私のブックマーク

機械学習における解釈性†1

原　聡（大阪大学産業科学研究所）

1. はじめに

近年の人工知能技術，特に機械学習の発展に伴い，これらの技術への社会的な期待が高まっている．しかし，このような期待の高まりと同時にこれら技術への不安も高まっている．特に，深層学習モデルをはじめとする機械学習モデルが複雑なブラックボックスであるがゆえに安易に信頼できないとする懸念の声が上がりはじめている．

これに対し，総務省はAIの利用の一層の増進とそれに伴うリスクの抑制のために「AI開発ガイドライン案」[1]を2017年に策定した．このガイドライン案では，上記のような懸念に対処するために以下のような「透明性の原則」および「アカウンタビリティ（説明責任）の原則」が盛り込まれている．

- 透明性の原則：開発者は，AIシステムの入出力の検証可能性および判断結果の説明可能性に留意する．
- アカウンタビリティの原則：開発者は，利用者を含むステークホルダに対しアカウンタビリティを果たすよう努める．

これらの原則は，機械学習モデルをブラックボックスとして運用することおよびそのリスクに対して一定の歯止めをかけることを目的としていると考えられる．

EUにおいては，同様の内容がGeneral Data Protection Regulation（GDPR）[2]として2018年5月25日より施行される予定である（特にGDPR-22 [3]が上記のガイドラインに対応）．

このような社会的な要請を背景に，特に2016年以降に機械学習モデルの解釈性・説明性といった研究への注目が増えてきている．本記事では，このような機械学習モデルの解釈性・説明性に関する近年の代表的な研究を紹介する．

2. 動向把握に有用な文献

個別の研究を紹介する前に，まず近年の研究動向を把握するのに有用な文献を紹介する．

- ICML 2017 tutorial on interpretable machine learning [4]
  解釈性における研究のトップランナーの一人，Google BrainのBeen Kimによるチュートリアル資料．
- Interpretable machine learning：A guide for making black box models explainable [5]
  解釈性に関する教科書的な資料．
- A survey of methods for explaining black box models [6]
  解釈性・説明性に関する近年の研究をまとめたサーベイ論文．p. 20のまとめ一覧は一見の価値あり．
- ワークショップの論文集（ICML'16 [7]，NIPS'16 [8]，ICML'17 [9]，NIPS'17 [10]）
  機械学習のトップ会議であるICML, NIPSで開かれた解釈性に関するワークショップ・シンポジウムの論文集．

3. 代表的な研究

以下では解釈性・説明性に関する近年の代表的な研究を紹介する．ここでは，研究を以下の4種類に大別して紹介する．

（1）大域的な説明
　複雑なブラックボックスモデルを可読性の高い解釈可能なモデルで表現することで説明とする方法．

（2）局所的な説明
　特定の入力に対するブラックボックスモデルの予測の根拠を提示することで説明とする方法．

---

†1 http://www.ai-gakkai.or.jp/my-bookmark_vol33-no3

（3）説明可能なモデルの設計
　そもそも最初から可読性の高い解釈可能なモデルをつくってしまう方法.
（4）深層学習モデルの説明
　深層学習モデル，特に画像認識モデルの説明法．アプローチとしては 2 の局所的な説明に該当.

(1) 大域的な説明

大域的な説明では，深層学習モデルやランダムフォレストのような決定木のアンサンブルなどの複雑なモデルを可読性の高いモデル，例えば単一の決定木やルールモデルで近似的に表現することでモデルの説明とする.

- Born again trees [11]
  ランダムフォレスの産みの親 Leo Breiman の論文．ニューラルネットなどのブラックボックスモデルをオラクルとして用いて追加の教師データを大量に生成し，追加データを使って決定木を学習する.
- Interpreting tree ensembles with intrees [12]［R 実装 inTrees [13]］
  ランダムフォレストに頻出するルールを主要なルールとして取り出し，モデルの近似的な説明とする.
- Node harvest [14]［R 実装 nodeHarvest [15]］
  ランダムフォレストを浅い決定木のアンサンブルで近似することで説明とする.
- Making tree ensembles interpretable：A Bayesian model selection approach [16]［Python 実装 defragTrees [17]］
  ランダムフォレストを確率的なモデルとみなして，ベイズ的モデル選択を用いて単純なルールモデルへと変換する.

(2) 局所的な説明

局所的な説明では，ある入力 x をモデルが y と予測したときに，その予測の根拠を説明として提示する.

- Why should i trust You?：Explaining the predictions of any classifier [18]［Python 実装 LIME [19]; R 実装 LIME [20]］
  KDD'16 論文．解釈性研究の代表例として扱われることが多い．線形モデルやルールモデルを用いた局所的な説明を生成する方法を提案．任意のモデルについて簡単に局所的な説明を生成できる点が優れている.
- A unified approach to interpreting model predictions [21]［Python 実装 SHAP [22]］
  NIPS'17 論文．上記の LIME を含むいくつかの局所的な説明法がゲーム理論の Shapley value の枠組みのもとで統一的に記述できることを示した.
- Understanding black-box predictions via influence functions [23]［Python 実装 influence-release [24]］
  ICML'17 ベストペーパ．予測結果に関連の深い訓練データを予測の根拠として提示する方法．ロバスト統計の影響関数を使った効率的な計算法を提案.

(3) 説明可能なモデルの設計

上記二つのアプローチはブラックボックスモデルを対象にそこから説明を生成することを目的としている．これに対し，この第三のアプローチでは最初から可読性の高い解釈可能なモデルをつくることを目的とする.

- Learning certifiably optimal rule lists for categorical data [25]［C++実装 corels [26]］
  KDD'17 論文．ルールリストという決定木の亜種を学習する方法を提案．組合せ最適化問題を各種探索の枝刈りを用いて高速化する.
- Interpretable decision sets：A joint framework for description and prediction [27]
  KDD'16 論文．ルールセットという決定木の亜種を学習する方法を提案．問題を劣モジュラ最大化に帰着して貪欲法で解く.
- Prototype selection for interpretable classification [28]［R 実装 protoclass[29]］
  分類問題の各カテゴリーを代表する訓練データを検出する方法を提案.
- Examples are not enough，learn to criticize! criticism for interpretability [30]
  NIPS'16 論文．各カテゴリーの代表的なデータだけでなく，例外的なデータをも提示することでユーザのデータ理解を深める方法を提案.

(4) 深層学習モデルの説明

深層学習モデルの説明は，特に画像認識の分野で数多く研究されている．基本的には，モデルが画像内のどの部分を認識しているかを特定してハイライトすることで説明とする.

- 勾配ベースのハイライト法
  出力ラベルに対する入力画像の勾配を計算する．ある特定の入力画素の微小変化が出力ラベルを大きく変化させる場合に，対象画素を認識対象であるとしてハイライトする．ただし，単純に勾配を計算するとノイズの多

いハイライトが生成されるので鮮明化させるために以下のような手法（カッコ内は手法名）が提案されている．［Python + Tensorflow 実装 saliency [31]; DeepExplain [32]］

- ○Striving for Simplicity：The All Convolutional Net [33]（GuidedBackprop）
- ○On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation [34]（Epsilon-LRP）
- ○Axiomatic Attribution for Deep Networks [35]（IntegratedGrad）
- ○SmoothGrad：Removing Noise by Adding Noise [36]（SmoothGrad）
- ○Learning Important Features Through Propagating Activation Differences [37]（DeepLIFT）

4. お わ り に

機械学習モデルの解釈性・説明性に関する代表的な研究について紹介した．なお，これらの研究はいまだ発展途上であり，本記事は 2018 年 3 月執筆時点における情報であることにご留意願いたい．本記事がこれらのトピックの理解の助けに，そしてゆくゆくは機械学習の社会応用への一助となれば幸いである．

最後に，以下の二点について言及して本記事を締めくくりたい．

- **実応用に基づく研究の必要性（文献 a [38]; 文献 b [39]）**
  現時点における解釈性・説明性の研究の多くは「こういった解釈・説明ができると便利だろう」という研究者各自の仮説に基づいている．今後は，より実応用に根ざした研究の必要性が求められている．具体的な問題に直面している産業界からの参入が待ち望まれる次第である．
- **解釈性・説明性への過度な信頼・期待への注意**
  - ○現段階の研究成果が手放しに使えるものではないことに注意する必要がある．特に深層学習モデルの説明において，生成される説明を意図的にミスリードするように変化させる Adversarial Example が生成できることが報告されている（文献 c [40]）．ときに“誤説明”に出合うリスクを考慮して実用前に適切に検証する必要がある．
  - ○解釈性・説明性はタダで手に入るものではないことに注意する必要がある．上記の“誤説明”のリスクに加えて，これらは必ず計算リソースや人間による判断・峻別を必要とする．解釈性・説明性を検討する際には，本当に解釈性・説明性が必要か，導入がコストに見合うと期待できるかを検討する必要がある．

[1] www.soumu.go.jp/main_content/000499625.pdf
[2] https://gdpr-info.eu/
[3] https://gdpr-info.eu/art-22-gdpr/
[4] http://people.csail.mit.edu/beenkim/icml_tutorial.html
[5] https://christophm.github.io/interpretable-ml-book/index.html
[6] https://arxiv.org/abs/1802.01933
[7] https://arxiv.org/html/1607.02531
[8] https://arxiv.org/html/1611.09139v1
[9] https://arxiv.org/html/1708.02666
[10] https://arxiv.org/html/1711.09889
[11] https://www.stat.berkeley.edu/users/breiman/BAtrees.pdf
[12] https://arxiv.org/abs/1408.5456
[13] https://cran.r-project.org/web/packages/inTrees/inTrees.pdf
[14] https://projecteuclid.org/euclid.aoas/1294167809
[15] https://cran.r-project.org/web/packages/nodeHarvest/nodeHarvest.pdf
[16] https://arxiv.org/abs/1606.09066
[17] https://github.com/sato9hara/defragTrees
[18] https://dl.acm.org/citation.cfm?id=2939778
[19] https://github.com/marcotcr/lime
[20] https://github.com/thomasp85/lime
[21] https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

[22] https://github.com/slundberg/shap
[23] http://proceedings.mlr.press/v70/koh17a.html
[24] https://github.com/kohpangwei/influence-release
[25] www.kdd.org/kdd2017/papers/view/learning-certifiably-optimal-rule-lists-for-categorical-data
[26] https://github.com/nlarusstone/corels
[27] www.kdd.org/kdd2016/subtopic/view/interpretable-decision-sets-a-joint-framework-for-description-and-predictio
[28] https://projecteuclid.org/euclid.aoas/1324399600
[29] https://cran.r-project.org/web/packages/protoclass/protoclass.pdf
[30] https://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability
[31] https://github.com/PAIR-code/saliency
[32] https://github.com/marcoancona/DeepExplain
[33] https://arxiv.org/abs/1412.6806
[34] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140
[35] https://arxiv.org/abs/1703.01365
[36] https://arxiv.org/abs/1706.03825
[37] http://proceedings.mlr.press/v70/shrikumar17a.html
[38] https://arxiv.org/abs/1702.08608
[39] https://arxiv.org/abs/1606.03490
[40] https://arxiv.org/abs/1711.00867