ELEG 5040: Homework #2

Due on Monday, March 30, 2015, 7:30pm

Xiaogang Wang

Problem 1

[**30 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 $Z = \sum_{\mathbf{x}} e^{-\mathcal{F}(\mathbf{x})}$

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 $E(\mathbf{x}, \mathbf{h})$.

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 $\mathcal{N}$?

Problem 2

[**20 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.

Problem 3

[**20 points**]

RBM can be represented as a product of experts

$$\begin{aligned} P(\mathbf{x}) &\propto e^{\mathbf{b}'\mathbf{x}+\sum_j \log \sum_{h_j} e^{h_j \mathbf{W}_{j\cdot}\mathbf{x}}} \\ &\propto \prod_j e^{f_j(x)} \end{aligned}$$

where $f_j = \log \sum_{h_j} e^{h_j \mathbf{W}_{j\cdot}\mathbf{x}}$

1. Prove that $f_j = \log(1 + e^{\mathbf{W}_{j\cdot}\mathbf{x}})$
2. Show that $f_j(\mathbf{x})$ is an attribute indicator and $\mathbf{W}_{j\cdot}\mathbf{x}$ can be treated as a linear classifier for the attribute. Hint: this classifier forms a hyperplane and partitions the space of $\mathbf{x}$ into two parts; consider how f_j changes when $\mathbf{W}_{j\cdot}\mathbf{x}$ is a large positive value, zero, and a large negative value.
3. Show that in a product of experts, each $f_j(\mathbf{x})$ represents a constraint, an input sample $\mathbf{x}$ can satisfy multiple constraints, and $P(\mathbf{x})$ is large only if all the constraints are (approximately) satisfied.
4. In a Gaussian mixture model, $P(\mathbf{x}) = \sum_{i=1}^{n} N(\mathbf{x}; \mu_i, \Sigma_i)P_i$, where μ_i is the mean vector, Σ_i is the covariance matrix, and P_i is the prior of the ith Gaussian mixture component. Please show that the constraint associated with an expert is an indication of belonging to a region which excludes the other regions; and that a mixture distribution can have high probability for a sample $\mathbf{x}$ if any of the single expert assigns high probability to that sample.

Problem 4

[**30 points**]

A recurrent neural network is shown in the figure below,

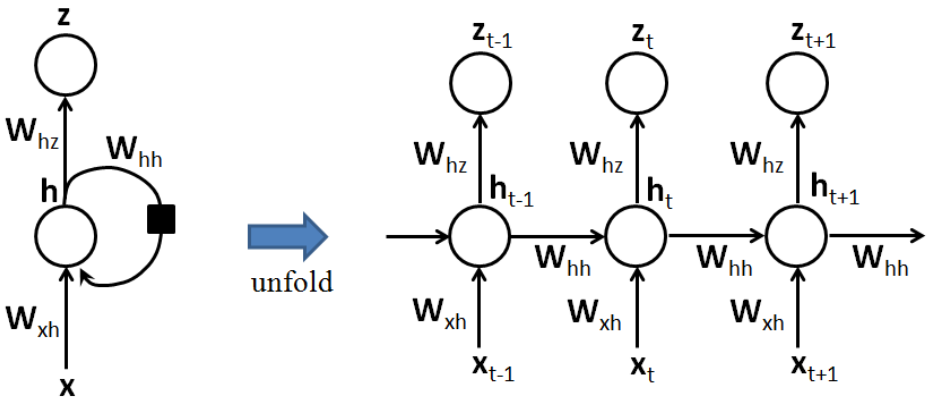
$$\mathbf{h}_t = \tanh(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{b}_h)$$

$$\mathbf{z}_t = \text{softmax}(\mathbf{W}_{hz}\mathbf{h}_t + \mathbf{b}_z)$$

The total loss for a given input/target sequence pair ($\mathbf{x}$, $\mathbf{y}$), measured in cross entropy

$$L(\mathbf{x}, \mathbf{y}) = \sum_t L_t = \sum_t -\log z_{t,y_t}$$

Problem 4 continued on next page...



In the lecture, we provide the general idea of how to calculate the gradients $\frac{\partial L}{\partial \mathbf{W}_{hz}}$ and $\frac{\partial L}{\partial \mathbf{W}_{hh}}$. Please provide the details of the algorithms and equations, considering the mapping and cost functions provided above.