Training Word Representation RBM (Confirmed by Hugo Larochelle)

Ji Ma

October 23, 2013

1 Training WRRBM

May I go directly to section 6.1 of the original paper? In their model, the energy function is given as (notations are consistent with the orignal paper):

$$E(v,h) = -c'h + \sum_{i=1}^{n} -b'v^{(i)} - h'\mathbf{U}^{(i)}\mathbf{D}v^{(i)} \tag{1}$$

Here, v and h denotes the visuable and hidden variables respectively. $v^{(i)}$ denotes the one-hot representation of the word in the i-th position. c and b are hidden and visuable bias respectively. I use c' to denote c transpose. $\mathbf{D}$ is the $D \times K$ word imbedding matrix where K is the size of our vocab and d is the length of word embedding. $\mathbf{U}^{(i)}$ corresponding to the $H \times D$ weight matrix for the i-th position.

My problem comes from how to update $\mathbf{U}$ and $\mathbf{D}$. Which is given by eq 8, eq 9 and eq 10. Remember that given a visuable data v^0, the Constive Divergence update rule is:

$$\sum_{\hat{h}^0} p(\hat{h}^0|v^0)\frac{\partial E(v^0,\hat{h}^0)}{\partial\theta} - \sum_{\hat{h}^1} p(\hat{h}^1|v^1)\frac{\partial E(v^1,\hat{h}^1)}{\partial\theta} \tag{2}$$

where v^1 and h^1 denotes the visuable and hidden variable of the negative or reconstructed sample, respectively.

To simplify the derivation, let me first take a closer look on the last term of eq 1. Suppose h is a vector of length H, i.e. $(h_1, h_2, ..., h_H)$, and the entry of row j, column k of $\mathbf{U}^{(i)}$ is $\mathbf{U}^{(i)}_{jk}$. Thus the vector of $h'\mathbf{U}^{(i)}$ is:

$$h'\mathbf{U}^{(i)} = (\sum_{j=1}^{H} h_j * U^{(i)}_{j1},\ \sum_{j=1}^{H} h_j * U^{(i)}_{j2},\ \ldots\ , \sum_{j=1}^{H} h_j * U^{(i)}_{jD}) \tag{3}$$

Suppose $v^{(i)} = e_k$, i.e., only the k-th element of $v^{(i)}$ is 1 and all others are 0. Thus $\mathbf{D}v^{(i)}$ selects the k'th column of $\mathbf{D}$.

$$\mathbf{D}v^{(i)} = (\mathbf{D}_{1k}, \mathbf{D}_{2k}, \ldots, \mathbf{D}_{dk})' \tag{4}$$

Combining eq 2 and eq 3 we have:

$$h'\mathbf{U}^{(i)}\mathbf{D}v^{(i)} = \sum_{d=1}^{D}\sum_{j=1}^{H} h_j * \mathbf{U}^{(i)}_{jd} * \mathbf{D}_{dk} \tag{5}$$

The derivative of E with respect to $\mathbf{U}^{(i)}_{jd}$ is:

$$\frac{\partial E(v,h)}{\partial \mathbf{U}^{(i)}_{jd}} = h_j * \mathbf{D}_{dk} \tag{6}$$

And *for the first posithon*, the derivative of E w.r.t $\mathbf{D}_{dk}$ is:

$$\frac{\partial E(v,h)}{\partial \mathbf{D}_{dk}} = \sum_{j=1}^{H} h_j * \mathbf{U}^{(i)}_{jd} \tag{7}$$

Combining eq 2 and eq 6, I got the update rule for $\mathbf{U}^{(i)}_{jd}$ is

$$\mathbf{U}^{(i)}_{jd} = \mathbf{U}^{(i)}_{jd} + \lambda * (p({h^0}_j = 1|v^0) * \mathbf{D}_{dk} - p({h^1}_j = 1|v^1) * \mathbf{D}_{dm}) \tag{8}$$

where the word at the first position of v^0 and v^1 are e_k and e_m, respectively.

For D, I have to break the update rule into 2 parts.

$$\mathbf{D}_{dk} = \mathbf{D}_{dk} + \lambda * \sum_{j}^{H} p(h^0_j = 1|v^0) * \mathbf{U}^{(i)}_{jd} \tag{9}$$

$$\mathbf{D}_{dm} = \mathbf{D}_{dm} - \lambda * \sum_{j}^{H} p(h^1_j = 1|v^1) * \mathbf{U}^{(i)}_{jd} \tag{10}$$