RESTRICTED BOLZMANN MACHINES WITH TWO VISIBLE LAYERS

Contents

This is a summary of the mathematical framework used in Restricted Boltzmann Machines. For the details regarding the two-layer symmetric model for pair association skip to Section 2.

1. Summary of Restricted Boltzmann Machines

Classical RBMs are used to approximate the probability distribution of a random variable V with values in (a subset of) $\mathbb{R}^p$, given a number of samples drawn from it. They do so by assuming that the law of V is in fact the marginal of a pair of random variables $(V, H) \in \mathbb{R}^p \times \mathbb{R}^q$, whose distribution has the form

$$dp(v, h) = \frac{1}{Z} e^{-E(v,h)} d\mu(v) d\nu(h)$$

where μ and ν are measures on $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, E is an energy function of the form

$$-E(v, h) = \sum_{i=1}^{p} \sum_{j=1}^{q} v_i w_{ij} h_j + \sum_{i=1}^{p} a_i v_i + \sum_{j=1}^{q} b_j h_j,$$

and Z is a normalization constant

$$Z = \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} e^{-E(v,h)} d\mu(v) d\nu(h).$$

Hereafter we always take ν to be the discrete measure

$$d\nu(h) = \prod_{j=1}^{q} \frac{1}{2} (d\delta_0(h_j) + d\delta_1(h_j)).$$

The choice of measure μ depends on the problem at hand. Typical choices are:

Binary units,

$$d\mu(v) = \prod_{i=1}^{p} \frac{1}{2}(d\delta_0(v_i) + d\delta_1(v_i)),$$

Gaussian units,

$$d\mu(v) = \prod_{i=1}^{p} \frac{1}{\sqrt{4\pi}} e^{-v_i^2} dv_i = N(0, I_p),$$

Others?

In general, provided $d\mu(x)$ is a measure on $\mathbb{R}^p$ for which $e^{\alpha\|x\|}$ is integrable, this model allows for explicit formulae for the conditional probabilities of V given H, and viceversa.

1.1. **Visible units.** For $A_i \in \mathcal{B}(\mathbb{R})$ and $B \in \mathcal{B}(\mathbb{R}^q)$ we have

$$\begin{aligned}
\mathbb{P}(V_1 \in A_1, ..., V_p \in A_p, H \in B) =& \mathbb{P}(V \in A_1 \times ... \times A_p, H \in B) \\
=& \frac{1}{Z} \int_B e^{(b,h)} \int_{A_1 \times ... \times A_p} e^{\sum_{i=1}^p (Wh+a)_i v_i} d\mu(v) d\nu(h) \\
=& \frac{1}{Z} \int_B e^{(b,h)} \left[\int_{A_1 \times ... \times A_p} \prod_{i=1}^{p} e^{(Wh+a)_i v_i} d\mu(v) \right] d\nu(h) \\
=& \frac{1}{Z} \int_B e^{(b,h)} \left[\prod_{i=1}^{p} \int_{A_i} e^{(Wh+a)_i v_i} d\mu_i(v_i) \right] d\nu(h).
\end{aligned}$$

Then, for any $i \in \{1, ..., p\}$ and $A \in \mathcal{B}(\mathbb{R})$, letting

$$\rho_i(h, A) = \frac{\int_A e^{(Wh+a)_i t} d\mu_j(t)}{\int_{\mathbb{R}} e^{(Wh+a)_i t} d\mu_i(t)},$$

we have

$$\begin{aligned}
\mathbb{P}(V_j \in A, H \in B) =& \frac{1}{Z} \int_B e^{(h,b)} \rho_j(h, A) \left[\prod_{i=1}^{p} \int_{\mathbb{R}} e^{(Wh+a)_i v_i} d\mu_i(v_i) \right] d\nu(h) \\
=& \frac{1}{Z} \int_B e^{(h,b)} \rho_j(h, A) \int_{\mathbb{R}^p} e^{(Wh+a,v)} d\mu(v) d\nu(h) \\
=& \int_{\mathbb{R}^p \times \mathbb{R}^q} \rho_j(h, A) I(B) d\mathbb{P}(v, h)
\end{aligned}$$

On the left hand side we have $\mathbb{E}(I(V_j \in A)I(B))$, and on the right hand side we have $\mathbb{E}(\rho_j(H, A)I(B))$. Since B is arbitrary, and $\rho_j(H; A)$ is clearly $\sigma(H)$ measurable, we conclude that

$$\mathbb{P}(V_j \in A | \sigma(H)) = \rho_j(H, A) = \frac{\int_A e^{(WH+a)_j t} d\mu_j(t)}{\int_{\mathbb{R}} e^{(WH+a)_j t} d\mu_j(t)}.$$

This conditional distribution can be given explicitly if μ is known. Some examples:

1.1.1. *Binary case:* If the i^{th} visible feature is binary we take $\mu_i = \frac{1}{2}(\delta_0 + \delta_1)$, so that

$$\mathbb{P}(V_i = 1|\sigma(H)) = \frac{e^{(WH+a)_i}}{1+e^{(WH+a)_i}} = \sigma((WH+a)_i).$$

Here $\sigma : \mathbb{R} \to (0,1)$ is the sigmoid function,

$$\sigma(x) = \frac{e^x}{1+e^x}.$$

Thus

$$V_i|H \sim \text{Bernoulli}(\sigma((WH+a)_i)).$$

1.1.2. *Gaussian case:* If the j^{th} visible feature is gaussian, say $\mu_j = N(m_j, \sigma_j^2)$, then

$$\begin{aligned}\int_A e^{(WH+a)_j t} d\mu_j(t) =& \frac{1}{\sqrt{4\pi\sigma_j^2}} \int_A e^{(WH+a)_j t - \frac{1}{2\sigma_j^2}(t-m_j)^2} dt \\ =& \frac{1}{\sqrt{4\pi\sigma_j^2}} e^{m_j(WH+a)_j + \frac{\sigma_j^2}{2}(WH+a)_j^2} \int_A e^{-\frac{1}{2\sigma_j^2}(t-m_j-\sigma_j^2(WH+a)_j)^2} dt\end{aligned}$$

and thus, after some cancellations,

$$\mathbb{P}(V_j|\sigma(H)) = N(m_j + \sigma_j^2(WH+a)_j, \sigma_j^2) = m_j + \sigma_j(\sigma_j(WH+a)_j + N(0,1)).$$

If we further take $m_j = 0$ and $\sigma_j = 1$, namely $\mu_j = N(0,1)$, then we get the much simpler expression

$$V_j|H \sim (WH+a)_j + N(0,1).$$

1.2. **Hidden units.** Since the hidden units are always binary we have

$$\mathbb{P}(H_j = 1|\sigma(V)) = \frac{e^{(W^tV+b)_j}}{1+e^{(W^tV+b)_j}} = \sigma((W^tV+b)_j) = 1 - \mathbb{P}(H_j = 0|\sigma(V)),$$

namely,

$$H_j|V \sim \text{Bernoulli}(\sigma((W^tV+b)_j)).$$

1.3. **Training.** Given a training set of N independent samples $\{v_1, ..., v_N\}$, the average log-likelihood is given by

$$\begin{aligned}\ell =& \frac{1}{N}\sum_{k=1}^N \log\left(\frac{1}{Z}\sum_h e^{-E(v_k,h)}\right) \\ =& \frac{1}{N}\sum_{k=1}^N \log\left(\sum_h e^{-E(v_k,h)}\right) - \log(Z)\end{aligned}$$

It follows that

$$\begin{aligned}\frac{\partial \ell}{\partial \theta} =& \frac{1}{N}\sum_{k=1}^N \frac{1}{\sum_h e^{-E(v_k,h)}} \sum_h \left(-\frac{\partial E}{\partial \theta}(v_k,h)\right) e^{-E(v_k,h)} \\ & - \frac{1}{Z}\int_{\mathbb{R}^p} \sum_h \left(-\frac{\partial E}{\partial \theta}(v,h)\right) e^{-E(v,h)} d\mu(v) \\ =& \mathbb{E}_{V\sim\mathbb{P}_{\mathcal{T}}}\left[\mathbb{E}_{H|V}\left(-\frac{\partial E}{\partial \theta}(V,H)\right)\right] - \mathbb{E}\left(-\frac{\partial \mathcal{E}}{\partial \theta}(V,H)\right).\end{aligned}$$

Here θ is just a generic parameter, and $\mathbb{P}_{\mathcal{T}}$ is the empirical distribution of V, namely, $\mathbb{P}_{\mathcal{T}}(V = v) = \frac{1}{N}\sum_{k=1}^{N} I(v_k = v)$.

So we see that the first term above is an expectation where V is drawn from the sample distribution $\mathbb{P}_{\mathcal{T}}$ (given), and H is drawn from its conditional distribution given $V = v_k$ (which we know how to compute).

The second term, on the other hand, is for all practical purposes impossible to compute (number of terms in the sum grows exponentially with q).

1.4. **k-step Contrastive Divergence.** The most crucial insight at this point is that we don't really need to maximize the log-likelihood, after all this is a somewhat arbitrary utility function among infinitely many others that yield, in the limit of infinite number of samples, a correct estimate of the true conditional probability.

Thus, when updating the values of the parameters we may as well be satisfied with using a direction that somewhat approximates the gradient of the log-likelihood but does not necessarily coincide with it. This is what Contrastive Divergence does.

Recall that the usual gradient ascent method consists on updating a parameter θ according to the rule

$$\theta^{(n+1)} = \theta^{(n)} + \epsilon(\nabla_\theta \ell)(\theta^{(n)})$$

where ϵ is a given learning rate.

k-step Contrastive Divergence follows the same procedure, except that it replaces the hard-to-compute direction $(\nabla_\theta \ell)|_{\theta=\theta^{(n)}}$ above by the more tractable

$$d_k(\theta) = \mathbb{E}_{V\sim\mathbb{P}_{\mathcal{T}}}\left[\mathbb{E}_{H|V}\left(-\frac{\partial \mathcal{E}}{\partial\theta}(V,H)\right)\right] - \mathbb{E}_{(V,H)\in\mathbb{P}_k}\left(-\frac{\partial \mathcal{E}}{\partial\theta}(V,H)\right)$$

Here the first term is the same as before (sample V from the empirical distribution, and then sample H from the conditional distribution), but the second term is obtained by sampling k times starting from an empirical draw.

More explicitly, $\mathbb{P}_k$ is obtained by recursive samplings: we let V_0 be a sample from the empirical distribution, and sample recursively $H_i \sim H|(V = V_i)$, $V_{i+1} \sim V|(H = H_i)$, $i = 0, .., k$.

2. Modified RBM

Using the same logic of a standard RBM, we look now for a probability distribution $\mathbb{P}$ on *pairs* of visible vectors, where order doesn't matter.

This amounts to seek for an energy functional of the form (names here are temporary),

$$-E(v_1, v_2, h) = (W_{1h}h, v_1)+(W_{2h}h, v_2)+(W_{12}v_1, v_2)+(W_{21}v_2, v_1)+(a_1, v_1)+(a_2, v_2)+(h, b)$$

with the additional constraint

$$E(v_1, v_2, h) = E(v_2, v_1, h), \qquad \forall v_1, v_2 \in \mathbb{R}^p,\ \forall h \in \mathbb{R}^q.$$

Simple algebraic considerations imply that such symmetry constraints hold if and only if

$$W_{1h} = W_{2h}, \quad a_1 = a_2, \quad W_{12} + W_{21}^t = W_{12}^t + W_{21}.$$

So we can write

$$-E(v_1, v_2, h) = (W^0 h, v_1 + v_2) + (W^I v_1, v_2) + (a, v_1 + v_2) + (h, b)$$

where W^0 is a matrix of size $p \times q$ matrix, and W^I is a symmetric matrix of size $p \times p$.

2.1. **Conditional sampling.** It can be verified that the components of V_1 are conditionally independent given H and V_2, with

$$\mathbb{P}(V_{1j}|\sigma(H, V_2)) = \frac{\int_A e^{(W_0 H + W_I V_2 + a)_j t} d\mu_j(t)}{\int_{\mathbb{R}} e^{(W_0 H + W_I V_2 + a)_j t} d\mu_j(t)},$$

the components of V_2 are conditionally independent given H and V_1, with

$$\mathbb{P}(V_{2j}|\sigma(H, V_1)) = \frac{\int_A e^{(W_0 H + W_I V_1 + a)_j t} d\mu_j(t)}{\int_{\mathbb{R}} e^{(W_0 H + W_I V_1 + a)_j t} d\mu_j(t)}$$

and the components of H are conditionally independent given V_1 and V_2, with

$$\mathbb{P}(H_j = 1|\sigma(V_1, V_2)) = \frac{e^{(W_0^t(V_1+V_2)+b)_j}}{1 + e^{(W_0^t(V_1+V_2)+b)_j}} = \sigma((W_0^t(V_1 + V_2) + b)_j)$$

2.2. **Training.** As before, the log-likelihood takes the form

$$\begin{aligned}\ell =& \frac{1}{N}\sum_{k=1}^{N} \log\left(\frac{1}{Z}\sum_h e^{-E(v_1^k, v_2^k, h)}\right) \\ =& \frac{1}{N}\sum_{k=1}^{N} \log\left(\sum_h e^{-E(v_1^k, v_2^k, h)}\right) - \log Z\end{aligned}$$

and

$$\frac{\partial \ell}{\partial \theta} = \mathbb{E}_{(V_1,V_2)\sim\mathbb{P}_{\mathcal{T}}}\left[\mathbb{E}_{H|(V_1,V_2)}\left(-\frac{\partial E}{\partial \theta}(V_1, V_2, H)\right)\right] - \mathbb{E}\left(-\frac{\partial E}{\partial \theta}(V_1, V_2, H)\right)$$

Again the first term above can be computed explicitly by making use of the conditional distribution of H given the visible units, but the second term involves exponentially many terms so it is computationally intractable.

2.3. **Generalizing Contrastive Divergence two the bivariate case.** As in the usual n-step Contrastive Divergence algorithm, we approximate the model expectation by a series of conditional samplings.

(1) Set (V_1, V_2) to a data point and $k = 0$.
(2) Sample $H|(V_1, V_2)$:

$$H_j|(V_1, V_2) \sim \text{Bernoulli}(\sigma(((W^0)^t(V_1 + V_2) + b)_j)), \qquad j = 1, ..., q.$$

(3) Using V_1 and H, sample $V_2'|(V_1, H)$.
In the binary case:

$$V_{2i}' \sim \text{Bernoulli}(\sigma((W^0 H + W^I V_1 + a)_i)), \qquad i = 1, ...p.$$

In the Gaussian$_\sigma$ case:

$$V_{2i}' \sim \sigma^2(W^0 H + W^I V_1 + a)_i + \sigma N(0, 1), \qquad i = 1, ...p.$$

(4) Using H and V_2', sample $V_1'|(H, V_2)$.
In the binary case:

$$V_{1i}' \sim \text{Bernoulli}(\sigma((W^0 H + W^I V_2' + a)_i)), \qquad i = 1, ...p.$$

In the Gaussian$_\sigma$ case:

$$V_{1i}' \sim \sigma^2(W^0 H + W^I V_2' + a)_i + \sigma N(0, 1), \qquad i = 1, ...p.$$

(5) Using V_1' and V_2', sample $H'|(V_1', V_2')$:

$$H_j'|(V_1', V_2') \sim \text{Bernoulli}(\sigma(((W^0)^t(V_1' + V_2') + b)_j)), \qquad j = 1, ..., q.$$

(6) Increment $k = k + 1$.
(7) If $k < n$, repeat steps 1–6 with $(V_1, V_2) = (V_1', V_2')$.
(8) The CD_n approximation is the law of (V_1', V_2', H').

Some notation: V_{kij} is the i^{th} component of the j^{th} visible layer in the k^{th} sample, and V'_{kij} is the i^{th} component of the j^{th} visible layer obtained after the n steps above starting from the k^{th} sample.

Then the updates for each of the variable are:

$$\frac{\partial \ell}{\partial W_{ij}^0} \approx \frac{1}{N} \sum_{k=1}^{N} \left[(V_{ki1} + V_{ki2}) H_{kj} - (V'_{ki1} + V'_{ki2}) H'_{kj} \right]$$

$$\frac{\partial \ell}{\partial W_{ij}^I} \approx \frac{1}{N} \sum_{k=1}^{N} \frac{1}{2} \left[V_{ki1} V_{kj2} + V_{kj1} V_{ki2} - (V'_{ki1} \mathbb{E} V'_{kj2} + V'_{kj1} \mathbb{E} V'_{ki2}) \right]$$

$$\frac{\partial \ell}{\partial a_i} \approx \frac{1}{N} \sum_{k=1}^{N} (V_{ki1} + V_{ki2}) - (V'_{ki1} + V'_{ki2})$$

$$\frac{\partial \ell}{\partial b_j} \approx \frac{1}{N} \sum_{k=1}^{N} (H_{kj} - H'_{kj})$$

The way symmetry is enforced in the update for W_{ij}^I might not be the best. We should definitely consider other options.

3. Theoretical considerations

3.1. **The cost functional.** The Kullback-Leiber divergence for discrete probability distributions is given by

$$D_{KL}(P\|Q) = -\sum_z P(z) \log \frac{Q(z)}{P(z)}$$

Assuming that P and Q are of the form

$$P(z) = \frac{1}{Z_P} e^{-E_P(z)}, \qquad Q(z) = \frac{1}{Z_Q} e^{-E_Q(z)},$$

where Z_P and Z_Q are appropriate normalizing factors, we get

$$\begin{aligned} D_{KL}(P\|Q) &= -\sum_z \frac{1}{Z_P} e^{-E_P(z)} \left(\log \frac{Z_P}{Z_Q} + E_P(z) - E_Q(z) \right) \\ &= \log Z_Q - \log Z_P + \frac{1}{Z_P} \sum_z e^{-E_P(z)} (E_Q(z) - E_P(z)). \end{aligned}$$

If E_P and E_Q are in addition functions of some parameter θ, then

$$\nabla_\theta Z_P = \sum_z (-\nabla_\theta E_P(z)) e^{-E_P(z)} = -Z_P \langle \nabla_\theta E_P \rangle_P$$

where $\langle \cdot \rangle_\mu$ denotes expecation with respect to a probability measure μ.

Thus

$$\begin{aligned}\nabla_\theta D_{KL}(P\|Q) =& -\langle\nabla_\theta E_Q\rangle_Q + \langle\nabla_\theta E_P\rangle_P \\ &+ \frac{1}{Z_P}\langle\nabla_\theta E_P\rangle_P \sum_z e^{-E_P(z)}(E_Q(z) - E_P(z)) \\ &- \frac{1}{Z_P}\sum_z e^{-E_P(z)}(E_Q(z) - E_P(z))\nabla_\theta E_P(z) \\ &+ \frac{1}{Z_P}\sum_z e^{-E_P(z)}(\nabla_\theta E_Q(z) - \nabla_\theta E_P(z)) \\ =&\langle\nabla_\theta E_Q\rangle_P - \langle\nabla_\theta E_Q\rangle_Q + \langle(E_P - E_Q)(\nabla_\theta E_P - \langle\nabla_\theta E_P\rangle_P)\rangle_P\end{aligned}$$

Then, letting

$$J(P_0, P_n, P_\infty) = D_{KL}(P_0\|P_n) - D_{KL}(P_n\|P_\infty)$$

where P_0 is independent of θ, and P_n, P_∞ are as Q and P above (and we replace P_m by m as a subindex to simplify the notation), we get

$$\nabla_\theta J(P_0, P_n, P_\infty) = \langle\nabla_\theta E_\infty\rangle_0 - \langle\nabla_\theta E_\infty\rangle_n + \langle(E_\infty - E_n)(\nabla_\theta E_n - \langle\nabla_\theta E_n\rangle_n)\rangle_n$$

In short, the issue here is that we know how to compute E_∞, but we can't sample from P_∞, and, conversely, we know how to sample from P_n, but we can't compute E_n. This means that the first two terms can always be computed, and P_n should be constructed so that the third term is always small.

One thing I would like to have a proof of is that if P_n is constructed in some recursive manner then as $n \to \infty$ the third term above always goes to 0 as $n \to \infty$, and maybe even get some estimates for that.