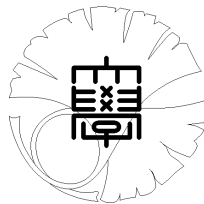


数理科学実践研究レター 2021-6 August 25, 2021

機械学習の表現能力について

by

高松 哲平



UNIVERSITY OF TOKYO
GRADUATE SCHOOL OF MATHEMATICAL SCIENCES
KOMABA, TOKYO, JAPAN

機械学習の表現能力について

高松哲平¹ (東京大学大学院数理科学研究科)

Teppei Takamatsu (Graduate School of Mathematical Sciences, The University of Tokyo)

概要

機械学習は、計算機に効率よく問題を解かせるためのアルゴリズムの研究である。特に、本研究では、オートエンコーダやランダムフォレストといった機械学習を考える。本研究の目的は、そうした機械学習の表現能力が、異なる手法や異なる構造を採用した際にどのように変化するかを考えることである。

1 はじめに

機械学習は、計算機に効率よく問題を解かせるためのアルゴリズムの研究であり、多くの手法が知られている。特に、深層学習の一部であるオートエンコーダや、ランダムフォレストといった手法は有名である。本研究では、まず、[LAG⁺20]の結果を一般化し、活性化関数がランプ関数よりも多項式的な状況で、オートエンコーダの表現能力を測る指標について考える。また、ランダムフォレストに対しても、[KA19]の与えた表現能力とノード数および木数との関係を述べる定理の変種を証明する。

2 オートエンコーダ

機械学習のアルゴリズムの一つであるオートエンコーダは、データを特徴づける量の次元を圧縮する手法である。現在では、データのノイズの除去や、生成モデルの構成に用いられている。本節では、オートエンコーダの表現効率について考える。

2.1 オートエンコーダの定義

定義 1 オートエンコーダとは、写像 $\phi: U \rightarrow V$ および $\psi: V \rightarrow W$ の族のことである。ここで、 U, V, W は有限次元 \mathbb{R} 線形空間で、 $n = \dim U = \dim W > m = \dim V$ をみたす。また、 ψ は変動するアファイン変換であり、 ϕ は変動するアファイン変換 ϕ' と固定された非線形写像 $\sigma: V \rightarrow V$ (活性化関数と呼ばれる) の合成 $\sigma \circ \phi'$ で与えられる。

実際のオートエンコーダでは、考える対象にあわせて ϕ' および ψ を調整していく。典型的な具体例が、以下のケースである。

例 2 U および W が写真全体の空間とする。また、 $\Sigma \subset U = W$ を、人間の写真のなす空間とする。このとき、オートエンコーダは、 $\psi \circ \phi$ が Σ 上恒等的になるように訓練される。

本研究で考えるのは、[LAG⁺20]で提唱された以下の問題である。

- 1) オートエンコーダの構造がどれだけよいのかを測る指標は存在するか？
- 2) 特徴量を得たい対象 Σ の複雑さを測る指標は存在するか？

[LAG⁺20]では、**ReLU DNN** (Rectified Linear Unit Neural Network) の場合、すなわち活性化関数 σ がランプ関数の場合に、正規化線形複雑性という概念を導入し、これらの問題に回答を与えた。彼らの結果は、言わば、 ϕ が区分的に線形な場合の結果であるが、本研究では、 ϕ が線形性を多項式に置き換えた場合を考える。

¹tepei@ms.u-tokyo.ac.jp

2.2 主結果

同型 $V \simeq \mathbb{R}^n$ を一つ固定する．定数でない多項式 $f(x) \in \mathbb{R}[X]$ (たとえば, $f(x) = x^c$) を固定し, 以下の活性化関数を考える．

$$\sigma : \mathbb{R}^n \rightarrow \mathbb{R}^n; x \mapsto \max\{0, f(x)\}.$$

以下が主結果である．

定理 3 活性化関数として上記 σ を採用したオートエンコーダ A を考える．また, $\Sigma \subset U$ を m 次元部分多様体とする．この時, $A = (\phi, \psi)$ および Σ の複雑性 $c(A) \in \mathbb{Z}$ および $c(\Sigma) \in \mathbb{Z}$ が定義され, 以下を満たす．

- 1) もしも ϕ が Σ 上位相同型を与えるならば, $c(\Sigma) \leq c(A)$ である．
- 2) $c(A)$ は U, V, W の次元のみによる上界を持つ．

証明 $c(A)$ を, 以下の条件を満たすような U の分解の, 成分の個数の最小値として定める．

(条件): U の超曲面による分解であり, 各成分上で $\psi \circ \phi$ は成分ごとに多項式になっている．

また, $c(\Sigma)$ は, 以下の条件を満たすような Σ のアトラス $(U_\alpha, \phi_\alpha)_{\alpha \in \mathcal{A}}$ についての $\#\mathcal{A}$ の最小値で定める．

(条件): $\phi_\alpha : U_\alpha \rightarrow \mathbb{R}^m$ は, 成分ごとに多項式写像であるような $\mathbb{R}^n \rightarrow \mathbb{R}^m$ に延長される．

このとき, 1) および 2) は [LAG⁺20] と同様の議論から従う．

本結果は, 上記の活性化関数によるオートエンコーダもまた, 表現能力に限界が存在することを示唆する．一方で, $c(A)$ の定義を [LAG⁺20] と比較してみると, 上記の活性化関数が, ランプ関数よりも表現能力に秀でている可能性も示唆している．

3 ランダムフォレスト

本節では, ランダムフォレストという機械学習アルゴリズムを考える．ランダムフォレストは, 決定木いくつかの多数決により結論を出力するものであり, 分類や回帰などに広く使われている．

3.1 ランダムフォレストの定義

以下が二分的なランダムフォレストの定義である．

定義 4 空間 \mathbb{F}_2^m を固定する． x_i を i 番目の射影 $\mathbb{F}_2^m \rightarrow \mathbb{F}_2$ とする．ランダムフォレスト RF とは, n 個の二分決定木 DT_1, \dots, DT_n の組である．ただし, n は正の奇数とする．

ここで, 二分決定木は, 二分木であり, 各分岐に条件 ($x_i = 0$ か $x_i = 1$ で分岐させることができる), 各葉にラベル (\mathbb{F}_2 の元) を与えられるものである．条件およびラベルが与えられた二分決定木 DT_j は, 条件に沿って二分木をたどり, 葉のラベルを対応付けることで, 写像 $DT_j : \mathbb{F}_2^m \rightarrow \mathbb{F}_2$ を誘導する．このとき, ランダムフォレスト RF は, 多数決により, 以下の写像を導く．

$$RF : \mathbb{F}_2^m \rightarrow \mathbb{F}_2 : x \mapsto \begin{cases} 0 & \#\{j | DT_j(x) = 0\} \geq (n+1)/2 \text{ の時} \\ 1 & \text{その他の時} \end{cases}$$

RF, RF' をランダムフォレストとする． RF の導く任意の写像 $\mathbb{F}_2^m \rightarrow \mathbb{F}_2$ を, RF' も導くとき, RF' は RF を表現するという．[KA19] は, ランダムフォレストの表現可能性と, ノード数および木数との関係を調べた．本研究の目標は, [KA19] の定理の変種を考えることである．

3.2 主結果

定理 5 n, T を正の奇数とする. RF を n 本の決定木, $O(n)$ 個の頂点からなるランダムフォレストとする. また, RF' を T 本の木からなるランダムフォレストとする. T, n を無限に飛ばした時, RF' が RF を表現するためには, RF' に少なくとも $\Omega(T(2^n/\sqrt{n})^{2/(T+1)})$ の頂点が必要である. ここで, Ω は, 定数倍の差を除いた漸近的な一致を意味する.

証明 T を固定して議論している [KA19] の証明の不等式を, 少しだけ強めることで証明される. 具体的には, [KA19, 定理 5] の議論により, 以下の [KA19, 補題 4] の変種を証明すれば十分である.

補助定理 6 $X \subset \mathbb{F}_2^m$ およびラベル $f: X \rightarrow \mathbb{F}_2$ が, 以下の条件を満足する時, X を T 本の木からなるランダムフォレストで識別するためには, 少なくとも $\Omega(T(\#X)^{2/(T+1)})$ 個の頂点が必要である. (条件): f を導く任意の (条件, ラベル付き) ランダムフォレスト RF に対して, もし $x \neq y \in X$ が $c := f(x) = f(y) \in \mathbb{F}_2$ を満たすならば,

$$\#\{RF \text{ の葉 } l \mid l \text{ のラベルは } c \text{ で, } x, y \text{ は決定木で } l \text{ に送られる}\} < (T+1)/2$$

が成立する.

補助定理の証明を述べる, X を識別する, T 本の木からなるランダムフォレスト $RF' = (DT_1, \dots, DT_T)$ をとる. X_0 を X の元でラベル 0 を持つものとする. L を RF' の葉全体の集合とし, L_0 をラベル 0 が与えられた葉の集合とする. また, L_0 の中で, DT_i に含まれる葉の集合を $L_{0,i}$ とおく. $l_0 := \#L_0$ とおく. 任意の X_0 の元 x は, RF' において $(T+1)/2$ 本以上の木でラベル 0 の葉に到達するので, そのうち $(T+1)/2$ 本を選ぶことで, 写像

$$X_0 \rightarrow \bigsqcup_{1 \leq i_1 < \dots < i_{(T+1)/2} \leq T} L_{0,i_1} \times \dots \times L_{0,i_{(T+1)/2}}$$

を得る. (条件) より, この写像は単射である. 一方右辺は, 簡単な計算より,

$$T C_{(T+1)/2}(l_0/T)^{(T+1)/2} \sim \frac{2^T}{\sqrt{T}} \left(\frac{l_0}{T}\right)^{(T+1)/2}$$

以下になることがわかる. したがって,

$$l_0 \geq T(\#X_0)^{2/(T+1)}$$

を得る. ラベル 1 についても同様の議論を行うと,

$$l \geq T(\#X)^{2/(T+1)}$$

を得て, 証明が完結する. この定理は, 特に $n = T$ の場合に最良の評価を与えていることから, いくらかのケースではよい評価を与えていると考えられる. しかし, 一般によりよい評価があるかどうかは, わかっていない.

4 終わりに

本論文では, オートエンコーダおよびランダムフォレストの表現能力の考察を行った. オートエンコーダに関する考察 (定理 3) においては, 本論文で導入した活性化関数が, 従来のランプ関数よりもよいものである可能性が示唆されていた. このことに実例を与えることが, 今後の大きな課題の一つといえる. また, その他の活性化関数を採用した場合にどういった指標が考えられるか, も考える価値のある問題である. 一方, ランダムフォレストの考察 (定理 5) については, 評価の精密化, および下からの評価を得ることが課題となる. 今後は, 適宜実装技術を学びながら, こうした課題を考えていきたい.

謝辞

本研究にあたり, 多くの有益な助言をくださった鮑先生と株式会社ニコンの皆様へ心より感謝申し上げます.

参考文献

- [KA19] S. Kumano and T. Akutsu, *On the trade-off between the number of nodes and the number of trees in random forest*, Japanese Society for Artificial Intelligence **JSAI2019** (2019).
- [LAG⁺20] N. Lei, D. An, Y. Guo, K. Su, S. Liu, Z. Luo, S.-T. Yau, and X. Gu, *A geometric understanding of deep learning*, Engineering **6** (2020), no. 3, 361 – 374.