h_i = 1, v)} \tag{14}$$

Since that $p(h_i = 0, v) = \frac{1}{Z} e^{-E(v, h_i = 0; \theta)}$ and $p(h_i = 1, v) = \frac{1}{Z} e^{-E(v, h_i = 1; \theta)}$. From equation (1), we can think $E(v, h; \theta) = h_i \gamma_i + \beta'$, where β' doesn't contain h_i. In this way, we have:

$$\begin{aligned} p(h_i|v) &= \frac{p(h_i, v)}{p(h_i = 0, v) + p(h_i = 1, v)} \\ &= \frac{exp(-h_i\gamma_i - \beta')}{exp(-\beta') + exp(-\gamma_i - \beta')} \\ &= \frac{exp(-h_i\gamma_i)}{1 + exp(-\gamma_i)} \end{aligned} \tag{15}$$

So, we have

$$p(h_i = 1|v) = \frac{exp(-\gamma_i)}{1 + exp(-\gamma_i)} = \frac{1}{1 + exp(\gamma_i)} \tag{16}$$

and

$$p(h_i = 0|v) = \frac{1}{1 + exp(-\gamma_i)} = 1 - \frac{1}{1 + exp(\gamma_i)} \tag{17}$$

From equation (1), we can see that $\gamma_i = -\sum_{j=1}^{m} w_{ij} v_j - c_i$, so we can see that:

$$p(h_i = 1|v) = sigmoid(-\gamma_i) = sigmoid(\textstyle\sum_{j=1}^{m} w_{ij} v_j + c_i) \tag{18}$$

and

$$p(h_i = 0|v) = 1 - sigmoid(-\gamma_i) = 1 - sigmoid(\textstyle\sum_{j=1}^{m} w_{ij} v_j + c_i) \tag{19}$$

From equation (18) and (19), we can see that:

$$\mathbb{E}(h_i|v, \theta) = 1 \cdot p(h_i = 1|v, \theta) + 0 \cdot p(h_i = 0|v, \theta) = sigmoid(\sum_{j=1}^{m} w_{ij} v_j + c_i) \tag{20}$$

In the vector format:

$$\mathbb{E}(h|v, \theta) = sigmoid(Wv + c) \tag{21}$$

3. Compute $\frac{\partial \mathcal{L}(D|\theta)}{\partial W_{ij}}$, $\frac{\partial \mathcal{L}(D|\theta)}{\partial b_j}$ and $\frac{\partial \mathcal{L}(D|\theta)}{\partial c_i}$

Now, suppose the given dataset is $D = \{v_1, v_2, \cdots, v_N\}$, then the log-likelihood can be calculated as:

$$\mathcal{L}(D|\theta) = \sum_{t=1}^{N} log(p(v_t|\theta)) \tag{22}$$

Since that

$$p(v_t|\theta) = \sum_{h_t} p(v_t, h_t|\theta) = \frac{1}{Z} \sum_{h_t} exp(-E(v_t, h_t; \theta)) \tag{23}$$

where $Z = \sum_{v_t} \sum_{h_t} exp(-E(v_t, h_t; \theta))$

we can get:

$$\begin{aligned} log(p(v_t|\theta)) &= log(\frac{1}{Z} \sum_{h_t} exp(-E(v_t, h_t; \theta))) \\ &= log(\sum_{h_t} exp(-E(v_t, h_t; \theta))) - log(\sum_{v_t} \sum_{h_t} exp(-E(v_t, h_t; \theta))) \end{aligned} \tag{24}$$

So, to calculate the derivation, we can get:

$$\begin{aligned}
\frac{\partial log(p(v_t|\theta))}{\partial \theta} &= \frac{\partial}{\partial \theta} log(\sum_{h_t} exp(-E(v_t, h_t; \theta))) - \frac{\partial}{\partial \theta} log(\sum_{v_t}\sum_{h_t} exp(-E(v_t, h_t; \theta))) \\
&= -\frac{1}{\sum_{h_t} e^{-E}} \sum_{h_t} e^{-E} \cdot \frac{\partial E}{\partial \theta} + \frac{1}{\sum_{v_t}\sum_{h_t} e^{-E}} \sum_{v_t}\sum_{h_t} e^{-E} \cdot \frac{\partial E}{\partial \theta} \\
&= -\frac{\sum_{h_t} e^{-E} \cdot \frac{\partial E}{\partial \theta}}{Z \cdot p(v_t|\theta)} + \sum_{v_t}\sum_{h_t} \frac{e^{-E}}{Z} \cdot \frac{\partial E}{\partial \theta} \\
&= -\sum_{h_t} p(h_t|v_t, \theta) \cdot \frac{\partial E}{\partial \theta} + \sum_{v_t}\sum_{h_t} p(v_t, h_t|\theta) \cdot \frac{\partial E}{\partial \theta} \\
&= \underline{-\sum_h p(h|v) \cdot \frac{\partial E(v,h)}{\partial \theta} + \sum_v \sum_h p(v,h) \cdot \frac{\partial E(v,h)}{\partial \theta}}
\end{aligned} \tag{25}$$

From equation (25), we can get:

$$\begin{aligned}
\frac{\partial \mathcal{L}(D|\theta)}{\partial W_{ij}} &= -\sum_h p(h|v) \cdot \frac{\partial E(v,h)}{\partial w_{ij}} + \sum_v \sum_h p(v,h) \cdot \frac{\partial E(v,h)}{\partial w_{ij}} \\
&= \sum_h h_i v_j \cdot p(h|v) - \sum_h \sum_v h_i v_j \cdot p(v,h) \\
&= \underline{p(h_i = 1|v) \cdot v_j - \textstyle\sum_v p(v) \cdot p(h_i = 1|v) \cdot v_j} \\
&= \underline{p(h_i = 1|v) \cdot v_j - \mathbf{E}[v_j h_i]}
\end{aligned} \tag{26}$$

In the similar way, we can have:

$$\begin{aligned}
\frac{\partial \mathcal{L}(D|\theta)}{\partial b_j} &= \sum_h v_j \cdot p(h|v) - \sum_h \sum_v v_j \cdot p(v,h) \\
&= \underline{v_j - \textstyle\sum_v p(v) \cdot v_j} \\
&= \underline{v_j - \mathbf{E}[v_j]}
\end{aligned} \tag{27}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}(D|\theta)}{\partial c_i} &= \sum_h h_i \cdot p(h|v) - \sum_h \sum_v h_i \cdot p(v,h) \\
&= \underline{p(h_i = 1|v) - \textstyle\sum_v p(v) \cdot p(h_i = 1|v)} \\
&= \underline{p(h_i = 1|v) - \mathbf{E}[h_i]}
\end{aligned} \tag{28}$$

4. Contrastive divergence

From equation (26), (27) and (28), we can find that $p(v, h)$ is actually a computationally intractable term. To solve this problem, we should consider 1-step contrastive divergence to solve this problem.

Considering Hinton approximation:

$$\mathbf{E}[v_j h_i] \simeq \mathbf{E}[v_j|h]\mathbf{E}[h_i|v] \tag{29}$$

And the k-step contrastive divergence is:

$$CD_k(\theta, v^{(0)}) = -\sum_h p(h|v^{(0)})\frac{\partial E(^{(0)}, h)}{\partial \theta} + \sum_h p(h|v^{(k)})\frac{\partial E(v^{(k)}, h)}{\partial \theta} \tag{30}$$

First, choose one visible data $v^{(0)}$ from the given training set, sample hidden nodes $h^{(0)}$ according to $p(h|v^{(0)})$. Then, calculate the probability $p(v|h^{(0)})$ using the sampled hidden nodes, and then reconstruct the visible nodes $v^{(1)}$ according to $p(v|h^{(0)})$. Finally, calculate the probability of $p(h^{(1)}|v^{(1)})$. Using these result to approximate and calculate the update rules for W, b and c.

Note: The above method is for 1-step CD sampling, if needed, this procedure can be repeated k times.

II. Code-from-Scratch

1. Methods

2. Results

III. TensorFlow

1. Methods

2. Results