A Unified Approach to Interpreting Model Predictions

https://arxiv.org/abs/1705.07874

LIME은, 개별적인 prediction에 대한 설명을 할 수 있는 방법론이다.

반면, SHAP은 모델 전체의 Feature Importance를 합리적인 방식으로 설명하는 방법론으로서 제안되었다.

Shapley Value 및 게임 이론에서 아이디어를 얻은 SHAP은, Model Interpretation 방면에서 가장 널리 쓰이고 있는 방법이며, 이전에 가장 각광받았던 LIME을 사장시켜버릴 정도로 아주 인기가 많다.

기존의 Feature Importance는 단순히 예측하는데 있어 변수가 어느 정도의 영향을 가지는지에 대한 정보만 준다. 반면, SHAP은 정, 부의 영향을 어느정도로 끼치는지까지의 정보를 제공한다.

Authors

ㄴ Scott M.Lundberg(Placement: MS Research)

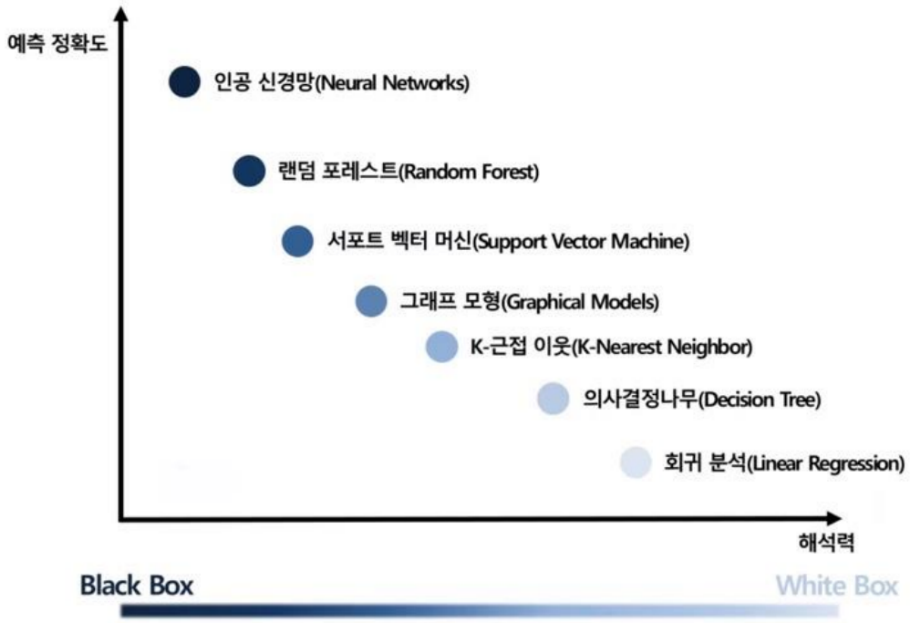
ㄴ Su-In Lee (Placement: Washington Univ.)

Introduction

How to ‘EXPLAIN’ a predictive model

- 예측에 대한 설명을 모델 자체, 즉 Explaination model로 정의
    - 설명 모델에 대한 새로운 분류인 Additive Feature를 제안, 이전의 XAI 방법론들이 분류에 포함되는 것을 보임
- Game Theory 기반, Additive Feature attribution method에는 오직 하나의 해만이 있음을 보이고, 새롭게 Feature Importance를 측정하는 방법인 SHAP을 제안
    - 사람의 직관과 잘 부합한다는 특징이 있음

Additive Feature Attribution Method이란?



예측력과 해석력에는 trade-off 관계가 존재한다. 현재 nn을 필두로 하는 여러 '복잡한' 모형의 경우, 전문가들 조차 해석이 어려워하는 경우가 존재한다. 이는 서비스 제공 차원에서도 엄청난 문제가 되는데, 왜냐하면

- 우리 모델 back testing이나 이런거 다 해봤어요 그니까 우리 모델 쓰세요

가 안된다는 것이다.

결국 의사결정을 하는 head들에게 이러한 결과를 '납득'시키고, 고객들에게 이러한 예측 결과를 '설명'하기 위해서는, 단순히 예측력이 높다고 applicable하지 않는다.

그렇기 때문에, 복잡한 구조의 모델을 설명하기 위해서는 보다 단순한 모델이 필요하며, 이렇게 활용하는 단순한 모델을

'EXPLAINATION MODEL'

이라 정의한다.

Explaination Model

이러한 ‘EXPLAINATION MODEL’에 대한 설명 이전, 먼저 여러 notation을 살펴보자.

- f: 설명이 필요한 원래 모델
- g: f를 설명하기 위한 단순화된 Explanation model
- x: f에 들어가는 원래 input; $f(x)$는 input x에 대한 output
- x': g의 input으로 들어가는 x의 단순화된 형태
- h_x: x'을 x로 매핑하는 함수; $x = h_x(x')$

출처: https://kicarussays.tistory.com/32

LIME과 같이 Local, 즉 각각의 예측 결과를 설명할 수 있는 Local Method들은,
특정 metric 하에서 x’와 가까운 z’에 대해,
설명 모델 g가

$$g(z') \approx f(h_x(z'))$$

를 만족하도록 해야할 것이다.

즉,
x를 설명하기 위한 모델이 g이므로,

$$f(x) \approx g(x')$$

를 만족한다.

따라서 x'와 가까운 z'들이

$$g(z') \approx f(h_x(z'))$$

를 만족해야 g가 국소적으로 합리적인 설명 모델이라고 해석할 수 있는것이다.

Definition 1.

Additive Feature Attribution method는 다음을 만족하는 이진 변수에 대한 '선형' Explaination model g를 갖는다.

왜 선형인가 하니, g는 해석가능한 간단한 모델이어야 하기 때문이다.

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

$$z' \in \{0,1\}^M$$

$$\phi_i \in \mathbb{R}.$$

input의 feature 수를 줄이는 것뿐만 아니라, 0이 아닌 feature들의 값을 모두 1로 만든 벡터 z′

이 설명 모델 g의 input이 되고, 이 g는 M 개의 변수를 갖는 선형식으로 구성되어있다.

결국 설명 방법론을 통해 추출한 Explanation model g를 통해 우리는 줄어든 변수 중에서 어떤 변수가 중요한지 판단할 수 있을 것이다.

Shapley regression value

다중공선성(multicollinearity)이 존재하는 선형 모델에서의 변수 중요도(feature importance)입니다. 다중공선성은 선형 모델에서 독립성 가정을 위배하는 성질이지만,

Shapley regression value는 다중공선성이 있는 선형 모델에서도 사용할 수 있는 값이라고 언급하고 있습니다.

Shapley regression value는 각 변수들이 학습에 포함되었을 때, 얼마나 모델의 성능에 영향을 미치는지에 따라 중요도(importance value)를 부여한다. 중요도를 계산하는 방식은 다음과 같다.

When,

F가 모든 변수의 집합

S는 F의 subset,

x_s와 f_s는 각각 S에 포함된 변수만 포함된 input과 학습 모델일 때,

Feature i의 중요도 sigma i는 다음과 같이 계산된다.

여기서 sigma i는 feature i가 포함되지 않은 F의 모든 subset에 대해, i를 추가했을 때의 output변화량의 평균을 측정한 것이다.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

Simple Properties Uniquely Determine Addtitive Feature Attributions / Game Theory

Explanation model을 산출하는 과정은 게임 이론과 매우 유사하다. 게임 이론은 다음과 같은 가정에서 유래하는데,

- 게임 내의 모든 플레이어들은 게임의 결과에 영향을 미칠 수 있음.
- Explanation model의 변수의 계수들은 예측 결과에 영향을 미칠 수 있음.
- Explanation model의 변수의 계수 = 게임 내의 모든 플레이어, 모델의 예측 결과 = 게임의 결과

Additive Feauter Attribution 문제가 게임 이론에 based하였을 떄, 이러한 additive feature attribution 문제, 즉 게임 이론 문제가 **오직 하나의 해를 갖기 위해서 만족해야 하는 네 가지 공리**(axiom)가 존재한다.

, 본 논문에서는 Additive feature attribution methods가 오직 하나의 해를 갖기 위해 가져야 할 조건들을 다음과 같이 제시한다.

조건 1.

Local Accuracy

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

$$x = h_x(x')$$

h_x는 mapping function이며, 이러한 mapping function을 이용해 x를 변환한 뒤 explanation model에 대입하는 것이며, 원래 모델 f로부터 생성된 output인 f(x)를 설명하기 위한 g가 Local Accuracy를 만족해야 한다는 조건이다.

h_x는 simplified input x'를 본래의 x로 변환해 주는 임의의 function이다.

Simplified input x'는, z'~x' 가정을 통해 x'가 아닌

$$z' \in 0, 1^M$$

로 표현된다.

즉, additive feature attribution은, 원래의 input이 아닌, simplified input z'를 통해

$$g(z') \approx f(h_x(z'))$$

를 만족하는 explanatory model g를 만드는 방법인 것이다.

이러한 additive feature attribution 방식에는 LIME, DeepLift, Shapely values 등이 존재하는데,

이전에 가장 많이 쓰였던 local한 prediction 결과의 '근거'를 제시하는 LIME의 경우에는 해가 오직 하나 존재하는 지 여부는 증명해내지 못했다. 하지만 본 논문에서는, 이 '해'의 유일

성을 밝힘으로서 그 contribution을 가지는 측면이 존재한다.

조건 2.

Missingness

$$x_i^{'} = 0 \Longrightarrow \phi_i = 0$$

단순화된 input인 x'에서 Feature i 에 해당하는 값이 0이라면,

해당 변수의 영향력(sigma i)도 0이라는 것이다.

조건 3.

Consistency

$$f_x(z') = f(h_x(z'))$$

그리고

$$z' \setminus i \Leftrightarrow z_i^{'} = 0$$

일 때, 모든 모델 f와 f',

모든 input

$$z' \in \{0,1\}^M$$

에 대하여,

$$f_i^{'}(z') - f_i^{'}(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \Longrightarrow \phi_i(f', x) \geq \phi_i(f, x)$$

Consistency는 만약 모델이 f에서 f' 으로 바뀌었을 때,

Feature i의 영향력이 더 커졌다면,

f′(x) 의 Explanation model의 Feature i 에 대한 계수가

f(x) 보다 더 크다는 뜻이다.

이러한 조건 3가지를 모든 만족할 때, 다음과 같은 정리가 성립한다.

Theorem)

Property 1, 2, 3을 만족할 때, Additive feature attribution methods(Definition 1)를 통해 나온 Explanation model g는 오직 하나 존재한다.

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

여기서 |z'|는 z'에서 0이 아닌 원소의 개수이다.

SHAP (SHapley Additive exPlanation)

SHAP는 유일한 Additive Feature Importance Measure로서 제시되었다.

Shapley values의 Conditional Expectation 버전으로 **Simplified Input을 정의하기 위해 정확한 f값이 아닌 f의 Conditional Expectation을 계산**한다

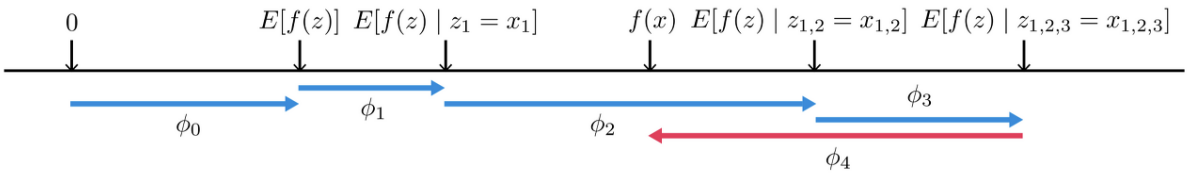


Figure 1: SHAP (SHapley Additive exPlanation) values attribute to each feature the change in the expected model prediction when conditioning on that feature. They explain how to get from the base value $E[f(z)]$ that would be predicted if we did not know any features to the current output $f(x)$. This diagram shows a single ordering. When the model is non-linear or the input features are not independent, however, the order in which features are added to the expectation matters, and the SHAP values arise from averaging the ϕ_i values across all possible orderings.

Tree SHAP)

Tree Ensemble Model의 Feature Importance를 측정하는 기존의 방법(ex. Gini, Split Count 등)들은 모델(or 개별 Tree)마다 일관되지 않은(Inconsitency) 한계점이 존재한다.

단적인 예로 RandomForest에서의 Feature Importance는 information gain에 따라 node를 나눈다.

허나, scikit-learn에서의 feature importance는 cardinality가 높은 경우에 대해 더 information gain이 높게 판단하려 하는 bias가 존재한다고 하여, 일반적으로는 A/B test 및 eli5의 permuation importance(이 또한 inconsistency가 존재)를 함께 사용한다.

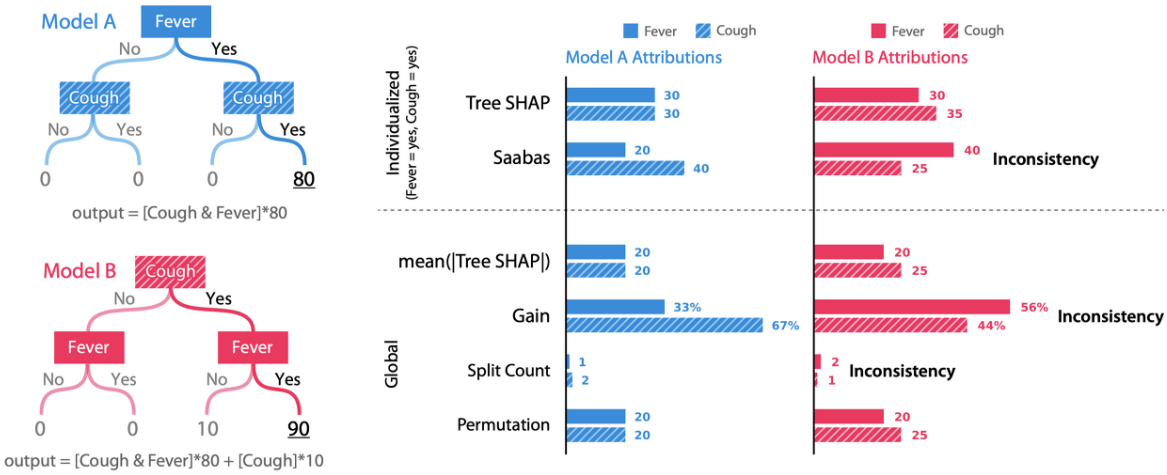


Figure 1: Two simple tree models that demonstrate inconsistencies in the Saabas, gain, and split count attribution methods: The Cough feature has a larger impact in Model B than Model A, but is attributed less importance in Model B. Similarly, the Cough feature has a larger impact than Fever in Model B, yet is attributed less importance. The individualized attributions explain a single prediction of the model (when both Cough and Fever are Yes) by allocating the difference between the expected value of the model's output (20 for Model A, 25 for Model B) and the current output (80 for Model A, 90 for Model B). The global attributions represent the overall importance of a feature in the model. Without consistency it is impossible to reliably compare feature attribution values.

위와 같이, 같은 데이터로 학습한 모델인데, 다른 importance를 뱉어낸다는 것은 모델 신뢰도에 있어 좋지 않은 역할을 준다.

하지만 Tree-SHAP은 split 순서와 무관하게 일관된 feature importance를 계산할 수 있다.

SHAP의 한계

항상 ML 모델을 해석한다는 차원에서 항상 유의해야 할 점이 SHAP에도 매한가지이다.

SHAP의 가장 큰 contribution은,

각각의 feature '값'에 따라 target y가 어떻게 변화하는 지를 설명할 수는 없었으나, 이에 대한 solution을 제공한 것이다.

하지만, 이것은 '현상'을 파악한 것에 불과하지,

'왜' 이러한 현상이 발생하는지까지는 알 수 없는 것이다.

SHAP의 공식 doc에서도 이와 같이 명시한다.

Be careful when interpreting predictive models in search of causal insights

https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be careful when interpreting predictive models in search of causal insights.html

Predictive model들이 Causal Question들에 대해 대답할 수 있는 경우에 대해서도 명시를 해 보았지만(Dendrogram적 접근 - 묶이는 경우에 대해서는 redundant하다는 결론, unobserved confounding에 대한 탐구 및 experimental 접근이 필요하다는 명시)

결국 SHAP 또한 Causal Inference까지는 해결할 수는 없다.

따라서 추가적인 casual-tree와 같은 알고리즘과 결부시켜 고도화시킴으로서 SHAP을 발전시킬 수 있을 것이라 기대한다.