ELEG 5040: Homework #2

April 9, 2015

PROBLEM 1

1. Proof: According to the definition of marginal distribution, we have

$$
\begin{aligned}
P(\mathbf{x}) &= \sum_{\mathbf{h}} P(\mathbf{x},\mathbf{h}) = \sum_{\mathbf{h}} \frac{e^{-E(\mathbf{x},\mathbf{h})}}{Z} \\
&= \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{x},\mathbf{h})}}{Z} = \frac{e^{\log \sum_{\mathbf{h}} e^{-E(\mathbf{x},\mathbf{h})}}}{Z} \\
&= \frac{e^{-(-\log \sum_{\mathbf{h}} e^{-E(\mathbf{x},\mathbf{h})})}}{Z} \\
&= \frac{e^{-\mathscr{F}(\mathbf{x})}}{Z}
\end{aligned}
$$

where $\mathscr{F}(\mathbf{x}) = -\log \sum_{\mathbf{h}} e^{-E(\mathbf{x},\mathbf{h})}$.

Because $\sum_{\mathbf{x}} P(\mathbf{x}) = 1$, we have

$$
1 = \sum_{\mathbf{x}} P(\mathbf{x}) = \sum_{\mathbf{x}} \frac{e^{-\mathscr{F}(\mathbf{x})}}{Z} = \frac{\sum_{\mathbf{x}} e^{-\mathscr{F}(\mathbf{x})}}{Z}
$$

Thus we get

$$
Z = \sum_{\mathbf{x}} e^{-\mathscr{F}(\mathbf{x})}
$$

2. Proof: First we derive the expression of negative log-likelihood,

$$
\begin{aligned}
-\log P(\mathbf{x}) &= -\log \frac{e^{-\mathscr{F}(\mathbf{x})}}{Z} \\
&= \log Z + \mathscr{F}(\mathbf{x})
\end{aligned}
$$

Then, we derive the partial grdient wrt θ,

$$
\begin{aligned}
-\frac{\partial \log P(\mathbf{x})}{\partial\theta} &= \frac{1}{Z}\frac{\partial Z}{\partial\theta} + \frac{\partial \mathscr{F}(\mathbf{x})}{\partial\theta} \\
&= \frac{1}{Z}\frac{\partial \sum_{\tilde{\mathbf{x}}} e^{-\mathscr{F}(\tilde{\mathbf{x}})}}{\partial\theta} + \frac{\partial \mathscr{F}(\mathbf{x})}{\partial\theta} \\
&= -\sum_{\tilde{\mathbf{x}}} \frac{1}{Z} e^{-\mathscr{F}(\tilde{\mathbf{x}})} \frac{\partial \mathscr{F}(\tilde{\mathbf{x}})}{\partial\theta} + \frac{\partial \mathscr{F}(\mathbf{x})}{\partial\theta} \\
&= -\sum_{\tilde{\mathbf{x}}} P(\tilde{\mathbf{x}}) \frac{\partial \mathscr{F}(\tilde{\mathbf{x}})}{\partial\theta} + \frac{\partial \mathscr{F}(\mathbf{x})}{\partial\theta}
\end{aligned}
$$

thus completes the proof.

3. To generate $\mathscr{N}$ from the model, we start with a training sample $\mathbf{x}$ (sampled from training samples), and do $|\mathscr{N}|$ Gibbs sampling steps:

$$
\begin{aligned}
\mathbf{x}_1 &\sim \hat{P}(\mathbf{x}) \\
\mathbf{h}_1 &\sim P(\mathbf{h}|\mathbf{x}_1) \\
\mathbf{x}_2 &\sim P(\mathbf{x}|\mathbf{h}_1) \\
\mathbf{h}_2 &\sim P(\mathbf{h}|\mathbf{x}_2) \\
\mathbf{x}_3 &\sim P(\mathbf{x}|\mathbf{h}_2) \\
\mathbf{h}_3 &\sim P(\mathbf{h}|\mathbf{x}_3) \\
&\vdots \\
\mathbf{x}_{|\mathscr{N}|} &\sim P(\mathbf{x}|\mathbf{h}_{|\mathscr{N}|-1})
\end{aligned}
$$

Each Gibbs sampling step in RBM only consists two substeps (sample $\mathbf{h}$ given current $\mathbf{x}$, and sample $\mathbf{x}$ given current $\mathbf{h}$), while in a fully-connected Boltzmann Machine, each step we have to sample every node in $\mathbf{x}$ and $\mathbf{h}$ given all the other nodes. This is because in RBM, there are no interactions within $\mathbf{x}$ or $\mathbf{h}$, so given $\mathbf{h}$, we can sample all the nodes of $\mathbf{x}$ without affecting each other, and vice versa. The sampling step reduce to two substeps.

Problem 2

Proof: First, let's define some notations:
Let $X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{n\times N}$ be the N samples (each is a n-dimension vector), $H = [h_1, h_2, \dots, h_N] \in \mathbb{R}^{p\times N}$ ($p < n$) be N outputs of hidden units and $Y = [y_1, y_2, \dots, y_N] \in \mathbb{R}^{n\times N}$ be the N outputs.
In auto encoder without non-linear transform, we have:

$$H = W_1 X + w_1 u^t \tag{0.1}$$

$$Y = W_2 H + w_2 u^t \tag{0.2}$$

where $W_1 \in \mathbb{R}^{p\times n}$ and $W_2 \in \mathbb{R}^{n\times p}$ are the two transformation matrices, $w_1 \in \mathbb{R}^p$ and $w_2 \in \mathbb{R}^n$ are the two bias terms, and $u = [1,1,\dots,1]^t \in \mathbb{R}^N$.
The target function we want to minimize is:

$$J = ||X - Y||^2$$

.

1. In the simplest case, we assume X and Y has zero means and the bias terms are also zeros:

$$\begin{aligned}\bar{x} &= \frac{1}{N} X u = \mathbf{0} \\ \bar{y} &= \frac{1}{N} Y u = \mathbf{0} \\ w_1 &= \mathbf{0} \\ w_2 &= \mathbf{0}\end{aligned}$$

Then the problem reduce to:

$$\begin{aligned}J &= ||X - W_2 H||^2 \\ &= ||X - W_2 W_1 X||^2\end{aligned}$$

where $W_2 W_1 \in \mathbf{R}^{n\times n}$ but rank $p < n$. This is equivalent to a typical PCA problem. Let suppose the SVD of X has the form of

$$X = U_n \Sigma_n V_n^t$$

where the columns of $U_n \in \mathbf{R}^{n\times n}$ are the eigenvectors of XX^t corresponding to eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \dots \geqslant \lambda_n$, $\Sigma_n = diag[\sigma_1, \sigma_2, \dots, \sigma_n]$ with $\sigma_i = \sqrt{\lambda_i}$ and columns of $V_n \in \mathbb{R}^{N\times n}$ are the eigenvectors of $X^t X$.

Then in a PCA problem we know the optimal solution is projecting samples into the space spanned by the first p eigenvectors:

$$W_2 W_1 X = U_p \Sigma_p V_p^t$$

with $\Sigma_p = diag[\sigma_1, \sigma_2, \dots, \sigma_p]$ and U_p are formed by the first p eigenvectors in U_n.

In general, suppose T is an arbitrary non-singular $p \times p$ matrix,

$$W_2 = U_p T^{-1}$$

and

$$H = T \Sigma_p V_p^t.$$

The intuition is that H is the p-dimension PCA projection of X plus an arbitrary transformation T in the p-dimensional space and W_2 is the inverse transformation of T plus the projection back to original space. Therefore, the space of hidden units is the same span of first p principal components.

2. In general cases,

$$
\begin{aligned}
J &= ||X - Y||^2 \\
&= ||X - W_2 H - w_2 u^t||^2 \\
&= \mathrm{tr}\,[(X - W_2 H - w_2 u^t)(X - W_2 H - w_2 u^t)^t] \\
&= \mathrm{tr}\,[(X - W_2 H)(X - W_2 H)^t] - 2 w_2^t (X - W_2 H) u + u^t u w_2^t w_2
\end{aligned}
$$

   To minimize J with respect to w_2, we have:

$$
\begin{aligned}
\frac{\partial J}{\partial w_2} &= -2(X - W_2 H)u + 2N w_2 = 0 \\
\hat{w}_2 &= \frac{1}{N}(X - W_2 H)u
\end{aligned}
$$

   Therefore, the original problem becomes:

$$
\begin{aligned}
J &= ||X - W_2 H - w_2 u^t||^2 \\
&= ||X - W_2 H - \frac{1}{N}(X - W_2 H) u u^t||^2 \\
&= ||(X - X\frac{uu^t}{N}) - W_2(H - H\frac{uu^t}{N})||^2 \\
&= ||X' - W_2 H'||^2
\end{aligned}
$$

   with $X' = X - X\frac{uu^t}{N} = X - \bar{x}u^t$, which is each sample subtracting the mean vector. So w_2 will ensure that X' has zero mean. In addition, it is easy to show that $\hat{w}_2 = \frac{1}{N}(X - W_2 H)u$ also ensures $\bar{y} = \frac{1}{N} Y u = \bar{x}$ and the problem has the same form as before.

PROBLEM 3

1. Proof: Since in RBM hidden nodes are binary, ie $h_j \in \{0, 1\}$, the proof is trivial

$$
\begin{aligned}
f_j &= \log \sum_{i_j} e^{h_j \mathbf{W}_j . \mathbf{x}} \\
&= \log (e^{0\ \mathbf{W}_j . \mathbf{x}} + e^{1\ \mathbf{W}_j . \mathbf{x}}) \\
&= \log (1 + e^{\mathbf{W}_j . \mathbf{x}})
\end{aligned}
$$

2. Let plot $f_j(\mathbf{x}) = \log(1 + e^{\mathbf{W}_j . \mathbf{x}})$ wrt $\mathbf{W}_j . \mathbf{x}$ in Fig 0.1. The function $f'(x) = \log(1 + e^x)$ is called a **softplus** function. When $\mathbf{W}_j . \mathbf{x}$ is a large positive value, f_j approaches x asymptotically, and approaches 0 when $\mathbf{W}_j . \mathbf{x}$ is a large negative value. When $\mathbf{W}_j . \mathbf{x}$ is close to zero, f_j has strong non-linearity.

   Since $P(\mathbf{x}) \propto \prod_j e^{f_j(\mathbf{x})}$, in each dimension j of hidden units, f_j indicates whether an attribute appears and $\mathbf{W}_j . \mathbf{x}$ classifies the samples into two parts ($\mathbf{W}_j . \mathbf{x} > 0$ and $\mathbf{W}_j . \mathbf{x} < 0$) in this dimension.

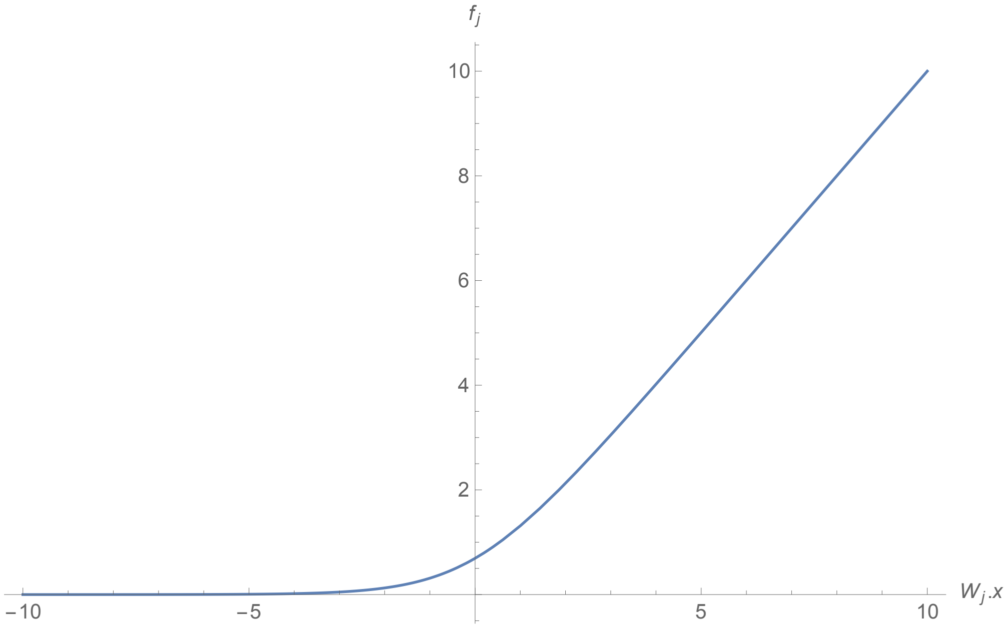

Figure 0.1: Softplus Function

3. Since $P(\mathbf{x}) \propto \prod_j e^{f_j(\mathbf{x})}$, f_j constrains the distribution of $\mathbf{x}$ along the projection direction $\mathbf{W}_{j\cdot}$. An input sample $\mathbf{x}$ satisfies a constraint f_i when $\mathbf{W}_{i\cdot}\mathbf{x} > 0$, and since the hidden units are independent (rows of $\mathbf{W}$ are independent), $\mathbf{x}$ can satisfy multiple constraints. $P(\mathbf{x})$ is product of all experts (constraints), so $P(\mathbf{x})$ is large only if all the constraint are satisfied.

4. In a mixture model (sum of experts), the probability is a weighted sum of mixture components (experts). For a Gaussian mixture model in particular, the constraints associated with an expert $N(\mathbf{x}; \mu_i, \Sigma_i)$ can be regarded as a local region measured in **Mahalanobis distance** $d_i = \sqrt{(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}$. Although probability within a certain Mahalanobis distance decreases as number of dimensionality increases, for a certain number of dimensionality, we could find a local region that contains the majority of samples. For example, in one dimensional example, the 3-sigma distance contains 99.7% of samples.

   Since probabilities are positive, a mixture distribution can have high probability for a sample $\mathbf{x}$ if any one of experts assign high probability to that sample.

PROBLEM 4

1. $\frac{\partial L}{\partial \mathbf{W}_{hz}}$

First we define some symbols and derive some expressions:

$$\frac{d\tanh(x)}{dx} = 1 - \tanh^2(x) \tag{0.3}$$

$$\delta_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

$$\delta_a = \begin{pmatrix} \delta_{1,a} \\ \delta_{2,a} \\ \vdots \\ \delta_{n,a} \end{pmatrix}$$

$$h_t = \tanh(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{0.4}$$

$$z_t = \text{softmax}(W_{hz}h_t + b_z) \tag{0.5}$$

$$\frac{\partial L_t}{\partial (z_t)_i} = -\frac{\partial \log(z_t)_{y_t}}{\partial (z_t)_i} = -\frac{1}{(z_t)_{y_t}}\frac{\partial (z_t)_{y_t}}{\partial (z_t)_i} = -\frac{1}{(z_t)_{y_t}}\delta_{y_t,i} \tag{0.6}$$

$$\frac{\partial (z_t)_i}{\partial (W_{hz})_{lk}} = (z_t)_i\delta_{i,l}(h_t)_k - (z_t)_i(z_t)_l(h_t)_k \tag{0.7}$$

Then,

$$\begin{aligned}
\frac{\partial L_t}{\partial (W_{hz})_{lk}} &= \sum_i \frac{\partial L_t}{\partial (z_t)_i}\frac{\partial (z_t)_i}{\partial (W_{hz})_{lk}} \\
&= \sum_i -\frac{1}{(z_t)_{y_t}}\delta_{y_t,i}[(z_t)_i\delta_{il}(h_t)k - (z_t)_i(z_t)_l(h_t)_k] \\
&= -\frac{1}{(z_t)_{y_t}}[(z_t)_{y_t}\delta_{y_t,l}(h_t)_k - (z_t)_{y_t}(z_t)_l(h_t)_k] \\
&= -\delta_{y_t,l}(h_t)_k + (z_t)_l(h_t)_k
\end{aligned}$$

Therefore,

$$\frac{\partial L}{\partial (W_{hz})_{lk}} = \sum_t -\delta_{y_t,l}(h_t)_k + (z_t)_l(h_t)_k$$

And in matrix form,

$$\frac{\partial L}{\partial W_{hz}} = \sum_t (z_t - \delta_{y_t})h_t^T$$

2. $\frac{\partial L}{\partial \mathbf{W}_{hh}}$

$$
\begin{aligned}
\frac{\partial(z_t)_i}{\partial(h_t)_p} &= (z_t)_i(W_{hz})_{ip} - (z_t)_i(W_{hz}^T)_p.z_t \\
\frac{\partial L}{\partial(z_t)_i} &= \frac{\partial L_t}{\partial(z_t)_i} = -\frac{1}{(z_t)_{y_t}}\delta_{y_t,i} \\
\frac{\partial(h_t)_r}{\partial(h_{t-1})_p} &= [1-(h_t)_r^2](W_{hh})_{rp} \\
\frac{\partial L}{\partial(h_t)_p} &= \sum_r \frac{\partial L}{\partial(h_{t+1})_r}\frac{\partial(h_{t+1})_r}{\partial(h_t)_p} + \sum_i \frac{\partial L}{\partial(z_t)_i}\frac{\partial(z_t)_i}{\partial(h_t)_p} \\
&= \sum_r \frac{\partial L}{\partial(h_{t+1})_r}[1-(h_t)_r^2](W_{hh})_{rp} + \sum_i -\frac{1}{(z_t)_{y_t}}\delta_{y_t,i}[(z_t)_i(W_{hz})_{ip} - (z_t)_i(W_{hz}^T)_p.z_t] \\
&= [\sum_r \frac{\partial L}{\partial(h_{t+1})_r}[1-(h_t)_r^2](W_{hh})_{rp}] + [(W_{hz}^T)_p.z_t - (W_{hz})_{y_t,p}] \\
\frac{\partial(h_t)_p}{\partial(W_{hh})_{mn}} &= [1-(h_t)_p^2]\delta_{m,p}(h_{t-1})_n
\end{aligned}
$$

In matrix form,

$$
\frac{\partial L}{\partial h_t} = W_{hh}^T(1-diag(h_t)^2)\frac{\partial L}{\partial h_{t+1}} + W_{hz}^T(z_t - \delta_{y_t})
$$

$$
\frac{\partial L}{\partial W_{hh}} = \sum_t \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial W_{hh}} = \sum_t (1-diag(h_t)^2)\frac{\partial L}{\partial h_t}h_{t-1}^T
$$