論文紹介
“Why Should I Trust You?” Explaining the Predictions of Any Classifier

東北大学経済学部2年
澤谷一磨

紹介する論文の概要

タイトル　“Why Should I Trust You?” : Explaining the Predictions of Any Classifier

提案手法　Local Interpretable Model-agnostic Explanations (LIME), Submodular Pick(SP)-LIME

著者　Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

発行年　2016/2

採択会議　KDD 2016

引用数　1649件 (google scholar, 2019/8/26時点)

紹介の流れ

- XAIの概要及び論文の位置づけ
- 提案手法の説明
- その後への影響

紹介の流れ

- **XAIの概要及び論文の位置づけ**
- 提案手法の説明
- その後への影響

解釈可能性のとらえ方

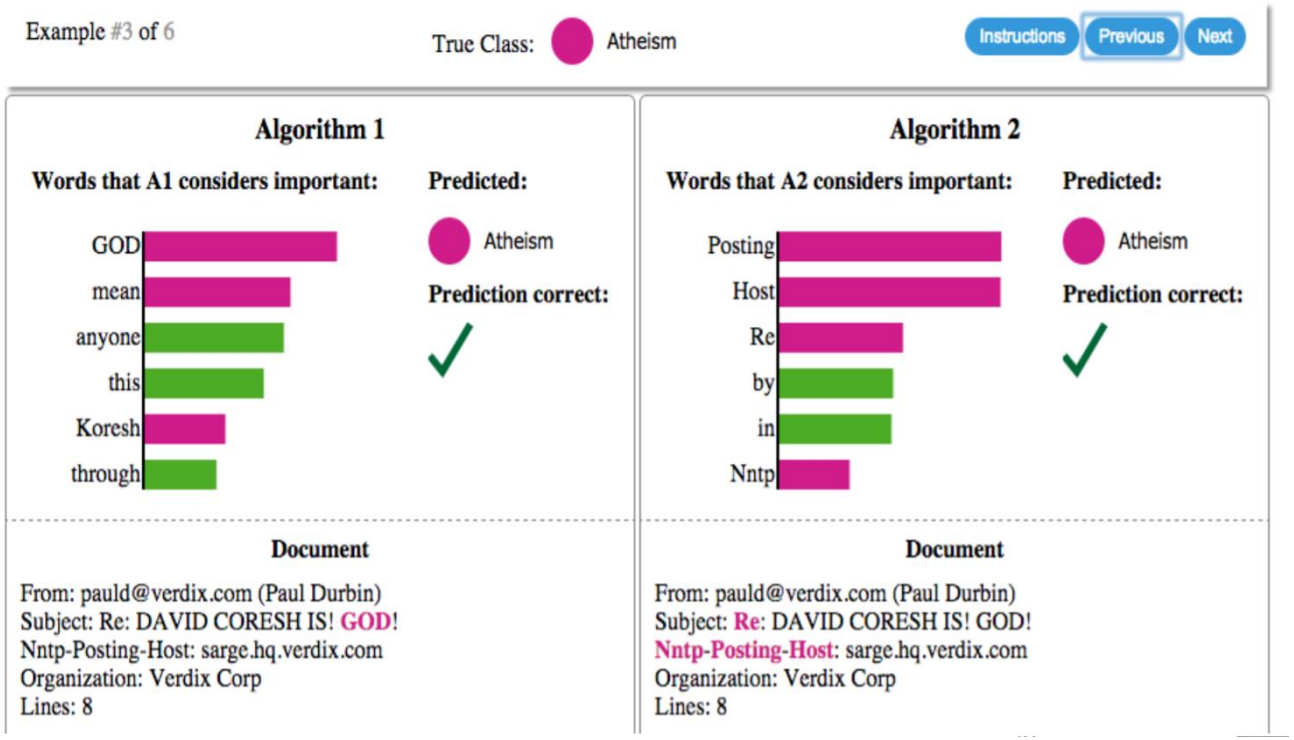
目的に応じたモデル選択

- 解釈を目的とする場合
  →線形回帰, 決定木, RuleList, etc.
- 高精度の予測を目的とする場合
  →DNN, SVM, RandomForest, Gradient Boosting Trees, etc.
- 高精度の予測をしながら解釈もしたい場合
  →Partial Dependence Plot, Feature Importance, LIME, ect.

精度偏重に対する糾弾

- 精度偏重の弊害
  - **Leakage**
  - **Dataset shift**

→解釈性の導入により回避可能



LIMEの位置づけ

前提：Model-Agnosticな解釈性を扱う場合、モデル内部の構造を見るのではなく、モデルの入力と出力のみから解釈を試みる

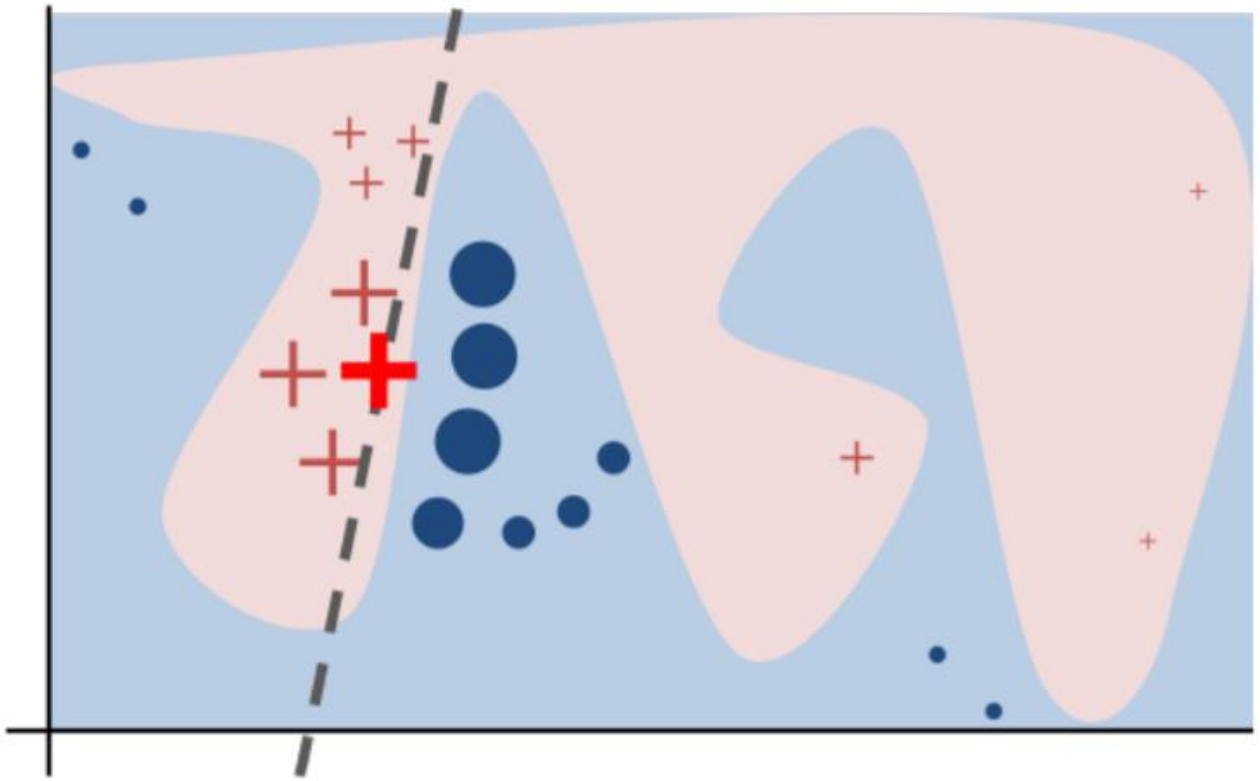
- ローカルな説明
  - 特定のデータ点に着目してそのregionについてのみ解釈可能にする
  - **LIME**, Shapley value, Anchor, [P.W. Koh, ICML ‘17]
- グローバルな説明
  - 決定木など解釈性の高いモデルへの近似である場合が多い
  - **SP-LIME**, [H Deng 2014], [N Meinshausen 2009]

紹介の流れ

- 論文の位置づけ
- **提案手法の説明**
- その後への影響

提案手法の概略

- LIME (Local Interpretable Model-agnostic Explanations)



あるデータ点について
解釈可能モデル（この場合線形回帰）を適用

※その際、ランダムにサンプリングした各点 z は、元の点との距離によって重み付けされる

LIME

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g) \qquad (1)$$

課題＝忠実性と解釈性のトレードオフ

- L： f と g の距離
- f ：元の複雑なモデル
- g ：解釈可能なモデル（e.g. 線形回帰モデル）
- π ：類似度カーネル（サンプル点の重み付け）
- Ω：複雑度

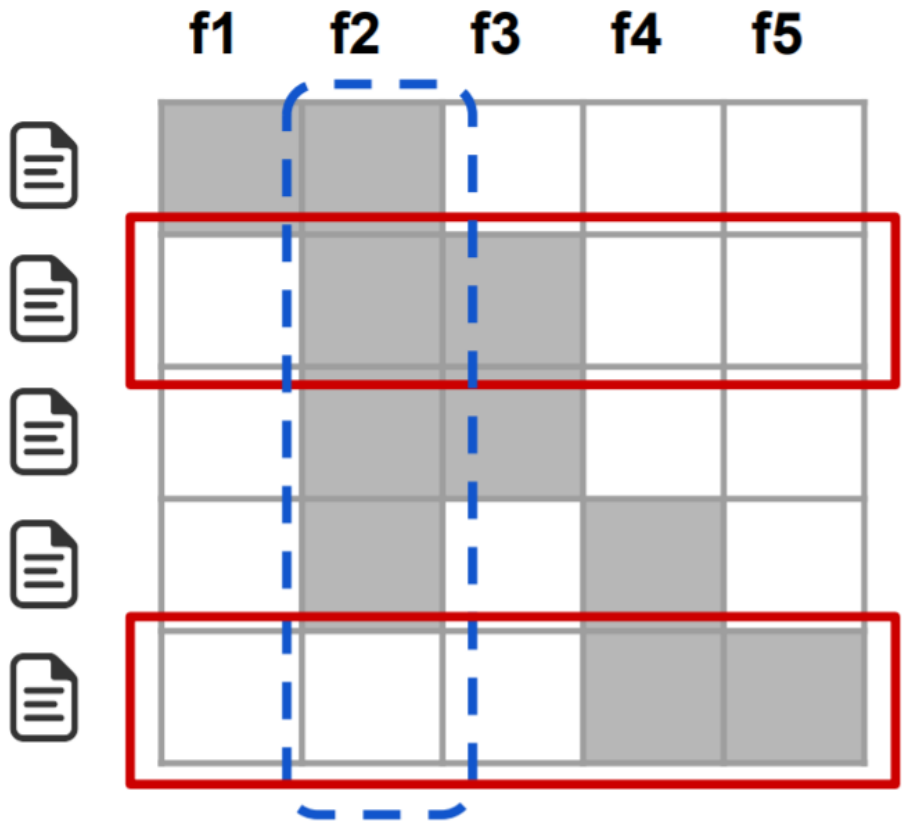
Algorithm 1 Sparse Linear Explanations using LIME

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

提案手法の概略

- SP(Submodular Pick)-LIME



LIMEの結果（インスタンスに対応）を統合してグローバルな解釈を導く

この際、全データ点を考えることはせず、Featuresの最大被覆を達成しつつ説明性の強いインスタンスをピックアップする

SP-LIME

$$c(V, \mathcal{W}, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : \mathcal{W}_{ij} > 0]} I_j \tag{3}$$

$$Pick(\mathcal{W}, I) = \underset{V, |V| \leq B}{\operatorname{argmax}} c(V, \mathcal{W}, I) \tag{4}$$

- I：Feature j の重要度（重みの合計）
- W：LIMEの結果各点が持つインスタンス

Algorithm 2 Submodular pick (SP) algorithm

Require: Instances X, Budget B

for all $x_i \in X$ **do**

$\mathcal{W}_i \leftarrow$ **explain**(x_i, x_i') ▷ Using Algorithm 1

end for

for $j \in \{1 \dots d'\}$ **do**

$I_j \leftarrow \sqrt{\sum_{i=1}^{n} |\mathcal{W}_{ij}|}$ ▷ Compute feature importances

end for

$V \leftarrow \{\}$

while $|V| < B$ **do** ▷ Greedy optimization of Eq (4)

$V \leftarrow V \cup \text{argmax}_i \, c(V \cup \{i\}, \mathcal{W}, I)$

end while

return V

紹介の流れ

- 論文の位置づけ
- 提案手法の説明
- **その後への影響**

その後の展開

- 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が示された[SM Lundberg NIPS, 2017]
  - Cf. SHAP (Shapley Additive exPlanations) [SM Lundberg 2016]
- 特定のデータ点周りで線形モデルではなく領域モデル（決定木の葉ノードのようなもの）を想定したAnchorを提案[MT Ribeiro AAAI, 2018]
- Saliency Masks

参考

- MT Ribeiro “Why Should I Trust You?” ：Explaining the Predictions of Any Classifier (2016) https://arxiv.org/abs/1602.04938
- Riccardo Guidotti, et al. “A Survey Of Methods For Explaining Black Box Models” (2018)
- Randy Goebel et al. “Explainable AI: the new 42?”(2018)
- Interpretable machine learning: A Guide for making black box models explainable https://christophm.github.io/interpretable-ml-book/agnostic.html
- KDD2016勉強会 資料https://www.slideshare.net/shima__shima/kdd2016
- https://sssslide.com/speakerdeck.com/bigseat/kdd2016du-mihui-zao-dao-tian
- https://qiita.com/fufufukakaka/items/d0081cd38251d22ffebf