A Unified Approach to Interpreting Model Predictions

이상용 / 2020-03-06

Computational Data Science LAB

A Unified Approach to Interpreting Model Predictions

Computational Data Science LAB	
목차	1. Introduction 2. Additive Feature Attribution Methods 3. SHAP (SHapley Additive exPlanation) Values 4. Experiments
논의사항 및 결정사항	
관련문서	Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *NIPS.* *Oral presentation NIPS workshop on Interpretable ML(2016)* – best paper award

CONTENTS

01 Introduction

- 다양한 분야에서 정확도 뿐만 아니라 모델이 어떤 이유로 특정한 예측을 했는지에 대한 이해가 중요해짐
- 많은 해석가능한 모델이 연구되고 있는데, 본 논문은 결과의 해석을 위한 unified framework인 SHAP (SHapley Additive exPlanation)을 제안함

- *Additive feature attribution methods*
- *Classic Shapely Value Estimation*

  위의 두 방법을 결합한 것이 SHAP

02 | Additive Feature Attribution Methods

- 모델에 대한 가장 좋은 해석은 해석가능한 간단한 모델을 만드는 것
  - ✓ f : original prediction model
  - ✓ g : explanation model
  - ✓ x' : simplified input
  - ✓ $x = h_x(x')$: mapping function
- Local method의 목적
  - ✓ $g(z') \approx f(h_x(z'))$

02 | Additive Feature Attribution Methods

- Definition Additive feature attribution methods

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

  - ✓ $z' \in \{0,1\}^M,\ \phi_i \in \mathbb{R}$
  - ✓ M은 simplified input features의 수

- 각 feature의 공헌도 ϕ_i를 구하여 모델 해석

- LIME 또한 Additive Feature Attribution Methods 중 하나의 방법

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

목적 : $g(z') \approx f\big(h_x(z')\big)$

$\pi_{x'}$: $local\ kernel$

Classic Shapely Value Estimation

- 게임이론을 바탕으로 하나의 특성에 대한 중요도를 알기 위해 여러 특성들의 조합을 구성하고 해당 특성의 유무에 따른 평균적인 변화를 통해 얻어낸 값

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}(v(S \cup \{i\}) - v(S))$$

✓ v – game

✓ N – all players

✓ S – subset of players

✓ i – specific player

- Example)

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}(v(S \cup \{i\}) - v(S))$$

$$\phi_i(v) = \frac{1}{|N|} \sum_{S \subseteq N \setminus \{i\}} \binom{|N|-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S))$$

$$problem \quad N = \{A, B, C, D\}$$
$$i = D$$

Classic Shapely Value Estimation

- Example)

1 $S \subseteq N \setminus \{i\}$

$$\begin{array}{cccc} & A & AB & \\ \varnothing & B & BC & ABC \\ & C & CA & \end{array}$$

2 $(v(S \cup \{i\}) - v(S))$

$$\begin{array}{cccc} & \Delta v_{A,D} & \Delta v_{AB,D} & \\ \Delta v_{\varnothing,D} & \Delta v_{B,D} & \Delta v_{BC,D} & \Delta v_{ABC,D} \\ & \Delta v_{C,D} & \Delta v_{CA,D} & \end{array}$$

3 $\binom{|N|-1}{|S|}^{-1}$

$$\begin{array}{cccc} & \frac{1}{3}\Delta v_{A,D} & \frac{1}{3}\Delta v_{AB,D} & \\ 1\Delta v_{\varnothing,D} & \frac{1}{3}\Delta v_{B,D} & \frac{1}{3}\Delta v_{BC,D} & 1\Delta v_{ABC,D} \\ & \frac{1}{3}\Delta v_{C,D} & \frac{1}{3}\Delta v_{CA,D} & \end{array}$$

03 | SHAP (SHapley Additive exPlanation) Values
Classic Shapely Value Estimation

- Example)

$$\phi_i(v) = \frac{1}{|N|} \sum_{S \subseteq N \setminus \{i\}} \binom{|N|-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S))$$

$$\phi_D(v) = \frac{1}{4} \sum \begin{pmatrix} & \frac{1}{3}\Delta v_{A,D} & \frac{1}{3}\Delta v_{AB,D} & \\ 1\Delta v_{\varnothing,D} & \frac{1}{3}\Delta v_{B,D} & \frac{1}{3}\Delta v_{BC,D} & 1\Delta v_{ABC,D} \\ & \frac{1}{3}\Delta v_{C,D} & \frac{1}{3}\Delta v_{CA,D} & \end{pmatrix}$$

Simple Properties Uniquely Determine Additive Feature Attributions

- SHAP

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

✓ f – ~~game~~ model

✓ x' – all ~~players~~ features

✓ S – subset of ~~players~~ features

✓ i – specific ~~player~~ feature

✓ x – instance being explained

SHAP (SHapley Additive exPlanation) Values

Simple Properties Uniquely Determine Additive Feature Attributions

- Property 1 (Local accuracy)
  - ✓ Explanation model $g(x')$은 original model $f(x)$와 같은 값을 반환함

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i'$$

- Property 2 (Missingness)
  - ✓ 특정한 simplified feature가 존재하지 않을 때, 해당 feature의 공헌도는 0

$$x_i' = 0 \implies \phi_i = 0$$

- Property 3 (Consistency)
  - ✓ Feature i의 영향이 모델 B 보다 A에서 많으면, feature i의 공헌도는 모델 B보다 A에서 항상 크거나 같음

$$f_x'(z') - f_x'(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i)$$

$$\textit{for all inputs } z' \in \{0,1\}^M, \textit{ then } \phi_i(f', x) \geq \phi_i(f, x)$$

03 | SHAP (SHapley Additive exPlanation) Values

- Linear SHAP
  - ✓ $f(x) = \sum_{j=1}^{M} w_j x_j + b$
  - ✓ $\phi_0(f, x) = b$
  - ✓ $\phi_i(f, x) = w_j(x_j - E[x_j])$

- Deep SHAP (DeepLIFT + Shapley values)
  - ✓ $slope = \frac{Y - Y^{baseline}}{x_j - x^{baseline}} \rightarrow \frac{Y - E[Y]}{x_j - E[x_j]}$

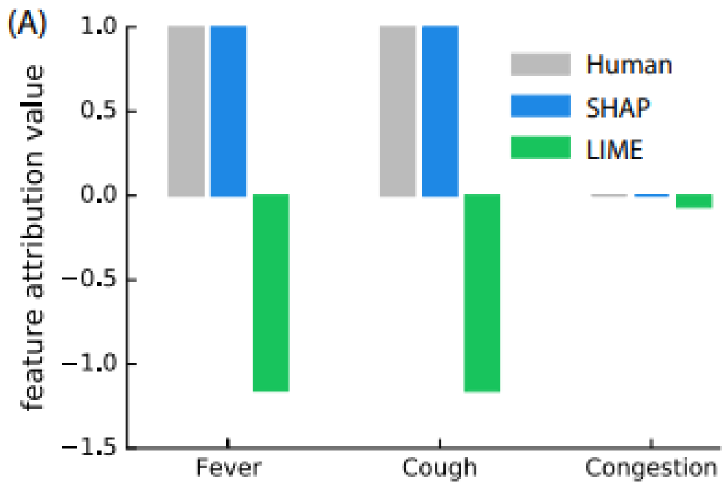
- Kernel SHAP (Linear LIME + Shapley values)

$$\Omega(g) = 0,$$

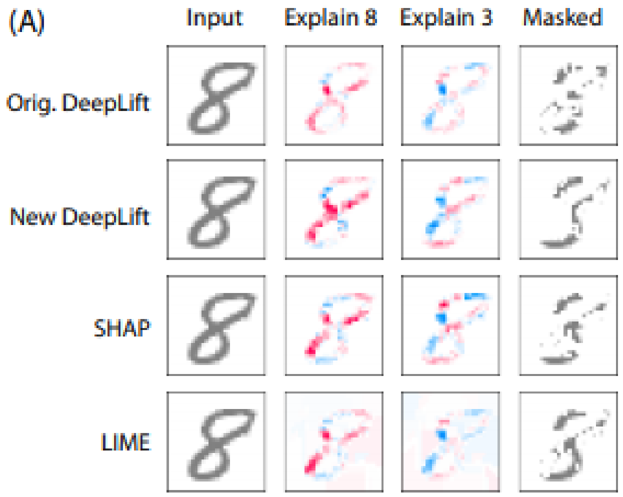
$$\pi_{x'}(z') = \frac{(M-1)}{(M\ choose\ |z'|)|z'|(M - |z'|)},$$

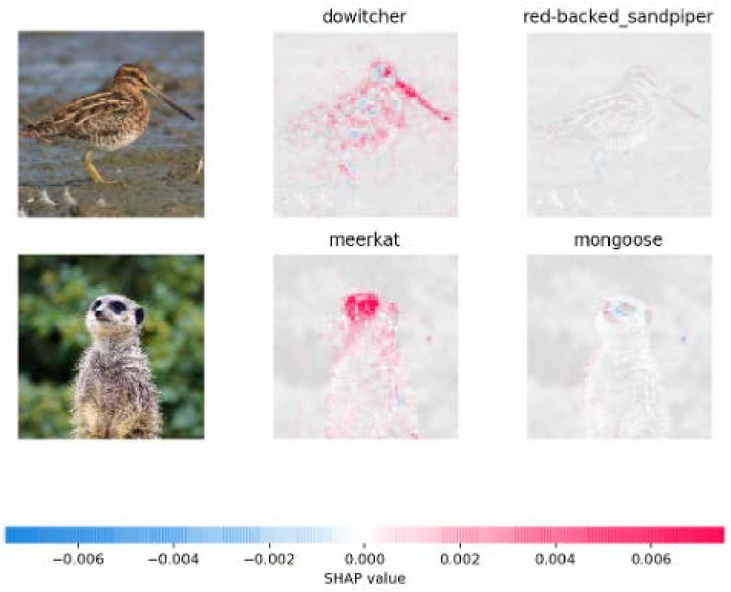
$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} \left[f(h_x(z')) - g(z')\right]^2 \pi_{x'}(z'),$$

- Sickness score



- Image data





Q&A

감사합니다.