2020/12/18 이유진 (youjin lee)

Paper Review

"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

KDD, 2016

arXiv:1602.04938v3

Abstract

널리 채택되었음에도 불구하고, 머신 러닝 모델들은 대부분 black boxes로 남아있다. 그러나 예측한 후 그 이유들을 이해하는 것은 그 결과를 토대로 조치를 취할지 아닐지 또는 새로운 모델로 쓸 만한지 판단하는 척도가 되는 'trust'를 평가하는데 있어 아주 중요하다. 예측에 대한 이해는 모델에 대한 insights를 제공하기 때문에 신뢰할 가치가 없었던 모델이나 예측 결과를 신뢰할 만한 것으로 변환하는데 사용된다.

위 논문에서 우리는 예측 결과를 중심으로 해석 가능한 모델을 학습하여 해석 가능하고 믿을 수 있는 방식으로 어떤 분류기의 예측 결과도 설명하는 새로운 방법인 LIME을 제안한다. 또한 해당 task를 하위 모듈의 최적화 문제로 틀을 잡아, 대표적인 각각의 예측 결과들과 그 설명을 유의미한 방식으로 제시함으로써 모델들을 설명하는 방법도 제안한다. 우리는 텍스트 모델 (e.g. random forests)와 이미지 분류(e.g. neural networks)들을 설명함으로써 이 방법들의 유연성에 대해 입증한다. 마지막으로 우리는 신뢰가 필요한 다양한 시나리오에서 시뮬레이션과 인간 피험자 실험들을 통해 설명의 유용성에 대해 보여준다: 예측 결과를 신뢰할지 결정하고, 모델들 중 선택하고, 신뢰할 수 없는 분류 모델을 개선하고, 왜 그 분류 모델을 신뢰할 수 없는지 확인한다.

Introduction

머신 러닝은 과학·기술에서 최근 많은 발전의 핵심이다. 안타깝게도 그 분야에서 사람의 중요한 역할은 간과되어 있다. 사람들이 도구로써 머신 러닝 분류기를 직접적으로 사용할지, 다른 제품 내에 모델들을 배치해 볼 것인지 간에 중요한 걱정은 남아있다: 만약 사용자들이 모델이나 예측 결과를 신뢰하지 않으면, 그들은 그것을 사용하지 않을 것이다. 여기서 우리는 신뢰 2가지 정의에 대해 구분하는 것이 중요하다: (1) 예측 결과를 신뢰하는 것, i.e. 사용자들이 개별적인 예측 결과를 충분히 신뢰하고 그것에 기반하여 행동하는 것, (2) 모델을 신뢰하는 것, i.e. 모델이 배포되었을 때, 사용자가 그 모델이 합리적인 방식으로 동작할 것을 믿는 것. 둘 모두 사람이 모델의 행동을 얼마나 이해하고 있는지에 직접적으로 영향을 받는다.

개별적인 예측 결과에 대한 신뢰를 결정하는 것은 그 모델에 의사 결정에 사용될 때 중요한 문제이다. 예를 들어 의료 진단이나 테러 탐지 목적으로 머신 러닝을 사용할 때, 그 결과에 따라 대재앙이 될 수 있기 때문에 예측 결과를 맹목적인 믿음에 따라 행동하기 어렵다.

또한 "실제 야생(현실)"에 배치해보기 전에 전체적으로 모델을 평가하는 것은 필요하다. 현실에 적용해도 되는지에 대한 결정을 내리기 위해서 사용자들은 관심 지표에 따라 모델이 실제 데이터에서도 좋은 성능을 발휘할 것이라는 신뢰성을 가져야 한다. 현재 모델들은 이용가능한 validation 데이터셋 내에서 정확도 지표로 평가되는데, 실제 데이터들은 훨씬 다르며 더 나아가 해당 평가 지표가 실제 제품 목표를 나타내지 않을 수도 있다. 그래

서 개별 예측과 그에 대한 설명을 검사하는 것은 가치 있는 해결책이며, 특히 데이터 셋이 클 때 검사할 예제를 제안함으로써 사용자를 돕는 것이 중요하다.

이 논문에서 우리는 "예측 결과 신뢰" 문제의 해결안으로써 개별적인 예측 결과에 대한 설명 제공과 "모델 자체에 대한 신뢰" 문제의 해결안으로써 다양한 예측 결과 선택을 제안한다. 우리의 주요한 연구 내용들은 아래와 같다.

- LIME : 해석가능한 모델로 locally하게 추정함으로써 모든 분류 모델 · 회귀 모델의 예측 결과를 설명하는 알고리즘
- SP-LIME : 하위 모듈 최적화를 통해 "모델 자체에 대한 신뢰" 문제를 해결 목적의 설명이 포함된 대표 예제 집합을 선택하는 방법
- 시뮬레이션과 사람 피험자 대상 실험을 실시하여 설명이 신뢰성 및 관련 업무에 미치는 영향을 평가한다. 우리 실험에서 LIME을 사용한 비전문가들은 어떤 분류 모델이 현실 세계에서 더 잘 일반화 되어있는지 선택할 수 있었다. 게다가 그들은 20개의 뉴스 그룹 데이터로 학습된 가치 없는 분류 모델들을 LIME을 사용한 feature engineering을 함으로써 엄청나게 향상시킬 수 있었다. 우리는 또한 이미지 관련 neural network의 예측 결과를 이해하는 것이 실무자 입장에서 언제, 왜 그 모델을 믿으면 안되는지를 알게 하는데 얼마나 도움이 되는지 보여준다.

The Case for Explanations

"예측 설명"이란 예제의 구성요소들(e.g. 문장 속의 단어, 이미지 속의 패치) 사이의 관계에 대한 질적인 이해를 제공하는 텍스트 또는 시각적 artifacts를 제시하는 것을 의미한다.

개별적인 예측 결과에 대한 설명 과정은 Figure1에 나와있다. 이 경우에 설명은 상대적인 가중치를 가진 증상 목록 - 예측에 기여하는 증상(녹색), 그것에 반대되는 증거(적색)이다. 사람들은 보통 예측 결과에 대한 추론 과정을 이해할 수 있게 된다면 그들이 받아들이거나(신뢰) 예측 결과를 거부하는데 사용될 도메인에 대한 사전 지식을 가지고 있다. (e.g. 의사의 경우에는 의학 지식을 기반으로 해당 모델을 평가할 것이다.)

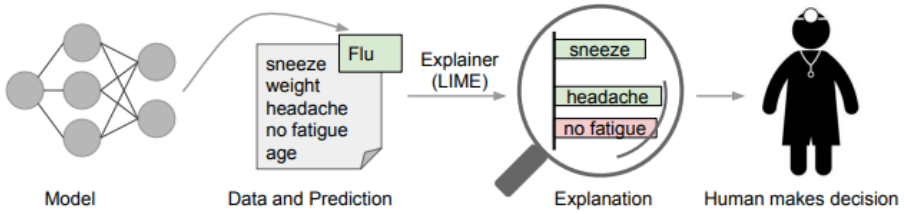


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.

또한 모든 머신 러닝의 적용은 모델에 대한 전반적인 신뢰 척도를 요구한다. 분류 모델의 개발 및 평가는 종종 주석 데이터 수집으로 구성되며, 그 중에서도 hold-out인 부분집합이 자동 평가에 사용되는데 이 방법은 유용하지만 validation 데이터에 대해 정확성이 과대 평가될 수 있기 때문에 전적으로 신뢰하기는 어렵다. 그래서 우리는 global한 이해를 제공하기 위한 방법으로 모델의 몇 가지 대표적인 개별 예측 설명을 제안한다.

모델이나 그 평가가 잘못될 수 있는 방법은 여러 가지가 있다. (1) 데이터 누출은 배치될 때는 보이지 않는 training · validation 데이터로 signal의 의도하지 않은 누출되는 것으로 정의되며 잠재적으로 정확도를 높인다. Kaufman이 인용한 도전적인 예제로는 환자 ID가 target 클래스와 크게 상관 관계가 있는 것으로 밝혀진

사례가 있다. 이는 원 데이터와 예측 결과 관찰만 가지고 알아내기에는 상당히 도전적이지만, Figure1처럼 설명에 제공된다면 훨씬 쉬워질 것이다. (2) 특히 탐지하기 어려운 문제로는 training 데이터와 test 데이터가 다른 데이터 셋 shift가 있다. (상세한 내용은 20개 뉴스 그룹 데이터에서 다루어 진다.) 설명으로 인해 얻은 insight는 신뢰할 수 없는 모델을 신뢰할 수 있는 모델로 바꾸기 위해 무엇을 해야 할지 알아내는데 특히 도움이 된다.

머신 러닝 실무자들은 종종 여러 대안들 중에서 모델을 선택해야 하며 두 개 이상의 모델 간 상대적 신뢰도를 평가해야 한다. Figure 2에서 우리는 정확도와 함께 모델들 중에서 선택하는데 개별 예측에 대한 설명을 어떻게 활용할 수 있는지 보여준다. 이 경우에 실제로 validation 정확도가 더 높은 알고리즘이 실제로는 더 좋지 않음을 보여주고 있는데 아래와 같은 설명이 없었더라면 알아차리기 힘들었을 것이다. 또한 우리가 계산하고 최적화할 수 있는 지표(e.g. 정확도)와 실제 사용자의 참여 · 유지와 같은 지표는 불일치할 수 있는데, 실무자는 모델의 정확성을 높이기 위해 "낚시 기사"와 같은 컨텐츠 추천보다는 사용자들이 꾸준히 사용하도록 정확도가 낮은 모델을 원할 수도 있다. 우리는 어떤 모델에 대해서도 이런 설명들을 생산할 수 있다면 이러한 시나리오에서 특히 유용하다는 것에 주목한다.

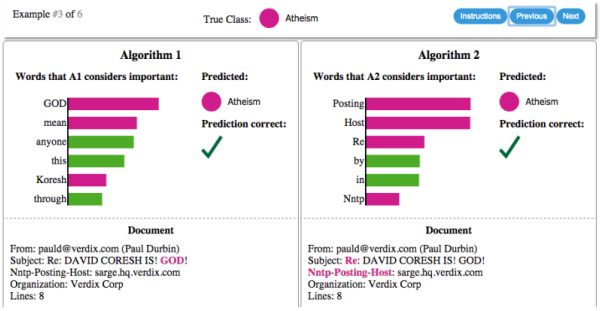


Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for "Christianity", magenta for "Atheism").

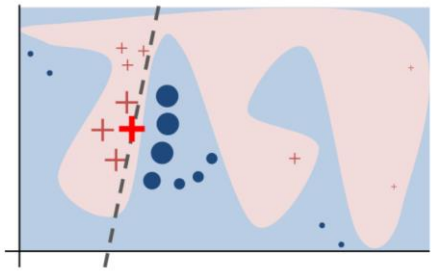


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

< Desired Characteristics for Explainers >

우리는 설명법에서 원하는 많은 특징들을 개략적으로 설명한다. (1) interpretable: 입력 변수들과 응답 사이에 질적인 이해 제공. 해석가능성은 사용자의 한계를 고려하여 이루어져야 하는데, 예를 들어 만약 수십만개 또는 수천개의 feature가 예측에 기여하는 경우, 개별 가중치를 확인할 수 있더라도 사용자가 예측에 대해 이해하기는 쉽지 않다. 또한 해석가능성의 개념은 대상 청중에 따라서도 달라짐을 알아야 한다. (2) local fidelity: 만약 모델 자체에 대한 완벽한 설명이 되지 않는다면 설명을 완벽히 신뢰하는 것은 불가능하지만, 그 설명은 적어도 locally faithful해야 한다. 예를 들어 예측되는 예제 부근에서 모델이 어떻게 작동하는 지와 일치한다. 여기서 local fidelity는 global fidelity를 의미하지만, global fidelity는 local fidelity는 의미함을 알고 있어야 한다. (3) model-agnostic: 반면에 모델 자체가 타고나서 해석 가능한 것들이 있는데, 이를 우리는 model-agnostic 하다고 한다. (4) 예측 결과를 설명하는 것 외에도 모델에 대한 신뢰를 확인하는데 'global perspective'를 제공하는 것은 중요하다.

Local Interpretable Model-Agnostic Explanations

우리는 이제 Local Interpretable Model-Agnostic Explanations (LIME) 소개한다. LIME의 목표는 분류 모델에 locally faithful한 해석 가능한 표현에 대해 해석 가능한 모델들을 식별하는 것이다.

Interpretable Data Representations

우리가 설명 체계에 대해 소개하기 전에 해석가능한 데이터 표현과 features을 구별하는 것은 중요하다. 해석가능한 설명은 해당 모델에 쓰인 실제 features과 관계없이 사람들이 이해할 수 있는 표현들을 사용하는 것이 필요하다. 예를 들어 실제로 분류 모델이 더 복잡한 features들을 사용했을지라도 텍스트 분류 목적의 해석가능한 표현은 이진 벡터로 해당 단어의 유무가 될 수 있다. 우리는 설명될 예제의 원래 표현이 $x \in \mathbb{R}^d$라 표현하고, 그것의 해석가능한 표현으로 이진 벡터인 $x' \in \{0,1\}^{d'}$을 사용한다.

Fidelity-Interpretability Trade-off

형식적으로 우리는 설명을 모델 $g \in G$로써 정의한다. G는 잠재적으로 해석 가능한 모델로 linear models, decision trees, falling rule lists 등을 의미하고, g는 $\{0,1\}^{d'}$로 해석가능한 구성 요소들의 유무를 의미한다. 물론 모든 $g \in G$들이 해석하기에 충분히 단순하지 않을 수도 있기 때문에, 우리는 $\Omega(g)$를 설명 $g \in G$ 복잡도(해석가능성의 반대)의 척도로 두었다. 예를 들어 Decision trees의 $\Omega(g)$는 나무의 깊이일 것이고, 반면 Linear models의 $\Omega(g)$는 0이 아닌 가중치들의 개수일 것이다. 설명되는 모델을 $f: \in \mathbb{R}^d \rightarrow \mathbb{R}$로 두고 보면, 분류에서 $f(x)$는 x가 특정 클래스에 속할 확률이다. 더 나아가 우리는 x 주변의 locality를 정의하기 위해서 예제 z에서 x사이의 근접 지표로써 $\pi_x(z)$를 사용한다. 결국 $\mathcal{L}(f,g,\pi_x)$는 π_x로 정의된 locality에서 f를 추정하는데 있어 g가 얼마나 충실하지 못했는지를 보는 측정 지표이다. 해석가능성과 local fidelity 모두를 보장하기 위해서는 사람들에게 해석될 수 있을 만큼의 $\Omega(g)$를 가지면서 결국 $\mathcal{L}(f,g,\pi_x)$를 최소화해야 한다. 위의 설명에 의하면 LIME의 공식은 아래와 같다:

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

Sampling for Local Exploration

우리는 설명이 model-agnostic되기를 바라기 때문에 f에 대한 어떠한 추정없이 $\mathcal{L}(f,g,\pi_x)$을 최소화하고 싶어한다. 따라서 해석가능한 입력이 다양할 때 f의 local한 행동을 학습하기 위해 π_x에 의해 가중된 샘플들을 만들면서 $\mathcal{L}(f,g,\pi_x)$을 근사한다. 우리는 x'의 0이 아닌 요소들을 무작위로 랜덤하게 만들어 x'주위의 예제들을 샘플링 한다. 변동된 $z' \in \{0,1\}^{d'}$가 주어지면 우리는 원래 표현 $z \in R^d$에서 샘플을 회수하고 설명 모델의 레이블로 사용되는 $f(z)$를 얻는다. 이와 같은 샘플의 데이터셋 z를 고려하여 설명 $\xi(x)$를 얻기 위해서 식 (1)을 최적화한다. LIME 기반의 주요한 intuition은 Figure 3에 설명되어 있는데 여기서 x 부근과 x 멀리 떨어진 곳에서 예제들을 샘플링 한다. 비록 기존 모델이 global하게 설명하기에는 너무 복잡하더라도, LIME은 locally faithful한 설명을 제시한다.

Sparse Linear Explanations

이 논문의 나머지 부분을 위해 우리는 G를 $g(z') = w_g \cdot z'$인 선형 모델의 클래스로 설정한다.

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

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 $i \in \{1,2,3,...,N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
  return w

알고리즘 1은 개별적인 예측 결과에 대한 설명들을 생성하기 때문에 그것의 복잡성은 데이터셋의 크기가 아닌 $f(x)$ 계산에 걸리는 시간과 샘플 N의 개수에 따라 달라진다. 실제로 N=5000인 1000개의 가지를 가지는 random forest의 경우 노트북에서도 3초 미만이 시간이 소요되었고, 이미지 분류를 위한 Inception network에서 각 예측을 설명하는데 10분 정도 소요된다.

해석 가능한 표현과 G의 선택은 몇 가지 필연적인 단점을 가진다. (1) 기본 모델은 black-box로 취급될 수 있지만, 특정 해석 가능한 표현들은 특정 행동들을 설명하기에 충분히 강력하지 않을 것이다. (2) G의 선택은 만약 기본 모델이 예측 결과의 locality에서도 매우 비선형적이라면 이는 충실한 설명이 없을 수도 있다는 것을 의미한다. 그러나 우리는 Z에 대한 설명의 신뢰성을 추정할 수 있고, 사용자들에게 이 정보를 제시할 수 있다. 우리는 선형적인 설명이 다수의 black-box형 모델에서도 꽤 잘 작용하기 때문에 이와 관련된 탐구를 미래 연구에 맡긴다.

Example

(Text classification with SVMs) Figure 2의 오른쪽을 보면 우리는 "기독교"와 "무신론"을 구별하기 위해 학습된 RBF 커널을 사용한 SVM의 예측 결과를 설명한다. 이 분류 모델의 정확도는 94%에 달해 이를 신뢰하고 싶은 생각이 들지만, 예제의 설명을 보면 제멋대로인 이유로 예측이 이루어짐을 확인할 수 있다. (단어 "포스팅", "게시자", "Re"는 기독교, 무신론 모두와 아무런 관련이 없다.) 설명을 통해 얻은 insight는 데이터셋에 심각한 문제가 있고 결과적으로 해당 모델은 신뢰할 수 없다는 결론을 얻을 수 있다.

(Deep networks for images) 이미지 분류 모델을 위한 sparse linear 설명을 사용했을 때 누군가는 왜 그 모델이 제시된 클래스라 생각하는지에 대한 직관을 주므로 특정 클래스에 대해 양의 가중치를 가진 super pixel을 강조하려고 할 수도 있다. Figure 4의 b, c, d는 top 3 예측 클래스에 대한 super pixel의 설명을 보여준다. 특히 4b가 일렉 기타라고 예측된 이유가 플랫보드 때문임을 알 수 있다. 이런 종류의 설명은 분류 모델에 대한 신뢰를 높여주는데 이는 그 모델이 비합리적인 방식으로 행동하지 않았다는 것을 보여주기 때문이다.

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Submodular Pick For Explaining Models

단일 예측 결과에 대한 설명은 사용자에게 분류 모델의 신뢰성에 대해 약간의 이해를 제공하지만 모델 전체에 대한 신뢰성을 평가하기에 충분치 않다. 우리는 개별적인 예제 집단의 설명함으로써 모델에 대한 전체적인 이해를 제안한다. 이 접근법은 여전히 model-agnostic하고 정확도와 같은 집약된 통계를 계산하는 것을 보완한다.

사용자들은 많은 설명들을 모두 조사하기에는 시간이 충분하지 않기 때문에 분별력 있게 예제 집단을 선택해야 한다. 여기서 우리는 사람이 모델이 이해하기 위해 기꺼이 살펴보려는 설명의 수를 예산 B라고 하고 이는 시간/인내를 나타낸다. 예제 집합 X가 주어졌을 때 사용자가 검사할 B 예제를 선택하는 작업을 pick step이라 정의한다. Pick step는 각 예측에 수반되는 설명을 고려해야 하고 사용자에게 보여줄 다양하고 대표적인 설명 집합을 선택해야 한다.

주어진 예제 집합 X ($|X| = n$)에 대한 설명을 고려할 때 우리는 각각 예제들을 위한 해석가능한 구성요소의 local 중요도를 나타내는 $n \times d'$인 설명 매트릭스 W를 구성한다. 게다가 W의 각 구성요소 j(column)에 대해 I_j는 설명 영역에서 해당 구성 요소의 global 중요도를 나타내도록 한다. (I_j 는 많은 예제들을 설명하는 feature

들이 더 높은 중요도 점수를 가질 수 있도록 한다.) Figure 5에서 우리는 이진이고 $n = d' = 5$ 형태인 W에 대한 예제를 볼 수 있는데 중요도 함수 I는 f1 보다 더 높은 f2에 점수를 줘야한다. 예제 집합 구성시에 중요한 구성 요소들을 포함하는 예제를 선택하고 싶겠지만 설명 집합의 중복은 좋지 않다. 따라서 figure 5를 보면 2번째 열을 선택한 다음에 3번째 열을 선택하는 것은 무의미하므로 겹치지 않으면서 더 많은 설명력을 가진 마지막 열을 선택하는 것이 옳다. 이 원리를 식으로 표현하면 아래와 같다.

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

$$Pick(W, I) = \underset{V, |V| \leq B}{argmax}\, c(V, W, I) \tag{4}$$

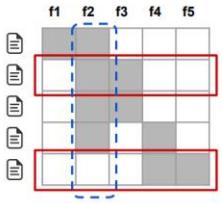


Figure 5: Toy example $\mathcal{W}$. Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

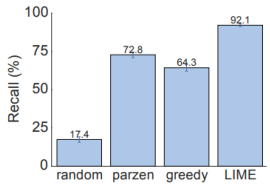
Algorithm 2 Submodular pick (SP) algorithm

Require: Instances X, Budget B
  for all $x_i \in X$ **do**
    $\mathcal{W}_i \leftarrow$ **explain**(x_i, x_i') ▷ Using Algorithm 1
  end for
  for $j \in \{1 \ldots d'\}$ **do**
    $I_j \leftarrow \sqrt{\sum_{i=1}^{n} |\mathcal{W}_{ij}|}$ ▷ Compute feature importances
  end for
  $V \leftarrow \{\}$
  while $|V| < B$ **do** ▷ Greedy optimization of Eq (4)
    $V \leftarrow V \cup \text{argmax}_i\, c(V \cup \{i\}, \mathcal{W}, I)$
  end while
  return V

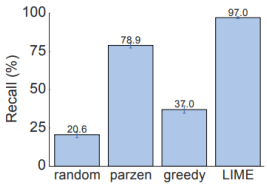
Simulated User Experiments

이번 파트에서 신뢰 관련 업무에서 설명의 유용성을 평가하기 위해 시뮬레이션 사용자 실험을 제안한다. 특히 다음과 같은 질문을 다룬다: (1) 그 설명이 모델에 충실한가? (2) 설명을 통해 사용자들이 예측에 대한 신뢰를 확인할 수 있는가? (3) 설명은 모델 전체를 평가하는데 유용한가?

우리는 책과 DVD에 대한 제품 리뷰를 긍정적 또는 부정적으로 분류하는 두 개의 데이터셋(각각 2000개씩)을 사용했고, decision tree(DT), logistic regression with L2(LR), nearest neighbors(NN), support vector machine with RBF 모델로 각각 학습시켰다. 개별적인 예측 결과를 설명하기 위해서 우리는 우리가 제안한 접근법(LIME)과 parzen 윈도우를 globally하게 black box 분류 모델을 추정하고 예측 확률 함수의 gradient를 이용하여 개별적인 예측결과를 설명하는 방법인 parzen을 비교했다.

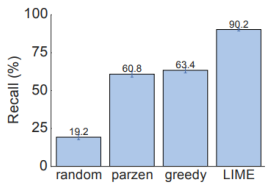


(a) Sparse LR

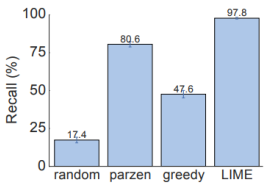


(b) Decision Tree

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.

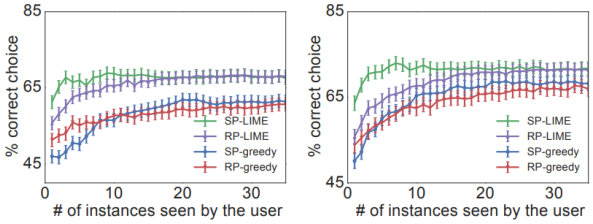
Figure 6과 7을 보면 greedy 접근법이 LR에서는 parzen과 비교될 수 있지만 DT에서는 현저히 나쁜 것으로 보인다. Parzen의 경우 전반적인 recall은 고차원 분류 모델을 추정하는데 어려움이 있으므로 약간 낮은 추세를

보이고 있고, 반면 LIME은 두 데이터셋, 분류 모델에서 지속적으로 90%이상의 recall을 보이므로 해당 설명이 모델에 충실함을 보여준다.

개별 예측 결과에 대한 신뢰를 시뮬레이션하기 위해 먼저 "신뢰할 수 없는" feature들의 25%를 무작위로 선택하고 사용자가 이러한 feature를 식별할 수 있고 신뢰하지 않을 것으로 가정했다. 더 자세히 말하자면 우리는 설명에 나타난 모든 신뢰할 수 없는 특징들이 제거될 때 linear approximation으로부터의 예측이 변할 경우 사용자들이 LIME과 parzen 설명으로부터 신뢰할 수 없는 예측을 한다고 가정했다. 이 설정 하에 우리는 Table 1에서 100번 이상 실시하여 평균값을 낸 각각의 설명에 대해 신뢰할 수 있는 예측 결과 F1을 기록했다. 그 결과 LIME이 두 데이터 셋, 분류 모델에서 다른 방법 대비 우세함을 확인할 수 있다.

Table 1: Average F1 of *trustworthiness* for different explainers on a collection of classifiers and datasets.

	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**



(a) Books dataset (b) DVDs dataset

Figure 8: Choosing between two classifiers, as the number of instances shown to a simulated user is varied. Averages and standard errors from 800 runs.

최종 시뮬레이션 사용자 실험에서 우리는 validation set에서 비슷한 정확도를 보이는 두 개의 경쟁 모델 사이에서 인간이 결정해야할 경우를 시뮬레이션 하여 설명이 모델 선택에 사용될 수 있는지에 대한 여부를 평가했다. 이를 위해 인공적으로 10개의 "noisy" features를 추가하여 실제처럼 유익한 feature뿐만 아니라 가짜 상호관계를 띄는 feature까지 쓰이는 상황을 재현했다. 또한 우리는 30그루 이상을 가진 분류 모델을 반복 훈련시켜 경쟁 모델 2개를 만들었고 이들의 test 정확도는 5% 이상이 차이가 나는 상태에서 실험을 진행했다. Figure 8은 우리가 정답인 분류 모델을 선택한 정확도를 800회 이상 시행하여 나온 평균값을 제시한다. 해당 그림에서 SP-parzen, RP-parzen의 경우에는 전혀 의미있는 설명을 만들지 못했기에 제외하였고, LIME은 여전히 pick 방법과 무관하게 일관적으로 greedy보다 좋은 결과를 보이고 있다.

Evaluation with Human Subjects

이번에 우리는 예측 결과와 모델에 대한 이해와 신뢰를 필요로 하는 머신 러닝에서의 3가지 시나리오를 재현했다. 특히 다음과 같은 세팅에서 우리는 LIME과 SP-LIME을 평가했다: (1) 사용자가 두 개의 분류 모델 중 더 잘 일반화되는 것을 선택할 수 있는가? (2) 설명 기반으로 사용자가 모델 개선을 위해 feature engineering을 수행할 수 있는가? (3) 사용자가 설명을 보고 분류 모델의 불규칙성을 식별하고 설명할 수 있는가?

위 실험을 위해 우리는 20개의 뉴스 그룹 데이터로부터 얻은 "기독교"와 "무신론" 문서를 사용했고, 이 데이터는 일반화되지 않은 feature들을 포함하여 validation 정확도가 실제 성능에 비해 상당히 과대평가되어 있는 상태이며, 우리는 실제 성능을 평가하기 위해서 종교 데이터를 만들었다. 그 다음 동일하게 SVM 모델을 2개의 다른 데이터셋(원데이터, 일반화되지 않은 feature들이 제거된 "정제된" 데이터)으로 학습시켰는데 test 정확도는 94.0%, 88.6%였지만 반대로 종교 데이터에 대한 정확도는 57.3%, 69.0%이 나왔다.

이 상태에서 우리는 AMT를 통해 종교적 지식은 있지만 머신 러닝 비전문가인 피험자를 채용했고, Figure2와 같은 설명을 보고 더 나은 알고리즘을 선택하는 방식으로 그들의 능력을 측정했다. 그리고 그 결과는 Figure 9에 나와있다. 모든 방법들이 더 나은 분류 모델을 식별함에 있어 뛰어나며 어떤 분류 모델을 신뢰해야 할지 결정하는데 설명이 유용하다는 것이 입증되었다.

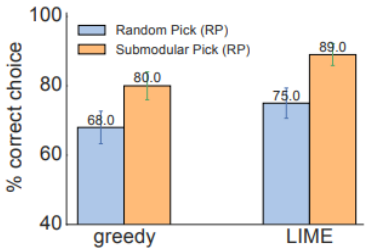


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

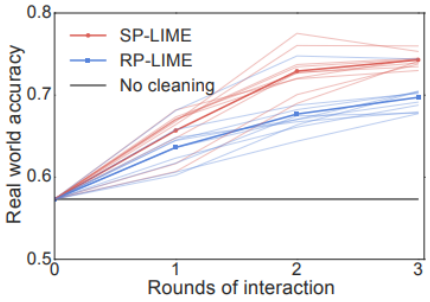
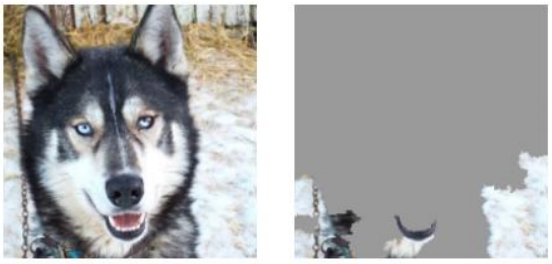


Figure 10: Feature engineering experiment. Each shaded line represents the average accuracy of subjects in a path starting from one of the initial 10 subjects. Each solid line represents the average across all paths per round of interaction.

설명은 특히 사용자가 일반화되어 있지 않다고 생각하는 feature를 제거하고 중요한 feature를 제시함으로써 모델 향상 개선에 도움이 될 수 있다. 우리는 이전 실험에 이어 사용자에게 더 나쁜 분류 모델을 위해 어떤 단어가 제거되어야 하는지 확인하도록 요청했다. 해당 결정을 위해 사용자에게 각 보여지는 설명과 예제는 SP-LIME 또는 LP-LIME에 의해 생성되었으며 우리는 Figure 10에서 원래의 10명의 피험자가 각각 시작한 경로에 대한 상호작용 단계에서 종교 데이터의 평균 정확도(음영선)를 보여주고, 모든 경로(실선)에 대한 평균을 보여준다. 이는 명백히 비전문가도 설명을 통해 중요하지 않은 feature를 제거하여 모델을 개선 가능함을 보여준다.



(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	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

Table 2: "Husky vs Wolf" experiment results.

종종 데이터 내의 artifact는 학습하는 동안 분류 모델들이 바람직하지 못한 상호 관계를 맺도록 유도한다. 우리는 의도적으로 모든 늑대 사진은 배경에 눈이 있는 것을 고르고 허스키의 경우 눈이 없는 것으로 골라 모델을 학습시켰다. 그랬더니 해당 분류 모델은 눈이 있을 경우에 "늑대"로 예측했고 그 외에는 동물의 색, 위치, 자세와 관계없이 "허스키"로 인식했다. 우리의 실험은 다음과 같이 진행된다: 먼저 설명없이 10개의 test 예측 결과들을 (눈이 없는 배경의 늑대 : 허스키로 예측, 눈이 있는 배경의 허스키 : 늑대로 예측, 을 포함한 10장) 제시한다. 그리고 피험자에게 3가지 질문을 했다: (1) 그들이 이 알고리즘을 신뢰할 것인지 (2) 왜 (3) 이 알고리즘이 어떻게 늑대와 허스키를 구별할 수 있다고 생각하는지. 그 다음 동일하게 이번에는 설명과 함께 사진을 보여주었고 질문을 제시했다. 그 결과 설명을 보기 전에는 전체의 3분의 1 이상이 분류 모델을 신뢰했고 어떠한 방식으로 예측하고 있는지 언급한 사람은 절반도 되지 않았다. 그러나 설명을 검토한 후 거의 모든 피험자들이 정확한 insight를 얻었고 분류 모델에 대한 신뢰도도 크게 떨어졌다.

Conclusion and Future Work

이 논문에서 우리는 머신 러닝 시스템과 사람 간의 효과적인 상호작용을 위해 신뢰도가 중요하며 각 예측에 대

한 설명은 신뢰도를 평가하는데 중요하다고 말해왔다. 우리는 해석 가능한 방식으로 어떤 모델의 예측이라도 충실히 설명해내기 위해 modular하고 확장가능한 방식인 LIME을 제안한다. 또한 사용자에게 모델에 대한 global한 관점을 제공하기 위하여 대표적이며 중복되지 않는 예측 결과를 선택하기 위한 SP-LIME도 소개한다. 우리의 실험은 설명이 text, 이미지 분야에서 trust-related 업무를 수행하는 다양한 모델들에 유용하다는 것을 입증했다. 이는 전문가 · 비전문가 모두에게 그러한데, 모델 간의 결정, 신뢰도 평가, 신뢰하라 수 없는 모델 개선, 예측 결과로부터 insights 획득 등을 예로 들 수 있다.

앞으로 우리가 더 진행해보고자 하는 향후 계획으로는 다음과 같다. 비록 우리가 sparse linear models 만을 가지고 설명했지만, 위 방법은 다양한 설명군(e.g. decision trees)에 적용해 볼 수 있으므로 실제 사용자들과 이들을 비교 연구하는 것도 흥미로울 것이다. 또한 이번 연구에서 언급하지 않았던 문제 중에 이미지 pick step을 어떻게 수행할 것인지가 있었는데 향후 이 한계점에 대해 다뤄보고 싶다. 또한 우리의 연구를 더 다양한 분야에 적용하여 사용성을 테스트해 볼 예정이며 마지막으로 적절한 샘플의 개수와 같은 이론적 특징과 병렬처리 · GPU 프로세싱과 같은 계산 최적화를 연구해 볼 것이다.