LIME

Mose Park

Department of Statistical Data Science
University of Seoul

Selective. Lab

April 9, 2024

Index

Overview

Why XAI

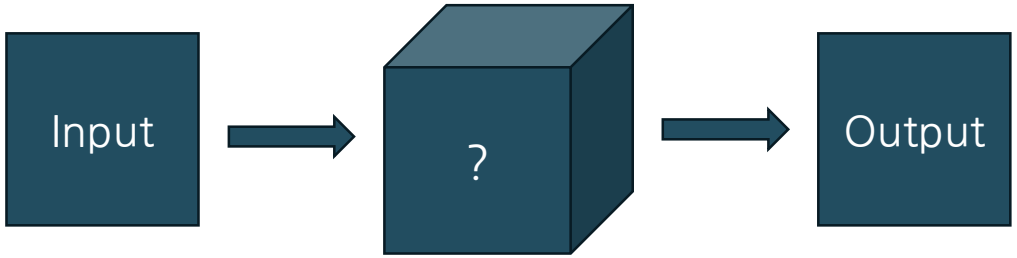
LIME

Experiment

1

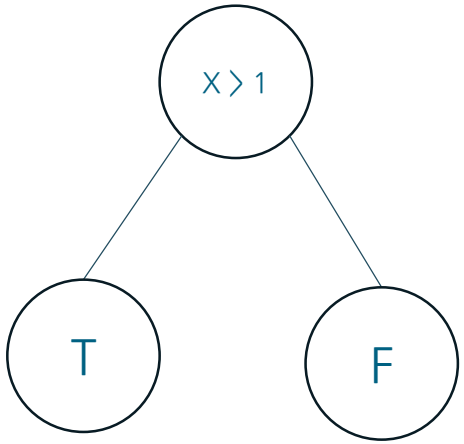
Introduction

Black box vs Interpretable



Black box model

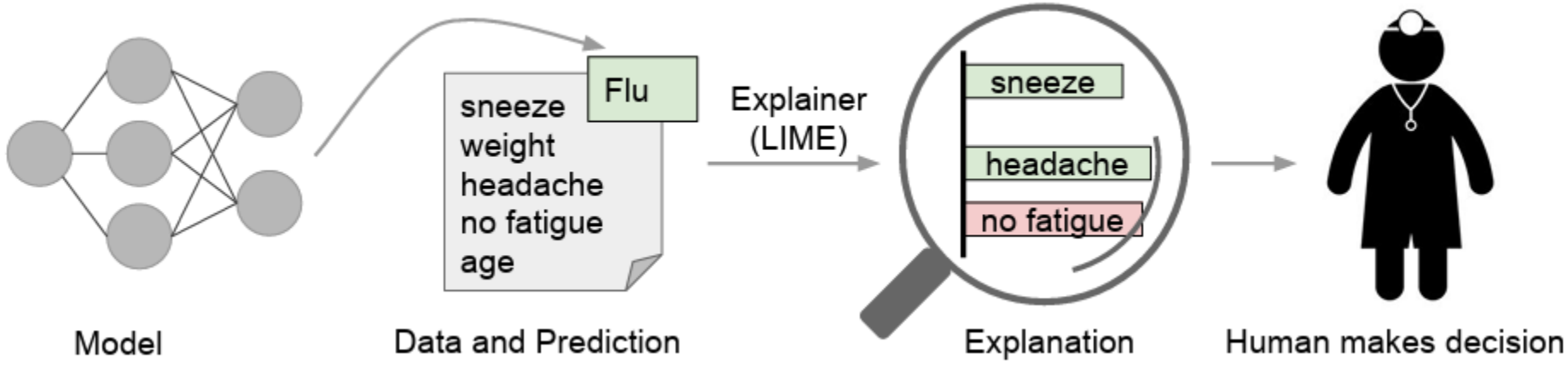
▶ 블랙박스는 데이터와 예측 결과와의 관계를 규명할 수 없음



Tree model

▶ 해석가능한 모델은 왜 그런 결과가 나왔는지 설명가능

Why ?



- black box model은 환자가 독감이라고 결정하는 것에서 끝남
- LIME은 과거 증상들이 무엇이었는지 해석할 수 있음
- 의사는 모델의 예측을 신뢰(trust)할지 결정함

2

LIME

Objective function

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

G : a class of interpretable models — ex) 'g'는 각각 선형 모델, 의사결정나무 등이 될 수 있음

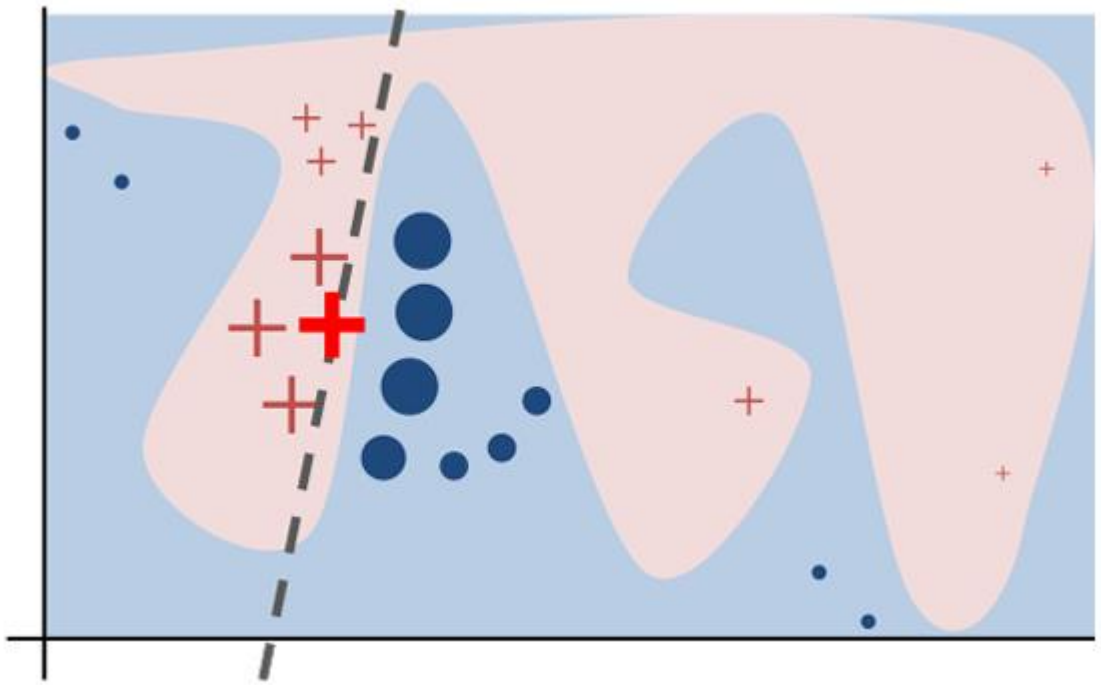
$\Omega(g)$: measure of complexity — ex) 0이 아닌 β의 개수, 트리의 깊이 등

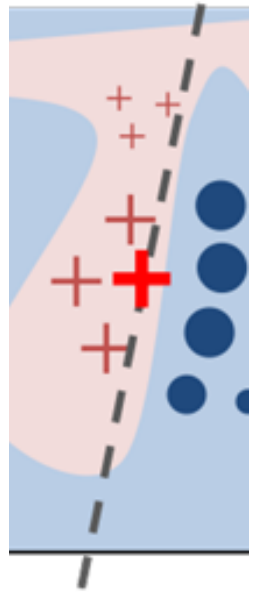
f : model being explained — ex) 설명되어야 할 블랙박스 모델

π_x : proximity measure between an instance z to x

- z가 무엇이고 π 의 역할은?

Sampling





Local

▶ 블랙박스 모델의 복잡한 결정 함수 f가 파란색/분홍색 배경으로 나뉨

▶+ 기호는 설명되고 있는 인스턴스 'z' 를 의미 ← 이것을 샘플링함

▶샘플링된 + 들은 인스턴스와의 거리를 통해 가중치를 부여, 상대적 크기가 근접성(거리)을 의미

Algorithm 1

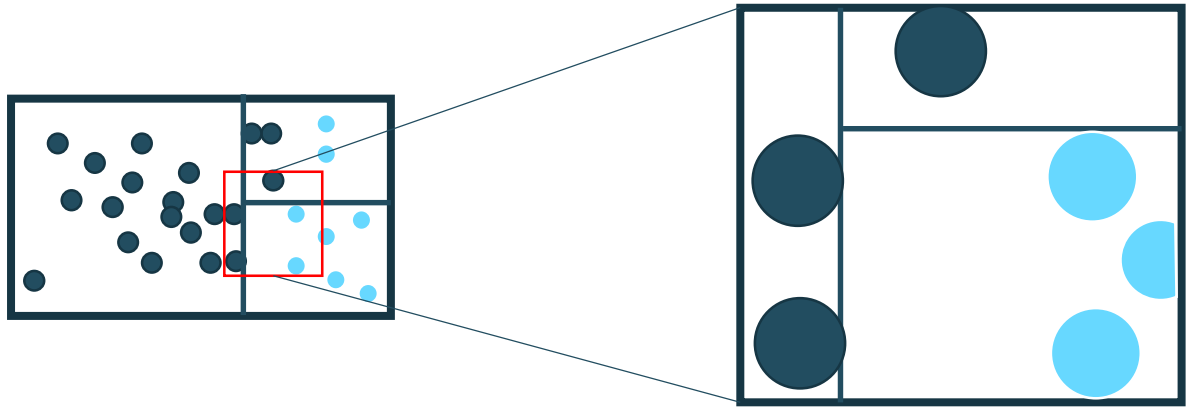
Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

1. Initialize Z.
2. For $i = 1$ to N do
   (a) Sample z_i' around x'.
   (b) Compute $Z = Z \cup \{(z_i', f(z_i), \pi_x(z_i))\}$.
3. End for.
4. Compute $w = \text{K-Lasso}(Z, K)$ with z_i' as features, $f(z_i)$ as target.
5. Return w.

3

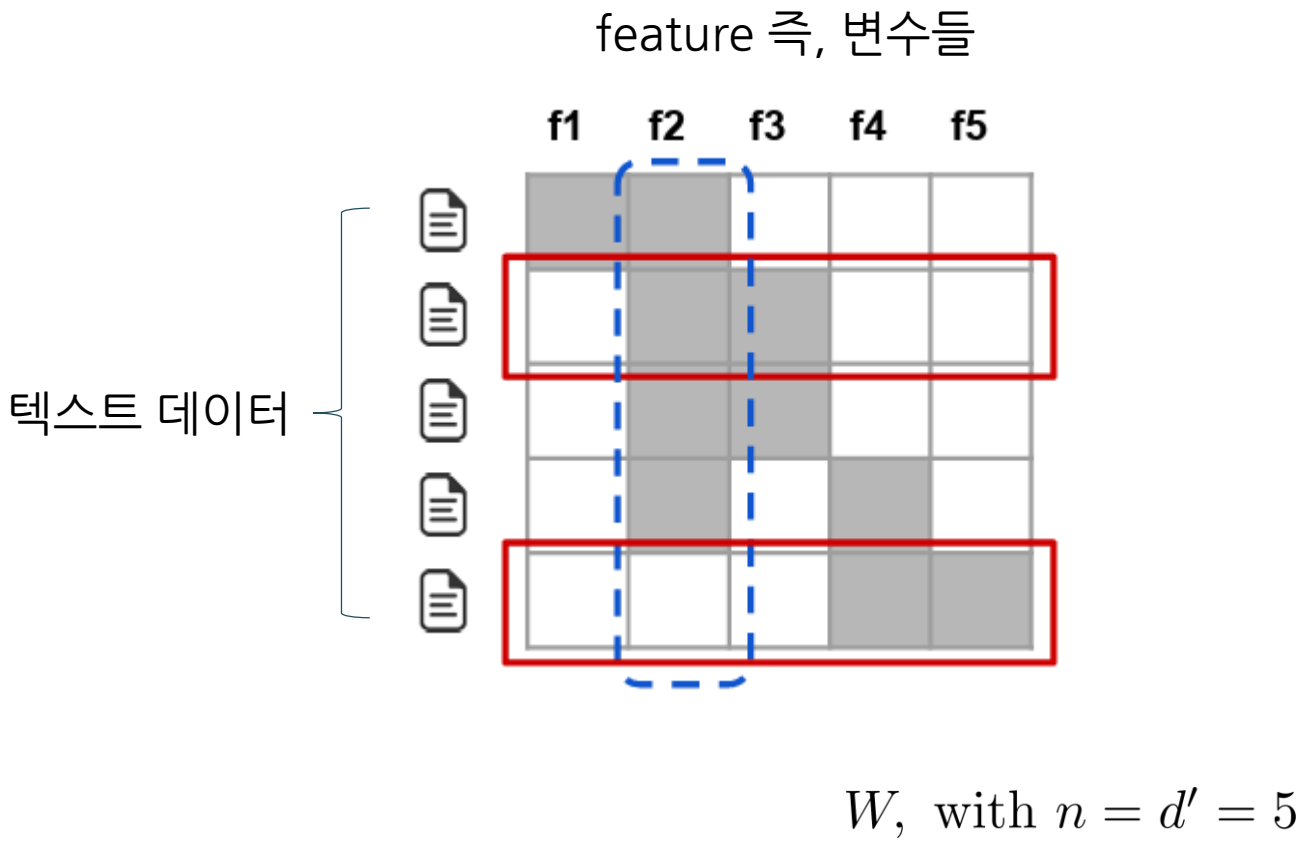
SP-LIME

Why SP-LIME?



- 단일 예측이기 때문에 모델 전체 해석의 일반화가 어려움
- 개별 인스턴스를 설명하는 것 대신에 인스턴스들의 집합을 설명하는 것을 목표

Expanation matrix



$$W, \text{ with } n = d' = 5$$

- 회색칸은 instance를 의미, f2가 f3에 비해 importance score가 높기 때문에 $I_2 > I_3$
- 2번째 행은 3번째 행과 유사한 설명을 가진 인스턴스기 때문에 동시에 고르면 안됨 2, 5과 최적

Pick Step

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

$$\text{Pick}(W, I) = \arg \max_{V : |V| \leq B} c(V, W, I) \tag{4}$$

- 식 (3)은 주어진 W와 I에 대해 집합 V 내의 적어도 한 인스턴스에 나타나는 특성들의 총 중요도를 계산하는 집합 함수 C
- 커버리지를 최대화하는 budget 내에 허락할 수 있는 index set V를 찾아주는 것이 목표

Algorithm 2

Require: Instances X, Budget B

- For all $x_i \in X$ do

  $W_i \leftarrow \text{explain}(x_i, \bar{x}_i)$ // Using Algorithm 1

  End for

- For $j = 1$ to d do

  $I_j \leftarrow \sum_{i=1}^{n} W_{ij}$ // Compute feature importances

  End for

- Initialize $V \leftarrow \emptyset$

- While $|V| < B$ do

  $V \leftarrow V \cup \{\arg\max_i c(V \cup \{i\}, W, I)\}$ // Greedy optimization of (4)

  End while

- Return V

NP hard 로 인한 수정
→ submodular pick (근사)

4

Experiment

Are Explanation faithful?

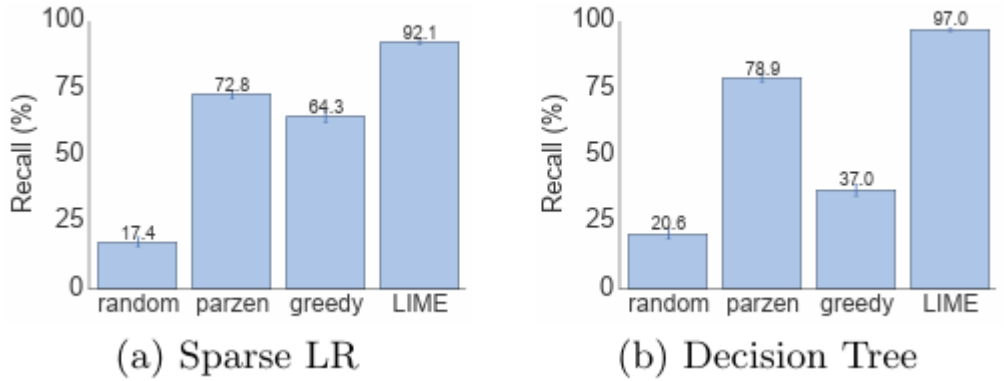


Fig. 6

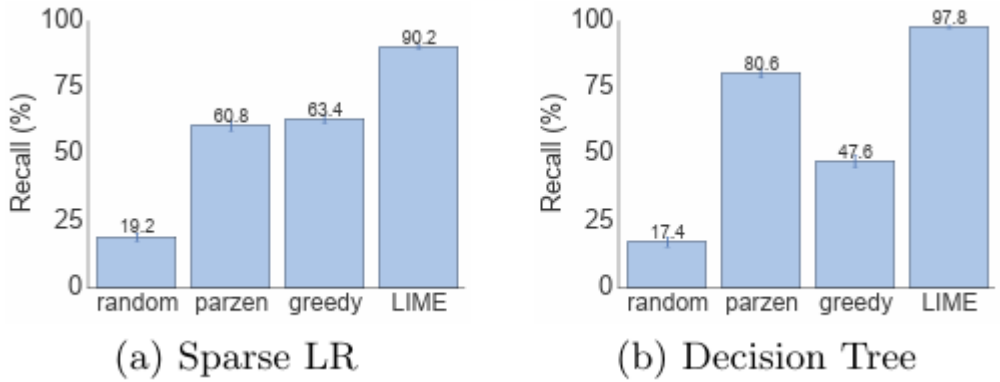


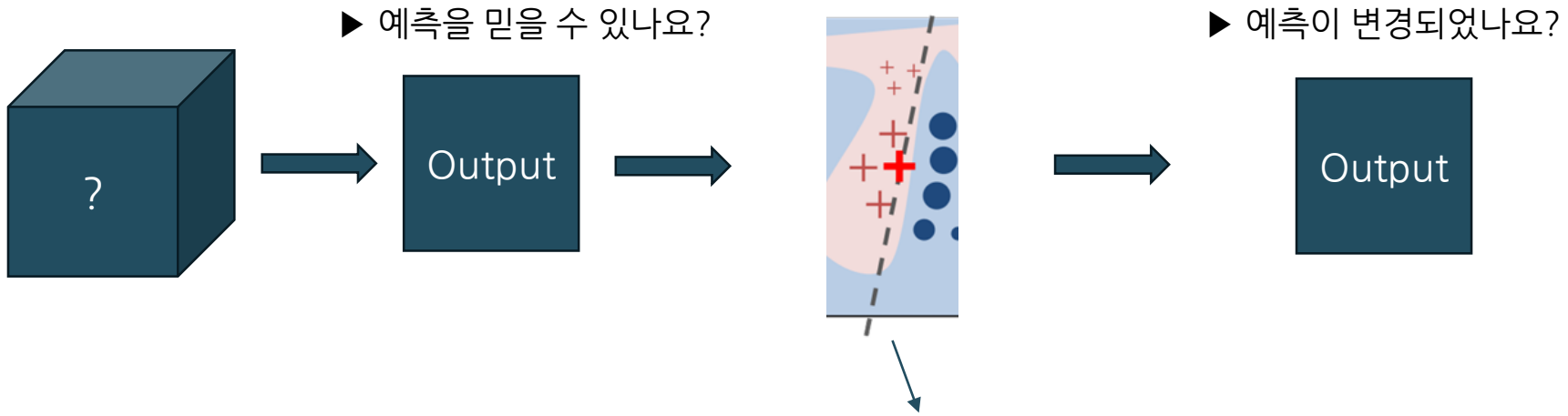
Fig. 7

▶ Recall은 LR과 DT에서 산출한 10개 feature들을 LIME에서 잘 Cover하는지에 대한 비율을 의미

gold set

▶LIME이 대부분 세팅에서 우수

Should I trust prediction?



변경 여부
- ○ :"untrustworthy" 즉, 신뢰할 수 없는 feature가 예측에 중요한 역할을 함
- X : "trustworthy" (labeling)

Table 1

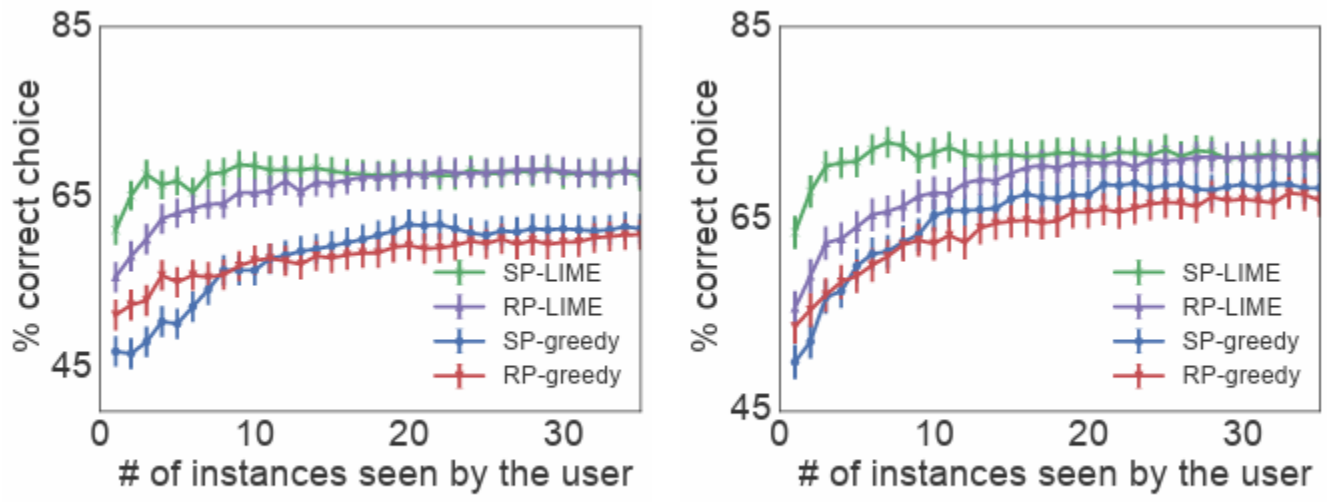
	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	**96.6**	**94.5**	**96.2**	**96.7**	**96.6**	**91.8**	**96.1**	**95.6**

▶ F1은 100회 실험 돌리고 평균값으로 계산

▶ Recall과 Precision 모두 LIME이 우수

Can I trust model?

Figure 8



▶ SP-LIME 을 통해 나온 instance 중 trustworthy한 instance의 비율