Generalised Classification Restricted Boltzmann Machines

Son N. Tran[a,1,*], Chadi Hajj[b,2], Artur Garcez[b,2], Tillman Weyde[b,2]

[a] *The Australian E-Health Research Centre, CSIRO*
[b] *Department of Computer Science, City University of London*

Abstract

Here goes the abstract

Keywords: Restricted Boltzmann machines, Classification

2010 MSC: 00-01, 99-00

1. Introduction

The restricted Boltzmann machine (RBM) is a generative latent-variable model which models the joint distribution of a set of input variables. It has gained popularity over the past decade in many applications, especially for pre-training deep neural network classifiers [1, 2]. One of its applications is as a standalone classifier, referred to as the Discriminative Restricted Boltzmann Machine (DRBM)[3]. As the name might suggest, the DRBM is a classifier obtained by carrying out discriminative learning in the RBM and it directly models the conditional distribution one is interested in for prediction. This by-passes one of the key problems faced in learning the parameters of the RBM generatively, which is the computation of the intractable *partition function*. In the DRBM this partition function is cancelled out in the expression for the conditional distribution thus simplifying the learning process.

*Corresponding author
Email address: `son.tran@csiro.au` (Son N. Tran)
[1] Level 5 UQ Health Sciences Building, Royal Brisbane and Women's Hospital, Herston, Queensland 4029 Australia
[2] College Building, Northampton Square, London, EC1V 0HB, United Kingdom

It is often the case that a new type of activation function results in an improvement in the performance of an existing model or in a new insight into the behaviour of the model itself. In the least, it offers researchers with the choice of a new modelling alternative. In fact, different type of units such as bipolar Bernoulli [4], Gaussian [5], Binomial [6] and rectified linear [7] have been studied. However, we observe that while effort has gone into enhancing the performance of a few other connectionist models by changing the nature of their hidden units, this has not been attempted with the DRBM. So in this paper, we first describe a novel theoretical result that makes it possible to generalise the model's cost function. The result is then used to derive two new cost functions corresponding to DRBMs containing hidden units with the Binomial and $\{-1, +1\}$-Bernoulli distributions respectively. These two variants are evaluated and compared with the original DRBM on the benchmark MNIST and USPS digit classification datasets, and the 20 Newsgroups document classification dataset. We find that each of the three compared models outperforms the remaining two in one of the three datasets, thus indicating that the proposed theoretical generalisation of the DRBM may be valuable in practice.

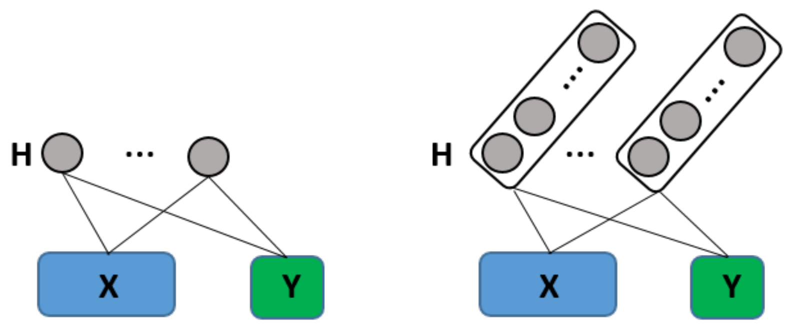
In the next Section, we explain the generalisation of the discriminative function in RBMs. It is followed by Section **??** that shows how to implement this idea. Experimental results are discussed in Section **??** and Section **??** presents a summary, together with potential extensions of this work

2. Related Work

3. Generalised Classification Restricted Boltzmann Machines

3.1. Model

The Restricted Boltmzann Machine (RBM) [8] is an undirected bipartite graphical model. In the case for classification, it contains a set of visible units $\mathbf{v} = \{\mathbf{x} \in \mathbb{R}^{n_x}, \mathbf{y} \in \mathbb{R}^{n_y}\}$, where $\mathbf{x}$ is the input vector, and $\mathbf{y}$ is the one-hot encoding of the class-label; and a set of hidden units $\mathbf{h} \in \mathbb{R}^{n_h}$. The two layers are fully inter-connected but there exist no connections between any two hidden



(a) Classification RBMs (b) Generalised classification RBMs

units, or any two visible units. Additionally, the units of each layer are connected to a bias unit whose value is always 1. The edge between the i^{th} input x_i and the j^{th} hidden unit h_j is associated with a weight w_{ij}. All these weights are together represented as a weight matrix $W \in \mathbb{R}^{n_x \times n_h}$. Similarly, $U \in \mathbb{R}^{n_y \times n_h}$ is the weight matrix between labels $\mathbf{y}$ and the hidden layer $\mathbf{h}$. The weights of connections between input and label units and the bias unit are contained in bias vectors $\mathbf{a} \in \mathbb{R}^{n_x}$, $\mathbf{b} \in \mathbb{R}^{n_y}$ respectively. Likewise, for the hidden units there is a hidden bias vector $\mathbf{c} \in \mathbb{R}^{n_h}$. The RBM is characterized by an energy function: $E(\mathbf{x}, \mathbf{y}, \mathbf{h}) = -\mathbf{a}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{y} - \mathbf{c}^\top \mathbf{h} - \mathbf{x}^\top W \mathbf{h} - \mathbf{y}^\top U \mathbf{h}$ to represent the joint probability of every possible pair of visible and hidden vectors as: $P(\mathbf{x}, \mathbf{y}, \mathbf{h}) = \frac{1}{Z} \mathrm{e}^{-E(\mathbf{x}, \mathbf{y}, \mathbf{h})}$ where Z is the partition function, $Z = \sum_{\mathbf{x}, \mathbf{y}, \mathbf{h}} \mathrm{e}^{-E(\mathbf{x}, \mathbf{y}, \mathbf{h})}$.

.... Normally, RBMs can have binary units or Gaussian units []. However, the latter is difficult to use for classification since different from the binary hidden units in RBMs with Gaussian hidden units $p(y|\mathbf{x})$ is intractable. We can ... in [] to extend binary hidden units to represent binomial distribution.

...

An RBM with binomial hidden units can be constructed by replicating each hidden unit N times []. Let us denote $h_j \in \{0, 1, ..., N\}$ as a binomial hidden unit and each h_j is represented by a group of binary hidden unit $h_j^{(1)}$, $h_j^{(2)}$, ..., $h_j^{(N)}$. The probability of activating a unit in this group is p_j where:

$$p_j = \sigma(\mathbf{x}^\top W + b_j) \tag{1}$$

Because the weights are shared between N replicas of hidden unit j the probability of n units are activated is:

$$p(h_j = n|\mathbf{x}) = \binom{N}{n} p_j^n (1 - p_j)^{(N-n)} = Bi(h_j = n, N, p_j) \tag{2}$$

So, one can say the group of N shared-weight hidden units represent the binomial distribution given the state of the other layer.

Let us consider the case of discriminative learning where the label is encoded as one hot vector $\mathbf{y}$, as in Figure **??**. The energy function will look like:

$$\begin{aligned} E(\mathbf{x}, y, \mathbf{h}) &= -\sum_{ijk} x_i w_{ij} h_j^{(k)} - \sum_{jk} u_{yj} h_j^{(k)} - \sum_i x_i a_i - b_y - \sum_{jk} c_j h_j^{(k)} \\ &= -\sum_j \sum_k h_j^{(k)} (\sum_i x_i w_{ij} - u_{yj} - c_j) - \sum_i x_i a_i - b_y \end{aligned} \tag{3}$$

For classification, learning is carried out by maximising a hybrid log-likelihood which combine the generative and discriminative functions:

$$\mathcal{L} = \mathcal{L}_{discriminative} + \alpha \times \mathcal{L}_{generative} \tag{4}$$

3.2. Generative Function

$$\mathcal{L}_{generative} = \frac{1}{N} \sum_n^N \log p(x^{(n)}, y^{(n)}) \tag{5}$$

3.3. Discriminative Function

In this paper, we are interested in the conditional function which is important for classification:

$$P(y|\mathbf{x}) = \frac{\sum_{\mathbf{h}} \exp\left(-E\left(\mathbf{x}, \mathbf{y}, \mathbf{h}\right)\right)}{\sum_{\mathbf{y}^*} \sum_{\mathbf{h}} \exp\left(-E\left(\mathbf{x}, \mathbf{y}^*, \mathbf{h}\right)\right)} \tag{6}$$

The denominator sums over all class-labels $\mathbf{y}^*$ to make $P(\mathbf{y}|\mathbf{x})$ a probability distribution. In the original RBM, $\mathbf{x}$ and $\mathbf{y}$ together make up the visible layer. The model is learned discriminatively by maximizing the log-likelihood function based on the expression of the conditional distribution above. Normally, such

RBMs have binary states $\{0,1\}$ for the hidden units. We will show how to extend the conditional distribution with different type of hidden units.

If an RBM whose hidden units have K states $\{s_k|k=1:K, K\in\mathbb{Z}\}$ then its conditional distribution in (6) can be computed analytically. We are interested in the conditional distribution

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x},y)}{\sum_{y^*} p(\mathbf{x},y^*)} \tag{7}$$

where

$$p(\mathbf{x},y) = \frac{1}{Z}\sum_{\mathbf{h}} \exp(-E(\mathbf{x},y,\mathbf{h}) \tag{8}$$

Apply (8) to (7) the partitional function Z will be cancelled out such that:

$$p(y|\mathbf{x}) = \frac{\sum_{\mathbf{h}} \exp(-E(\mathbf{x},y,\mathbf{h}))}{\sum_{y^*}\sum_{\mathbf{h}} \exp(-E(\mathbf{x},y^*,\mathbf{h}))} \tag{9}$$

In order to compute the conditional function in (9) we need to find $\sum_{\mathbf{h}} \exp(-E(\mathbf{x},y,\mathbf{h}))$. Here,

$$\sum_{\mathbf{h}} \exp(-E(\mathbf{x},y,\mathbf{h})) = \exp(\sum_{h_1=0}^{N} ... \sum_{h_J=0}^{N} (\sum_j \sum_k h_j^{(k)}(\sum_i x_i w_{ij} + u_{yj} + c_j) + \sum_i x_i a_i + b_y)) \tag{10}$$

Let us consider the term:

$$\sum_{\mathbf{h}} \exp(-E(\mathbf{x},y,\mathbf{h})) = \exp(\sum_{h_1=0}^{N} ... \sum_{h_J=0}^{N} (\sum_j \sum_k h_j^{(k)}(\sum_i x_i w_{ij} + u_{yj} + c_j)) \tag{11}$$

Note that if $h_j = n$ then there will be n units in the group $h_j^{(1)}, .., h_j^{(N)}$ activated and therefore

$$\sum_k h_j^{(k)}(\sum_i x_i w_{ij} + u_{yj} + c_j) = n\alpha_j \tag{12}$$

where $\alpha_j = (\sum_i x_i w_{ij} + u_{yj} + c_j)$

$$\begin{aligned}
\sum_{\mathbf{h}} \exp(-E(\mathbf{x}, y, \mathbf{h})) &= \exp(\sum_{h_1=0}^{N} \sum_{h_2=0}^{N} ... \sum_{h_J=0}^{N} \sum_{j} h_j \alpha_j) \\
&= \prod_{h_1=0}^{N} \prod_{h_2=0}^{N} ... \prod_{h_J=0}^{N} \prod_{j} \exp(h_j \alpha_j)) \\
&= \prod_{j} \sum_{h_j=0}^{N} \exp(h_j \alpha_j)
\end{aligned} \tag{13}$$

Now we denote s_k as a state of the hidden unit such that $s_0 = 0$, $s_1 = 1$, ... the product $\prod_j \sum_{h_j=0}^{N} \exp(h_j \alpha_j)$ can replaced by $\prod_j \sum_{k=0}^{N} \exp(s_k \alpha_j)$.

where s_k is each of the k states that can be assumed by each hidden unit j of the model. The last step of (**??**) results from re-arranging the terms after expanding the summation and product over $\mathbf{h}$ and j in the previous step respectively. The summation $\sum_{\mathbf{h}}$ over all the possible hidden layer vectors $\mathbf{h}$ can be replaced by the summation $\sum_k$ over the states of the units in the layer. The number and values of these states depend on the nature of the distribution in question. The result in (**??**) can be applied to (**??**) and, in turn, to (6) to get the following general expression of the conditional probability $P(y|\mathbf{x})$:

$$P(y|\mathbf{x}) = \frac{\exp(b_y) \prod_j \sum_k \exp(s_k \alpha_j)}{\sum_{y^*} \exp(b_{y^*}) \prod_j \sum_k \exp\left(s_k \alpha_j^*\right)}$$

4. Model Instances

4.1. DRBM

The $\{0, 1\}$-Bernoulli DRBM corresponds to the model originally introduced in [3]. In this case, each hidden unit h_j can either be a 0 or a 1, i.e. $s_k = \{0, 1\}$. This reduces $P(y|\mathbf{x})$ in (**??**) to

$$P_{\text{ber}}(y|\mathbf{x}) = \frac{\exp(b_y) \prod_j (1 + \exp(\alpha_j))}{\sum_{y^*} \exp(b_{y^*}) \prod_j \left(1 + \exp\left(\alpha_j^*\right)\right)} \tag{14}$$

which is identical to the result obtained in [3].

4.2. Bipolar DRBM:

A straightforward adaptation to the DRBM involves replacing its hidden layer states by $\{-1, +1\}$ as previously done in [4] in the case of the RBM. This is straightforward because in both cases the hidden states of the models are governed by the Bernoulli distribution, however, in the latter case each hidden unit h_j can either be a -1 or a $+1$, i.e. $s_k = \{-1, +1\}$. Applying this property to (**??**) results in the following expression for $P(y|\mathbf{x})$:

$$P_{\text{bip}}(y|\mathbf{x}) = \frac{\exp(b_y)\prod_j(\exp(-\alpha_j) + \exp(\alpha_j))}{\sum_{y^*}\exp(b_{y^*})\prod_j\left(\exp\left(-\alpha_j^*\right) + \exp\left(\alpha_j^*\right)\right)} \tag{15}$$

4.3. Binomial DRBM:

It was demonstrated in [6] how groups of N (where N is a positive integer greater than 1) stochastic units of the standard RBM can be combined in order to approximate discrete-valued functions in its visible layer and hidden layers to increase its representational power. This is done by replicating each unit of one layer N times and keeping the weights of all connections to each of these units from a given unit in the other layer identical. The key advantage for adopting this approach was that the learning algorithm remained unchanged. The number of these "replicas" of the same unit whose values are simultaneously 1 determines the effective integer value (in the range $[0, N]$) of the composite unit, thus allowing it to assume multiple values. The resulting model was referred to there as the Rate-Coded RBM (RBMrate).

The intuition behind this idea can be extended to the DRBM by allowing the states s_k of each hidden unit to assume integer values in the range $[0, N]$. The summation in (**??**) would then be $S_N = \sum_{s_k=0}^{N}\exp(s_k\alpha_j)$, which simplifies as below:

$$S_N = \sum_{s_k=0}^{N}\exp(s_k\alpha_j) = \frac{1-\exp((N+1)\alpha_j)}{1-\exp(\alpha_j)} \tag{16}$$

in (**??**) to give

$$P_{\text{bin}}(y|\mathbf{x}) = \frac{\exp(b_y)\prod_j \frac{1-\exp((N+1)\alpha_j)}{1-\exp(\alpha_j)}}{\sum_{y^*}\exp(b_{y^*})\prod_j \frac{1-\exp\left((N+1)\alpha_j^*\right)}{1-\exp\left(\alpha_j^*\right)}} \ . \tag{17}$$

5. Experiments

5.1. Methodology

5.2. MNIST handwritten digit recognition

5.3. USPS handwritten digit recognition

5.4. 20 newsgroup document classification

6. Conclusions

References

[1] G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 15271554. `doi:10.1162/neco.2006.18.7.1527`.
URL `http://dx.doi.org/10.1162/neco.2006.18.7.1527`

[2] A.-R. Mohamed, G. Dahl, G. Hinton, Acoustic Modeling using Deep Belief Networks, IEEE Transactions on Audio, Speech, and Language Processing 20 (1) (2012) 14–22.

[3] H. Larochelle, Y. Bengio, Classification using discriminative restricted Boltzmann machines, in: International Conference on Machine Learning, ACM Press, 2008, pp. 536–543.

[4] Y. Freund, D. Haussler, Unsupervised Learning of Distributions on Binary Vectors using Two Layer Networks, in: Advances in Neural Information Processing Systems, 1992, pp. 912–919.

[5] M. Welling, M. Rosen-Zvi, G. Hinton, Exponential Family Harmoniums with an Application to Information Retrieval, in: Advances in Neural Information Processing Systems, 2004, pp. 1481–1488.

[6] Y. W. Teh, G. Hinton, Rate-Coded Restricted Boltzmann Machines for Face Recognition, Advances in Neural Information Processing Systems (2001) 908–914.

[7] V. Nair, G. Hinton, Rectified Linear Units Improve Restricted Boltzmann Machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.

[8] P. Smolensky, Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1, MIT Press, 1986, Ch. Information Processing in Dynamical Systems: Foundations of Harmony Theory, pp. 194–281.