連載

5分で分かる!? 有名論文ナナメ読み

石井一夫（公立諏訪東京理科大学）

Marco T. Ribeiro et al. : “Why Should I Trust You?” : Explaining the Predictions of Any Classifier

Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD'16

説明（解釈）可能な AI（Explainable AI）

深層学習など機械学習による予測モデルの構築では，勾配降下法などのアリゴリズムに基づいて，パラメータを最適化し，予測の精度を上げることが行われる．しかし，最適化されたパラメータが予測結果に対してどのように寄与しているのかを知ることができない，いわゆる，「AI のブラックボックス問題」は，しばしば問題視されてきた．特に，医療や金融，ビジネスの現場において，説明（解釈）のできない AI を使用して意思決定を行うことは実用に耐えないため，AI に説明可能性が求められてきた．2016 年ごろから，重要な特徴量や情報を提示して，AI に説明性を付与する説明可能な AI（explainable AI）のアイディアが，たくさん発表されてきた．紙面の都合で，個々の論文を引用することはしないが，たとえば，LIME, SHAP, Anchor, Influence, Concept Vector などがある（余力のある方は，これらをキーワードに Web などで検索され，原著の論文にあたってほしい）．

本稿では，この説明可能な AI の中でも最も初期に提案された LIME を取り上げ，説明可能な AI の活用法，つまり，ブラックボックスである AI に，いかに解釈性を持たせ活用すべきか，ということについて議論したい．とはいえ，本論文の LIME は，explainable AI の例としてあまりに有名で，インターネットで検索するとたくさんの解説記事がヒットする．Google Colab を用いたデモも紹介されており，Web 上で動作を確認できる．したがって，詳細はそちらを拝見していただくことにし，本稿では概要だけを述べる．

LIME（Local Interpretable Model-agnostic Explanations）

Explainable AI として，予測モデルに説明（解釈）可能性を持たせる方法として，説明困難なモデルを説明可能なモデルで近似するという手法が用いられる．本論文の提案する LIME という手法では，予測するデータの周辺を局所的にサンプリングし，これを線形回帰で近似することでどの特徴量が予測結果に寄与しているかを表現する．

問題となっている説明困難な予測モデルを $f(x)$ とし，これを近似した説明可能な線形回帰モデルを $g(x')$ とする（ここで，データ x というベクトルは人が解釈しやすいように x' というバイナリベクトル（0 と 1 からなる数字のベクトル）に変換されている）．このとき，LIME による f を求めるための損失 $\xi(x)$ は，以下で表される．

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

ここで，π_x はカーネル関数などの距離尺度で，$\pi_x(z)$ において x が z とどれくらい近いかを表す．$L(f, g, \pi_x)$ は，モデル f に対する g の近似誤差であり（つまり，f と g の解離を表す尺度である），次のように表せる．

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

$\mathcal{Z}$は，fとgへの入力データであるzとz'を含む入力データ全体の集合を示す．$\xi(x)$式の，$\Omega(g)$は，モデルgの複雑度である．$\xi(x)$は，gの複雑さを低く保ちつつfをどれほどよく近似できるかを表し，すなわち，モデルfの解釈可能性と考えてよい．

LIMEでは，予測するデータの周辺を局所的にサンプリングし，これを線形回帰で近似することでどの特徴量が予測結果に寄与しているかを表現する．結果的に，$g(x') = w \cdot x'^{\mathrm{T}}$のような解釈可能な線形回帰の近似モデルが得られる（図-1）．

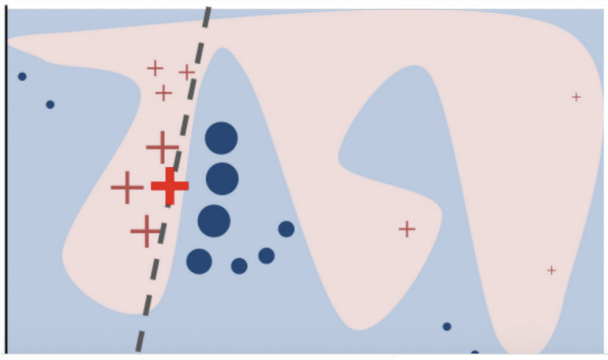

図-1　LIMEで局所的に解釈可能な線形モデルの模式的な図（文献より転載）

図中のピンクとグレーの色が分類される2つの領域を示し，それを識別するその局所的な線形回帰の近似モデルを点線で示している．青い丸とピンクの＋は，局所領域を分ける説明変数を示し，その大きさは局所的な線形回帰の寄与の大きさを示している．

LIMEを用いてスパース線形回帰モデルに解釈性を持たせるアルゴリズムの例をアルゴリズム1に示す．

本論文では，LIMEの実施例としてサポートベクトルマシンを用いたテキスト分類と，画像解析におけるディープラーニングの例が示されている．LIMEは，RやPythonで実装されており，Rパッケージlime[1]やPythonのライブラリlime[2]で簡単に試すことができる．

説明（解釈）可能AIへの批判と，筆者らの使い方

しかし，このようなexplainable AIに対して，近似モデルはあくまで近似であってモデルそのものではないという，批判的な意見も出てきている[3]．場合によっては，誤った解釈を誘発するという指摘すらもある[4]．使い方に注意を要するのである．一

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
  return w

アルゴリズム1

方で，explainable AI は，ブラックボックスである AI に特徴抽出の手段を提供するものという捉え方もできる．

実際，AI のブラックボックス性に関し否定的な意見を耳にすることもよくある．特に，生物統計学や医療統計学の分野では，統計的因果推論の考え方があり，機械学習のようなアルゴリズムに基づいて自動的にパラメータを最適化するのとは違ったアプローチを用いた数理モデリングを行うケースも少なくない．一般化線形モデル以外には，ベイジアンモデリングも因果推論という点ではしばしば使われる．

機械学習には，大まかに教師なし学習と，教師あり学習の 2 種類があり，クラスタリングや主成分分析などの教師なし学習は，教師あり学習への説明変数を提供する手法と捉えることができる．Explainable AI は，これら教師なし学習に似た性格を持つことから，教師あり学習への説明変数を提供する手法として用いることができる．可能ならば，分析デザインを考えるときに，因果推論的な一般化線形モデルやベイジアンモデリングと機械学習をうまく組み合わせた使い方を行うことをお勧めする．自然言語処理や画像診断において，その結果と因果推論数理モデリングを組み合わせるようなやり方である．

参考文献

1) https://cran.r-project.org/web/packages/lime/index.html
2) https://colab.research.google.com/github/arteagac/arteagac.github.io/blob/master/blog/lime_image.ipynb, https://cran.r-project.org/web/packages/lime/index.html
3) Rudin, C. : Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, Nat Mach Intell 1, pp.206-215 (2019).
4) Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S. and Tapp, A. : Fairwashing : The Risk of Rationalization, Proceedings of the 36th International Conference on Machine Learning, in Proceedings of Machine Learning Research 97:161-170 Available from http://proceedings.mlr.press/v97/aivodji19a.html (2019).

(2021 年 7 月 19 日受付)

石井一夫
kishii@rs.sus.ac.jp
公立諏訪東京理科大学工学部教授，久留米大学医学部内科学講座客員准教授．専門は，計算機統計学，データサイエンス．2015 年度情報処理学会優秀教育賞受賞．社会課題（少子高齢化，地球温暖化）の克服に向けた医療ビッグデータ，環境／農業ビッグデータの研究教育に従事．