Restricted Boltzmann Machines

Jason Liang, Julian Sia

***Abstract*—Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.**

***Keywords*—IEEEtran, journal, LaTeX, paper, template.**

I. Introduction

Restricted Boltzmann Machines are a type of undirected graphical model consisting of two sets of nodes, a set of hidden units and a set of visible units. Given the graph $G = (V, E)$ that describes the Restricted Boltzmann Machine, $\forall v_{vis} \epsilon V_{vis} \subset V, \exists e'$, such that $e' = (v_{vis}, v_{hid}), \forall v_{hid} \epsilon V_{hid} \subset V$. Restricted Boltzmann Machines are a variant of the Boltzmann Machine posed by Hinton and Sejnowski, with the restriction that there are neither edges connecting any visible units nor edges connecting any hidden units. This restriction guarantees that the visible units are conditionally independent of other visible units; and hidden units, conditionally independent of hidden units. Restricted Boltzmann Machines today are commonly used in classification tasks, dimensionality reduction, and feature learning tasks.

A. History

Restricted Boltzmann Machines were created as a variant of Boltzmann Machines after practical difficulties were encountered in training Boltzmann machines, namely the exponential training time required to learn weights in a Boltzmann Machine and sequential nature of training particular to the Boltzmann Machine. Boltzmann Machine themselves were motivated primarily by limitations of Hopfield Networks, primarily their poor storage capacity and tendency to converge at local energy minima.

In the early days of Boltzmann machines, Hinton and Sejnowski calculated the gradient of the log likelihood by "[fixing] a training vector on the visible units, initializing the hidden units to random binary states, and using sequential Gibbs sampling of the hidden units" (Salakhutdinov and Hinton 2012), using simulated annealing to speed up the convergence process. Other attempts were made by Neal with persistent Markov chains and Peterson and Anderson with mean-field methods (instead of Gibbs Sampling) to speed up the learning process. After Restricted Boltzmann Machines were posed as a simplifying restriction by Smolensky, Hinton discovered a way to speed up the learning process, coined contrastive divergence, which is described in this report.

B. Learning Procedures

Given a set of training vectors, V, to train a Restricted Boltzmann, one aims to maximize the average log probability, $p(v), v \epsilon V$, where

$$p(v) = \frac{1}{Z} \sum_{h} e^{-E(v,h)},\ Z = \sum_{v,h} e^{-E(v,h)} \tag{1}$$

One can do accomplish this by taking the derivative of the log probability of $p(v)$ with respect to the weights defined in the energy function, $E(v, h)$ defined as

$$\begin{aligned} E(v,h) &= -\sum_{i \epsilon visible} a_i v_k - \sum_{j \epsilon hidden} b_j h_j - \sum_i \sum_j v_i w_{i,j} h_j \\ &= -b'v - c'h - h'Wv \end{aligned} \tag{2}$$

Differentiating with respect to $w_{i,j}$, we find that the partial derivative reduces to the following:

$$\frac{\partial p(v)}{\partial w_{i,j}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \tag{3}$$

With this reduction the following learning rule with ϵ learning rate can be derived

$$\Delta w_{i,j} = \epsilon \langle \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \rangle \tag{4}$$

Alternatively, using free energy notation, the partial derivative can be defined as

$$-\frac{\partial p(v)}{\partial \Theta} = \frac{\partial \mathcal{F}(v)}{\partial \Theta} - \sum_{\tilde{v}} p(\tilde{v}) \frac{\partial \mathcal{F}(\tilde{v})}{\partial \Theta} \tag{5}$$

where we define free energy as

$$\mathcal{F}(v) = -log \sum_{h} e^{-E(v,h)} \tag{6}$$

Because the computation of the second term in the difference that defines the partial derivative is intractable, we can approximate the second term using a fixed number of samples from the model, such that

M. Shell is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

J. Doe and J. Doe are with Anonymous University.

Manuscript received April 19, 2005; revised January 11, 2007.

$$-\frac{\partial p(v)}{\partial \Theta} = \frac{\partial \mathcal{F}(v)}{\partial \Theta} - \frac{1}{|\mathcal{N}|} \sum_{\tilde{v} \epsilon N} p(\tilde{v}) \frac{\partial \mathcal{F}(\tilde{v})}{\partial \Theta} \tag{7}$$

In the case of a binary Restricted Boltzmann Machine, the free energy function reduces to the following:

$$\mathcal{F}(v) = -b'v - \sum_i log \sum_{h_i} e^{h_i(c_i + W_i v)} \tag{8}$$

Using the free energy function defined in (8) and the derivative defined in (7), we can define the following update equations for the Restricted Boltzmann Machine's log likelihood gradient:

$$\begin{aligned} -\frac{\partial p(v)}{\partial W_{i,j}} &= E_v[p(h_i|v) \cdot v_j] - v_j^{(i)} \cdot sigm(W_i \cdot v^{(i)} + c_i) \\ -\frac{\partial p(v)}{\partial c_i} &= E_v[p(h_i|v)] - v_j^{(i)} \cdot sigm(W_i \cdot v^{(i)}) \\ -\frac{\partial p(v)}{\partial b_j} &= E_v[p(h_i|v)] - v_j^{(i)} \end{aligned} \tag{9}$$

To obtain samples from p(x), one can use Gibbs sampling until a Markov chain converges or near converges. Because of the conditional independence assumptions of the visible and hidden nodes such that

$$\begin{aligned} p(h|v) &= \prod_i p(h_i|v) \\ p(v|h) &= \prod_j p(v_j|h) \end{aligned} \tag{10}$$

we can define the following activation functions:

$$\begin{aligned} p(h_i = 1|v) &= sigm(c_i + W_i v) \\ p(v_j = 1|j) &= sigm(b_j + W_i' h) \end{aligned} \tag{11}$$

Given the conditional independence assumptions of the Restricted Boltzmann Machine, we can perform block Gibbs Sampling with the following update rules:

$$\begin{aligned} h^{(n+1)} &\sim sigm(W' v^{(n)} + c) \\ v^{(n+1)} &\sim sigm(W h^{(n+1)} + b) \end{aligned} \tag{12}$$

As $t->, (v^{(t)}, h^{(t)})$, the samples of $p(v, h)$ will be accurate. Nevertheless, one can further speed up the process by using the "contrastive divergence," initialize the Markov chain with a training example $v \epsilon V$ and running the Markov chain for k (often with $k = 2$.

II. Conclusion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Appendix A
Proof of the First Zonklar Equation

Some text for the appendix.

Acknowledgment

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

John Doe Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.