



“why should I trust you?” Explaining the predictions of any classifier

이상용 / 2020-03-06



Computational Data Science LAB



“why should I trust you?” Explaining the predictions of any classifier

Computational Data Science LAB

목차

1. INTRODUCTION
2. LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS
3. SIMULATED USER EXPERIMENTS



논의사항 및
결정사항

관련문서

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

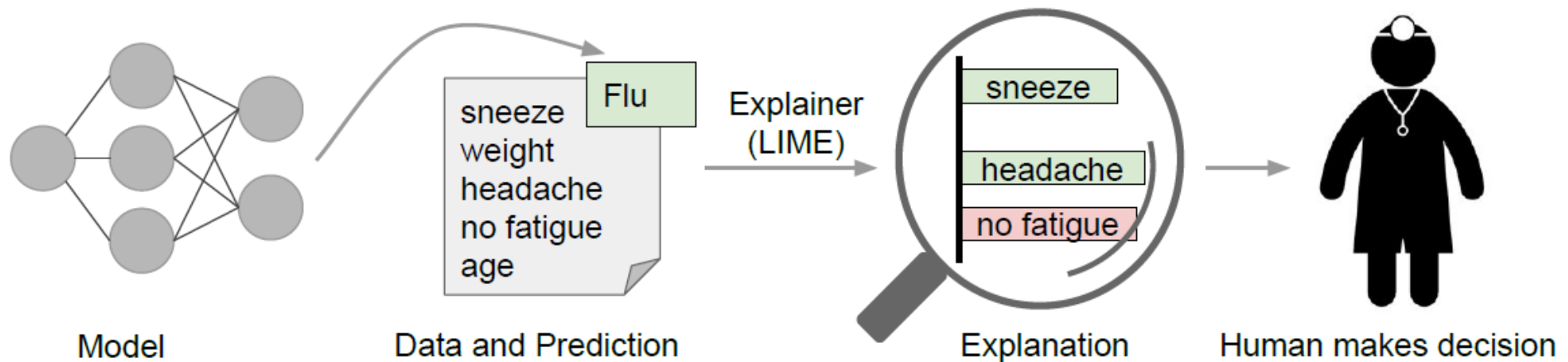


CONTENTS

1. INTRODUCTION
 2. LOCAL INTERPRETABLE MODEL-AGNOSTIC
EXPLANATIONS
 3. SIMULATED USER EXPERIMENTS
- 
- 

01 | INTRODUCTION

- 의사결정을 위해 여러분야에서 머신러닝 기법이 활용되지만, 대부분은 black-box 모형이기에 해석이 어렵지만 최근 해석을 위한 연구가 활발히 진행 중
- 본 논문에서는 결과에 대해 신뢰하는 것을 ‘trusting a prediction’, ‘trusting a model’로 정의하며, 두 정의의 explanation을 제공하는 Local Interpretable Model-agnostic Explanation (LIME)을 제안함



02 | LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

Interpretable Data Representations

- Black-box 모델을 해석하기 위해 feature 와 interpretable data representations 을 나눔
 - ✓ 실제 모델에 투입되는 것 feature
 - ✓ 인간이 해석할 수 있는 것 interpretable data representations
- Text data
 - ✓ 단어의 유무를 나타내는 binary vector
- Image data
 - ✓ 이미지 상에서 비슷한 부분 (super-pixel, segment)의 유무를 나타내는 binary vector
- Binary vector는 “presence” or ”absence”를 나타내며 interpretable data representations을 위해 d 차원의 원래 데이터 $x \in \mathbb{R}^d \rightarrow x' \in \{0,1\}^{d'}$ 로 나타냄

02 | LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

Fidelity-Interpretability Trade-off

- Explanation을 위한 본질적인 기준으로 fidelity와 interpretability가 존재함
 - ✓ Fidelity : 데이터 공간의 전체에서 모델을 설명하는 것은 어렵지만, 국소적인 데이터 공간에서는 의미 있는 모델로 설명 가능
 - ✓ Interpretability : 입력변수와 반응의 정량적인 이해를 의미
 - ✓ 국소 공간에서의 중요한 feature가 전체 공간에서는 중요하지 않을 수 있고, 그 반대도 가능함
 - ✓ 국소 공간의 explanation은 전체 공간을 설명하기 어렵지만, 전체 공간의 explanation은 국소 공간 설명 가능 but, 모델의 복잡성이 높아져서 interpretability가 떨어짐 (Fidelity-Interpretability Trade-off)

02 | LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

Fidelity-Interpretability Trade-off

- Local Fidelity와 Interpretability 반영하는 explanation은 다음과 같이 수식화 할 수 있음
 - ✓ $f : \mathbb{R}^d \rightarrow \mathbb{R}$, f 는 학습한 모델 (black-box model)
 - ✓ $g : X' \rightarrow \mathbb{R}$, $g \in G$, prediction에 대한 **explanation**을 위한 모형 (linear model, decision trees, ...)
 - ✓ $\Omega(g)$: 모형의 복잡도 (나무의 depth, non-zero 계수의 개수, ...)
 - ✓ $\pi_x(z)$: 유사도 측도로써, 설명하고자 하는 데이터 x 와 다른 데이터 z 간의 유사도 측도
 - ✓ Local Fidelity와 Interpretability 반영하는 minimize 문제를 정의

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

- ✓ G 는 explanation families, \mathcal{L} 은 fidelity functions(local-aware loss), Ω 는 complexity measures

02 | LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

Sampling for Local Exploration

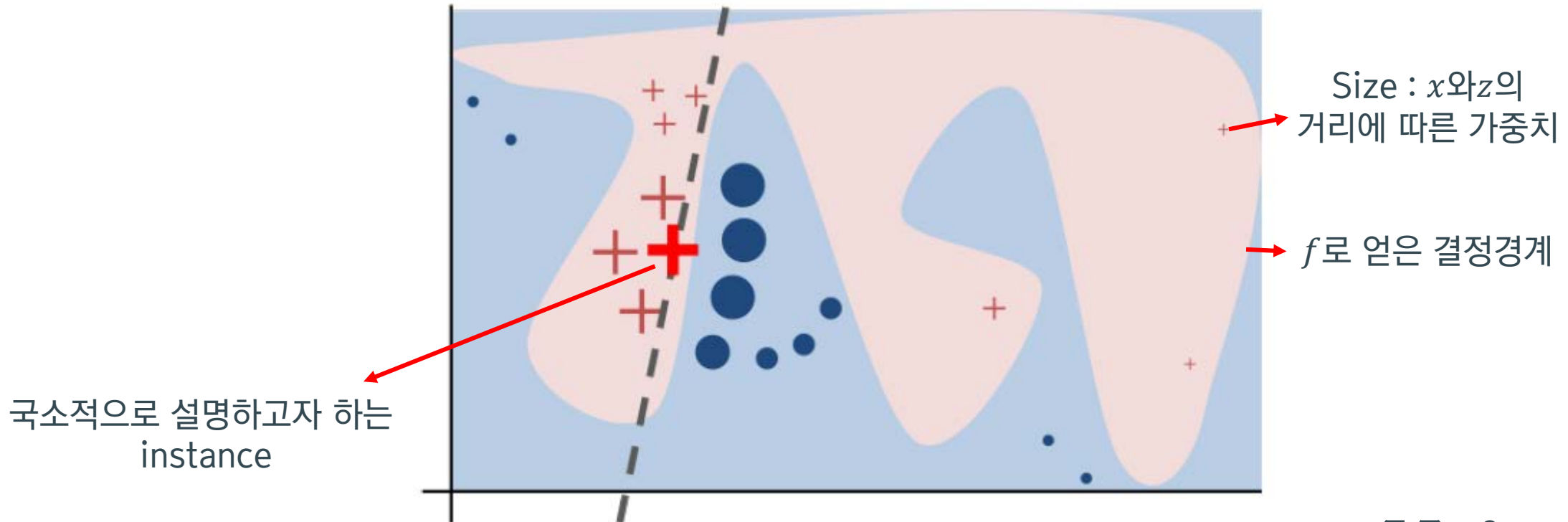
- Local explanation을 생성할 때, 설명하고자 하는 데이터에서 약간의 변화를 주고 어떤 feature가 변했을 때 결과가 크게 변하면, 모형에 영향을 많이 끼친다는 것을 의미함
- f 를 국소적으로 설명하기 위해 다음과 같은 과정을 거침
 - x 의 interpretable representation x' 을 랜덤하게 샘플링 함. 샘플링시 고려되는 '선택되는 representation' + representation의 수' 모두 **uniformly random sampling**, 여기서 얻어진 샘플은 **perturbed sample**, $z' \in \{0,1\}^{d'}$ 으로 표현
 - z' 을 f 에 넣기 위해 feature의 형태를 z 로 바꿈
 - $f(z)$ 를 계산하고 Local-aware loss를 계산

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

02 | LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

Sampling for Local Exploration

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



02 | LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

Sparse Linear Explanation

- G 를 linear model family로 고려, $g(z') = w_g z'$ 형태
- 유사도 척도는 cosine distance를 사용한다고 할 때, $\pi_x(z) = e^{-\frac{D(x,z)^2}{\sigma^2}}$
- 이때 locality-aware loss는 다음과 같음

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

- 이때 interpretable representation의 수를 조절하여 모형 복잡도 $\Omega(g)$ 를 조절 가능
- $\Omega(g) = \infty \mathbb{I} \left[\|w_g\|_0 > K \right]$, L0 norm 사용 (논문에서는 K-LASSO라고 부름)

02 | LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

Sparse Linear Explanation

- 따라서, 최적화 문제는 다음과 같음

$$\xi(x) = \operatorname{argmin}_{w_g} \left[\sum_{i=1}^N e^{-\frac{D(x_i, z_i)^2}{\sigma^2}} (f(x_i) - g(z'_i))^2 + \infty \mathbb{I} \left[\|w_g\|_0 > K \right] \right]$$

- pseudocode

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 $i \in \{1, 2, 3, \dots, N\}$ **do**

$z'_i \leftarrow \text{sample_around}(x')$

$\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$

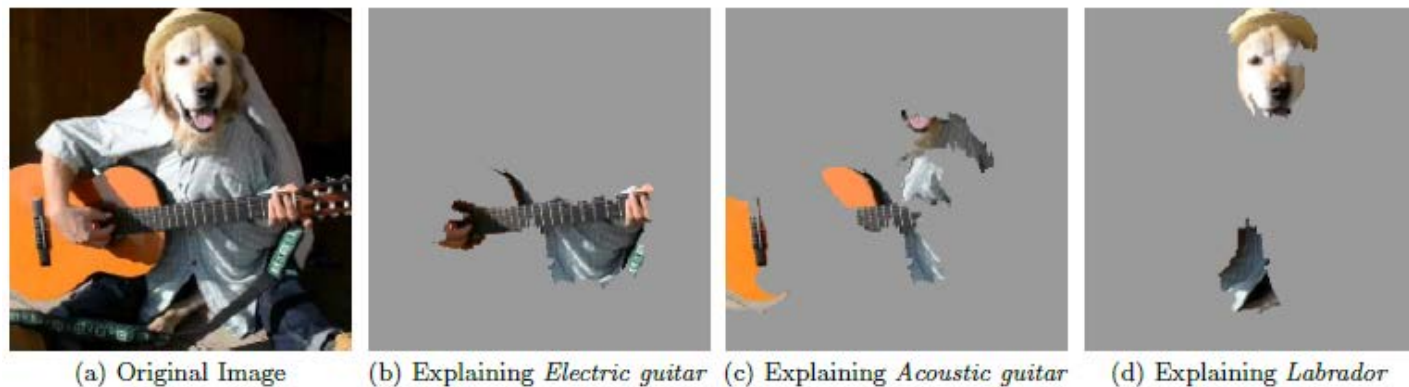
end for

$w \leftarrow \text{K-Lasso}(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target

return w

03 | SIMULATED USER EXPERIMENTS

- Explaining an image classification prediction (Google's Inception neural network)



- sentiment analysis datasets (books and DVDs, 2000 instances each)

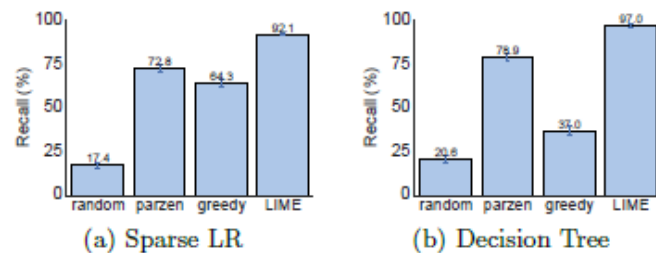


Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.

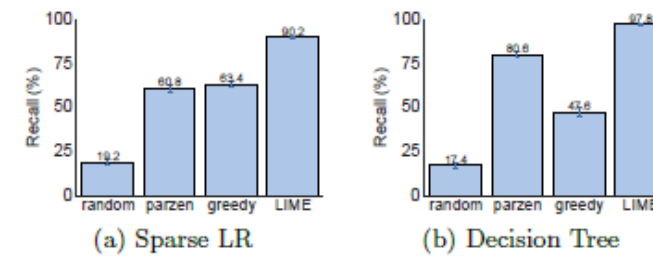


Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

Q&A

감사합니다.