Young Scientists Meeting on Statistical Physics and Information Processing 2010

Deep Learning and Restricted Boltzmann Machine

安田　宗樹
東北大学情報科学研究科　応用情報科学専攻
muneki@smapip.is.tohoku.ac.jp

背景

近年の情報科学の研究・応用の場において機械学習理論によるデータ学習の手法は大きな位置を占めるようになってきている．本講演では得られたデータの “分布” を学習する確率的機械学習モデルに注目する．確率的機械学習モデルは例えば Support Vector Machine のような通常のカーネルマシンなどとは異なりモデルそのものが確率として表現されているため，一般にはそれに応じた統計的な最適化法を必要とする．

確率的機械学習モデルは分類器というよりもむしろベイズの枠組みに則った自動認識や画像処理などのための事前確率モデルとして多く用いられる．例えばノイズや落書き欠損等で劣化した画像をベイズの枠組みで修復する問題を考える．劣化する前の画像を $\mathbf{I}$ とし，劣化した受信画像を $\mathbf{D}$ とするとベイズの枠組みから修復画像 $\mathbf{R}$ は $\mathbf{R} = \arg\max_{\mathbf{I}} P_{\mathrm{LH}}(\mathbf{D} \mid \mathbf{I}) P_{\mathrm{prior}}(\mathbf{I})$ により定式化される．ここで $P_{\mathrm{LH}}(\mathbf{D} \mid \mathbf{I})$ は尤度に相当する（例えば○○%の確率で画素が劣化するなどの）劣化過程の確率モデルであり，扱うシステムに応じて適宜モデル化される．$P_{\mathrm{prior}}(\mathbf{I})$ が事前確率であり，言わば，“世の中の画像が従う確率分布” をあらわしており，これの適切な設計が困難なのは想像に難くないであろう．これまでこの事前確率の設計には非常に多くの取り組みがなされてきているが，最近著しい発展を遂げているのが確率的機械学習モデルによる事前分布の設計である．膨大な量の画像データをモデルに学習させ，現実の画像に則した事前確率を設計しようとするこの取り組みは実際従来の手法に比べて高品質な修復を与えている [1].

パラメトリックな確率的機械学習モデル $P(\mathbf{v} \mid \mathbf{\Theta})$ の教師あり学習の一つは以下のように定式化される．M 個の観測データ $\{\mathbf{D}^1, \cdots, \mathbf{D}^M\}$ を得たとするとこのデータセットに対するヒストグラムは

$$P_D(\mathbf{v}) \triangleq \frac{1}{M} \sum_{\mu=1}^{M} \delta\left(\mathbf{v} - \mathbf{D}^{\mu}\right)$$

とあらわされる．この分布はしばしば “経験分布” とよばれる．このデータに対して

$$\mathbf{\Theta}^* = \arg\max_{\mathbf{\Theta}} \sum_{\mathbf{v}} P_D(\mathbf{v}) \ln P(\mathbf{v} \mid \mathbf{\Theta}) = \arg\min_{\mathbf{\Theta}} \sum_{\mathbf{v}} P_D(\mathbf{v}) \ln \frac{P_D(\mathbf{v})}{P(\mathbf{v} \mid \mathbf{\Theta})}$$

が（対数）尤度最大化または Kullback-Leibler Divergence 最小化による学習の定式化である．これよりこの学習則は “モデルと経験分布を尤も近くするようなパラメータ $\mathbf{\Theta}$ を求めること” と解釈できる．とくにモデルを指数分布族にすると上記の学習則はモデルの十分統計量のモーメントマッチングの問題となり，この観点から（分布から統計量を求める問題を順問題とすれば）逆問題とよばれることもある．本講演では指数分布族のパラメトリックな機械学習モデルの一つであるボルツマンマシンを扱う．ボルツマンマシンは統計力学でいうところの 2 体相互作用をもつスピン系の物理モデルと等価な数理をもっており，平均場理論などの物理学由来の方法論が大きく効果を発揮する機械学習モデルの一つである．

Restricted Boltzmann Machines and Deep Belief Networks

先に述べた通り，ボルツマンマシンは指数分布族の統計モデルであるため，尤度最大化による学習はモーメントマッチングの問題となる．つまりボルツマンマシン（スピンモデル）の統計量を計算しなくてはならないが，この課題は一般には計算困難な問題となっている．このような背景からボルツマンマシンの実用はあまり現実的ではないと考えられてきた．

しかし近年になり restricted Boltzmann machine (RBM) とよばれる特殊な構造をもつボルツマンマシンが注目され，その事情が少々変わりつつある．RBM は次のように定式化される．

$$P_{\mathrm{RBM}}(\mathbf{v},\mathbf{h}\mid\boldsymbol{\Theta}) \triangleq \frac{1}{Z(\boldsymbol{\Theta})}\exp\left(\sum_{i\in V}\alpha_i v_i+\sum_{j\in H}\beta_j h_j+\sum_{i\in V}\sum_{j\in H}W_{ij}v_i h_j\right),\quad \boldsymbol{\Theta}=\{\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{W}\}.$$

ここで確率変数 $\mathbf{h}$ は隠れ変数 (hidden variable) であり，$\{V,H\}$ はそれぞれ可視変数（データに対応する変数），隠れ変数のラベルの集合としている．$P_{\mathrm{RBM}}(\mathbf{v},\mathbf{h}\mid\boldsymbol{\Theta})$ は図 1 に示す通り可視変数の層と隠れ変数の層からなる 2 部グラフの構造をとる．ただし層内の結合はない．RBM の顕著な性質の一つとして，片方の層をある状態に固定するともう片方の層に含まれる変数は互いに独立となる（条件付き独立）といった性質がある．RBM は products of experts (PoE) とよばれる積混合分布モデルの一つである．PoE に対する条件付き独立の性質を巧みに用いた contrastive divergence (CD) learning とよばれる効率的な学習アルゴリズムの発見 [2] により RBM の実際のアプリケーションへの利用が一挙に現実性を帯びた．また条件付き独立の性質は学習のみならず推論のフェーズにおいても有効に効く．

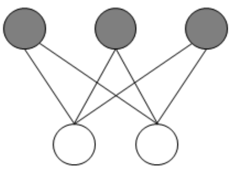

図 1: RBM の例．可視層（下段）と隠れ層（上段）の 2 層構造である．

隠れ変数の存在はより複雑な確率モデルを構築する目的において非常に重要である．学習において興味ある分布はあくまで可視変数の分布（データの分布）であり，RBM の学習の目的は隠れ変数について周辺化した分布の尤度を最大にすることである：

$$\arg\max_{\boldsymbol{\Theta}}\sum_{\mathbf{v}}P_D(\mathbf{v})\ln\sum_{\mathbf{h}}P_{\mathrm{RBM}}(\mathbf{v},\mathbf{h}\mid\boldsymbol{\Theta}). \tag{1}$$

RBM の結合確率は 2 体相互作用までのエネルギー関数で記述されているが，この周辺化によりシステムは 2 体相互作用を超える “高体相互作用の寄与” を含むことができる．これが RBM における隠れ変数のもつ意味である（PoE としてみると隠れ変数の数はそのままコンポーネント数に対応しており，より隠れ変数の意義が明確である.）．

隠れ変数の数を増やせばそれだけ分布の表現能力は増す．したがって RBM において隠れ変数の数を増やしていくのも一つの方針であるが，それとは別に RBM を積み重ね多層のネットワークを構成し，表現能力を向上させるといった方針もある．この方針により構築されたのが deep belief net (DBN) である [3]．DBN は最下層がデータ変数の層（$\mathbf{v}$ の層）であり，残りはすべて隠れ変数で構成される層で構成される．

本講演では確率的機械学習モデルの学習の基礎から最近注目されている RBM・DBN までを物理的・情報科学的立場から概観する．

参考文献

[1] S. Roth and M. J. Black: Fields of experts, *International Journal of Computer Vision*, vol.82, no.2, pp.205–229, 2009.

[2] G.E. Hinton: Training products of experts by minimizing contrastive divergence, *Neural Computation*, vol.14, pp.1771–1800, 2002.

[3] G.E. Hinton, S. Osindero and Y.W. Teh: A fast learning algorithm for deep belief nets. *Neural Computation*, vol.18, pp.1527–1554, 2006.