Notes and Questions on RBM

Ji Ma

December 19, 2013

1 Energy-based model, training set log-likelihood and gradient of it

Before this section, shall we add some notes about Markov Random Field and Gibbs distribution?

Since RBM is a special case of energy-based models, so before going into the details of RBM, we first briefly review energy-based model.

1.1 Energy-based model

The energy-based model defines a distribution as:

$$P(x) = \frac{e^{-\text{Energy}(x)}}{Z} \tag{1}$$

Here, Z is called the partition function or normalizing factor which is to guarantee the property of probability distribution.

Some times (RBM), latent or hidden variables h are introduced so as to make the model more flexible. In such cases, the joint probability distribution over the observed and hidden variables becomes:

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

And the marginal distribution over x is:

$$P(x) = \sum_h \frac{e^{-\text{E}(x,h)}}{Z} \tag{3}$$

where $\sum$ range all possible values of h.

1.2 Log-Likelihood, empirical distribution

For the model given by equation 3, the log-likelihood of an observed data x is:

$$\log \mathcal{L}(x) = \log \sum_h e^{-\text{E}(x,h)} - \log(Z) \tag{4}$$

Here we use $\mathcal{L}(x)$ to denote the likelihood of x (which is actually $p(x)$). Suppose the we have a training set S which consists of n samples $\{x^1, x^2, \ldots, x^n\}$, the log likelihood of S is:

$$\begin{aligned}\log \mathcal{L}(S) &= \log \mathcal{L}(x^1) * \mathcal{L}(x^2) * \ldots * \mathcal{L}(x^n) \\ &= \sum_{i=1}^{n} \log \mathcal{L}(x^i) \\ &= \sum_{i=1}^{n} \log \sum_h e^{-\mathrm{E}(x^i,h)} - \log Z \end{aligned} \tag{5}$$

There is also another way to express the log-likelihood of S using the empirical distribution. Suppose that S contains m unique assignments of x, e.g., $x_1, x_2, \ldots, x_m$. Each x_i occurs n_i times, therefore:

$$\sum_{i}^{m} n_i = n \tag{6}$$

And with these notations, equation 5 can be written as:

$$\begin{aligned}\log \mathcal{L}(S) &= \log \mathcal{L}(x_1)^{n_1} * \mathcal{L}(x_2)^{n_2} * \ldots * \mathcal{L}(x^m)^{n_m} \\ &= \sum_{i=1}^{m} n_i * \log \mathcal{L}(x_i) \\ &= \sum_{i=1}^{m} n_i * (\log \sum_h e^{-\mathrm{Energy}(x_i,h)} - \log Z) \end{aligned} \tag{7}$$

Remember that the empirical distribution is defined as:

$$\hat{p}(x_i) = \frac{n_i}{n} \tag{8}$$

I saw many times terms like "average the log-likelihood over the training set" which is:

$$\begin{aligned}\frac{1}{n} * \log \mathcal{L}(S) &= \sum_{i=1}^{m} \frac{n_i}{n} * \log \mathcal{L}(x_i) \\ &= \sum_{i=1}^{m} \hat{p}(x_i) * \log \mathcal{L}(x_i) \\ &= < \log \mathcal{L}(x) >_{\hat{p}(x)} \end{aligned} \tag{9}$$

the RHS of the last line in (9) is the expectation of $\log L$ w.r.t the empirical distribution of the training set.

1.3 Training, MLE, KL distance

Regarding training the models mentioned above, the only thing I come up with is MLE. Given S, MLE select the parameters that maximize $\log L(S)$. This can be done by compute the gradient and set them to zero. Let θ to be the model parameters, the gradient is (also see Fischer and Igel (2012)):

$$\frac{\partial ln\mathcal{L}(x|\theta)}{\partial\theta} = -\sum_{h} p(h|x)\frac{\partial E(x,h)}{\partial\theta} + \sum_{x,h} p(x,h)\frac{\partial E(x,h)}{\partial\theta} \tag{10}$$

Obviously, the difficulty in computing the gradient is that the two terms on the RHS of (10) either involves enumerating all possible values of h or (v,h) which is intractable. We will talk about how to deal with it later on. Here one thing about (10) should be explained more clearly.

1.3.1 Make the gradient easier to understand

I always confused with the right hand side of 10 in that I cannot tell what x really mean. Is it the training sample identical to that on the LHS of 10? Let's rewrite (10) to make it easy to understand:

$$\frac{\partial ln\mathcal{L}(x|\theta)}{\partial\theta} = -\sum_{\hat{h}} p(\hat{h}|x)\frac{\partial E(x,\hat{h})}{\partial\theta} + \sum_{\hat{x},\hat{h}} p(\hat{x},\hat{h})\frac{\partial E(\hat{x},\hat{h})}{\partial\theta} \tag{11}$$

Now we can see that x in the first term on RHS of (11) denotes the training sample. The second term on RHS of (11) has nothing to do with the training sample x. *It is the expectation of* $\frac{\partial E(\hat{x},\hat{h})}{\partial\theta}$ *under the model's distribution which is indeed a constant, denoted by* C*, w.r.t any training sample (given the current model parameters fixed).*

Thus, if we could (1) compute the first term (RBM) and (2) remove the second (contrastive divergence), we can learn the model parameters.

Shall we go a step further to see the gradient of the average log-likelihood of the training set? It is another highlight of this note:

$$\begin{aligned}
\frac{\partial\frac{1}{n} * \log\mathcal{L}(S)}{\partial\theta} &= \frac{\partial\sum_{i=1}^{m}\hat{p}(x_i) * \log\mathcal{L}(x_i)}{\partial\theta} \\
&= \sum_{i=1}^{m}\hat{p}(x_i) * \frac{\partial\log\mathcal{L}(x_i)}{\partial\theta} \\
&=< \frac{\partial\log\mathcal{L}(x)}{\partial\theta} >_{\hat{p}(x)} \\
&= -\sum_{\hat{p}(x)}\sum_{\hat{h}} p(\hat{h}|x)\frac{\partial E(x,\hat{h})}{\partial\theta} + \sum_{\hat{p}(x)}\sum_{\hat{x},\hat{h}} p(\hat{x},\hat{h})\frac{\partial E(\hat{x},\hat{h})}{\partial\theta} \\
&= -<\sum_{\hat{h}} p(\hat{h}|x)\frac{\partial E(x,\hat{h})}{\partial\theta} >_{\hat{p}(x)} + C
\end{aligned} \tag{12}$$

Always remember that the two difficult factors (DIFFICULT TWO) in computing the gradient: (1) computing $\sum_{\hat{h}} p(\hat{h}|x)\frac{\partial E(x,\hat{h})}{\partial \theta}$; (2) computing C. The empirical expectation of $\sum_{\hat{h}} p(\hat{h}|x)\frac{\partial E(x,\hat{h})}{\partial \theta}$ can be achieved by go though all samples of the training set S.

1.3.2 KL distance and MLE

Maximize the log-likelihood of the training set corresponds minimize the KL distance between the model's distribution p and the empirical distribution of the training set $\hat{p}$. This is because that the KL distance is

$$KL(\hat{p}\|p) = \sum_{x\in S} \hat{p}(x) \ln \frac{\hat{p}(x)}{p(x)} = \sum_{x\in S} \hat{p}(x) \ln \hat{p}(x) - \sum_{x\in S} \hat{p}(x) \ln p(x) \tag{13}$$

note that the first term of RHS (entropy of $\hat{p}$) is a constant and the second term is right the average log-likelihood given by equation 9. Thus, minimize the KL distance corresponds to maximize 9 (7 and 9 only differs by a constant factor). In addition, the gradient of 9 which is given by 12 can be used to derive the gradient of the KL-distance is:

$$\frac{\partial KL(\hat{p}\|p)}{\partial \theta} = < \sum_{\hat{h}} p(\hat{h}|x)\frac{\partial E(x,\hat{h})}{\partial \theta} >_{\hat{p}(x)} - C \tag{14}$$

2 RBM

The energy function of RBM is:

$$E(x,h) = -b'x - c'h - h'Wx \tag{15}$$

The independence assumption of RBM states that given x, the h_i is independent of the rest of the hidden variables h_{-i}. Due to this property, the marginal distribution over x can be easily computed. The details are to be completed. Right now, just refer to Fischer and Igel (2012) and (Bengio, 2009).

A important property of RBM is that the first factor of DIFFICULT TWO can be easily computed, take w_{ij} for example (see Fischer and Igel(2012) for the derivation):

$$\sum_{\hat{h}} p(\hat{h}|x)\frac{\partial E(x,\hat{h})}{\partial w_{ij}} = p(h_i = 1|x)x_j \tag{16}$$

In the next section, we will see how Contrastive Divergence deals with the second factor of DIFFICULT TWO.

3 Contrastive Divergence learning

The basic idea of $k-$step Contrastive Divergence learning is to minimize

$$KL(\hat{p}\|p) - KL(\hat{p}_k|p) = \sum_{x\in S} \hat{p}(x) \ln p(x) - \sum_{x'\in S_k} \hat{p}_k(x') \ln p(x') \tag{17}$$

where $\hat{p}_k(x)$ is empirical distribution over S_k which is the $k-$step MCMC samples of S(also see the references for the theoretical reason behind).

The advantage is of minimizing (17) is that the its gradient is tractable (the second factor of the DIFFICULT TWO cancels out):

$$\begin{aligned}\frac{\partial(KL(\hat{p}\|p) - KL(\hat{p}_k|p))}{\partial\theta} =& - < \sum_{\hat{h}} p(\hat{h}|x) \frac{\partial E(x,\hat{h})}{\partial\theta} >_{\hat{p}(x)} \\ &+ < \sum_{\hat{h}} p(\hat{h}|x') \frac{\partial E(x',\hat{h})}{\partial\theta} >_{\hat{p}_k(x')}\end{aligned} \tag{18}$$

Combining (16) and (18), given a training sample x, we have the following rule to update w_{ij} (an on-line version) for RBM

$$w_{ij} = w_{ij} + \varepsilon(p(h_i = 1|x)x_j - p(h'_i = 1|x')x'_j) \tag{19}$$

Here ε is the learning rate, h' and x' is the $k-$step MCMC sample starting from x.

Dear Shujie and Nan, 19 is consistent with Fischer and Igel (2012) but different from Bengio (2009), why? How the update rule in Bengio(2009) is derived? Thanks very much for the help.