

深層学習 3章後半

3章 事前学習とその周辺

復習3.3 自己符号化器による内部表現の学習

事前学習←自己符号化器の学習に用いる

自己符号化器を層ごとに貪欲学習して事前学習

- 3.4 確定的なモデル
- 3.5 確率的なモデル
- 3.6 Product of Experts の学習法としてのCD法
CD法 (Contrastive Divergence法) を中心

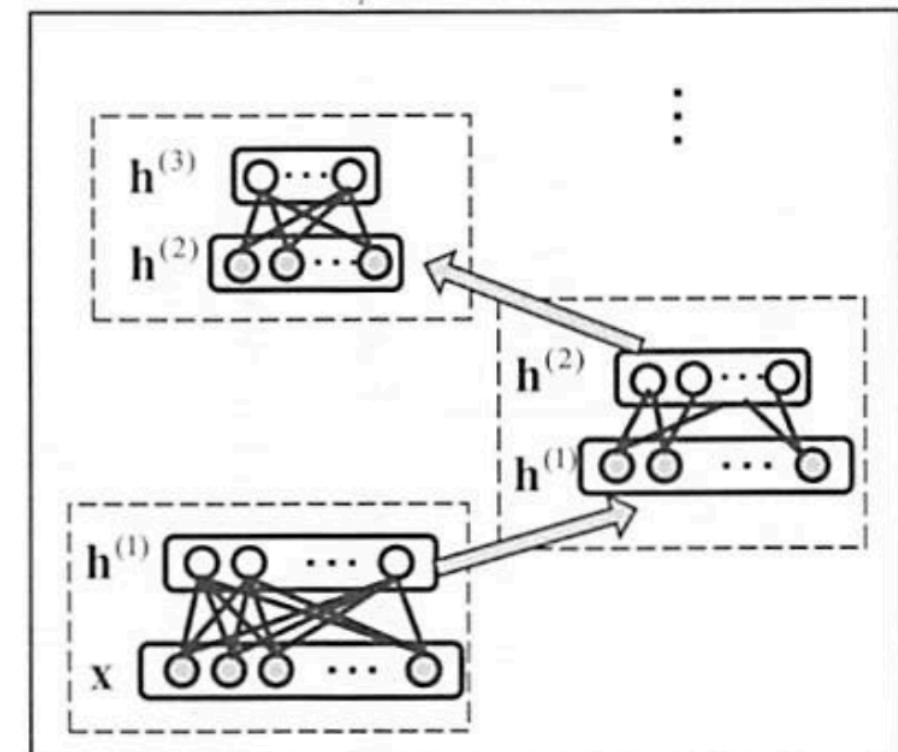
事前学習

- ・多層NNを隣接二層ごとに分解
- ・二層間でRBMなど小規模モデルを構成
- ・下層から順に学習

$p(X)$ の学習

各隣接層でEFH(RBM)

$$p(X = \mathbf{x}_t) = \sum_{\mathbf{h}_t} p(X = \mathbf{x}_t, H = \mathbf{h}_t) \text{ の学習}$$



3.4 確率的なモデルを用いた事前学習

確率的なモデルの一種: 制限ボルツマンマシン (RBM)

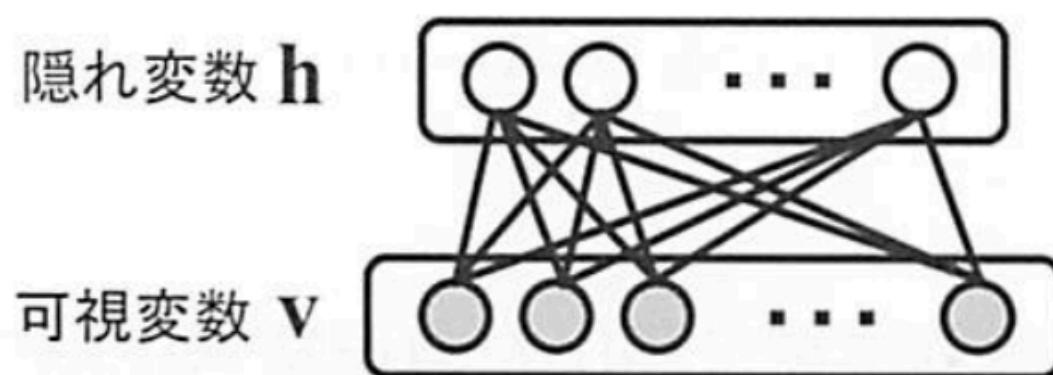


図 3.4 RBM や EFH のグラフィカルモデルによる表現

可視変数と隠れ変数との間にのみ結合があり、可視変数同士、隠れ変数同士には結合が存在しない。

3.4.1 RBM (制限ボルツマンマシン)

$$\mathbf{v} \in \{0, 1\}^m, \mathbf{h} \in \{0, 1\}^n$$

$$p(\mathbf{v}, \mathbf{h}|\theta) = \frac{1}{Z(\theta)} \exp \left(\mathbf{v}^\top \mathbf{W} \mathbf{h} + \mathbf{b}^{(1)\top} \mathbf{v} + \mathbf{b}^{(2)\top} \mathbf{h} \right)$$

$$\theta = \{\mathbf{W}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)}\}$$

θ を最適化する

RBM

規格化定数 v と h に依存しない

$$Z(\theta) = \sum_{v,h} \exp \left(v^\top Wh + b^{(1)^\top} v + b^{(2)^\top} h \right)$$

$$\sum_{v,h} p(v, h | \theta) = 1$$

となるように規格化定数を定める

RBM

- $b(1), b(2)$ の各要素
 - 正 : v, h の対応する要素が 1 を取りやすくする
 - 負 : 0を取りやすくする
- 結合重みの行列 W は v と h の相関関係を定める
 - W_{ij} が正 : v_i と h_j が同時に 1 を取りやすくなり, 負であれば取りにくくなる
 - W_{ij} がゼロならグラフ表記において v_i と h_j 間に線が無い

RBMの条件付き確率

- v同士やh同士が独立であることから

$$\begin{cases} p(\mathbf{v}|\mathbf{h}, \theta) &= \prod_{i=1}^m p(V_i = v_i | \mathbf{h}, \theta) \\ p(\mathbf{h}|\mathbf{v}, \theta) &= \prod_{j=1}^n p(H_j = h_j | \mathbf{v}, \theta) \end{cases}$$

さらにそれぞれの要素の条件付き確率分布は、

$$\begin{cases} p(V_i = 1 | \mathbf{h}, \theta) &= \text{sig} \left(\sum_{j=1}^n w_{ij} h_j + b_i^{(1)} \right) \\ p(H_j = 1 | \mathbf{v}, \theta) &= \text{sig} \left(\sum_{i=1}^m w_{ij} v_i + b_j^{(2)} \right) \end{cases}$$

RBMのhの期待値

期待値のベクトル表記

$$\bar{\mathbf{h}} = E_{p(\mathbf{h}|\mathbf{v})}[\mathbf{h}] = \text{sig}(\mathbf{W}^\top \mathbf{v} + \mathbf{b}^{(2)}), \bar{\mathbf{v}} = E_{p(\mathbf{v}|\mathbf{h})}[\mathbf{v}] = \text{sig}(\mathbf{W}\mathbf{h} + \mathbf{b}^{(1)})$$

3.3.1項で述べられた自己符号化器に用いられた多層NNの
決定論的関数と対応

3.4.2 指数型ハーモニウム族(EFH)

RBMは条件付き分布が独立
条件付き分布が独立になる分布を一般にEFHといい以下で表す

$$p(\mathbf{v}, \mathbf{h}|\theta) = \frac{1}{Z(\theta)} \exp(-\Phi(\mathbf{v}, \mathbf{h}|\theta))$$
$$\Phi(\mathbf{v}, \mathbf{h}|\theta) = -\sum_{i,j} \phi_{ij}(v_i, h_j | w_{ij}) - \sum_i \alpha_i(v_i | b_i^{(1)}) - \sum_j \beta_j(h_j | b_j^{(2)})$$

エネルギー関数

$$Z(\theta) = \sum_{\mathbf{v}, \mathbf{h}} \exp(-\Phi(\mathbf{v}, \mathbf{h}|\theta))$$

ギプスサンプリング

ギプスサンプリングをk回繰り返した後の分布を左辺として

$$p^{(k)}(\mathbf{v}, \mathbf{h}|\theta) = \sum_{\mathbf{v}'} \sum_{\mathbf{h}'} p(\mathbf{h}|\mathbf{v}, \theta) p(\mathbf{v}|\mathbf{h}', \theta) p^{(k-1)}(\mathbf{v}', \mathbf{h}'|\theta)$$

この式による分布の更新をギプスサンプリングという

3.4.3 指数型ハーモニウム族のCD法による学習

CD法 (Contrastive Divergence法)

- EFHの θ に関する勾配計算にEFHのモデル分布の期待値計算が必要→ギプスサンプリングで近似→計算時間が掛かり過ぎる
- ギプスサンプリングをk回（典型的にk=1）だけ行う
- CD-k法では以下の近似導関数を用いてパラメータ更新

$$\Delta w_{ij} \propto E_{p(h_j|v, \theta)q(v)} \left[\frac{\partial \phi_{ij}}{\partial w_{ij}} \right] - E_{p^{(k)}(v, h|\theta)} \left[\frac{\partial \phi_{ij}}{\partial w_{ij}} \right] \quad (3.13)$$

3.4.4 CD法が最適化している損失関数

- ・ギプスサンプリングは導関数を用いて対数尤度を最大化
- ・CD法も導関数の級数展開を打ち切った形なので対数尤度の最急勾配の良い近似
- ・CD法を損失関数を最適化する手法と解釈
ギプスサンプリングをk回行ったCD-k法の損失関数

$$F_{\text{CD-}k} \equiv D_{\text{KL}} \left(q(\mathbf{v}) \parallel p^{(\infty)}(\mathbf{v}|\boldsymbol{\theta}) \right) - D_{\text{KL}} \left(p^{(k)}(\mathbf{v}|\boldsymbol{\theta}) \parallel p^{(\infty)}(\mathbf{v}|\boldsymbol{\theta}) \right)$$

Detailed balance learning法(DBL法)

- RBMを含むEFHの学習法として利用できる
- マルコフ連鎖を定義する条件付き分布 $p(h|v, \theta)$ と
- そのパラメータ θ を学習する
- マルコフ連鎖の定常分布 $p(\infty)(v, h | \theta)$ が真の分布 $q(v)$ に近づくよう
にする

3.4.5 CD法と類似した学習則を与えるアルゴリズム

最小確率流法(MPF法)

1. 連續時間tのマルコフ過程を考える
2. 確率過程の定常分布を経験分布に近づける損失関数を導入

$$C(\theta) = D_{KL} \left(q(v) \parallel r^{(\epsilon)}(v|\theta) \right)$$

3. 詳細釣り合い条件と類似の制約下で損失関数を最小化

$$\Gamma_{ji}(\theta)p_i(\theta) = \Gamma_{ij}(\theta)p_j(\theta)$$

MPF法の利点

- ・損失関数の最小化から導かれるため
 - ・収束性がある
 - ・適切な学習率の調整が可能
- ・損失関数 $C(\theta)$ が凸関数で観測データ数に比例するオーダーで評価
- ・→学習が高速
- ・CD法では尤度が最大化されずに途中で収束してしまう問題が起こる場合があった

3.4.6 CD法から派生した学習則

- 繙続的CD法
 - 近似導関数(3.13)を対数尤度の勾配の式(3.12)に近づける
- CD法
 - 式(3.12)の第二項の期待値計算は解析的な計算が困難
 - ギプスサンプリング無限回で近似してサンプルを定常分布に収束
 - 近似導関数(3.13)をギプスサンプリングk回で代用
 - パラメータ更新のたびに初期分布を観測データからの経験分布に戻す
- 繙続的CD法
 - 更新後にサンプルで構成される経験分布からギプスサンプリング
 - 繙続的CD-1法 \geq CD-10法
 - 計算量が少ない

CD法から派生した学習則

- ・パラレルテンパリング=交換モンテカルロ法
 - ・マルコフ連鎖においてサンプル集団の偏りがなかなか解消されない
 - ・広い状態空間を探索させるため、複数の互いに異なる乱雑さをもつマルコフ連鎖を並列して用いる
- ・継続的CD法と併用することで、パラメータを上手く設定すれば継続的CD法より優れた学習性能
- ・欠点：設定するパラメータが増える

3.4.7 確率的なモデルの事前学習と自己符号化器の学習の関係

- 事前学習:
 - 後の教師あり学習の良い初期値を与えるため一般的なモデルで用いる
 - 多層NNを隣接二層ごとに分解
 - 二層でRBMなど小規模モデルを構成
 - 下層から順に学習
- EFHは同時分布 $p(v, h | \theta)$ が与えられる
- 式(3.21)の詳細釣り合い条件を満たすように、マルコフ連鎖の条件付き分布 $p(v|h, \theta)$ と $p(h|v, \theta)$ と θ を学習すれば、確率的自己符号化器が得られることがわかる

$$\forall v, v', h, p(v'|h, \theta)p(h|v, \theta)q(v) = p(v|h, \theta)p(h|v', \theta)q(v') \quad (3.21)$$

変分自己符号化器

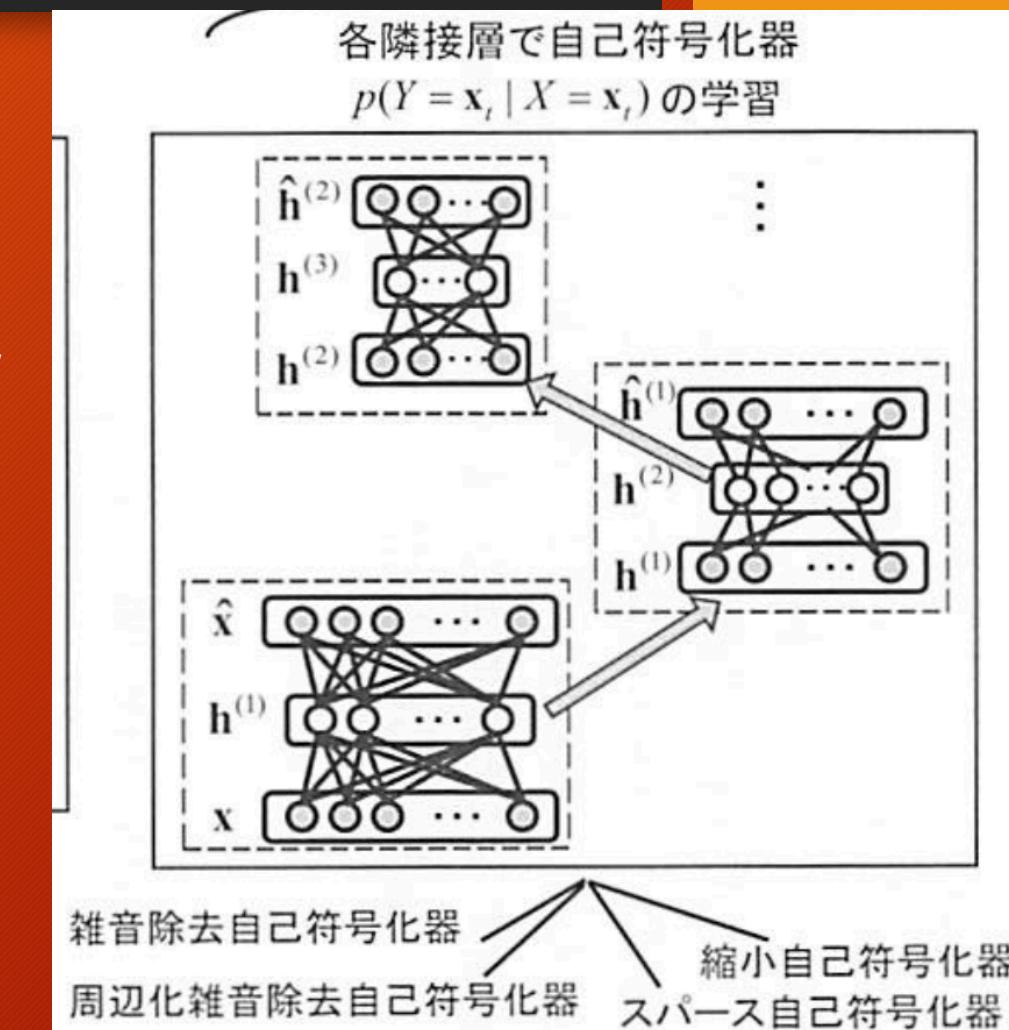
- ・事後分布 $p(h|v, \theta)$ が深層NNで近似
- ・符号化器の分布の近似が変分法による近似と形式的に同じ近似

3.5 確定的なモデルを用いた事前学習

- ・事前学習は次元削減によって後の深層NNの過学習を防ぐ
- ・確率的なモデルEFHは、決定論的な教師あり学習の損失関数とは間接的関係
- ・確定的なモデルは直接的関係があり、それを用いた事前学習がある
 - ・3.5.1 教師なし学習
 - ・3.5.2 教師あり学習

3.5.1 教師なし学習による確定的なモデルの学習

- ・ 積層自己符号化器による事前学習
 - ・ 二層ごとに自己符号化器を学習
 - ・ EFHの対数尤度の勾配法による学習則として、CD法と比べると、展開が一つ少なく、平均場近似が適用されておらず、劣る
 - ・ 計算機実験でもCD法と同等か少し劣る



雑音除去自己符号化器による事前学習

- ・自己符号化器の入力：観測データ x にノイズを加えた $x\sim$
- ・教師信号： x
- ・ x を復元するように学習→より効果的な特徴

ノイズ

- ・欠落雑音：ランダムに値をゼロにする
- ・ガウスノイズ
- ・ランダムノイズは観測データを擬似的に増やせる
→その分の計算コストが増大

周辺化雑音除去自己符号化器

- ・ノイズの計算コストの増大を解決する手法
- ・ノイズの加わった入力 $x\sim$ の平均の周りで損失関数の二次の泰ラ一展開をとる→解析的な近似計算
- ・不要になるもの
 - ・ノイズデータを多数作らなくて良い
 - ・入力データによるコスト関数やそのパラメータに関する導関数の平均の計算
- ・性能が従来の雑音除去自己符号化器と同等か優れている
- ・多層に積み重ねるときの性能向上が小さくなる

その他の自己符号化器による事前学習

- スパース自己符号化器
 - 自己符号化器において中間層の次元を高めると、ゼロが多くスパースに
- 縮小自己符号化器
 - 正則化項として入力 v から特徴 h を出力する関数の滑らかさを促進する項を付与→雑音除去符号化器と同等
- 極端学習機械(ELM)

極端学習機械(ELM)

- ・多層にしないことで損失関数を凸関数にし学習（最適化）を容易に（局所解が無い）
- ・一層の隠れ変数の層
- ・入力と隠れ変数の間の重みWやバイアスbはランダムに生成
- ・Wやbを推定する問題が、単に二乗誤差の損失関数の下で線形回帰に
→解析的に求められる
- ・隠れ変数のノードを増やして、任意の有界な区分連続関数を近似できる
- ・SVMと比べて大量のデータでも学習がはやすく、学習後の分類器の性能が良い
- ・層ごとに自己符号化器による事前学習を行って高性能化

3.5.2 教師あり学習による確定的なモデルの学習

- ・ ここまで入力をそのまま出力する教師なし学習をする自己符号化器
- ・ →教師あり学習を自己符号化器に使おう（音声分野の識別的事前学習）
- ・ 画像認識において良くない性能
 - ・ 三層NNで特徴が表現できなければ学習できないためか
- ・ 音声認識ではDBNによる事前学習と同等

3.6 PoEの学習法としてのCD法

- PoE(Product of Experts) エキスパート関数fの積

- 尤度

$$p(\mathbf{v}|\theta) \propto \prod_{i=1}^M f_i(\mathbf{v}|\theta_i)$$

- 前の式で定義されたEFHはPoEの一種