Complex RBMs

Andi Gu

November 2020

1 Classical RBMs

1.1 Background

Classical restricted Boltzmann machines (RBM) are graphical models that encode a probability distribution $P(\mathbf{v})$ over binary strings $\{0,1\}^n$. They are a bipartite graph structure, with a number $n_v = n$ of 'visible' units and a number n_h of 'hidden' units. Each layer has an associated bias vector that acts as a local magnetic field 'biasing' each layer towards a particular configuration: for the visible, it is some $a \in \mathbb{R}^{n_h}$ and for the hidden it is $b \in \mathbb{R}^{n_v}$. The two layers of the graph also interact with some interaction term $W \in \mathbb{R}^{n_v \times n_h}$. Formally, there is an energy associated with any given configuration (some setting of the visible $\mathbf{v}$ and hidden $\mathbf{h}$ units) of the RBM:

$$E(\mathbf{v}, \mathbf{h} \mid W, a, b) = -a^T \mathbf{v} - b^T \mathbf{h} - \mathbf{v}^T W \mathbf{h} \tag{1}$$

For brevity, we write energy as just $E(\mathbf{v}, \mathbf{h})$, and the dependence upon W, a, b is implicitly assumed. Then, the probability of any given configuration is simply the Boltzmann term (normalized appropriately) associated with this energy:

$$P(\mathbf{v}, \mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z} \tag{2}$$

$$Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \tag{3}$$

The conditional probabilities for each of the individual visible units, in terms of the logistic function $\sigma(x) \equiv \frac{1}{1+e^{-x}}$, is:

$$P(\mathbf{v}_i = 1 \mid \mathbf{h}) = \sigma(a_i + (W\mathbf{h})_i) \tag{4}$$

$$P(\mathbf{h}_j = 1 \mid \mathbf{v}) = \sigma\big(b_j + (W^T\mathbf{v})_j\big) \tag{5}$$

These relations make clear what the effect of the bias vectors are: a positive a_i increases the probability that $\mathbf{v}_i$ will be 1 (and likewise for the hidden units).

1.2 Sampling

The closed form marginal $P(\mathbf{v})$ is generally intractable for an RBM. In practice, we find it with Monte Carlo Markov Chain (MCMC) sampling. Since we have the conditional distributions, it is

practical to use Gibbs sampling. We start with some random vector $\mathbf{v}^{(0)} \in \{0,1\}^{n_v}$, and then take a sample $\mathbf{h}^{(0)}$ from the hidden units, using $\mathbf{v}^{(0)}$ as the given according to (5). We then generate another sample $\mathbf{v}^{(1)}$ using $\mathbf{h}^{(0)}$ as the given, according to (4). Typically, under 5 iterations of this process is enough to reach stationarity.

1.3 Training

Given some dataset of N, v-length binary strings $\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\right\}$, the goal is simply to find W, a, b that maximizes the probability of this dataset. We formulate this in terms of log-likelihood:

$$\max_{W,a,b} \sum_{i=1}^{n} \log P(\mathbf{x}^{(i)}; W, a, b) \tag{6}$$

We define the free energy of some visible vector to be $\mathcal{F}(\mathbf{v}) \equiv -\log \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} = -\log(Z \cdot P(\mathbf{v}))$. Luckily, we can find a closed form for $\mathcal{F}(\mathbf{v})$:

$$\mathcal{F}(\mathbf{v}) = \log e^{a^T \mathbf{v}} \sum_{\mathbf{h}} e^{(b+W^T\mathbf{v})^T \mathbf{h}}$$

This sum can be split nicely, since every entry of $\mathbf{h}$ is binary.

$$\begin{aligned} &= -a^T\mathbf{v} - \log \prod_{j=1}^{n_h} (1 + e^{(b+W^T\mathbf{v})_j}) \\ &= -a^T\mathbf{v} - \sum_{j=1}^{n_h} \log\left(1 + e^{(b+W^T\mathbf{v})_j}\right) \end{aligned}$$

Then the objective can be reformulated as:

$$\max_{W,a,b} \sum_{i=1}^{n} -\mathcal{F}(\mathbf{x}^{(i)}) - \log Z \tag{7}$$

$\log Z$ is still intractable – however, its gradient can be approximated:

$$\begin{aligned} \boldsymbol{\nabla}(\log Z) &= \frac{\boldsymbol{\nabla} Z}{Z} \\ &= \frac{\sum_{\mathbf{v}} \boldsymbol{\nabla} e^{-\mathcal{F}(\mathbf{v})}}{Z} \\ &= -\frac{\sum_{\mathbf{v}} e^{-\mathcal{F}(\mathbf{v})} \boldsymbol{\nabla} \mathcal{F}(\mathbf{v})}{Z} \end{aligned}$$

Note that $P(\mathbf{v}) = \frac{e^{-\mathcal{F}(\mathbf{v})}}{Z}$.

$$\begin{aligned} &= -\sum_{\mathbf{v}} P(\mathbf{v}) \boldsymbol{\nabla} \mathcal{F}(\mathbf{v}) \\ &= -\mathbb{E}(\boldsymbol{\nabla} \mathcal{F}(\mathbf{v})) \end{aligned}$$

That is, the gradient $\boldsymbol{\nabla} \log Z$ is simply the expectation of the gradient of free energy – this we can find, because we can approximate the expectation by Gibbs sampling.

2 Complex Generalizations

We again parameterize our RBM with weights $a \in \mathbb{C}^N$, $b \in \mathbb{C}^{n_h}$, and $W \in \mathbb{C}^{N \times n_h}$. We allow the hidden units to be binary strings $\mathbf{h} \in \{0,1\}^{n_h}$ and the visible units now represent the projection of the wavefunction onto the z-basis: $\mathbf{v}_j = \sigma_j^z$. Let us label the basis functions in the z-basis $\mathbf{v}_1 = |000\ldots0\rangle$, $\mathbf{v}_2 = |000\ldots1\rangle$, $\mathbf{v}_{2^N} = |111\ldots1\rangle$. The wavefunction encoded explicitly by the RBM is $\Psi_{RBM} : \{0,1\}^N \to \mathbb{C}$.

$$\begin{aligned}\Psi_{RBM}(\mathbf{v}) &= \sum_{\mathbf{h}\in\{0,1\}^{n_h}} e^{a^H\mathbf{v} + b^H\mathbf{h} + \mathbf{v}^H W\mathbf{h}} \\ &= e^{a^H\mathbf{v}} \sum_{\mathbf{h}} e^{(b+W^H\mathbf{v})^H\mathbf{h}}\end{aligned}$$

Again, this factors:

$$= e^{a^H\mathbf{v}} \prod_{i=1}^{n_h} \left(e^{(b+W^H\mathbf{v})_i^*} + 1\right) \tag{8}$$

Define:

$$\mathcal{F}(\mathbf{v}) \equiv -\log \Psi_{RBM}(\mathbf{v}) = -a^H\mathbf{v} - \sum_{i=1}^{n_h} \log\left(e^{(b+W^H\mathbf{v})_i^*} + 1\right) \tag{9}$$

We can extend the domain of the wavefunction beyond binary strings to $\tilde{\Psi} : \mathbb{C}^{2^N} \to \mathbb{C}$:

$$\tilde{\Psi}(\boldsymbol{\sigma}) = \sum_{i=1}^{2^N} \langle \mathbf{v}_i|\boldsymbol{\sigma}\rangle\, \Psi_{RBM}(\mathbf{v}_i) \tag{10}$$

$$P(\boldsymbol{\sigma}) = \frac{\left|\tilde{\Psi}_{RBM}(\boldsymbol{\sigma})\right|^2}{\sum_{i=1}^{2^N} \left|\Psi_{RBM}(\mathbf{v}_i)\right|^2} \tag{11}$$

For shorthand, define $\vec{\boldsymbol{\Psi}}$ to be a vector that is the result of evlauating Ψ_{RBM} on every basis vector. We define $\vec{\boldsymbol{\sigma}}$ to be the vector representation of $|\sigma\rangle$ in the z-basis.

$$\vec{\boldsymbol{\Psi}} = \begin{bmatrix} \Psi_{RBM}(\mathbf{v}_1) \\ \Psi_{RBM}(\mathbf{v}_2) \\ \vdots \\ \Psi_{RBM}(\mathbf{v}_{2^N}) \end{bmatrix} \tag{12}$$

$$\vec{\boldsymbol{\sigma}} = \begin{bmatrix} \langle \mathbf{v}_1|\boldsymbol{\sigma}\rangle \\ \langle \mathbf{v}_2|\boldsymbol{\sigma}\rangle \\ \vdots \\ \langle \mathbf{v}_{2^N}|\boldsymbol{\sigma}\rangle \end{bmatrix} \tag{13}$$

Then, the general wavefunctions and probabilities reduce to:

$$\begin{aligned}\tilde{\Psi}(\boldsymbol{\sigma}) &= \vec{\boldsymbol{\sigma}}^T \vec{\boldsymbol{\Psi}} \\ P(\boldsymbol{\sigma}) &= \frac{\left|\vec{\boldsymbol{\sigma}}^T \vec{\boldsymbol{\Psi}}\right|^2}{\left\|\vec{\boldsymbol{\Psi}}\right\|^2} \\ &= \frac{\vec{\boldsymbol{\Psi}}^H \vec{\boldsymbol{\sigma}}^* \vec{\boldsymbol{\sigma}}^T \vec{\boldsymbol{\Psi}}}{\left\|\vec{\boldsymbol{\Psi}}\right\|^2}\end{aligned}$$