Derivation of Restricted Boltzmann Machine s(RBM)

We assume the number of visible units to be m and the number of hidden units to be n
The energy of the configuration of the visible units **v** and hidden units **h** is given by

$$E(v,h) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}h_i v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i$$

Where **b** and **c** are the biases for the visible and the hidden units respectively and $w_{ij},\ b_j,\ c_i$ are the parameters . Also we assume that the training data D has l data points from $x_1 .... x_l$. Let's assume the original distribution of the data to be q , with RBMs our aim is to learn a distribution p which best approximated q .

Maximum Likelihood Formulation

Calculating KL divergence of $q\ w.r.t\ p$ on a finite state space S

$$KL(q\|p) = \sum_{x\,\varepsilon\,S} q(x)\ ln\ \frac{q(x)}{p(x)} = \sum_{x\,\varepsilon\,S} q(x)\ ln\ q(x) - \sum_{x\,\varepsilon\,S} q(x)\ ln\ p(x)$$

KL is non - negative and 0 if p = q
Therefore minimizing KL corresponds to maximizing the likelihood of x in the training data.
Thus our objective becomes as follows

$$L(\theta \mid D) = \prod_{k=1}^{l} p(x_k \mid \theta)$$

Or equivalently , maximizing the log likelihood given by

$$ln\ L(\theta \mid D) = \sum_{k=1}^{l} ln\ p(x_k \mid \theta)$$

We use gradient ascent to maximize our objective. Therefore we will need to calculate the following gradient

$\frac{d}{d\theta}\left(\sum_{k=1}^{l} ln\ p(x_k \mid \theta) \right)$ where θ are our parameters $w_{ij},\ b_j,\ c_i$

Calculating conditional probabilities

Since this is an energy based model, the joint distribution is given by

$p(v,\ h) = e^{-E(v,\ h)}/\ Z$ where $Z = \sum_{v}\sum_{h} e^{-E(v,\ h)}$

In RBM, the connections are only between the visible and hidden units, therefore

$$p(h \mid v) = \prod_{i=1}^{n} p(h_i \mid v)$$

$$p(v \mid h) = \prod_{i=1}^{m} p(v_i \mid h)$$

Log Likelihood Calculation

$p(v) = \sum_h p(v, h) = \frac{1}{Z}\sum_h e^{-E(v, h)}$

Writing the log likelihood function $ln\, p(x \mid \theta)$ using the above equation

$ln\, p(x \mid \theta) = ln\ \frac{1}{Z}\sum_h e^{-E(x, h)} = ln\ \sum_h e^{-E(x, h)} - ln\ \sum_{x,h} e^{-E(x, h)}$

where θ are our parameters w , b , c

The next step is to calculate the derivative of this log likelihood. For that we would need $p(h \mid v)$

$p(h \mid v) = p(v, h) / p(v) = e^{-E(v, h)} / \sum_h e^{-E(v, h)}$

Gradient of log likelihood

$\frac{d}{d\theta} ln\, p(v \mid \theta) = \frac{d}{d\theta} ln\ \sum_h e^{-E(v, h)} - \frac{d}{d\theta} ln\ \sum_{v,h} e^{-E(v, h)}$

$= \frac{1}{\sum_h e^{-E(v, h)}} \sum_h e^{-E(v, h)} \frac{d}{d\theta} E(v, h) + \frac{1}{\sum_{v,h} e^{-E(v, h)}} \sum_{v,h} e^{-E(v, h)} \frac{d}{d\theta} E(v, h)$

$= -\sum_h p(h \mid v). \frac{d}{d\theta} E(v, h) + \sum_{v,h} p(v, h). \frac{d}{d\theta} E(v, h)$

where θ are our parameters w_{ij}, b_j, c_i

Calculating $\frac{d}{dw_{ij}} E(v, h) = -h_i v_j$, $\frac{d}{db_j} E(v, h) = -v_j$, $\frac{d}{dc_i} E(v, h) = -h_i$

Taking average of the gradient across the training set

$\frac{1}{l} \sum_{v\, \varepsilon\, D} [-\sum_h p(h \mid v). \frac{d}{d\theta} E(v, h) + \sum_{v,h} p(v, h). \frac{d}{d\theta} E(v, h)]$

For $\theta = w_{ij}$

$= \frac{1}{l} \sum_{v\, \varepsilon\, D} [\sum_h p(h \mid v). h_i v_j - \sum_{v,h} p(v, h). h_i v_j]$

$= \frac{1}{l} \sum_{v\, \varepsilon\, D} [\sum_h p(h \mid v). h_i v_j - \sum_{v,h} p(v, h). h_i v_j]$

$= \frac{1}{l} \sum_{v\, \varepsilon\, D} [E_{p(h \mid v)}\, h_i v_j - E_{p(v, h)}. h_i v_j]$

$= < h_i v_j >_{data} - < h_i v_j >_{model} = \frac{d}{dw_{ij}} ln\, p(v)$

The first term is the expectation of $h_i v_j$ when the visible vector is fixed, and the second term is the expectation of v and h if we sample from the model.

Now calculating the expression for $p(v_k = 1|h)$

Let $\gamma(v_{-k}, h) = -\sum_i \sum_{j \neq k} w_{ij} h_i v_j - \sum_{j \neq k} b_j v_j - \sum_i c_i h_i$

$\eta_k(h) = -\sum_{i=1}^{n} w_{ij} h_i - b_k$

We can see that $E(v, h) = \gamma(v_{-k}, h) + v_k \eta_k(h)$

Since the visible units are independent

$$p(v_k = 1 | h) = p(v_k = 1 | v_{-k}, h) = \frac{p(v_k = 1, v_{-k}, h)}{p(v_{-k}, h)} = \frac{e^{-E(v_k = 1, v_{-k}, h)}}{e^{-E(v_k = 1, v_{-k}, h)} + e^{-E(v_k = 0, v_{-k}, h)}}$$

$$= \frac{e^{-\gamma(v_{-k}, h) - 1.n_k(h)}}{e^{-\gamma(v_{-k}, h) - 1.n_k(h)} + e^{-\gamma(v_{-k}, h) - 0.n_k(h)}} = \frac{e^{-\gamma(v_{-k}, h)}.e^{-\eta_k(h)}}{e^{-\gamma(v_{-k}, h)}.e^{-\eta_k(h)} + e^{-\gamma(v_{-k}, h)}}$$

$$= \frac{e^{-\gamma(v_{-k}, h)}.e^{-\eta_k(h)}}{e^{-\gamma(v_{-k}, h)}.(e^{-\eta_k(h)} + 1)} = \frac{e^{-\eta_k(h)}}{(e^{-\eta_k(h)} + 1)} = \frac{1}{1 + e^{\eta_k(h)}} = \sigma(-\eta_k(h))$$

$$= \sigma(\sum_{i=1}^{n} w_{ik} h_i + b_k)$$

By Symmetry

$$p(h_k = 1 | v) = \sigma(\sum_{j=1}^{m} w_{kj} v_j + c_k)$$

This represents the probability of h_k when v is clamped

Now, **Simplifying the first term from the gradient of the log likelihood**

$$-\sum_h p(h | v).\frac{d}{d\theta} E(v, h)$$

For $\theta = w_{ij}$

$= \sum_h p(h | v).h_i v_j = \sum_{h_i} \sum_{h_{-i}} p(h_i | v).p(h_{-i} | v) h_i v_j$ since h_i is independent of other h_i

h_{-i} denotes all the hidden units except i.

$= \sum_{h_{-i}} p(h_{-i} | v). \sum_{h_i} p(h_i | v) h_i v_j = 1. \sum_{h_i} p(h_i | v) h_i v_j$ as other units other than i are considered to be constant

$= p(h_i = 1 | v).1.v_j + p(h_i = 0 | v).0.v_j$ h_i = {0, 1}

$= p(h_i = 1 | v).v_j$

For $\theta = b_j$

$= v_j$ since $\frac{d}{db_j} E(v, h) = -v_j$ *and* v_j is not dependent on h

For $\theta = c_i$

$= p(h_i = 1 | v)$ since $\frac{d}{dc_i} E(v, h) = -h_i$

Simplifying the **second term of the log likelihood gradient**

$$\sum_{v,h} p(v, h).\frac{d}{d\theta} E(v, h)]$$

For $\theta = w_{ij}$

$$-\sum_{v,h} p(v,\ h).h_i v_j \ = \ -\sum_{v} p(v) \sum_{h} p(h\,|v).h_i v_j$$

Simplifying $\sum_{h} p(h\,|v).h_i$ as above from the first term calculation

$$= -\sum_{v} p(v).p(h_i = 1|v).v_j$$

Now instead of summing over all possible visible vectors, we use contrastive divergence to estimate the expected v

we use contrastive divergence for 1 step
Starting from the training vector $v^{(0)}$
calculate vector $h^{(0)}$ samping from $p(h|\ v^{(0)})$
calculate vector $v^{(1)}$ sampling from $p(v|\ h^{(0)})$

So our final equation for the second term

For $\theta = \ w_{ij}$
$= p(h_i = 1\ |\ v^{(1)})\ .\ v_j{}^{(1)}$

For $\theta = b_j$
$= v_j{}^{(1)}$

For $\theta = c_i$
$= p(h_i = 1|\ v^{(1)})$

Final Parameter Update Rules

$\Delta w_{ij} = p(h_i\ =\ 1|\ v).v_j\ -\ p(h_i\ =\ 1|\ v^{(1)}).v_j{}^{(1)}$
$\Delta b_j =\ v_j\ -\ v_j{}^{(1)}$
$\Delta c_i =\ p(h_i\ = 1|\ v\ )\ -\ p(h_i\ =\ 1|\ v^{(1)})$