Notes on Boltzmann Machines and Contrastive Divergence Process

Nils Hallerfelt

October 2022

Boltzmann Machines

A Boltzmann machine is an energy based model consisting of N nodes connected to each other, where each node can take a binary value. We denote each nodes random variable as X_i, for $i \in [1, N]$. Thus the state of the Boltzmann machine can be expressed by the random vector $\mathbf{X}$. The network has parameters $\theta = (\mathbf{W}, \mathbf{b})$, where $\mathbf{W}$ are the weights and $\mathbf{b}$ are biases. As mentioned the Boltzmann machine is an energy based model, where the energy function is defined by:

$$E_\theta = -\mathbf{b}^T\mathbf{x} - \mathbf{x}^T\mathbf{W}\mathbf{x} \tag{1}$$

From the energy function, the Boltzmann machines defines a probability distribution over the binary states:

$$p_\theta(\mathbf{x}) = \frac{\exp -E_\theta(\mathbf{x})}{\sum_{\mathbf{x}'} \exp -E_\theta(\mathbf{x}')} = \exp -E_\theta(\mathbf{x}) - \log Z(\theta) \tag{2}$$

where the sum is over all possible states and Z is the partition function. This shows the entropy interpretation that the idea of a Boltzmann machine is built on: the higher the energy is of a state $\mathbf{x}$, the less likely it is to occur.

Introducing Latent Nodes

A Boltzmann machine with N visible units have $\frac{N(N+1)}{2}$ parameters. This model is used to model N-bit binary patterns, that is 2^N possible states. Thus the number of parameters of the the Boltzmann machine is significantly smaller than the number of parameters needed to model a distribution of 2^N possible states. One way to make the model more expressive is to introduce extra latent nodes to the model: we will denote these nodes by $\mathbf{h}$ and say that we have M hidden units.

Introducing Restrictions and Real Valued Inputs

In order to represent real valued inputs, i.e. $\mathbf{x} \in \mathbb{R}^N$, while having binary hidden nodes $\mathbf{h} \in [0,1]^M$, we can define the energy function in another way which also restricts the Boltzmann machine:

$$E_\theta(\mathbf{x}, \mathbf{h}) = \sum_{i=1}^{2} \frac{(x_i - b_i^v)^2}{2\sigma^2} - \sum_{i=1}^{2}\sum_{j=1}^{M} \frac{x_i w_{i,j} y_j}{\sigma^2} - \sum_{j=1}^{M} y_j b_j^h. \tag{3}$$

Where we have chosen $N = 2$. By defining the energy function in this way we can show that the (visible) input nodes are independent between each other and Gaussian given any state of the hidden nodes:

$$p(\mathbf{x}|\mathbf{h}) = p(x_1|\mathbf{h})p(x_2|\mathbf{h}) \quad p(x_i|\mathbf{h}) = \mathcal{N}(b_i^v + \sum_{j=1}^{M} w_{i,j} h_j, \sigma_i^2). \tag{4}$$

Similarly the conditional of $\mathbf{h}$ given $\mathbf{x}$ are internally independent, but are Bernoulli instead of Gaussian:

$$p(\mathbf{h}|\mathbf{x}) = \prod_{j=1}^{M} p(h_j|\mathbf{x}) \quad p(h_i|\mathbf{x}) = \mathcal{B}(\text{sigmoid}(b_j^h + \sum_{i=1}^{2} \frac{x_i w_{i,j}}{\sigma_i^2})) \tag{5}$$

Gaussian-Bernoulli RBM

By equation (3), (4), and (5) we have defined a Gaussian-Bernoulli RBM with the joint probability distribution:

$$p_\theta(\mathbf{x}, \mathbf{h}) = \frac{\exp -E_\theta(\mathbf{x}, \mathbf{h})}{\sum_{\mathbf{x}', \mathbf{h}'} \exp -E_\theta(\mathbf{x}', \mathbf{h}')} = \exp\left(-E_\theta(\mathbf{x}, \mathbf{h}) - \log Z(\theta)\right) \tag{6}$$

As we will want to maximize the log-probability of the input, we marginalize over the latent variables.

$$p_\theta(\mathbf{x}) = \frac{\exp -c_\theta(\mathbf{x})}{\int_{\mathbf{x}'} \exp -c_\theta(\mathbf{x}')}, \quad c_\theta(\mathbf{x}) = \sum_{j=1}^{M} \log[1 + \exp(b_j^h + \sum_{i=1}^{2} \frac{x_i w_{i,j}}{\sigma_i^2})] - \sum_{i=1}^{2} \frac{(x_i - b_i^v)^2}{2\sigma^2} \tag{7}$$

where $c_\theta(\mathbf{x})$ is the log partition function of $p_\theta(\mathbf{h}|\mathbf{x})$.

Training

In order to train the model we define the loss function as:

$$\mathcal{L}(\theta, \mathbf{x}) = -\log p(\mathbf{x}). \tag{8}$$

the gradient with respect to the parameters θ is defined as:

$$\nabla_\theta \mathcal{L}(\theta, \mathbf{x}) = \nabla_\theta c_\theta(\mathbf{x}) - \mathbb{E}_{x' \sim p_\theta(\mathbf{x})}[\nabla_\theta c_\theta(\mathbf{x}')] \tag{9}$$

As the expectation term is intractable so we need to use some estimation technique. As both conditional distributions are easy to sample from, it can be estimated taking samples from the distribution via Gibbs MCMC sampling. However, we will instead use contrastive divergence where only one step in the Gibbs process is taken.

Contrastive Divergence

We can rewrite the objection function of maximizing the log probability as a minimization of the KL-divergence between a target distribution and a parameterized distribution.

$$D_{KL}(p_{\text{target}}||p_\theta) = \sum_{\mathbf{x}'} p_{\text{target}} \log p_{\text{target}} - \sum_{\mathbf{x}'} p_{\text{target}} \log p_\theta \tag{10}$$

As the first term does not depend on θ we can overlook it, and then maximize the negated second term, having it as the objective:

$$f(\theta) = \sum_{\mathbf{x}'} p_{\text{target}} \log p_\theta \tag{11}$$

As we saw earlier the gradient of the $\log p_\theta(\mathbf{x})$ has a term that is intractable. So we remove as by the following.

Consider a Gibbs-sampler that is initialized by a point from the data, this will define the target distribution $p_{\text{target}} = p_0$ by definition. In terms of KL-divergence this results in the following:

$$D_{KL}(p_0||p_\theta) = \sum_{\mathbf{x}'} p_0(\mathbf{x}') \log p_0(\mathbf{x}') - \sum_{\mathbf{x}'} p_0(\mathbf{x}') \log p_\theta(\mathbf{x}') \tag{12}$$

As the Gibbs sampler progresses, we denote the k:th step as p_k^θ, and by definition p_k^θ will converge to p_θ as $k \to \inf$. Thus we have that:

$$p_{\inf}^\theta(\mathbf{x}) = \frac{\exp c_\theta(\mathbf{x})}{\int_{\mathbf{x}'} \exp c_\theta(\mathbf{x}'}$$

The gradient of (12) can be derived to be:

$$\sum_{x'} p_0(\mathbf{x}') \nabla c_\theta(\mathbf{x}') - \sum_{x'} p_{\inf}^\theta(\mathbf{x}') \nabla c_\theta(\mathbf{x}') \tag{13}$$

To get rid of the intractable term, we can proceed one step in the Gibbs sampling, getting p_1^θ, and define the contrastive divergence as:

$$CD = D_{KL}(p_0||p_{\text{inf}}^\theta) - D_{KL}(p_1^\theta||p_{\text{inf}}^\theta) \tag{14}$$

The gradient of the CD term is:

$$\nabla_\theta CD(\theta) = \sum_{x'} p_0(\mathbf{x}')\nabla c_\theta(\mathbf{x}') - \sum_{x'} p_1^\theta(\mathbf{x}')\nabla c_\theta(\mathbf{x}') - \epsilon(\theta), \quad \epsilon(\theta) \approx 0 \tag{15}$$

Since we are having a RBM it the log partition c_θ is tractable, and thus both these terms are tractable, where the second term is using a single step in Gibbs sampling algorithm.

Notes

During these notes I have used the paper "Boltzmann machines and energy-based models" by Takayuki Osogami extensively.