ELEG 5491 HW3

Zhizhong Li
Apr 2017
Information Engineering, CUHK
`1155070507, lz015@ie.cuhk.edu.hk`

1 Problem 1

1.1

$$
\begin{aligned}
P(\mathbf{x}) &= \sum_h \frac{e^{-E(\mathbf{x},\mathbf{h})}}{Z} \\
&= \exp\left(\log \sum_h \frac{e^{-E(\mathbf{x},\mathbf{h})}}{Z}\right) \\
&= \exp\left(\log \sum_h e^{-E(\mathbf{x},\mathbf{h})} - \log Z\right) \\
&= \frac{\exp\left(\log \sum_h e^{-E(\mathbf{x},\mathbf{h})}\right)}{Z} \\
&= \frac{e^{-\mathcal{F}(\mathbf{x})}}{Z}.
\end{aligned}
\tag{1}
$$

1.2

$$
\begin{aligned}
-\frac{\partial \log p(\mathbf{x})}{\partial \theta} &= -\frac{\partial}{\partial \theta} \log\left(\frac{e^{-\mathcal{F}(\mathbf{x})}}{Z}\right) \\
&= -\frac{\partial(-\mathcal{F}(\mathbf{x}))}{\partial \theta} + \frac{\partial \log Z}{\partial \theta} \\
&= \frac{\partial \mathcal{F}(\mathbf{x})}{\partial \theta} + \frac{1}{Z}\frac{\partial}{\partial \theta} \sum_{\tilde{\mathbf{x}}} e^{-\mathcal{F}(\tilde{\mathbf{x}})} \\
&= \frac{\partial \mathcal{F}(\mathbf{x})}{\partial \theta} - \sum_{\tilde{\mathbf{x}}} \frac{\partial \mathcal{F}(\tilde{\mathbf{x}})}{\partial \theta} \frac{e^{-\mathcal{F}(\tilde{\mathbf{x}})}}{Z} \\
&= \frac{\partial \mathcal{F}(\mathbf{x})}{\partial \theta} - \sum_{\tilde{\mathbf{x}}} \frac{\partial \mathcal{F}(\tilde{\mathbf{x}})}{\partial \theta} p(\tilde{\mathbf{x}})
\end{aligned}
\tag{2}
$$

1.3

The samples can be generated by MCMC. One approach in MCMC is Gibbs sampling. First generate an initial sample $\mathbf{x}_0$, Then generate $\mathbf{h}_1$ according to conditional distribution $P(\mathbf{h}|\mathbf{x}_0)$. Then

ELEG 5491 Homeworks [1].

generate $\mathbf{x}_1$ from conditional distribution $P(\mathbf{x}|\mathbf{h}_1)$. So on and so on. Then the series of samples $\mathbf{x}_0, \mathbf{x}_1, \dots \mathbf{x}_{|\mathcal{N}|}$ is the wanted stuff.

RBM does not model the interactions of variables of the same layer. So the mathematical expression for energy is simplified, and the conditional distributions used in Gibbs sampling have closed forms. This helps the generation process.

2 Problem 2

Suppose there are m training samples $x_1, x_2, \dots, x_m$ with sample mean $\bar{x}$. Assemble them into a matrix X whose i-th column is x_i. Let $\mathbb{1} \in \mathbb{R}^m$ is a all-one vector. Then the total loss is

$$\begin{aligned} L(W, W', b, b') &= \|W'(WX + b\mathbb{1}^T) + b'\mathbb{1}^T - X\|_F^2 \\ &= \|W'W(X - \bar{x}\mathbb{1}^T) + W'W\bar{x}\mathbb{1}^T + W'b\mathbb{1}^T + b'\mathbb{1}^T - (X - \bar{x}\mathbb{1}^T) - \bar{x}\mathbb{1}^T\|_F^2 \\ &= \|W'W\tilde{X} - \tilde{X} + (W'W\bar{x} + W'b + b' - \bar{x})\mathbb{1}^T\|_F^2 \\ &= \|W'W\tilde{X} - \tilde{X} + v\mathbb{1}^T\|_F^2. \end{aligned} \tag{3}$$

where $\tilde{X} := X - \bar{x}\mathbb{1}^T$ is the centered data matrix, $W'W$ is a matrix with rank no more than k, and $v := W'W\bar{x} + W'b + b' - \bar{x}$. Let v_i denote the i-th column of matrix $(W'W\tilde{X} - \tilde{X})$, we have

$$\sum_{i=1}^{m} v_i = 0. \tag{4}$$

Then the loss is equivalent to

$$L = \sum_{i=1}^{m} \|v_i - v\|^2 \tag{5}$$

To minimize loss, the best v is 0. So we set $v = 0$ and get

$$L = \|W'W\tilde{X} - \tilde{X}\|_F^2. \tag{6}$$

Since the column space of $W'W\tilde{X}$ is contained in the column space of W'. So the best we can get for L is let $W'W\tilde{X}$ be the projection of column vectors of $\tilde{X}$ onto the column space of W' and W be the projection operator. Suppose an SVD of $\tilde{X}$ is $U\Sigma V^T$, then let W be the first k columns of U, and let $W' = W^T$, we obtain the minimum loss L. Look at the whole process then, it is equivalent to a PCA.

References

[1] X. Wang. Assignments, 2017. URL http://dl.ee.cuhk.edu.hk/.