ELEG 5491: Homework #3

Due on Thursday, April 6, 2017, 4:30pm (in class)

Xiaogang Wang

Problem 1

[**70 points**]

An energy based model with hidden units is defined as below

$$P(\mathbf{x}, \mathbf{h}) = \frac{e^{-E(\mathbf{x},\mathbf{h})}}{Z}$$

where $\mathbf{x}$ is the observed data, $\mathbf{h}$ is a vector of hidden variables, $E()$ is an energy function, and Z is the normalization factor. The marginal distribution of the observed data is

$$P(\mathbf{x}) = \sum_{\mathbf{h}} P(\mathbf{x}, \mathbf{h}) = \sum_{\mathbf{h}} \frac{e^{-E(\mathbf{x},\mathbf{h})}}{Z}$$

1. If the free energy of $P(\mathbf{x})$ is defined as $\mathcal{F}(\mathbf{x}) = -\log \sum_{\mathbf{h}} e^{-E(\mathbf{x},\mathbf{h})}$, prove that $P(\mathbf{x})$ can be written as $P(\mathbf{x}) = \frac{e^{-\mathcal{F}(\mathbf{x})}}{Z}$ with $Z = \sum_{\mathbf{x}} e^{-\mathcal{F}(\mathbf{x})}$. [**20 points**]
2. Prove that the negative data log-likelihood gradient has the form

$$-\frac{\partial \log p(\mathbf{x})}{\partial \theta} = \frac{\partial \mathcal{F}(\mathbf{x})}{\partial \theta} - \sum_{\tilde{\mathbf{x}}} p(\tilde{\mathbf{x}}) \frac{\partial \mathcal{F}(\tilde{\mathbf{x}})}{\partial \theta}$$

   where θ is the parameter vector of $E(\mathbf{x}, \mathbf{h})$. [**30 points**]
3. When computing the gradient on a training sample $\mathbf{x}$, it is easy to calculate the first term. However, the second term usually has no closed-form solution. To make the computation tractable, we can estimate the second term by using a fixed number of samples $\mathcal{N}$, which are generated according to P.

$$-\frac{\partial \log p(\mathbf{x})}{\partial \theta} = \frac{\partial \mathcal{F}(\mathbf{x})}{\partial \theta} - \frac{1}{|\mathcal{N}|} \sum_{\tilde{\mathbf{x}} \in \mathcal{N}} \frac{\partial \mathcal{F}(\tilde{\mathbf{x}})}{\partial \theta}$$

Please explain how to generate $\mathcal{N}$ if the model is RBM. Compared with Boltzmann Machine, how does the special structure of RBM help to generate $\mathcal{N}$? [**20 points**]

Problem 2

[**30 points**]

Please prove that in autoencoder, if there is one linear hidden layer and the mean squared error criterion is used to train the network, the k hidden unites learn to project the input in the span of the first k principal components of data obtained by PCA.