Are Restricted Boltzmann Machines Universal?

David J.C. MacKay
Cavendish Laboratory, Cambridge, CB3 0HE.
mackay@mrao.cam.ac.uk

Marcus Frean
Cavendish Laboratory, Cambridge, CB3 0HE.
marcus@mrao.cam.ac.uk

Philip Sterne
Cavendish Laboratory, Cambridge, CB3 0HE.
pjs67@cam.ac.uk

April 1, 2007 – Draft 1.0

Abstract

This is an incomplete research note.

1 Restricted Boltzmann Machines

Hinton's deep belief networks have shown exciting capabilities as generative models of handwritten digits, images, and human gait. Restricted Boltzmann Machines are the main component used in building up these multilayer networks. A Restricted Boltzmann Machine is a two-layer Boltzmann machine where connections exist between all units in different layers, but there are no within-layer connections. We'll call the visible layer state $\mathbf{x} = \{x_i\}_{i=1}^{I}$ and the hidden layer state $\mathbf{h} = \{h_j\}_{j=1}^{J}$. The units are all binary, and we'll take the states to be ± 1. The weight between hidden unit j and visible unit i is called w_{ji}. We will assume that each layer includes a single unit (x_0 or h_0 respectively) which is defined to be pinned to state $+1$ at all times. The bias of visible unit i is w_{0i}. The bias of hidden unit j is w_{j0}.

The aim of this note is to better understand the capabilities of restricted Boltzmann machines. Are restricted Boltzmann machines universal? That is, given a sufficiently large number J of hidden units, can the distribution $P(\mathbf{x})$ over the visible units be made to

approximate and desired distribution $P_D(\mathbf{x})$ arbitrarily closely, or are there classes of distribution that can't be captured?

We wouldn't want our interest in universality to be misunderstood. People often view a proof that a model is universal as justifying the use of that model – 'it can learn anything!' – but we do not share that view. Rather, we are interested in understanding what probability distributions a model is naturally matched to; to describe this implicit distribution over distributions, the first step is to work out what domain of distributions the model can capture.

If restricted Boltzmann machines are not universal, we would like to better understand what happens in their hidden layers when we attempt to learn a distribution not in the class of accessible distributions. Do the hidden acitivities preserve information content that subsequent layers in a deep belief network can exploit?

2 Free energy viewpoint

We write the probability distribution

$$P(\mathbf{x} \,|\, \mathbf{W}) = \frac{1}{Z(\mathbf{W})} \sum_{\mathbf{h}} \exp(\mathbf{h}^{\mathsf{T}} \mathbf{W} \mathbf{x}) \tag{1}$$

in terms of a visible-state-dependent free energy, $F(\mathbf{x}; \mathbf{W})$, defined by

$$P(\mathbf{x} \,|\, \mathbf{W}) = \frac{1}{Z(\mathbf{W})} \exp(-F(\mathbf{x}; \mathbf{W})), \tag{2}$$

that is,

$$F(\mathbf{x}; \mathbf{W}) = -\ln \sum_{\mathbf{h}} \exp(\mathbf{h}^{\mathsf{T}} \mathbf{W} \mathbf{x}). \tag{3}$$

The task of matching any desired distribution $P_D(\mathbf{x})$ with $P(\mathbf{x} \,|\, \mathbf{W})$ is equivalent to the task of matching any function $F_D(\mathbf{x})$ with $F(\mathbf{x}; \mathbf{W})$.

The factor in the sum in (3) is separable, so we can write:

$$F(\mathbf{x}; \mathbf{W}) = -\ln \sum_{\mathbf{h}} \exp\left(\sum_{i=1}^{I} w_{0i} x_i \right) \prod_{j=1}^{J} \left[\exp\left(h_j \sum_{i=0}^{I} w_{ji} x_i \right) \right] \tag{4}$$

$$F(\mathbf{x}; \mathbf{W}) = -\sum_{i=1}^{I} w_{0i} x_i - \sum_{j=1}^{J} \ln \cosh \left[\sum_{i=0}^{I} w_{ji} x_i \right] + \text{const.} \tag{5}$$

$[e^x + e^{-x} = 2 \cosh x;\ \ln(e^x + e^{-x}) = \ln \cosh x + \ln 2.]$

Thus the question about universality of the restricted Boltzmann machine is the same as the question:

Is the function $F(\mathbf{x};\mathbf{W}) = -\sum_{i=1}^{I} w_{0i}x_i - \sum_{j=1}^{J} f\left[\sum_{i=0}^{I} w_{ji}x_i\right]$ universal (within an additive constant), where $f \equiv -\ln\cosh$?

This is similar to the question for three-layer feedforward networks,

Is the function $y(\mathbf{x};\mathbf{W},\omega) = \theta + \sum_{j=1}^{J} \omega_j f\left[\sum_{i=0}^{I} w_{ji}x_i\right]$ universal, where $f \equiv \tanh$?

to which the answer is 'yes', so we might sniff the scent of a universality result in the offing; but there are a few differences.

1. The standard three-layer network includes output-weights ω which weight the nonlinear hidden-layer functions f, and which are permitted to have any sign. In (5), in contrast, the equivalent weights are all fixed to $+1$.
2. The standard nonlinear functions f for neural networks are logistic functions with both concave and convex regions. Here $f \equiv -\ln\cosh$ is a strictly concave function.
3. Whereas, with three-layer feedforward networks, the visible variables $\mathbf{x}$ can take on any real values, here they take on only binary values.

3 Conclusion

So, what's the answer? I don't know yet.

Acknowledgments

We thank Geoff Hinton for helpful discussions. Marcus Frean's sabbatical visit to Cambridge is supported by the Gatsby Charitable Foundation. DJCM is supported by the Gatsby Charitable Foundation.