Approximate Inverse Model Explanation(AIME)の概要とその展開

［論文］

Approximate Inverse Model Explanation(AIME)の概要とその展開

Overview of Approximate Inverse Model Explanation (AIME) and its application

中西　崇文

概要

近年 Artificial Intelligence(AI)，機械学習モデルを用いた応用が，様々な分野に展開されつつある．これらの AI，機械学習モデルの精度を上げるために，様々なアルゴリズムが提案されている一方，これらの内部の挙動や，これらのモデルから導出された推定結果がなぜ導出されたのかを理解することが難しくなってきている．我々は，単なる精度のみによる AI，機械学習モデルの評価から脱却し，なぜ，このモデルがこのような推定結果を導出したかを確認しながら，意思決定だけでなく，AI，機械学習モデルのチューニングを行なっていくべきであると考える．そのような背景の中，説明可能な AI(XAI)という技術が注目を集めている．XAI は，対象となるモデルに依存して説明を導出する model-specific 手法と対称となるモデルの別のモデルにより説明を導出する model-agnostic 手法がある．我々はこれまで，着目する AI，機械学習モデルを対象として，近似逆作用素を導出し，モデルの挙動(大局的特徴重要度(global feature importance))，モデルから導出される代表的なそのクラスの値(代表推定インスタンス)，各データポイントにおける推定結果の説明(局所的特徴重要度(local feature importance))，データ群自体の分割のしやすさの可視化(代表インスタンス類似度分布プロット)を同時に導出することが可能とする approximate inverse model explanations(AIME)を提案している．これまでの LIME や SHAP に比べ，計算コストが低く，局所的特徴重要度のみではなく，様々な説明を導出可能で，さらに LIME や SHAP に比べ，導出される説明がシンプルで解釈しやすいことがわかっている．本論文では，我々の AIME の概要を述べながら，その応用について述べる．

１.はじめに

近年，Artificial Intelligence(AI)，機械学習モデルは，自動運転，医療診断補助，金融取引など多様な分野の進歩を推進する上で極めて重要な役割を果たしており，その存在こそが現代の意思決定に多くの寄与をしている．しかしながら，AI，機械学習モデルが，重要な意思決定への影響力が増していることを考えると、予測や推定がどのように導き出され，どのデータ特徴が結果に最も影響を与えるかを理解することが極めて重要である．これらの洞察には，不公正なバイアスの特定，モデルの信頼性の確認，法的・倫理的観点からの説明責任の評価などが

含まれる．この文脈において，ディープラーニングやブラックボックスモデルのような複雑なモデルは、その内部挙動を直感的に理解することが難しいため問題となる．また，G20(20 カ国財務大臣・中央銀行総裁会議)で承認された「人間中心の AI 社会原則」における「公平性、説明責任及び透明性の原則」においては，Fairness，Accountability，Transparency の重要性を特に強調されており，AI，機械学習モデルにおいて，これらを実現する技術的手法の確立が重要となってきている．

我々はこれらを実現する技術的手法の一つとして，説明可能な AI(XAI)[1]が挙げられると考えられる．XAI は，モデルの挙動やデータインスタンスに対する推定結果の説明を導出するものである．XAI というキーワードが注目され出したのは，2017 年における米国の DARPA（Defense Advanced Research Projects Agency：国防高等研究計画局）[1]が主導する研究プロジェクトが発端であるが，AI，機械学習モデルにおける説明可能性については，古くは 40 年前のルールベールのころから，文献[2][3]のようなルールベースにおける説明可能性についての議論があった．確かにルールが複雑になればなるほど，人間がルールベールエンジンによって導出した結果を直感的に把握することが難しくなるであろう．それから 2000 年以降，ニューラルネットワークを応用した，機械学習，特に深層学習が主流になったころ，文献[4]によって，ディープラーニングはニューラルネットワーク自体によっても、外部の説明要素によっても説明することの難しさを指摘している．その前後で，XAI 界隈では非常に有名となってきている LIME[5]，SHAP[6]が発表されている．

SHAP の論文[6]において，興味深い Additive feature-attribution methods と呼ばれる，現在の XAI の手法のほとんどがこれに当てはまるだろうという次の式を提示している．

$$g'(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

ここで，gはオリジナルモデルfの説明モデル，$z' \in \{0,1\}^M$，Mは簡略化された入力特徴の数，$\phi_i \in R$である．元のモデルの出力$f(x)$は，各特徴に効果ϕ_iを帰属させ，全ての特徴帰属の効果を合計することにより近似される．この式は，説明変数の一つとなるx_j を z'_i.に変換した後，ϕ_i（実数）とz'_i.（{0,1}）の乗算を単純に加算する．この式は，ほとんどの model-agnostic method が，例えば LIME や SHAP がこの式に当てはまるということを示している．これをもう少し，深く解釈するならば，人間にとって説明可能にするためには，上記の式のように単純な線形回帰の式で表現できるモデルにサロゲートモデル(元のオリジナルモデルの説明モデル)を設定しなければならないということを示している．LIME はここでいうオリジナルモデル，つまり着目する AI，機械学習モデルのサロゲートモデルを求めるために，着目するデータインスタンスの近傍に着目し，線形回帰を推定している．また，SHAP は，協力ゲームを利用し，それぞれの貢献を分配するという形で，説明を導出している．しかしながら，これらの手法では，主に局所的重要特徴しか抽出する機能しか持っておらず(SHAP は全てのデータインスタンスに対して計算をすることにより，大局的重要特徴を抽出する機能を持っているが，可視化としては解釈が難しい)，さらに，これらの手法の計算量が大きいことが問題であった．

これらのことから，我々は，オリジナルモデルfの近似逆演算子$\widehat{f^{-1}}$を導出することにより，

モデルの挙動(大局的特徴重要度(global feature importance))，モデルから導出される代表的なそのクラスの値(代表推定インスタンス)，各データポイントにおける推定結果の説明(局所的特徴重要度(local feature importance))，データ群自体の分割のしやすさの可視化(代表インスタンス類似度分布プロット)を同時に導出することが可能とするapproximate inverse model explanations(AIME) [7]を提案している．着目するオリジナルモデル，つまり，AI，機械学習モデルの簡単な近似逆演算を求める意味のエッセンスは，説明変数(X)と目的変数(Y)とを比較した場合，多くの場合において，目的変数(Y)のほうが単純な構造になっていることに気づくだろう．そうなれば，目的変数から説明変数を求める近似逆演算子$\widehat{f^{-1}}$を導出することは，オリジナルモデルよりもより簡素でかつ，人間に解釈しやすいモデルを導出することが可能であるという考えに基づく．

本論文では，我々が提案するAIMEの概要を述べるとともに，AIMEによって導出された例を挙げることにより，AIMEもとよりXAIの考え方が，今後のAI，機械学習の発展に不可欠であることも述べていく．我々はXAIが現在着目されているLLMsなどの生成AIと同等のインパクトを持つ意思決定を行う際に重要な技術になると考えている。

２.関連研究

Speith[8]は，XAIの種類として大きくante-hocとpost-hocの手法で分けられるとしている。ante-hocとは，事前にAI，機械学習モデル自体に説明できる機能を持たせることを意味している．つまり，線形回帰モデル，決定木モデル，kNNモデル，ルールベースモデルなど，一般に説明可能であると考えるいわゆるグラスボックスモデルを指すことが多い．それに対して，post-hocとは，いわゆるブラックボックスモデルに適用する手法であり，model-specificとmodel-agnosticの2つにさらに分類できる。model-specificは，特定のAI，機械学習モデルに限定した説明導出手法で，例えばConvolutional neural network(CNN)では，Grad-CAM[9]がそれにあたる。それに対して，model-agnosticは，モデルに依存せず汎用的に説明を導出できる手法である．

我々は，model-agnostic手法を大きく3つに分けている．1つ目は，着目するAI，機械学習モデル(ブラックボックスモデル)自体を使ってモデルの振る舞いを観察し，説明を導き出す方法である．partial dependence plots(PDP)[10][11]，individual conditional expectations (ICE)[12]などがある．2つ目は，ブラックボックスモデルと同じ順方向で別の簡単な手法で説明を導出ものであり,ここまで説明してきたLIME[5]，SHAP[6]が挙げられる．3つ目は，ブラックボックスモデルと逆方向で説明を導出する手法であり，これが我々の提案するAIME[7]にあたる．

ちなみにこの分類は，我々がややこしい計算問題の検算する手法と似ている．1つ目の場合は，同じ手法を使って解いていると考えて良い．２つ目の場合は，別解で解いていると考えて良い．3番目は逆から検算していると考えればよい．

３.AIME の概要

AIME は，LIME，SHAP などの従来の手法と比べて計算コストすくなく，解釈性の高い説明を導出することが文献[7]に示している．また AIME は，次節で示すように，目的変数を中心とした式となるため，説明変数が持つ多重共線性でロバストである点が挙げられる．また，AIME 一つで，モデルの挙動(大局的特徴重要度(global feature importance))，モデルから導出される代表的なそのクラスの値(代表推定インスタンス)，各データポイントにおける推定結果の説明(局所的特徴重要度(local feature importance))，データ群自体の分割のしやすさの可視化(代表インスタンス類似度分布プロット)を同時に導出する手法はこれまでない．図 1 に AIME の強みである 4 つの説明について示す．

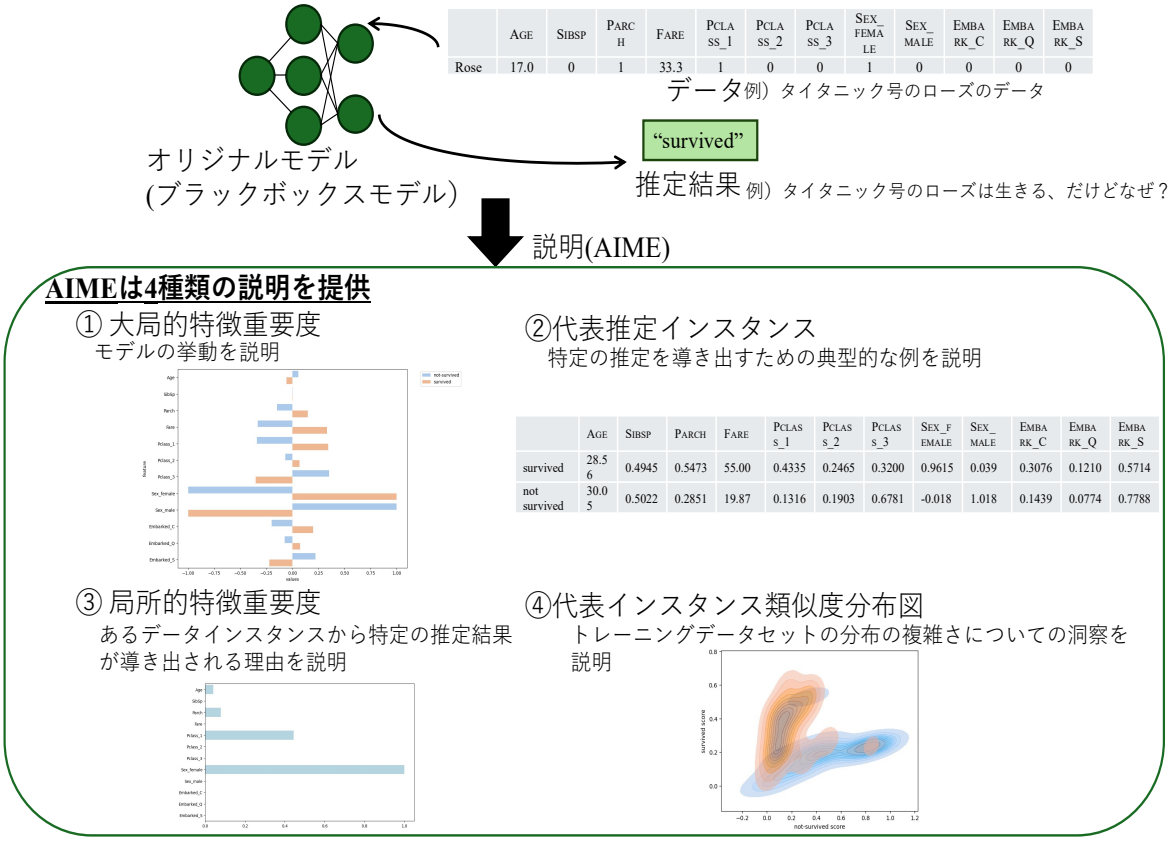

図 1　AIME の 4 つの説明表現[7]

図 1 では，タイタニックデータを何らかの機械学習モデル(図中オリジナルモデル/ブラックボックスモデル)に学習させ，映画「タイタニック」のローズのデータをブラックボックスモデル入れたときを想定している．この場合，ブラックボックスモデルから「survived」と導出される．しかしながら，このブラックボックスモデルからなぜローズが「survived」と推定されたのかを説明することは難しい．そこで，AIME を用いて，このブラックボックスモデルの近似逆演算子を訓練データから導出し，①大局的特徴重要度，②代表推定インスタンス，③局所的特徴重要度，④代表インスタンス類似度分布図の 4 つの説明を導出することができる．大局的特徴重要度とは，モデル全体として，各特徴が各クラスにどのように影響を及ぼすかを導出している．例えばタイタニックデータの場合，「not survived」は青，「survived」はオレンジで表現

され，絶対値の値が大きいほど，そのクラスに貢献していると考えることができる．代表推定インスタンスは，survived，not survived の典型的な人の数値を推定している．局所的特徴重要度はローズのデータのようなデータインスタンスについて，「survived」と推定に貢献した特徴に値が大きくなるようにしている．代表インスタンス類似度分布は，訓練データセットの分布として，「not survived」と「survived」をそれぞれ代表推定インスタンスとの相関を取り，プロットしたものである．これで重なりが大きければ，識別することが難しいことをしめしている．図 1 の例の場合，一部「not survived」と「survived」の識別の難しい箇所は存在するが，識別は可能であると説明できる．

4．AIME の定式化

図 2 に AIME の定式化のための様々な数式記号や入出力を整理している．

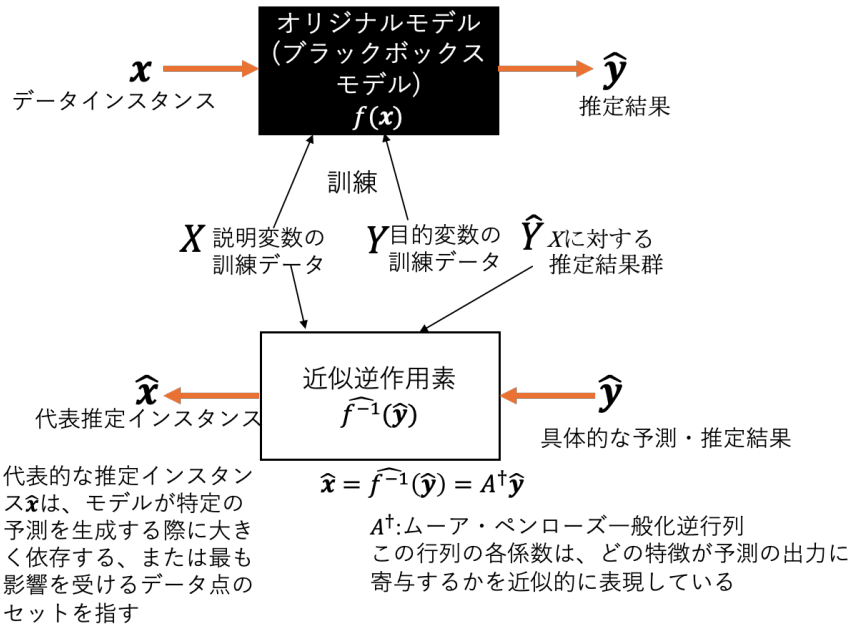

図 2　AIME の定式化[7]

まず，訓練データとしてX,Yを準備し，オリジナルモデル(ブラックボックスモデル)を訓練する．これにより，ブラックボックスモデル，つまり AI，機械学習モデルが学習され，使える状態となる．$\boldsymbol{x}$というデータを入れれば，ブラックボックスモデルは$\hat{\boldsymbol{y}}$を推定結果として導出する．しかしながら，このブラックボックスモデルは，外からはどのように$\hat{\boldsymbol{y}}$を推定結果として導出したのか分からない．そこで，訓練データXを 1 データずつブラックボックスモデルに入力させ，推定値を導出し，$\hat{\boldsymbol{Y}}$を求める．その上で，X，$\hat{Y}$を用いて，近似逆作用素$\widehat{f^{-1}}$を導出する．$\hat{\boldsymbol{x}} = \widehat{f^{-1}}(\hat{\boldsymbol{y}})$というふうに，目的変数から典型的な説明変数を推定することが可能となる．ここで，X，$\hat{Y}$を用いて，近似逆作用素$\widehat{f^{-1}}$をどのように導出するか示す．

4.1　近似逆作用素$A^{\dagger}$の導出

AIME[7]では，逆作用素f^{-1}を線形近似したものを行列$A^{\dagger}$(近似逆作用素)として以下のように

定義する．

$$X = A^{\dagger}\hat{Y}$$

行列$A^{\dagger}$は行列Aのムーア・ペンローズの一般逆行列[14][15]を意味する．ムーア・ペンローズの一般逆行列は正則ではない行列に対して擬似的な逆行列を求める手法である．近似逆作用素は以下のように導出される．

$$X = A^{\dagger}\hat{Y}$$
$$X\hat{Y}^{\intercal} = A^{\dagger}\hat{Y}\hat{Y}^{\intercal}$$
$$X\hat{Y}^{\intercal}(\hat{Y}\hat{Y}^{\intercal})^{-1} = A^{\dagger}(\hat{Y}\hat{Y}^{\intercal})(\hat{Y}\hat{Y}^{\intercal})^{-1}$$
$$A^{\dagger} = X\hat{Y}^{\intercal}(\hat{Y}\hat{Y}^{\intercal})^{-1} = X\hat{Y}^{\dagger}$$

ムーアペンローズの一般逆行列において下記が成立する．

$$\hat{Y}^{\intercal}(\hat{Y}\hat{Y}^{\intercal})^{-1} = \hat{Y}^{\dagger}$$

ここで，上記の式から，$A^{\dagger}$は$\hat{Y}^{\intercal}(\hat{Y}\hat{Y}^{\intercal})^{-1}$に大きく依存するため，説明変数における多重共線性がロバストであると考えてもよい．もちろん，Xをかけていることから，全く影響がないわけではないが，順方向から求めるよりも，Xの影響力が小さいため，他の XAI 手法に比較して説明変数における多重共線性がロバストと考えることができる．

4.2 大局的特徴重要度

4.1 節で導出した，一般化逆行列$A^{\dagger}$は，行数と列数がそれぞれデータセットXの特徴数と考慮クラス数に等しい行列である．この一般化逆行列$A^{\dagger}$は、その成分が各特徴の特定のクラスへの影響を反映するため、大域的特徴重要度[7]を示す。具体的には，一般化逆行列$A^{\dagger}$は，各クラスに対する各特徴の寄与度を示す．$A^{\dagger}$の特定の成分（例えば、行 i の列 j の成分）は、j 番目のクラスの出力に対する i 番目の特徴の寄与を表します。これにより、各特徴がモデルの全体的な動作にどの程度影響するかを全体的に理解することができます。

モデルの予測結果に対する各特徴の影響は、これにより、一般化逆行列 $A^{\dagger}$の構成要素が特定の予測結果にどのように寄与するかを明確に理解することができる。

4.3 代表推定インスタンス

代表推定インスタンス[7]は，モデルの理解に従って特定のクラスに分類できるインスタンスの本質的な特徴をカプセル化した理論的構成要素である．

より正式な用語では，代表推定インスタンス$\boldsymbol{x}^{*}$は，特定のクラスの完全な予測を表すベクトル$\boldsymbol{y}^{*}$に近似逆演算子$A^{\dagger}$を適用して得られる解である．代表的な推定インスタンス$\boldsymbol{x}^{*}$は以下の式を用いて導出される：

$$\boldsymbol{x}^{*} = A^{\dagger}\boldsymbol{y}^{*},$$

ここで，$\boldsymbol{y}^{*}$は特定のクラスに正確に対応する予測値のベクトルで，そのクラスに対応する成分のみが 1 であり，他の成分は 0 であり，$\boldsymbol{x}^{*}$はそのようなベクトルに近似逆演算子$A^{\dagger}$を適用することによって得られる．ここで、$\boldsymbol{y}^{*}$はそのクラスの代表推定インスタンスであり，クラスの数であるm次元ベクトルである．特定のクラスkを仮定すると・そのクラスに対するベクトル$\boldsymbol{y}^{*}$

の各要素y_i^*は，以下の式を用いて評価することができる：

$$y_i^* = \begin{cases} 1 & (if\ i = k) \\ 0 & (otherwise) \end{cases}$$

代表推定インスタンスは，モデルが特定のクラスに属すると期待する「理想的な」または「典型的な」インスタンスを表す．これらのインスタンスを詳しく見ることで，特定のクラスにおける特徴の全体的な重要性をよりよく理解することができる．

代表推定インスタンスを調べることで，クラス・メンバシップに最も近い特徴に関するユニークな洞察を得ることができる．したがって，これらのインスタンスは，モデルの予測を解釈し検証する際に極めて重要である．

4.4 局所的特徴重要度

AIME における局所特徴重要度[7]は，各特徴が特定のデータインスタンスの分類にどの程度寄与しているかを導出する．これは，データインスタンスがモデルによって特定のクラスに分類される理由を理解するのに役立つ．特定のインスタンスの分類に直接影響する特徴を明らかにすることができる．これらの局所的な特徴の重要度の値は，特定のデータインスタンスの予測において特徴の影響がどのように変化するかを明確に理解し，予測が特定の特徴に依存する程度を示すことができる．局所的特徴重要度は、モデルが特定のインスタンスに対して重要であると見なす特徴を特定し，その理解をモデルの解釈に統合する上で極めて重要な要素となる．

データインスタンス$\boldsymbol{x}$、クラスな予測を表すベクトル$\hat{\boldsymbol{y}}$、一般化逆行列$A^{\dagger}$を組み込むことで、以下の式を使用して局所特徴重要度ベクトル$\boldsymbol{l}$を導出することができる：

$$\boldsymbol{l} = A^{\dagger}\hat{\boldsymbol{y}} \circ \boldsymbol{x},$$

ここで，局所特徴重要度ベクトル $\boldsymbol{l}$ は，全特徴の次元数で構成され，各値はその特徴の重要度を表し，∘はアダマール積を表す。

ベクトル$\hat{\boldsymbol{y}}$は$\boldsymbol{x}$に対する推定値を表し，$A^{\dagger}$はこの予測を特徴空間にマップする近似逆演算子である．これは，特定のクラスにおける特徴の重要性の全体的な表現を提供した.データインスタンス$\boldsymbol{x}$と$A^{\dagger}\hat{\boldsymbol{y}}$における各特徴の値は,各特徴の値がその重要性に影響する程度を反映するために,要素ごとに乗算される．これはアダマール積∘によって実現される．この操作により，$\boldsymbol{x}$の特定の特徴が大きな値を持つ場合，その重要性が強調される．しかし、$\boldsymbol{x}$の特定の特徴の値が小さかったり，ゼロであったりすると，その重要性は減少する．これにより，特定のデータインスタンス$\boldsymbol{x}$の特定のクラスへの分類に対する各特徴の寄与を具体的に表現することができる．これがAIME における局所特徴の重要性である．

ここで，データインスタンス$\boldsymbol{x}$はデータセットXと同じパラメータで正規化する必要があり，これは一般化逆行列$A^{\dagger}$を得るために使用される．

4.5 代表インスタンス類似度分布プロット

代表インスタンス類似度分布プロット[7]は，4.3 節で述べた代表推定インスタンスとターゲットデータセット内の各インスタンスとの類似度分布を可視化する．代表インスタンス類似度分布プロットは，モデルが特定の推定値を出力したときの挙動と，その推定値が対応する学習

データの領域を視覚的に理解することができる．具体的には，モデルによって推定された出力に対応する代表推定インスタンスと，データセット全体の各インスタンスとの類似度分布をプロットする．これは，特定の推定値がデータセットのどの部分から引き出される可能性が最も高いかを理解するのに役立つ．さらに，このプロットは，特定の推定値を出力するときのモデルの挙動と，その推定値に関連する複雑さに関する具体的な視覚的情報を提供する.これにより，モデルの挙動をよりよく理解し，その信頼性を推定することができる.

まず，各クラスに対応する代表推定インスタンスを導出する．次に，各クラスについて，データセットXの各データインスタンス$\boldsymbol{x}$と代表推定インスタンス$\boldsymbol{x}_k^*$との類似度を計算する．本研究では，類似度の計算に RBF カーネル K を導入し，次式のように計算する：

$$K(\boldsymbol{x}_k^*, \boldsymbol{x}) = exp(-\gamma \|\boldsymbol{x}_k^* - \boldsymbol{x}\|^2),$$

ここで，γは RBF カーネルの広がりを制御するパラメータである．最適値はデータセットの特性によって異なる可能性がある．従って，γは，結果の可視化を調べたり，他のモデル選択基準を採用したりして，慎重に選択すべきである.

代表インスタンス類似度分布プロットにおけるカーネル密度推定は，特定の推定値の代表インスタンスとデータセット全体の他のインスタンスとの間の類似度分布を可視化するために使用された.類似度スコアにカーネル密度推定を適用することで滑らかな分布を得ることができ，さらに特定の推定値が最も密接に関連しているデータセット全体の部分について，より直感的な理解を提供する.

５.AIME 適用例 1

機械学習で単純な例であるタイタニックデータで，AIME の出力結果を導出していく．この例の詳細は文献[7]に示されている.

図 3 にタイタニックデータによる大局的特徴重要度を示す．タイタニックデータの各特徴について,″not survived ″と ″survived″は,それぞれ予測と推定に対するモデルの寄与度を示している．図 3 から，Sex_female は ″survived″として寄与し，Sex_male は ″not survived ″として寄与している．Pclass_1，つまり最初のクラスは，″survived ″に正に寄与する。Fare は，″survived″に正に寄与し，Class_3 は ″not survived″に正に寄与する．これらの結果は，AIME 大局的特徴重要量が，タイタニック号のデータの傾向とモデルの振る舞いを直感的に説明することを示している.

図４に映画「タイタニック」のローズのデータを入力し，″survived″と認識した説明を局所的特徴重要度として示し,図５に映画「タイタニック」の Jack のデータを入力し,″not survived″と認識した説明を局所的特徴重要度として示している.図 4 の場合、″Sex_female″、″Pclass_1″、″Parch″、″Age″ の順に正の寄与をしていることが出力される．この結果はシンプルで解釈しやすい．図 5 の場合、″Sex_male″、″Pclass_3 ″の順にマイナスに寄与し、″Age ″がプラスに寄与している。この結果はシンプルで解釈しやすい.

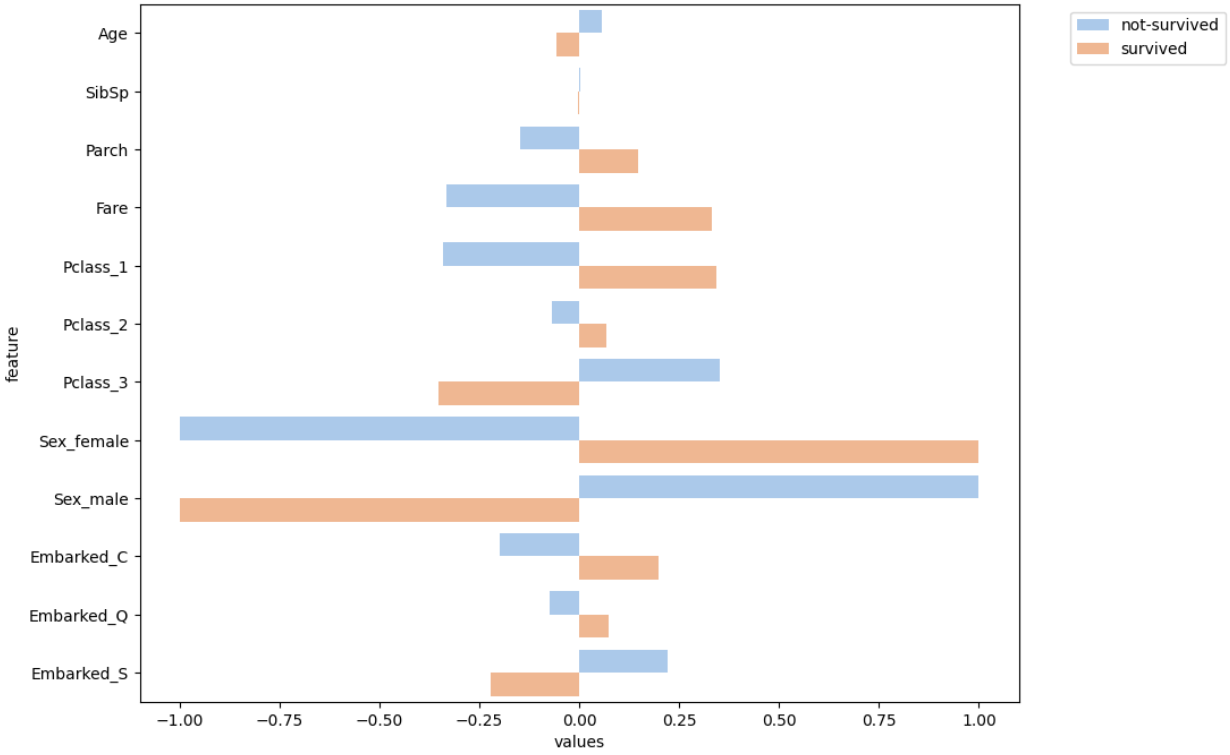

図 3　タイタニックデータにおける AIME の大局的特徴重要度

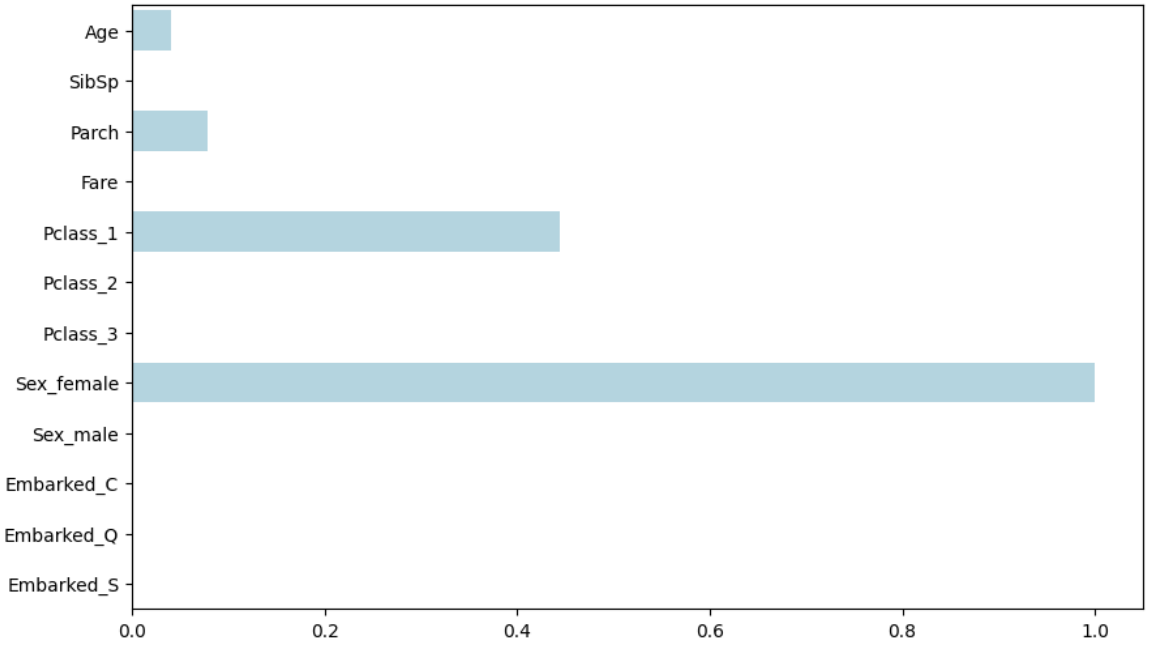

図 4　タイタニックデータにおける AIME の映画「タイタニック」のローズを想定した局所的特徴重要度

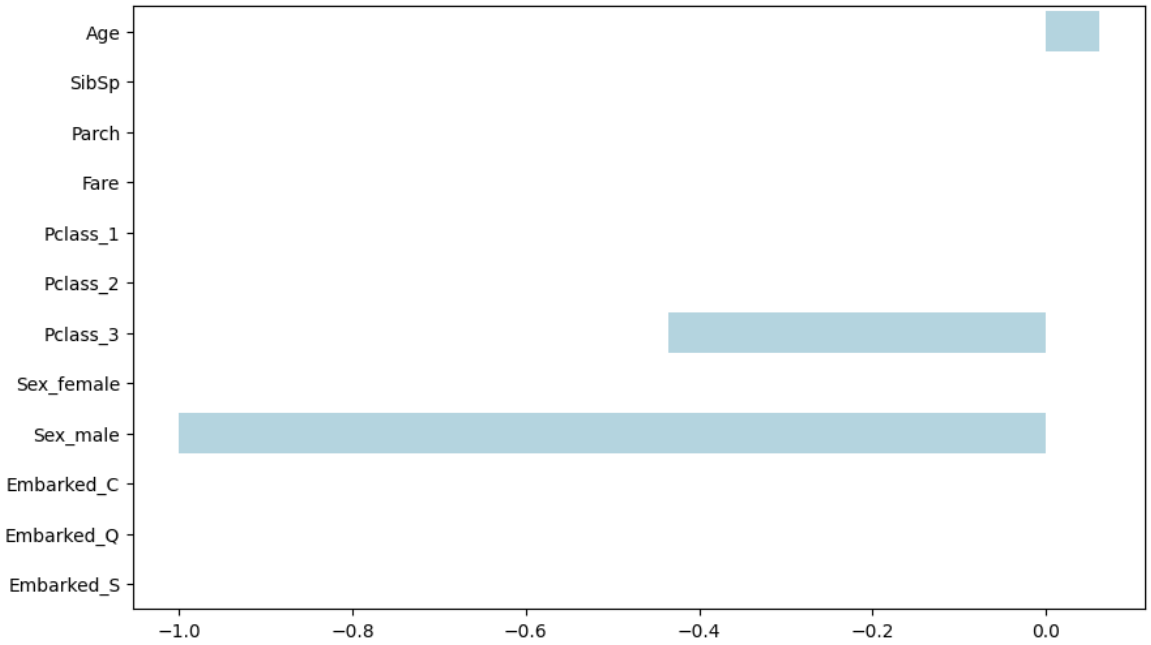

図 5　タイタニックデータにおける AIME の映画「タイタニック」のジャックを想定した局所的特徴重要度

タイタニックデータの LightGBM に対して AIME を用いて得られた代表インスタンスの類似度分布プロットを図 6 に示す．これはデータセットXの各クラスの分布を示し，"not survived"代表インスタンスとの類似度が高いデータインスタンスは横軸に沿って広がり，"survived"代表インスタンスとの類似度が高いデータインスタンスは縦軸に沿って広がっている．この分布は，多くのインスタンスがタイタニック・データセットの対応する代表インスタンスによってよく表現されていることを示唆している．言い換えれば，タイタニックのデータは"not survived"と"survived 比較的簡単に分類できる．

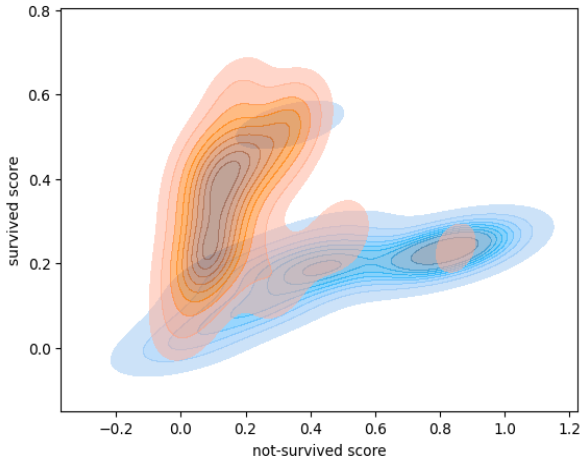

図 6　タイタニックデータにおける AIME の代表インスタンスの類似度分布プロット

６．AIME 適用例２

AIME は画像メディアコンテンツの認識の説明についても用いることができる．ここでは実装方法は書かず，その例のみ示すことにする．今回認識，つまりブラックボックスモデルとして使用したのは Inception-3 であり，データセットは STL-10 を用いて学習している．

例えば，図７の左図は飛行機の写真であり，Inception-3 でも飛行機と認識される．その説明として AIME を使って可視化をすると，図７の右図となり，飛行機の頭，翼，尾翼を捉えていることが見えてくる．

これに対して，図８の左図は鳥の写真であり，Inception-3 でも鳥と認識される．その説明として AIME を使って可視化をすると，背景部分を見て鳥と認識していることが分かる．

これらのことから，認識エンジンの精度が良かったとしても，うまく学習ができていない部分を AIME で見つけることが可能となる．この例は，これまで，AI，機械学習が精度の良さを KPI としてチューニングしてきたが，それだけでは不十分であり，どのように学習され，どのように推定されているかを可視化しながらチューニング，運営していくことが重要であることを示唆している．

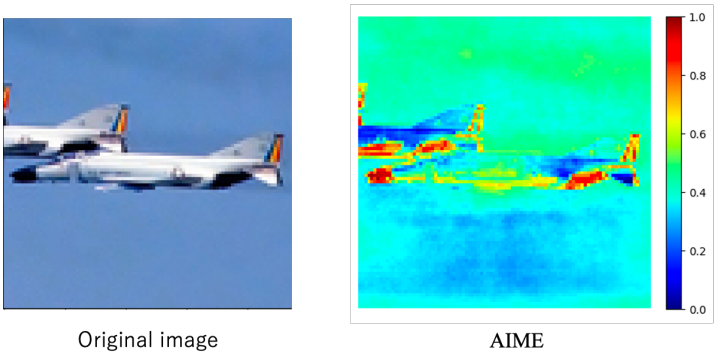

図７　画像認識における AIME の局所的特徴重要度可視化例１

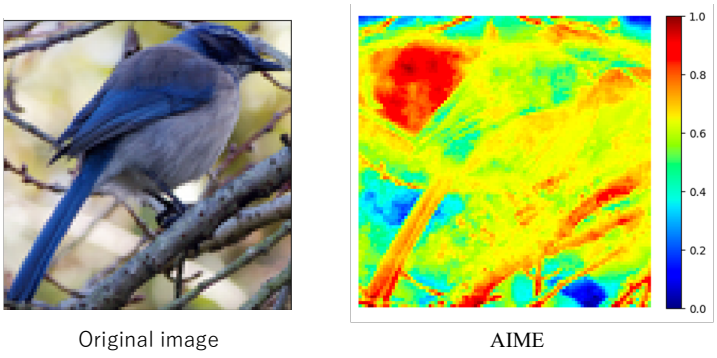

図８　画像認識における AIME の局所的特徴重要度可視化例２

7.おわりに

本論文では，文献[7]で示したAIMEの概要とその展開について示した．AIMEは，これまでのXAI手法とは異なり，近似逆作用素を導出することにより，説明を生み出すため．計算量が比較的軽く，解釈しやすい説明を導出し易い．

また，AIMEを用いた展開例についても示した．AIMEのみならずXAIはこれまでのAI，機械学習モデルの精度主義を考え直すきっかけになると考える．これまで見過ごされていたAI，機械学習モデルがなぜこのような推定を導出するかの説明を見ることにより，AI，機械学習のチューニング等の運用の方法も変わってくるのではないかと考える．

さらに，本稿では画像認識を対象とし，認識結果は正解であるが，その認識結果の導出根拠が想定したものと違う例を示した．これらのことから，これからのAI，機械学習モデルの構築では，精度が上がるだけがよいわけではなく，その手法がなぜよいのか，どのように働いているのかを科学的に検証するフェーズに入ったと考えてもよいだろう．

我々は，この論文を通じて，なぜその結果を導出しているか分からないAI，機械学習モデル(生成AIを含む)の解析手法を確立し，AI，機械学習モデルの科学的検証を行っていくことが重要であると考える．

なお本論文は，文献[7]の成果に基づき，AIMEの仕組みの基本をまとめ直したものである．

参考文献

[1] D. Gunning, "Explainable artificial intelligence (XAI)," Defense Advanced Research Projects Agency (DARPA), vol. 2, no. 2, p. 1, 2017. https://www.darpa.mil/program/explainable-artificial-intelligence

[2] A. C. Scott, W. J. Clancey, R. Davis, and E. H. Shortliffe, "Explanation capabilities of production-based consultation systems," American Journal of Computational Linguistics, vol. 62, 1977.

[3] W. R. Swartout, "Explaining and justifying expert consulting programs," in Computer-Assisted Medical Decision Making, New York, NY: Springer New York, 1985, pp. 254-271.

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM Computing Surveys (CSUR), vol. 51, no. 5, pp. 1-42, 2018.

[5] M. T. Ribeiro, S. Singh, C. Guestrin, " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining , pp. 1135-1144, 2016.

[6] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[7] T. Nakanishi, "Approximate Inverse Model Explanations (AIME): Unveiling Local and

Global Insights in Machine Learning Models," in IEEE Access, vol. 11, pp. 101020-101044, 2023, doi: 10.1109/ACCESS.2023.3314336.

[8] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in: Association for Computing Machinery, 2022; New York, NY, USA, pp. 2239-2250.

[9] R. R. Selvaraju, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in: Proceedings of the 2017 Conference, pp. 618-626.

[10]J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189-1232, 2001, doi: 10.1214/aos/1013203451.

[11]Q. Zhao, T. Hastie, "Causal interpretations of black- box models," J. Bus. Econ. Stat., vol. 39, no. 1, pp. 272-281, 2017.

[12]A. Goldstein, et al., "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," J. Computat. Graph. Stat., vol. 24, no. 1, pp. 44-65, 2015, doi: 10.1080/10618600.2014.907095.

[13]E. H. Moore, "On the reciprocal of the general algebraic matrix," Bull. Am. Math. Soc., vol. 26, pp. 294-300, 1920.

[14]R. Penrose, "A generalized inverse for matrices" in Mathematical Proceedings of the Cambridge Philosophical Society, Cambridge: Cambridge University Press, vol. 51, no. 3, 1955, pp. 406-413. DOI: 10.1017/S0305004100030401.