Restricted Boltzmann Machine の導出に至る自分用まとめ

齋藤 真樹

2012 年 8 月 18 日

1 問題設定

可視素子 (visible node) の集合 $\boldsymbol{v} = (v_1, \cdots, v_L), \forall v_i \in \{0, 1\}$，隠れ素子 (hidden node) の集合 $\boldsymbol{h} = (h_1, \cdots, h_M), \forall h_j \in \{0, 1\}$ から成る結合確率 $p(\boldsymbol{v}, \boldsymbol{h})$ を以下のように定義する．

$$p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} \exp(-E(\boldsymbol{v}, \boldsymbol{h})) \tag{1}$$

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_i b_i v_i - \sum_j c_j h_j - \sum_i \sum_j v_i W_{ij} h_j \tag{2}$$

このような $p(\boldsymbol{v}, \boldsymbol{h})$ を一般に Restricted Boltzmann Machine(RBM) と呼ぶ．b_i, c_j, W_{ij} はそれぞれ可視素子のバイアス，隠れ素子のバイアス，ウェイトパラメータと呼ばれている．Z は正規化定数あるいは分配関数と呼ばれており，

$$Z = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h})) \tag{3}$$

の関係を持つ．

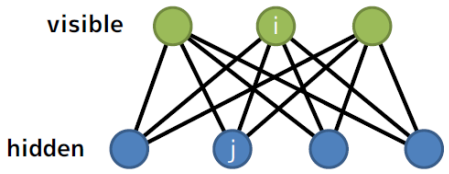

図 1　RBM のグラフィカルモデル．計算上の理由から可視素子間，隠れ素子間の結合は存在しない．

ここで，我々は手元に可視素子に関しての N 個の観測データ $\{\boldsymbol{v}^1, \boldsymbol{v}^2, \cdots, \boldsymbol{v}^N\}$ を手に入れているが，隠れ素子に関しては手に入れていないものとする．このような制約の中からどうにかして，我々が求めるべきパラメータである b_i, c_j, W_{ij} を推測したい．

この場合，典型的な戦略としてはまず $p(\boldsymbol{v}, \boldsymbol{h})$ を $\boldsymbol{h}$ に関して積分を行うことによって，周辺分布 $p(\boldsymbol{v}) = \sum_{\boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h})$ を得る．この分布は $\boldsymbol{h}$ に依存しておらず，$\boldsymbol{v}$ それ自身によってのみ決定される分布である．この分布を用いて最尤推定を行うことで，パラメータを推測する戦略をとることにしよう．

2 計算の概要

以降は具体的な計算について示す．まず，最尤推定では対数尤度を，パラメータ θ に関しての最大化を行うことで表現する．

$$J \equiv \left\langle \ln \sum_{\boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h}) \right\rangle_q = \left\langle \ln \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v}, \boldsymbol{h})} \right\rangle_q - \ln Z \tag{4}$$

ここで，$\langle \cdot \rangle_q$ のような形を確率分布 $q(\boldsymbol{v})$ の期待値として表現する．つまり

$$\langle f(\boldsymbol{v}) \rangle_q = \sum_{\boldsymbol{v}} f(\boldsymbol{v}) q(\boldsymbol{v}) \tag{5}$$

である．$q(\boldsymbol{v})$ は観測データに関しての確率分布 $q(\boldsymbol{v})$

$$q(\boldsymbol{v}) = \frac{1}{N} \sum_k \delta(\boldsymbol{v} - \boldsymbol{v}^k) \tag{6}$$

であり，$\delta(x)$ はディラックのデルタ関数である．このようにして定義された J を最大化するような b_i, c_j, W_{ij} を求める戦略が最尤推定である．

パラメータの推定値を求めるために，勾配法を用いて最適化を行うこととする．そのためにはまず，パラメータに関しての微分 $\partial J / \partial \theta$ が必要である．ゆえに，任意のパラメータ θ に関して微分を取ることによって下記を得る．

$$\frac{\partial J}{\partial \theta} = -\left\langle \frac{1}{\sum_{\boldsymbol{h}} e^{-E}} \sum_{\boldsymbol{h}} \frac{\partial E}{\partial \theta} e^{-E} \right\rangle_q + \frac{1}{Z} \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \frac{\partial E}{\partial \theta} e^{-E} \tag{7}$$

ここで，条件付き確率の定義

$$p(\boldsymbol{h}|\boldsymbol{v}) = \frac{e^{-E(\boldsymbol{v}, \boldsymbol{h})}}{\sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v}, \boldsymbol{h})}} \tag{8}$$

並びにマルコフ確率場の定義を用いることによって，上式は以下のように簡略化できる．

$$\frac{\partial J}{\partial \theta} = -\sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \frac{\partial E}{\partial \theta} p(\boldsymbol{h}|\boldsymbol{v}) q(\boldsymbol{v}) + \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \frac{\partial E}{\partial \theta} p(\boldsymbol{v}, \boldsymbol{h}) \tag{9}$$

$$\equiv -\left\langle \frac{\partial E}{\partial \theta} \right\rangle_{\text{data}} + \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{\text{model}} \tag{10}$$

ここで，$\langle \cdot \rangle_{\text{data}}, \langle \cdot \rangle_{\text{model}}$ は，それぞれ確率分布 $p_{\text{data}}(\boldsymbol{v}, \boldsymbol{h}) = p(\boldsymbol{h}|\boldsymbol{v}) q(\boldsymbol{v})$ と $p_{\text{model}}(\boldsymbol{v}, \boldsymbol{h}) = p(\boldsymbol{v}, \boldsymbol{h})$ に対しての期待値を表す．この表記を用いることで，パラメータ θ に関しての微分操作の本質的意味合いをより簡潔に表現できるようになった．

一番初めの定義へと戻ろう．我々が現在持っている観測データは $\boldsymbol{v}^k$ のみであり，隠れ素子についての観測データは持ち合わせていない．そのため，$\boldsymbol{v}$ に関しての周辺化を行うことによって隠れ素子の変数を消去し，その枠組の中で最大化を行うこととした．これは，本来得られてることが望ましい観測データの結合確率 $\frac{1}{N} \sum_k \delta(\boldsymbol{v} - \boldsymbol{v}^k) \delta(\boldsymbol{h} - \boldsymbol{h}^k)$ を用いるのではなく，現在のパラメータの元で推測した結合確率 $p(\boldsymbol{h}|\boldsymbol{v}) q(\boldsymbol{v})$ を用いて計算を行なっていることと等価である．

幸いにも，Restricted Boltzmann Machine に関しては可視素子間，隠れ素子間についての結合が存在していないため，条件付き確率 $p(\boldsymbol{v}|\boldsymbol{h}), p(\boldsymbol{h}|\boldsymbol{v})$ の厳密解は容易に計算できる．しかも，条件付き確率下での各引

数は独立であり ($p(\boldsymbol{v}|\boldsymbol{h}) = \prod_i p(v_i|\boldsymbol{h}), p(\boldsymbol{h}|\boldsymbol{v}) = \prod_j p(h_j|\boldsymbol{v})$)，このことも RBM の計算の省力化に貢献している．

それぞれの条件付き確率を計算することによって下記を得る．

$$p(v_i = 1|\boldsymbol{h}) = \sigma(b_i + \sum_j W_{ij} h_j) \tag{11}$$

$$p(h_j = 1|\boldsymbol{v}) = \sigma(c_j + \sum_i v_i W_{ij}) \tag{12}$$

ここで，$\sigma(x) = 1/(1 + \exp(-x))$ はシグモイド関数である．

3 Contrastive Divergence

上式の第一項 $\langle\cdot\rangle_{\text{data}}$ は比較的簡単に計算できるが，一般に第二項 $\langle\cdot\rangle_{\text{model}}$ に関しての計算は非常に困難である場合が多い．何故なら，第二項はあらゆる $\boldsymbol{v}, \boldsymbol{h}$ に関しての和を取らなければならないため，ノード数の増加によって状態量が指数的に爆発してしまうためである．

このような直接的に計算不可能な期待値の計算を行う際，大まかに分けて 2 種類の方法が存在する．まず一つ目はマルコフ連鎖モンテカルロ (MCMC) を用いることによって，力技で計算を行う手法と，平均場近似などを用いて近似的に解を求めるという手法である．一般に前者は計算速度，後者は近似精度が問題となる場合が多いが，今回用いる Contrastive Divergence(CD) では MCMC の手法を用いてはいるものの高速に計算を行うことができ，しかもその精度は経験的にも理論的にも良好であることが確認されている．

具体的な手法を下記に示す．まず，MCMC では任意の初期値 $\boldsymbol{v}_0, \boldsymbol{h}_0$ から，緩和に至るまでに十分な回数だけ遷移を行うことによって，$p(\boldsymbol{v}, \boldsymbol{h})$ に従ったサンプルを無数に生成し，その平均を取ることによって近似的に期待値を得る．しかしこの手法は非常に多くの回数の遷移を行う必要があるため，その計算速度は比較的遅いことが欠点であった．

一方，Contrastive Divergence では同じように MCMC を用いて遷移を行っているが，2 つの点が MCMC とは異なっている．まず，遷移させる回数は無数ではなく，A 回の非常に少ない遷移で構わない．大抵の場合は $A = 1$ 回で十分である．これは隠れ素子，可視素子間の結合が存在していないため，MCMC における緩和時間が比較的高速であるからである．次に，初期値には N 個のサンプルである $\{\boldsymbol{v}^1, \cdots, \boldsymbol{v}^N\}$ を用いる．この僅かなルールの付け足しによって，CD は高速かつ高精度に RBM の勾配を求めることができる．

なお，条件付き確率を解析的に求められるという利点から，MCMC による遷移は一般に Gibbs Sampling を用いて行われる．また，遷移によって用いるサンプルは $\boldsymbol{v}$ についてのみ用うことで確率分布 $\hat{q}(\boldsymbol{v}) = \frac{1}{N}\sum_k \delta(\boldsymbol{v} - \hat{\boldsymbol{v}}^k)$ を生成し，これと $\boldsymbol{h}$ についての条件付き確率 $p(\boldsymbol{h}|\boldsymbol{v})$ の積によって結合確率 $\hat{q}(\boldsymbol{v})p(\boldsymbol{h}|\boldsymbol{v})$ を表現している．

Contrastive Divergence の更新式

$A = 1$ の場合，Contrastive Divergence を用いた第二項の更新は下記の通りとなる．

1. 初期値に $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N\}$ をセットする．
2. 条件付き確率 $p(\boldsymbol{h}|\boldsymbol{v})$ を用いて，隠れ素子の状態 $\{\hat{\boldsymbol{h}}^1, \cdots, \hat{\boldsymbol{h}}^N\}$ を求める．
3. 2. で得た隠れ素子の状態を用いて，条件付確率 $p(\boldsymbol{v}|\boldsymbol{h})$ を用いて可視素子の状態 $\{\hat{\boldsymbol{v}}^1, \cdots, \hat{\boldsymbol{v}}^N\}$ を求める．
4. 3. で得た可視素子の状態と条件付確率から結合確率

$$p_{\text{model}}(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{N} \sum_k \delta(\boldsymbol{v} - \hat{\boldsymbol{v}}^k) p(\boldsymbol{h}|\hat{\boldsymbol{v}}^k) \tag{13}$$

を求め，計算を行うことによって近似的に第二項を得る．

4 パラメータの計算

以降の章では具体的なパラメータ b_i, c_j, W_{ij} についての計算を行い，実際に導出された式について書き下す．
まず，W_{ij} についての計算を行う．$\partial E/\partial W_{ij} = -v_i h_j$ より，

$$\frac{\partial J}{\partial W_{ij}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}} \tag{14}$$

ここで，第一項について，

$$\langle v_i h_j \rangle_{\text{data}} = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} v_i h_j q(\boldsymbol{v}) p(\boldsymbol{h}|\boldsymbol{v}) \tag{15}$$

$$= \sum_{\boldsymbol{h}} \left\{ \frac{1}{N} \sum_k v_i^k h_j p(\boldsymbol{h}|\boldsymbol{v}^k) \right\} = \frac{1}{N} \sum_k v_i^k \sum_{h_j} h_j p(\boldsymbol{h}|\boldsymbol{v}^k) \tag{16}$$

$$= \frac{1}{N} \sum_k v_i^k \sigma(c_j + \sum_{i'} v_{i'}^k W_{i'j}) \tag{17}$$

第二項も同様にして，

$$\langle v_i h_j \rangle_{\text{model}} = \frac{1}{N} \sum_k \hat{v}_i^k \sigma(c_j + \sum_{i'} \hat{v}_{i'}^k W_{i'j}) \tag{18}$$

以上より，W_{ij} についての勾配は

$$\frac{\partial J}{\partial W_{ij}} = \frac{1}{N} \sum_k v_i^k \sigma(c_j + \sum_{i'} v_{i'}^k W_{i'j}) - \frac{1}{N} \sum_k \hat{v}_i^k \sigma(c_j + \sum_{i'} \hat{v}_{i'}^k W_{i'j}) \tag{19}$$

によって表現できる．

次に，b_i について．$\partial E/\partial b_i = -v_i$ より，

$$\frac{\partial J}{\partial b_i} = \langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{model}} \tag{20}$$

ここで，

$$\langle v_i \rangle_{\text{data}} = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} v_i q(\boldsymbol{v}) p(\boldsymbol{h}|\boldsymbol{v}) = \sum_{\boldsymbol{v}} v_i q(\boldsymbol{v}) = \frac{1}{N} \sum_k v_i^k \tag{21}$$

ゆえに，b_i についての勾配は下記の通りとなる．

$$\frac{\partial J}{\partial b_i} = \frac{1}{N}\sum_k v_i^k - \frac{1}{N}\sum_k \hat{v}_i^k \tag{22}$$

最後に，c_j について．$\partial E/\partial c_j = -h_j$ より，

$$\frac{\partial J}{\partial c_j} = \langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{model}} \tag{23}$$

ここで，

$$\langle h_j \rangle_{\text{data}} = \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} h_j q(\boldsymbol{v}) p(\boldsymbol{h}|\boldsymbol{v}) = \sum_{\boldsymbol{h}} \left\{ \frac{1}{N} h_j \sum_k p(\boldsymbol{h}|\boldsymbol{v}^k) \right\} \tag{24}$$

$$= \frac{1}{N}\sum_k \sum_{h_j} h_j p(h_j|\boldsymbol{v}^k) = \frac{1}{N}\sum_k \sigma(c_j + \sum_{i'} v_{i'}^k W_{i'j}) \tag{25}$$

ゆえに，c_j についての勾配は下記の通りとなる．

$$\frac{\partial J}{\partial c_j} = \frac{1}{N}\sum_k \sigma(c_j + \sum_{i'} v_{i'}^k W_{i'j}) - \frac{1}{N}\sum_k \sigma(c_j + \sum_{i'} \hat{v}_{i'}^k W_{i'j}) \tag{26}$$

最後に，これらの計算結果をまとめることにしよう．

二値の Restricted Boltzmann Machine についての更新式

可視素子 v_i が二値 $\{0, 1\}$ しか取り得ない場合の，RBM のパラメータに関しての更新式は下記の通りとなる．

$$\frac{\partial J}{\partial W_{ij}} = \frac{1}{N}\sum_k v_i^k \sigma(c_j + \sum_{i'} v_{i'}^k W_{i'j}) - \frac{1}{N}\sum_k \hat{v}_i^k \sigma(c_j + \sum_{i'} \hat{v}_{i'}^k W_{i'j}) \tag{27}$$

$$\frac{\partial J}{\partial b_i} = \frac{1}{N}\sum_k v_i^k - \frac{1}{N}\sum_k \hat{v}_i^k \tag{28}$$

$$\frac{\partial J}{\partial c_j} = \frac{1}{N}\sum_k \sigma(c_j + \sum_{i'} v_{i'}^k W_{i'j}) - \frac{1}{N}\sum_k \sigma(c_j + \sum_{i'} \hat{v}_{i'}^k W_{i'j}) \tag{29}$$

ここで，N はサンプル数，$\hat{v}_i^k$ は可視素子のサンプル v_i^k を Gibbs Sampling によって A 回だけ遷移させた場合についてのサンプルを表している．通常の場合，A は 1 回だけで良い．

5 可視素子が連続的な値を取る場合の RBM

前節では可視素子が二値という離散的な場合における RBM の更新式を導出したが，可視素子が連続的な値を取る場合においては以前のエネルギー関数をそのまま用いてしまうと不都合が生じる．そのため，可視素子が連続的な値を取っても問題のないよう修正した RBM も存在する．その場合のエネルギー関数は下記の通りとなる．

$$E(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{2}\sum_i v_i^2 - \sum_i b_i v_i - \sum_j c_j h_j - \sum_i \sum_j v_i W_{ij} h_j \tag{30}$$

この場合，条件付き確率 $p(\boldsymbol{v}|\boldsymbol{h}), p(\boldsymbol{h}|\boldsymbol{v})$ は下記の通りである．

$$p(v_i|\boldsymbol{h}) = N(v_i; b_i + \sum_j W_{ij} h_j, 1) \tag{31}$$

$$p(h_j = 1|\boldsymbol{v}) = \sigma(c_j + \sum_i v_i W_{ij}) \tag{32}$$

ここで，$N(x; \mu, \sigma^2)$ はガウス関数である．$p(\boldsymbol{v}|\boldsymbol{h})$ がシグモイド関数でなく，正規分布の確率密度関数として変化したことに注意されたい[*1]．

幸いにもこの場合におけるパラメータの更新式は，Contrastive Divergence を行う際に用いる条件付き確率が正規分布に変化したことを除いて，二値の RBM と全く同等である．ただし，仮定している生成モデルが異なるため，連続値の場合において推測されたパラメータは二値のそれとは本質的に異なる．

[*1] これは観測したデータの分布がガウシアンであることを強制するものではない．観測データが従う分布は周辺分布 $p(\boldsymbol{v}) = \sum_{\boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h})$ それ自身であり，条件付き確率ではない．