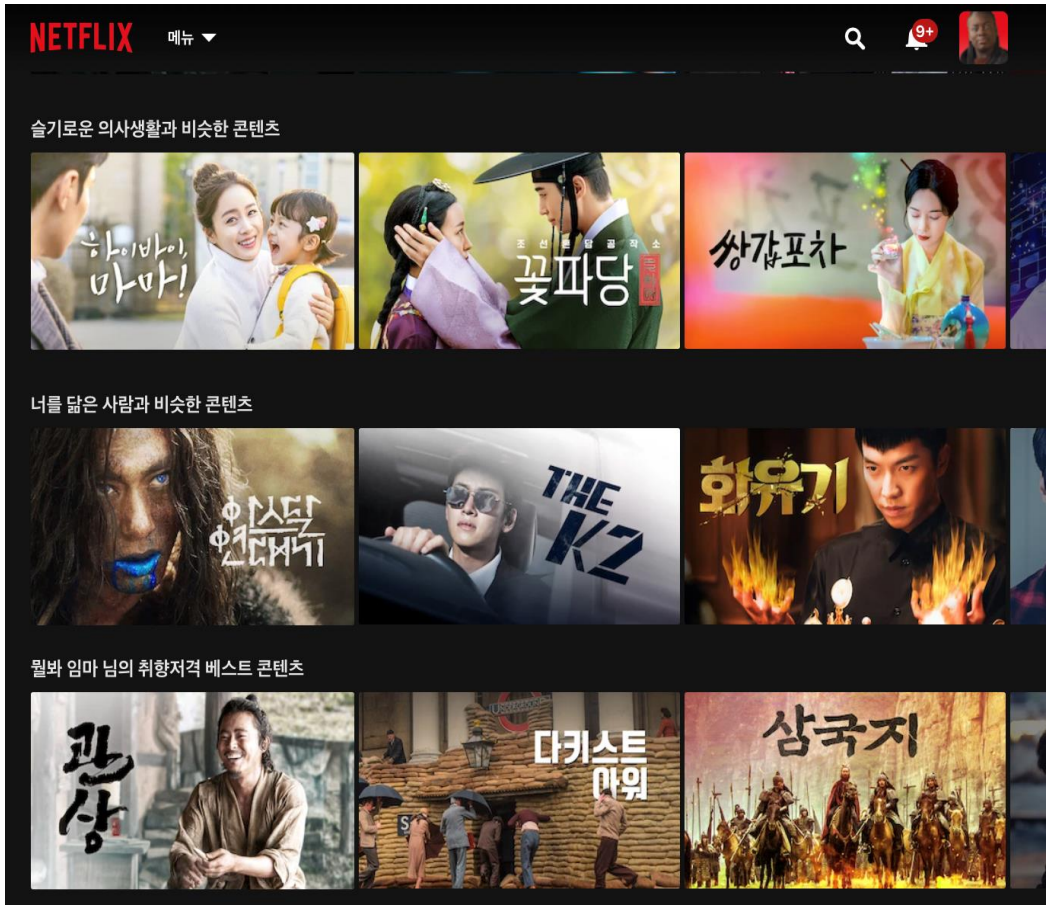
“Why Should I Trust You?”
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

KDD 2016

INTRO



OTT의 추천 시스템

이 사람들은 존재하지 않습니다. 모두 GAN이 만든 가상의 인물입니다.

ChatGPT

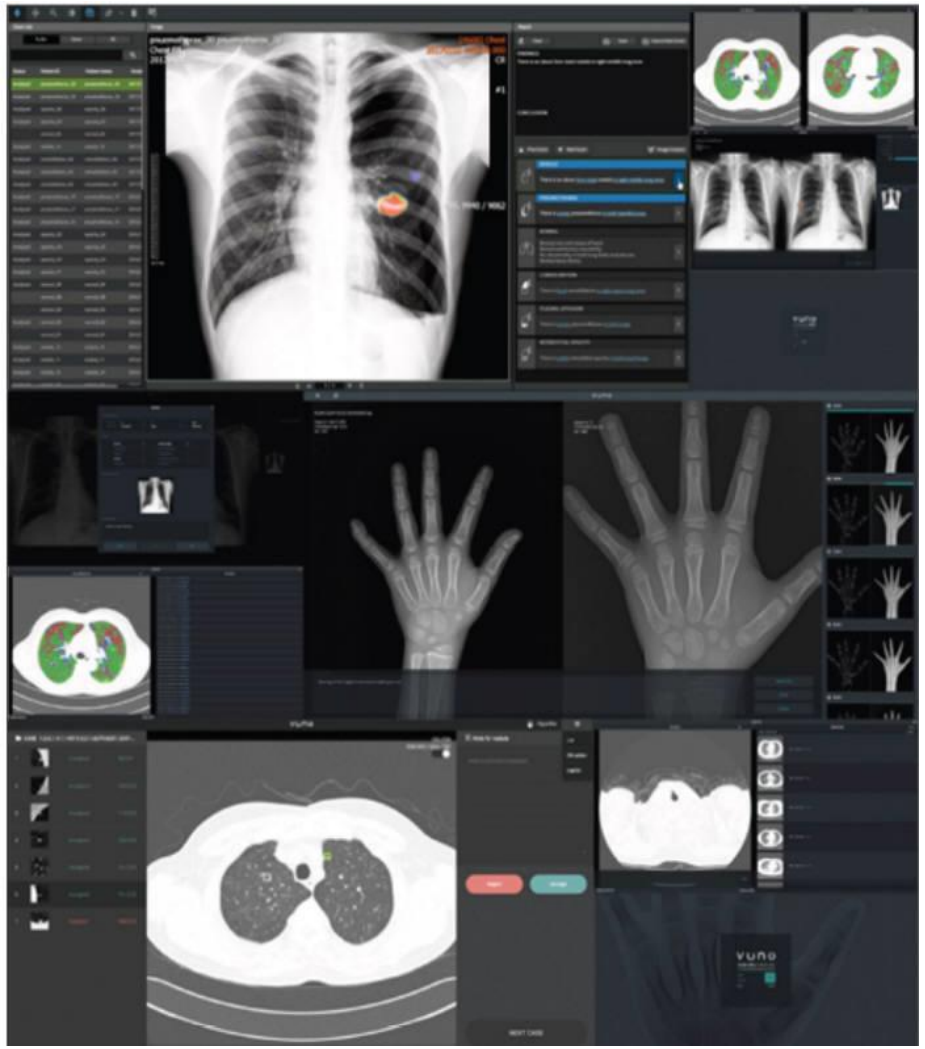
창작 영역에서의 AI

INTRO

Deep Learning 이 발전
모델을 믿을 수 있을 것인가가 중요

타이완에선 AI가 판사...범죄별 '맞춤 모델' 적용까지

작성 2023.02.08 17:29 조회 1,191 프린트



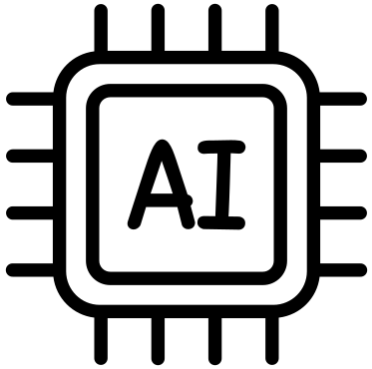
DEFINITION

모델을 신뢰한다?

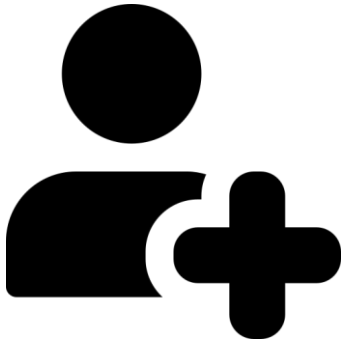
1. 개별 예측을 믿는다
2. 모델 자체를 믿는다

DEFINITION

의료 기록



암 여부 예측 모델



1. 개별 예측에 대한 신뢰
(추가 검사, 치료 권장 등)

2. 모델 자체에 대한 신뢰
(자동화 된 암 검진 시스템으로 사용)

DEFINITION

모델을 신뢰한다?

1. 개별 예측을 믿는다 → 개별 예측에 대한 설명 제공: LIME
2. 모델 자체를 믿는다 → 대표적인 instance 집합 구성: SP-LIME

METHOD

LIME (Local Interpretable Model-agnostic Explanation)

목표 : 모든 Classifier에 대해,

interpretable + local faithful 한 설명 제공

→ 특정 예측 값 근처에서 해석 가능한 모델 학습

METHOD

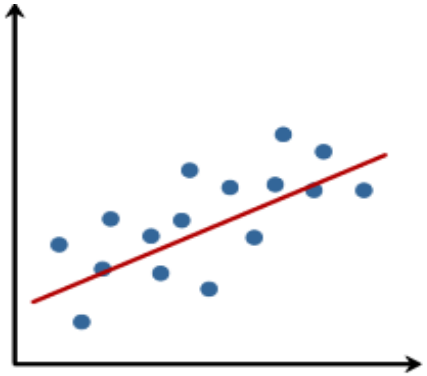
LIME (Local Interpretable Model-agnostic Explanation)

목표 : 모든 Classifier에 대해,

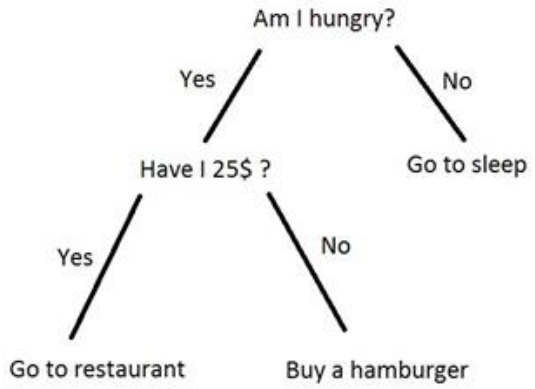
interpretable + local faithful 한 설명 제공

→ 특정 예측 값 근처에서 해석 가능한 모델 학습

Linear model

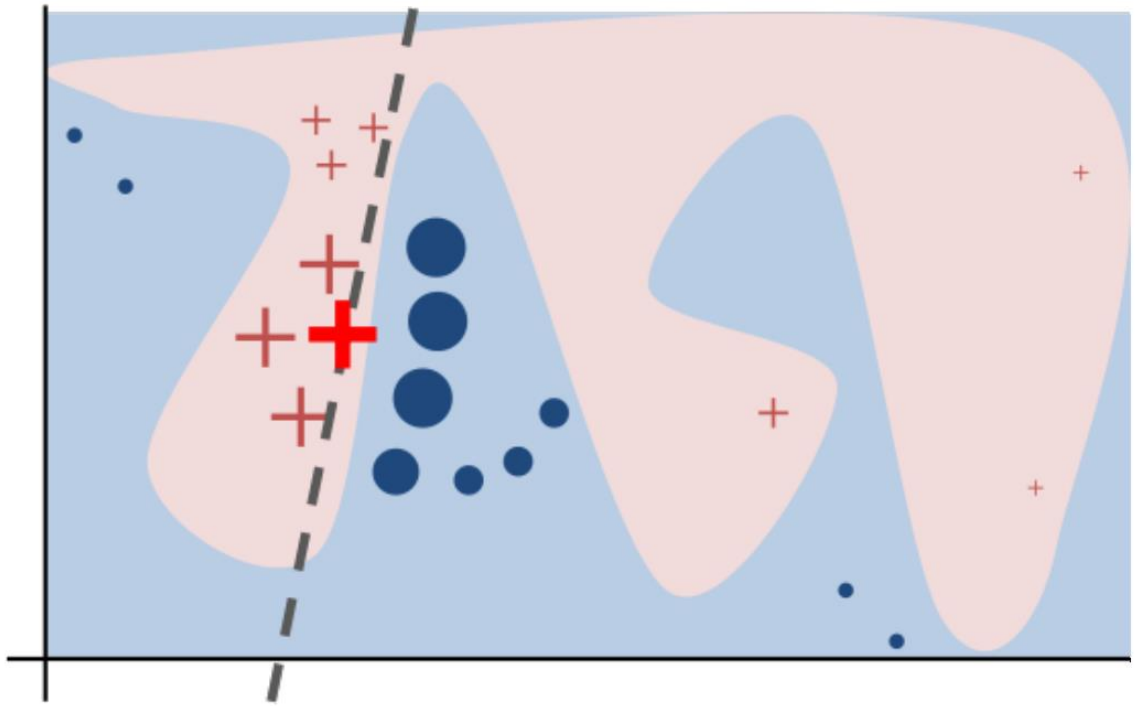


Decision tree



METHOD

LIME (Local Interpretable Model-agnostic Explanation)



- blue/pink 배경 : Classifier Class
- x, y 축 : 예측에 영향 주는 feature

- 진한 빨간 십자
  : 해석하고자 하는 단일 예측
- 검은 점선 : 설명 모델

METHOD

Interpretable Data Representation

Blackbox 모델 해석 → 사람이 해석할 수 있는 데이터 형태로 표현

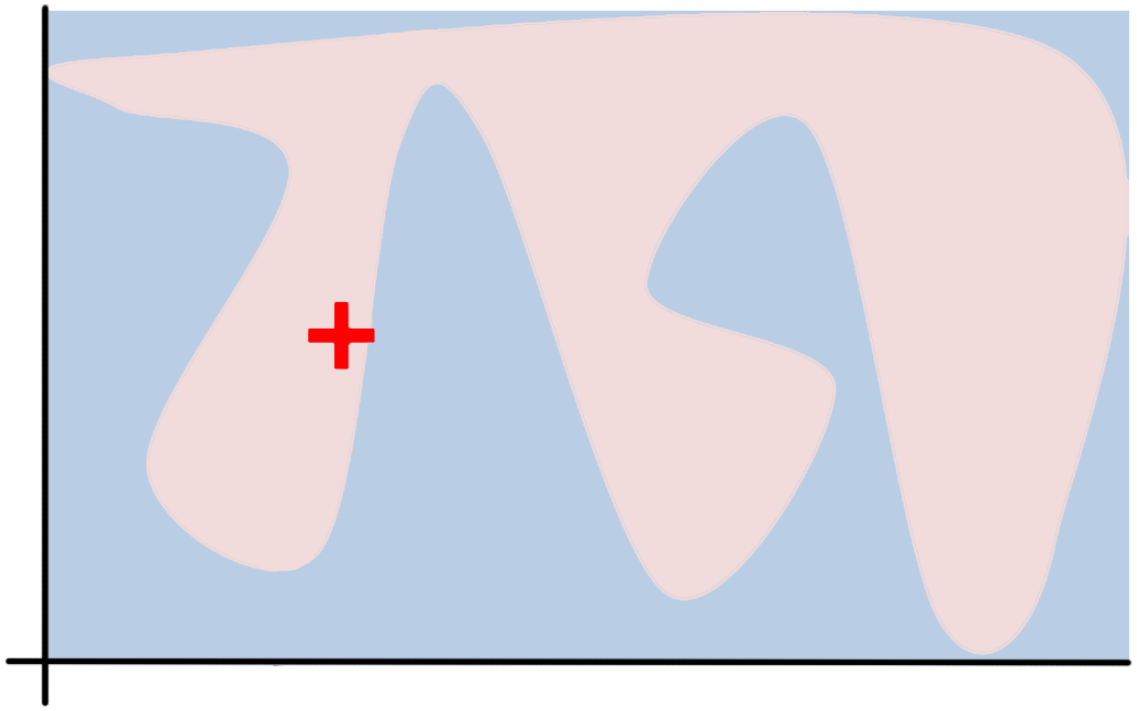
Ex)

텍스트 데이터(단어 유무 나타내는 binary vector)

이미지 데이터(특정 패턴 여부 나타내는 binary vector)

METHOD

Sampling for Local Exploration



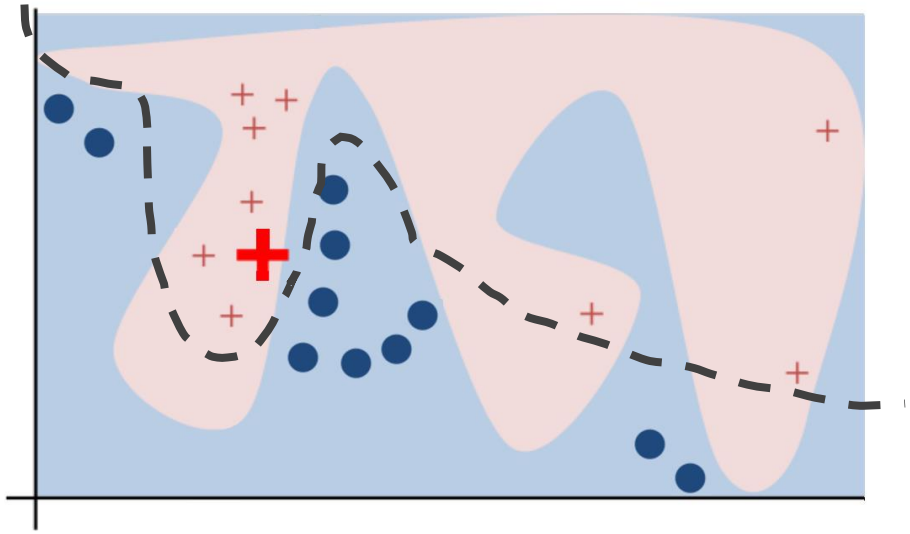
1. 입력 데이터를 변형한 값 랜덤 선택
→ Original 모델에 넣어 예측값 얻음

변형 Ex)
텍스트 데이터 : 특정 단어 제거
이미지 데이터 : 특정 super-pixel 제거

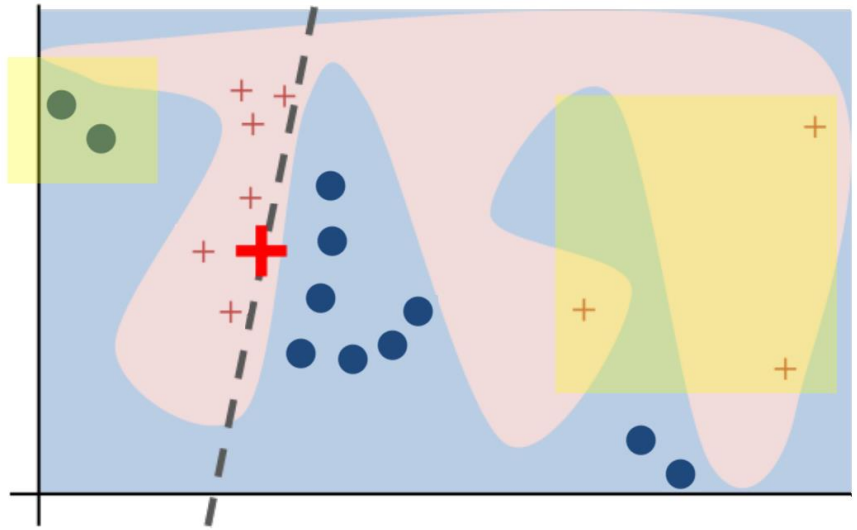
METHOD

Fidelity-Interpretability Trade-off

모든 인스턴스 학습(모델 정확도 ↑)
→ 해석 모델이 복잡해짐



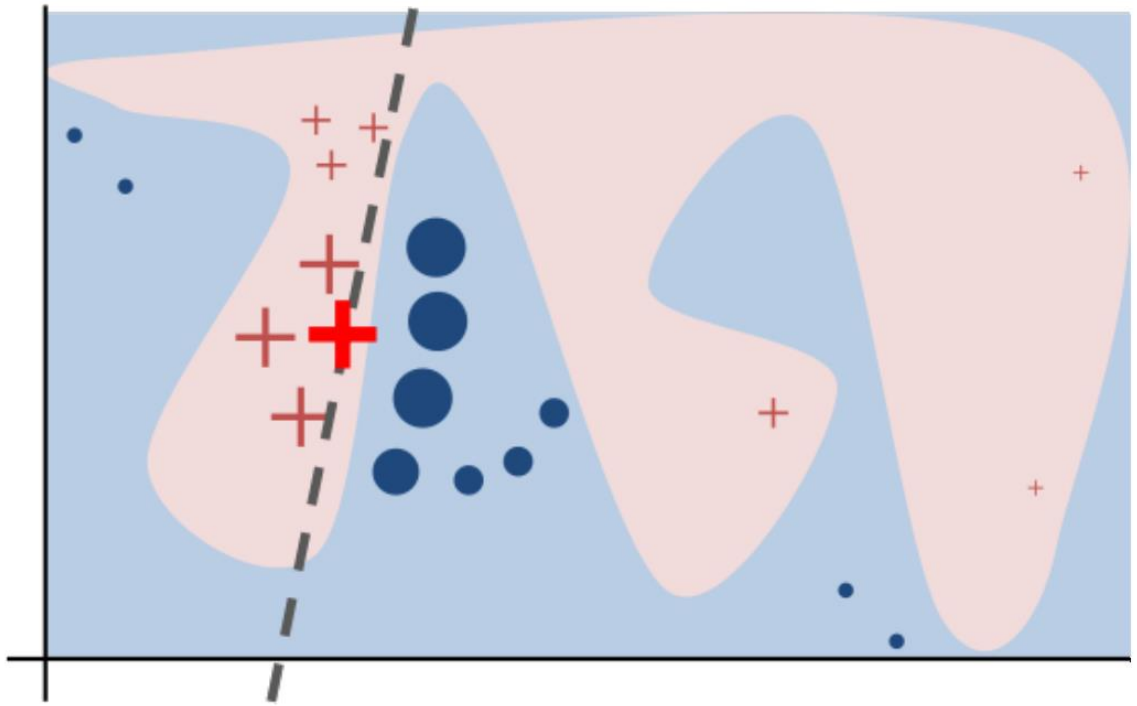
단순한 모델 추구(해석 가능성 ↑)
→ 모델 정확도가 떨어짐.



Local Fidelity와 Interpretability를 모두 충족하는 최적화 문제로 정의됨

METHOD

Sparse Linear Explanations



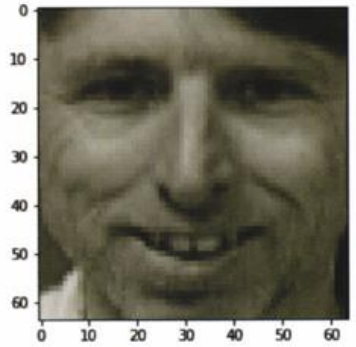
2. Instance를 해석하고자 하는
   예측값과의 거리에 따라 가중치 부여

→ 가중치가 높은 몇몇 Instance를
  사용하여 Linear한 모델로 근사화

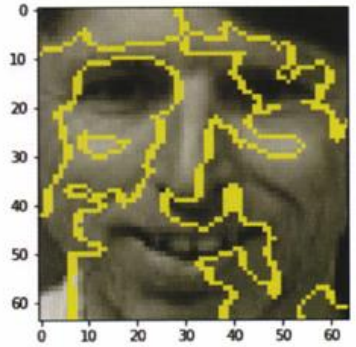
Feature의 중요도 = 해석 모델의 계수

METHOD

Original Input



Interpretable
Data Representation



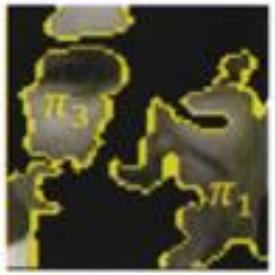
Sampling for
Local Explanation

$p(Person_{38}) = 0.623$

$p(Person_{38}) = 0.021$

$p(Person_{38}) = \mathbf{0.717}$

Sparse Linear
Explanation



Best Explanation

METHOD

SP-LIME (Submodular Pick-LIME)

- 모델 전체에 대한 신뢰도 평가 단일 예측만으로는 불충분

→ 단일 사례들의 집합 구성

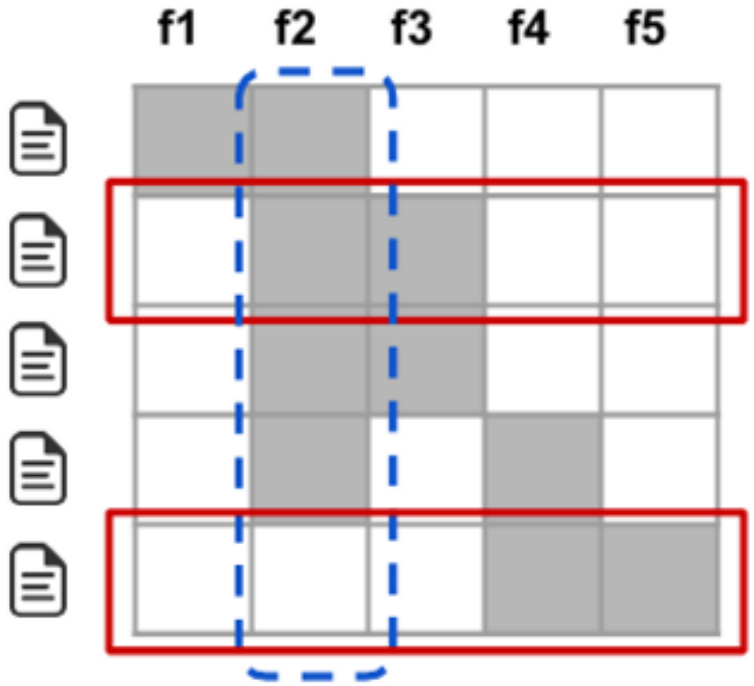
- 단일 사례들을 모두 확인하기엔 cost 많이 소모

→ 모델의 행동을 대표할 수 있는 다양한 설명 집합 선택

METHOD

SP-LIME (Submodular Pick-LIME)

설명하고자 하는 데이터들의 집합 생성 → Local 중요도를 나타내는 설명 행렬 사용
→ 표현된 행렬을 보았을 때 많은 instance를 설명할수록 global 중요도가 높다



- F2 feature가 가장 중요한 feature
- 설명의 중복을 피하기 위해
  2, 5번 instance를 대표 instance로
  선택하는 것이 좋다

EXPERIENCE

설명이

1. 모델에 충실한가
2. 신뢰를 평가하는 데 도움이 되는가
3. 사람에게 유용한가

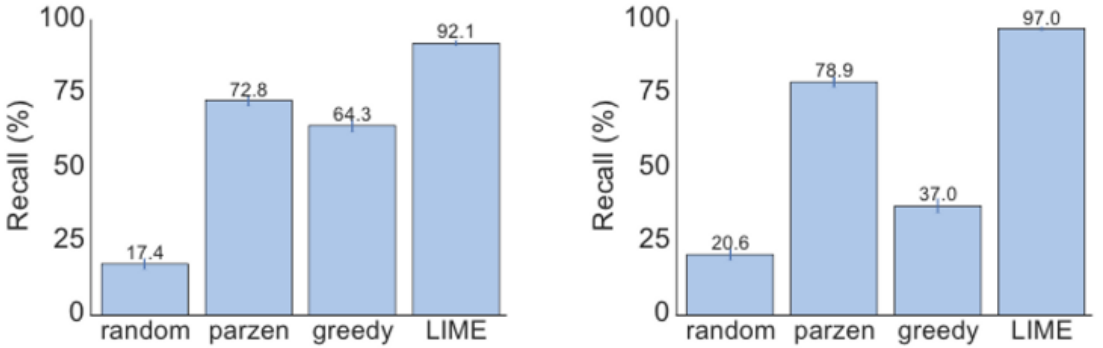
EXPERIENCE

실험 세팅

- Books, DVD 후기 데이터로 감정 분석 실시 (긍정/부정)
- Train : 1600개 / Test : 400개
- Word2vec 활용하여 임베딩 된 bag of words 를 feature 로 사용

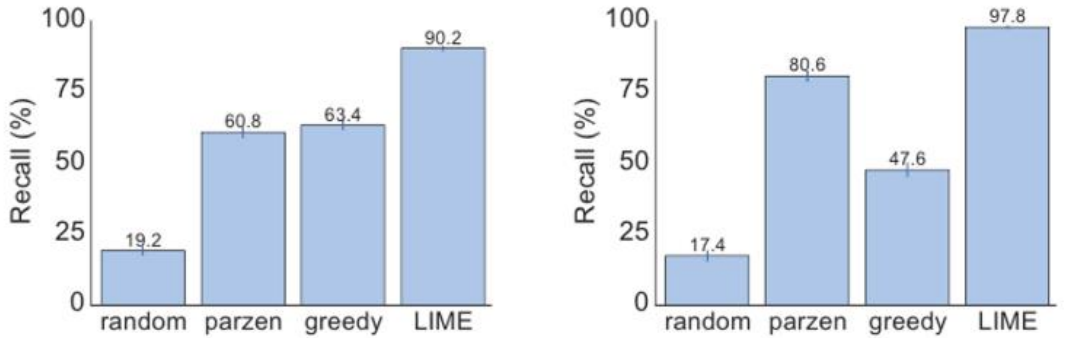
EXPERIENCE

설명이 모델에 충실한가



(a) Sparse LR (b) Decision Tree

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

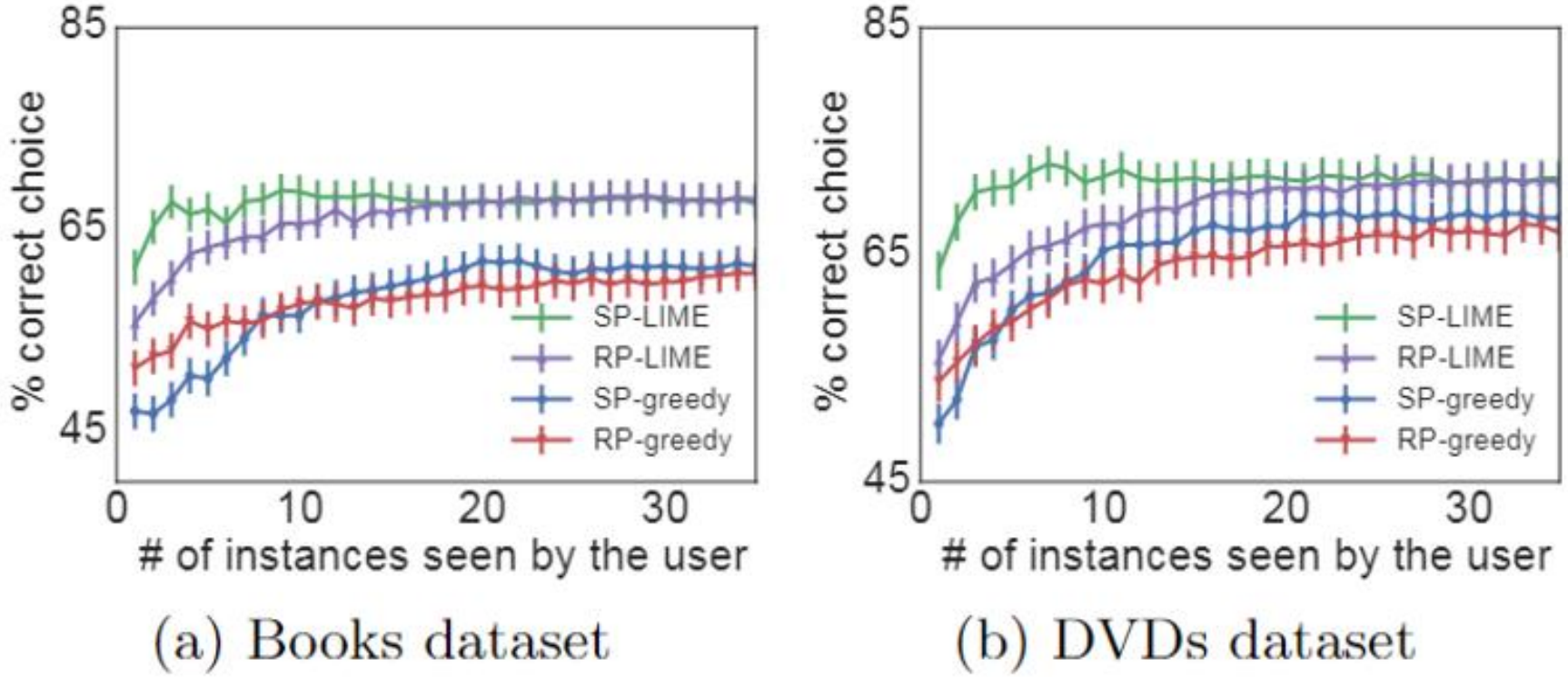


(a) Sparse LR (b) Decision Tree

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

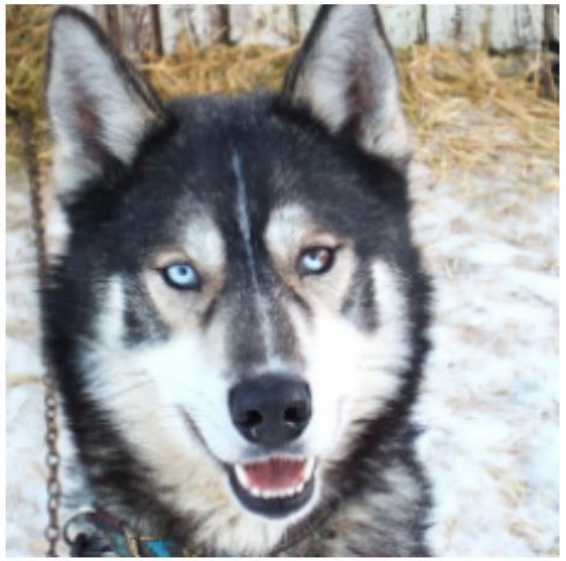
EXPERIENCE

설명이 신뢰를 평가하는데 도움이 되는가



EXPERIENCE

설명을 통해 사람이 좋은 모델을 고를 수 있는가



(a) Husky classified as wolf

(b) Explanation

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

CONCLUSION

LIME의 장단점

장점

- Model-agnostic
- 비교적 적은 계산량

단점

- 데이터 분포가 매우 비선형적인 경우 설명력에 한계