

テンソルツリーによる生成モデル構築のスキーム

数値物性科学研究施設 川島 直輝

1. はじめに

人工知能流行りである。猫も杓子も AI だ。もちろん、これは単なる流行ではなく、社会が劇的に変化していく最初の様子を目撃しているわけである。100 年後、1000 年後にも歴史というものに意味があるとなると、現在は人類史上の大転換点という位置付けになりそうだ。昨今の人工知能は、おおむねネットワーク型の機械、あるいは、それをエミュレートする機械による機械学習であり、ホップフィールドモデルやボルツマンマシンなどのスキームは、統計力学で研究されるスピングラスモデルと多くの共通点をもっていることから、機械学習の理論的研究においては、統計力学で開発された方法や概念が有用である。AI 研究の進展と比較すると、統計力学の方法論の進化はかなりゆったりとしたものであるが、我々に近い研究サークルではそのゆったりした進化にのって、最近、テンソルネットワーク (TN) と呼ばれる手法をさまざまな統計力学モデルや、物性理論モデルに対して開発し、応用してきている。そのような研究においては、システム内の物理的相関を一種の情報伝達として捉えるくせが自然と身についてくる。その自然な流れから、TN を物性科学に限らない一般の機械学習にも応用してみたらどうだろうか、という好奇心に駆動され、杓子ぐらいにはなれるかもしれないということで、TN の機械学習への応用研究 [1]、[2] を行った。

2. 方法

使った方法は上記のように TN 法である。TN は、単に統計力学モデルを多数のテンソルからなるテンソル積の縮約だとみなす、というものの見方にすぎないので、これ自体は方法ではない。そのようにみなすことの利点は、データ (統計力学モデルの場合は自由度) の圧縮近似をテンソルの特異値分解として自然に表現できることである。さまざまな圧縮操作を利用することで、大自由度系からスケール次元などの有用な情報をシステムティックに抽出できるようになった。一般の固体物理では、ガラス系などを除くと、空間構造にそれほど複雑なものは登場しないため、ネットワーク構造自体を最適化する必要はあまりないが、一般のデータには構造がある場合が多い。そこで、[1] では、ネットワークの構造自体を相互情報量の観点から

自動的に最適化する手法を考案した。構造最適化がしやすいように、ネットワークとしてはループをもたないツリーとし、適応型テンソルツリー (ATT) と呼ぶことにした。[1]で行ったのは、多自由度システムの状態出現確率を ATT で表現すること、つまり、いわゆる ATT で生成モデルを構成することであった。確率分布なので値が非負である必要があるが、非負であれという条件のもとでの最適化には技術的な困難が伴う。そこで、[1] では、ATT 自体に拘束条件をつけなくても良いように、ATT が表す数が直接確率分布を表すのではなく、その二乗が確率分布になるような条件で最適化を行なった。そのように構成した学習機械を一般にボルンマシンと呼んでいる。その結果、たとえば、アメリカ株式銘柄 S&P500 などの株価の騰落パターンをうまく学習することができ、その副産物として業種によるクラスタリングなどの構造が再現された。しかし、この研究で問題だったのは、学習の結果得られたネットワークを因果関係のネットワークと解釈できないことだった。全体は二乗するために正になるが、部分的なネットワークを取り出すと負になってしまうため、それを条件付き確率と解釈することができず、自由度間の局所的な関係を導くことができなかったからである。そこで、[2] では、局所的にも正であるという条件つきで最適化を行う方法を提案した。この修正によって、最終的に得られたネットワークを因果関係のネットワークと解釈できるようになった。

3. 具体例

バイオインフォマティクスから得られた実データを用いて、食肉目 (犬、猫、および関連種) の分類群に属する $n (=16)$ 種の生物の系統樹を構成した。そのために、長さ $L (=1,140)$ のミトコンドリア DNA 配列を、 L セットのデータと見る。つまり、対象とする第 s の種 ($s = 1, 2, \dots, n$) の DNA 配列で、 i 番目 ($i = 1, 2, \dots, L$) の位置にあるヌクレオチドの種類を $x_i^{(s)} = 1, 2, 3, 4$ として、 $X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)})$ を i 番目のデータセットと見なす。その結果 L 個のデータセットが得られるが、これを生成モデルにおける学習データとして学習させる。学習機械に与えるのはそれだけで、それぞれのノード s が、どのような種であるかに関する情報は与えない。図は、その結果得られたネットワークで

