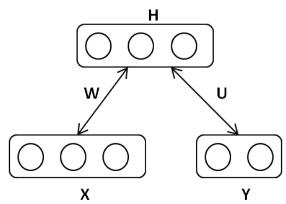
Label Sparsity Constrained RBM For Classification

Son N. Tran - 2013

This note presents the idea of applying sparsity constraints into label softmax units in RBM for classification. We show that if the constraint is in L1-norm, the RBM is similar as hybrid-discriminative RBM proposed by Hugo []. In



general, an RBM with label encoded by softmax units as shown in Figure **??** has energy function:

$$\mathbf{E}(x,y) = -\sum_{ij} x_i w_{ij} h_j - \sum_j U_{yj} h_j - \sum_i a_i x_i - b_y - \sum_j c_j h_j \tag{1}$$

with x is a vector for input data and y is a scalar representing the label (class). The RBM classification is trained by maximizing the log-likelihood and minimizing the total activation of softmax units.

$$\Gamma = \sum_{x,y} \log P(x,y) - \lambda \sum_{x,y} \|1 - P(y|x)\|_l \tag{2}$$

The classification RBM can be trained using gradient ascent in which the total gradient is the combination of the log-likelihood's gradient $\Delta_{\mathcal{L}}$ and the sparsitks gradient $\Delta_{\mathcal{S}}$.

$$\Delta\theta = \Delta_{\mathcal{L}}\theta + \lambda\Delta_{\mathcal{S}}\theta \tag{3}$$

The gradient of log-likelihood can be computed approximately using CD [] or PCD [] methods. For example, with CD:

$$\begin{aligned}
\Delta_{\mathcal{L}} w_{ij} &= \langle v_i h_j \rangle_0 - \langle v_i h_j \rangle_\kappa \\
\Delta_{\mathcal{L}} u_{kj} &= \langle \mathbb{I}(k=y) h_j \rangle_0 - \langle \mathbb{I}(k=y) h_j \rangle_\kappa \\
\Delta_{\mathcal{L}} a_i &= \langle v_i \rangle_0 - \langle v_i \rangle_\kappa \\
\Delta_{\mathcal{L}} b_k &= \langle \mathbb{I}(k=y) \rangle_0 - \langle \mathbb{I}(k=y) \rangle_\kappa \\
\Delta_{\mathcal{L}} c_j &= \langle h_j \rangle_0 - \langle h_j \rangle_k
\end{aligned} \tag{4}$$

1 L1-norm

If $l = 1$, then

$$\begin{aligned}
\Gamma &= \sum_{x,y} \log P(x,y) - \lambda \sum_{x,y} (1 - P(y|x)) \\
&= \sum_{x,y} \log P(x,y) + \lambda \sum_{x,y} P(y|x) - \psi
\end{aligned} \tag{5}$$

As being shown in the transformation above, the cost function of RBM classification with L-1 sparsity constrant is similar to the hybrid-discriminative RBM in [].

The gradients of L-1 sparsity constraint are (See DiscriminativeRBM note for more mathematic details):

$$\begin{aligned}
\Delta_{\mathcal{S}} w_{ij} &= x_i P(h_j|y,x) - x_i \sum_k P(h_j|k,x)P(k|x)) \\
\Delta_{\mathcal{S}} u_{kj} &= \mathbb{I}(k=y)P(h_j|k,x) - P(k|x)P(h_j|x,k) \\
\Delta_{\mathcal{S}} a_i &= 0 \\
\Delta_{\mathcal{S}} b_k &= \mathbb{I}(k=y) - P(k|x) \\
\Delta_{\mathcal{S}} c_j &= P(h_j|x,y) - \sum_k P(h_j|x,k)P(k|x)
\end{aligned} \tag{6}$$

2 L2-norm

If $l = 2$ then,

$$\Gamma = \sum_{x,y} \log P(x,y) - \frac{\lambda}{2} \sum_{x,y} (1 - P(y|x))^2 \tag{7}$$

The gradients of L-2 sparsity constraint are shown below(See DiscriminativeRBM note for more mathematic details of computing conditional probability derivatives):

$$\begin{aligned}
\Delta_{\mathcal{S}} w_{ij} &= (1 - P(y|x)\Big(x_i P(h_j|y,x) - x_i \sum_k P(h_j|k,x)P(k|x))\Big) \\
\Delta_{\mathcal{S}} u_{kj} &= (1 - P(y|x)\Big(x_i P(h_j|y,x)\mathbb{I}(k=y)P(h_j|k,x) - P(k|x)P(h_j|x,k)\Big) \\
\Delta_{\mathcal{S}} a_i &= 0 \\
\Delta_{\mathcal{S}} b_k &= (1 - P(y|x)\Big(x_i P(h_j|y,x)\mathbb{I}(k=y) - P(k|x)\Big) \\
\Delta_{\mathcal{S}} c_j &= (1 - P(y|x)\Big(x_i P(h_j|y,x)P(h_j|x,y) - \sum_k P(h_j|x,k)P(k|x)\Big)
\end{aligned} \tag{8}$$