第2章 業務に必要な判断の根拠を示す

大規模言語モデルを使った説明可能AI

山崎 貴史

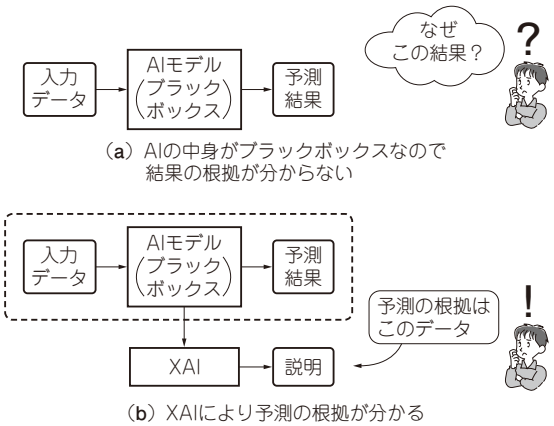

図1 AIによるブラックボックスな予想とXAIによる説明

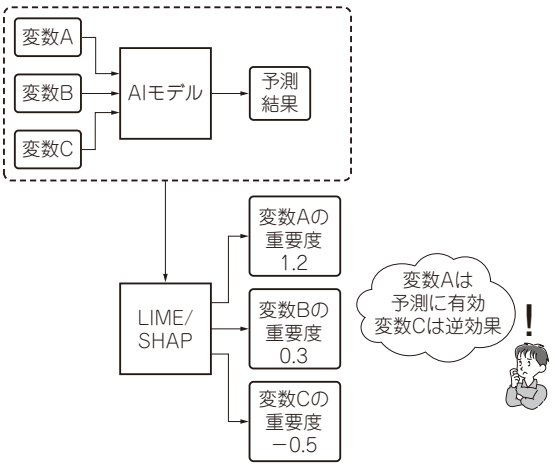

図2 変数ごとの重要度を数値化するLIME/SHAP

XAI（説明可能AI）とは何か

● AIのブラックボックス性を解決するXAI

ディープ・ラーニングは理論的背景の分からないデータを処理するために発展した手法です．データ同士の関係を直接学習することで高精度の予測が可能になります．

その一方で，どのような予測結果が出力されるのか分からないという性質（ブラックボックス性）があり，結果の分析やモデルの改良を妨げる要因として問題視されています．

そのような問題に対処するため，AIの処理内容や予測結果に人間が解釈可能な説明を与える技術が長年研究されており，それらの手法はXAI（eXplainable AI；説明可能AI）と呼ばれています（**図1**）．

近年ではLLM（大規模言語モデル）の発展により新たなXAIの在り方が期待されています．ここではXAIの基本から始めてLLMを使ったAIモデルの振る舞いを説明する方法について解説します．

● XAIは定量的と定性的な説明がある

XAIは説明対象となるAIモデルの入出力や中間の計算結果に処理を加えて，解釈しやすい形に変換する手法です．大きく分けて重要度などを数値として算出する定量的説明と，画像や言語による直観的に理解しやすい情報を提示する定性的説明の2種類があります．

▶定量的説明手法LIME/SHAP

LIME[(1)]，SHAP[(2)]と呼ばれる手法は定量的説明を与えるXAI手法です（**図2**）．入力データの各数値が予測結果にどれだけ影響を与えているかを数値化し，予測の根拠となる変数や，逆に予測に反するように働いている変数を検出するのに有用です．

▶定性的説明手法GradCAM

定性的説明を行う代表的な手法としてはGradCAM[(3)]があります．画像用のディープ・ラーニング・モデルであるCNN（畳み込みニューラル・ネットワーク）に適用できるXAIとして知られています．

GradCAMは入力画像の各ピクセルの重要度を数値化するという意味では定量的説明の手法です．重要度をヒートマップとして可視化することで，画像の重要な領域を視覚的に示して直観的な解釈が可能になります．**図3**はGradCAMを使った反実仮想説明と呼ばれる例です．例えば**図3（b）**は，画像内の猫らしくない部分が赤く（ここでは黒く）強調されています．