シャープレイ値の公理と
深層ニューラルネットワークへの応用

本田あおい

九州工業大学 情報工学研究院

深層ニューラルネットワークは現在の第三世代 AI ブームを牽引する人工知能技術として重要な役割をはたしている。しかしながら実用化の障壁となるようないくつかの問題があり、その一つが推定の根拠の説明や解釈ができないという、いわゆるブラックボックス問題である。例えば医療分野や人事評価のように、いくら精度の良い推定ができてもその推定の根拠や理由が説明できないのであれば社会実装が難しい分野が多々ある。推定の根拠を説明することはモデルの原理を解明することでもあるので、より高性能な新技術への開発にもつながる。このような要請のため、近年では深層学習ネットワークを解釈するための様々な手法が提案されている。

シャープレイ値は協力ゲームから算出される統計量で自然な公理に特徴づけられた最も合理的な解概念であると言える。近年, 機械学習の分野でシャープレイ値が再発見され、シャープレイ値を使って機械学習モデルを解釈する SHAP(SHapley Additive exPlanations) と呼ばれる技術が大注目されているところである。

本講演では、ニューラルネットワークの解釈手法としてのシャープレイ値の利用例を 2 つ紹介する。一つ目はファジィ積分型深層ニューラルネットワークに対してモデルの重みから直接影響度を読み取る Model-Specific Approach な手法である。ネットワーク自体が積分を表現したものになっているので、重み＝ファジィ測度から直接シャープレイ値を算出することができ、これによって入力変数の貢献度を測るものである。もう一方は前述の Lundberg らが提案した SHAP(SHapley Additive exPlanations) である [1]。入出力の関係から影響度を算出する Model-Agnostic Approach な手法である。Lundberg らのこの論文は被引用件数 5000 を超える (2022 年 6 月現在) 大ヒット論文となっている。どちらもシャープレイ値の計算手法を利用して、入力の各成分が出力に与える影響を成分間の交互作用を考慮した形で算出するものであり、前者は比較的計算量が少ない、後者はニューラルネットワークのモデルに依存せず他の機械学習手法にも適用できるといった長所を持つ。それぞれの逆が短所である。

積分型ニューラルネットワークに用いる包除積分や、これを利用した機械学習モデルは、筆者らが提案したもので、こちらも合わせてご紹介します [2, 3]。

以下、本講演で用いる記号や定義を示す。プレイヤーや説明変数の集合を $X = \{1, 2, \ldots, n\}$ 、これらの提携集合全体を 2^X、そしてこの上に定義される協力ゲー

ムやファジィ測度 (協力ゲームに単調性の仮定を加えたもの) を v で表す。

Definition 1 (シャープレイ値) 協力ゲーム $v : 2^X \to \mathbb{R}$ に対して次で定義される n 次元ベクトル $\Phi(v) := (\phi_i(v), \ldots, \phi_n(v))$ をシャープレイ値と呼ぶ.

$$\phi_i(v) := \sum_{A \subset X \setminus \{i\}} \frac{(n - |A| - 1)!|A|!}{n!} v(A \cup \{i\}) - v(A).$$

包除積分はファジィ積分の 1 種で、ルベーグ積分やショケ積分を特別な場合として含む。定義が包除原理の離散構造に基づいており、コンピュータアルゴリズムとの親和性が高い。

Definition 2 (包除積分) $v : 2^X \to [0, \infty)$ をファジィ測度、 $x = \{x_1, x_2, \ldots, x_n\}$ を n 次元実数値ベクトルとする。x の v に関する包除積分は次で定義される.

$$\otimes \int x dv := \sum_{A \subset X} \left(\sum_{B \supset A} (-1)^{|B \setminus A|} \bigotimes_{i \in B} x_i \right) v(A),$$

ただし $\bigotimes$ は X 上の掛け算型の多項演算。

包除積分は v のメビウス変換 m^v を用いて

$$\otimes \int x dv = \sum_{A \subset X} \left(\bigotimes_{i \in B} x_i \right) m^v(A)$$

と、交互作用項つきの重回帰分析と類似した形に別表現できる。数理モデルではこちらを利用する。ただし、メビウス変換は

$$m^v(A) := \sum_{B \setminus A} (-1)^{|A \setminus B|} v(B)$$

で定義される集合関数である。

参考文献

[1] Scott Lundberg, Su-In Lee, A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems, 2017.

[2] A. Honda,Y. Okazaki, Theory of inclusion-exclusion integral, Information Sciences, 376, 136–147, 2017.

[3] A. Honda, Y. Kamata, S. James, Representation and Interpretability of IE Integral Neural Networks, Lecture Notes in Computer Science, to appear.