XAI_03. 설명가능한 인공지능에 대해서

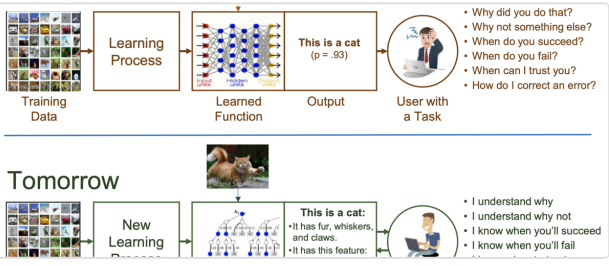
생성일	@2022년 8월 30일 오후 4:03
유형	머신러닝/딥러닝
작성자	동훈 오

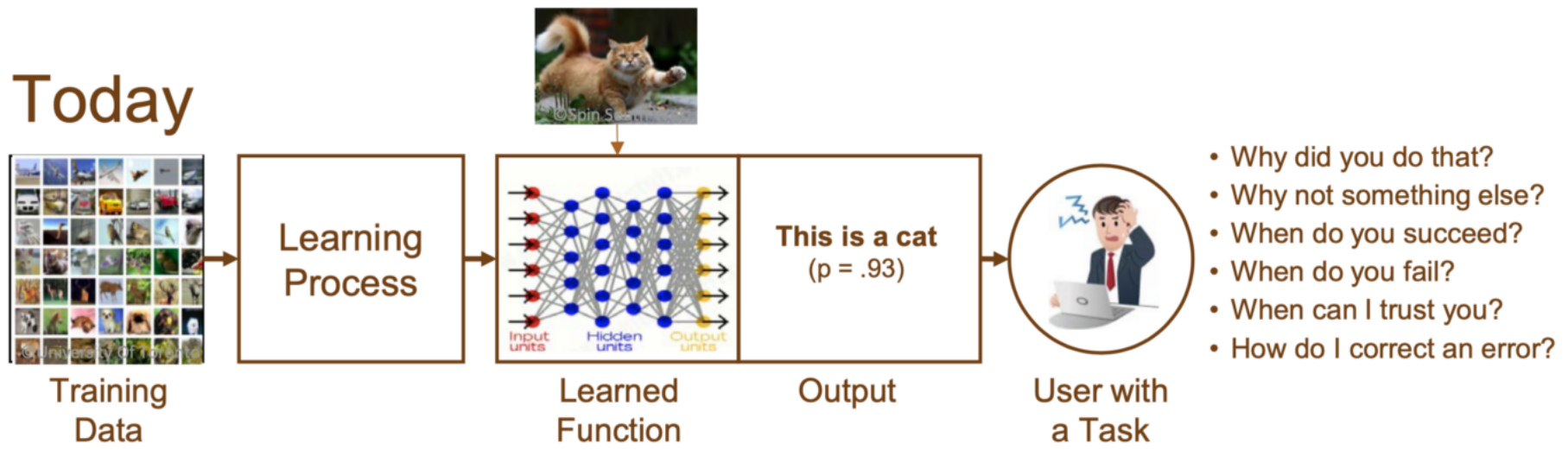
레퍼런스 1 : '설명가능한 인공지능이란?', medium 자료

설명가능한 인공지능(eXplanable AI, XAI)이란?

인공지능, AI란 단어가 일상 속에 침투한 이래로 인공지능은 학계에서, 비즈니스 영역에서, 일상 생활에서 수많은 성취를 거두었습니다. 불과 10년도 지나지 않아 머신러닝(Machine Learning)은 컴퓨터비전, 시계열예측, 분류, 회귀분석, 음성인식, 문자인식 등에서 성공적으로 적용되었고 실시간에 가까운 속도로 빠르게 발전하고 있습니다. 머신러닝 알고리즘 또

https://medium.com/daria-blog/%EC%84%A4%EB%AA%85%EA%B0%80%EB%8A%A5%ED%95%9C-%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5-explanable-ai-xai-%EC%9D%B4%EB%9E%80-4e51d9e7bb59

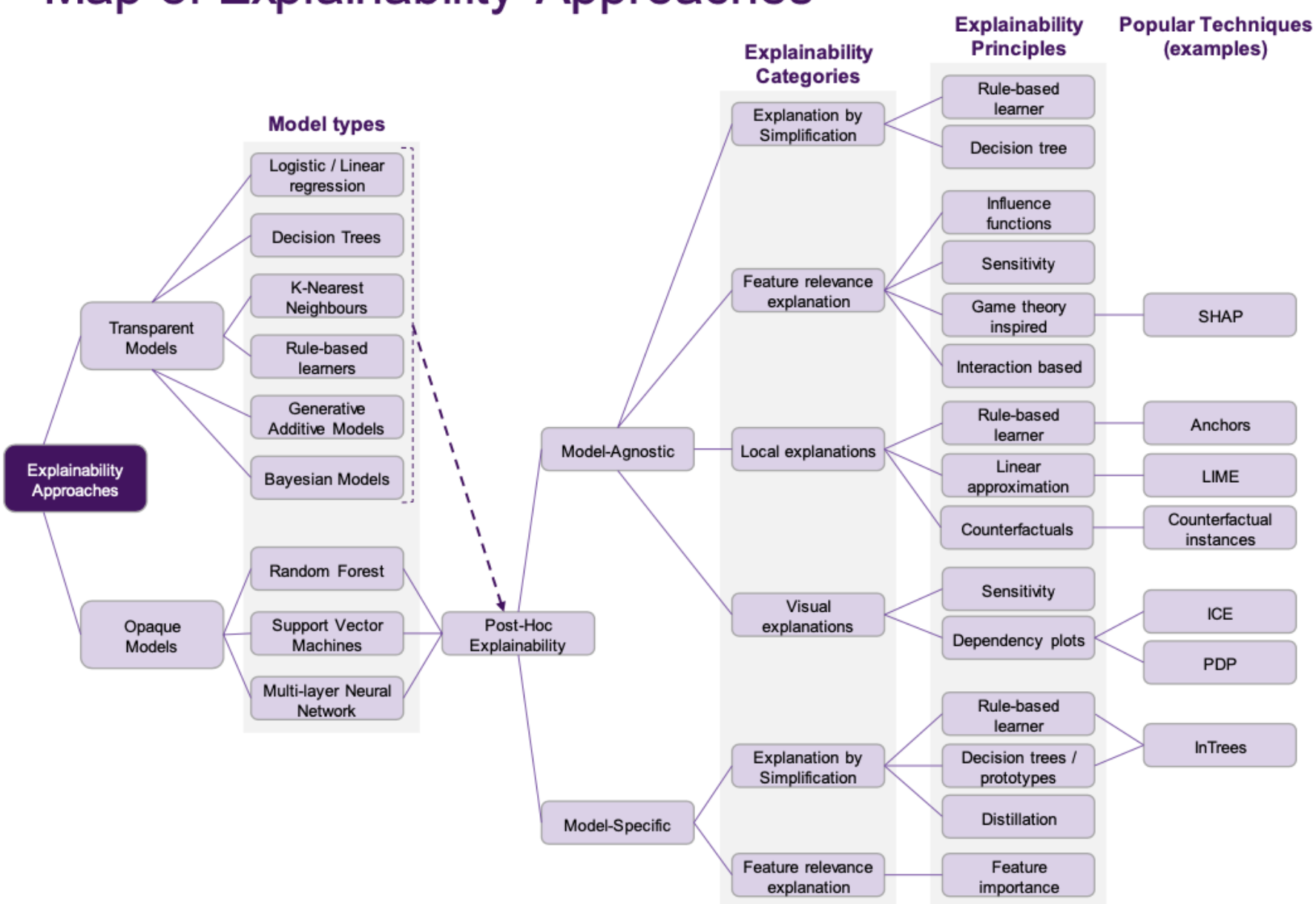




특정 사진에 고양이가 존재하는지 아닌지를 알고자 한다고 가정해보자. 머신러닝 모델은 사진에 존재하는 객체가 고양이인지 아닌지에 대해서 계속해서 학습하여 정답을 찾을 확률을 높여간다. 많은 데이터를 오랫동안 학습하여 마침내 모델이 테스트 데이터에 대해 정답을 맞추게 된다. 평가 데이터셋에 대한 모델의 정확도는 99%가 나왔다고 가정해보자. 과연 우리는 딥러닝 모델의 선택을 믿을 수 있을까?

사용자의 입장에서 어떤 의사결정을 위해서 모델의 정확도 이외의 추가적인 정보가 필요하다. 즉, 모델을 신뢰하기 위한 정보가 필요하다는 것. 모델이 왜 사진에 고양이가 있다고 판단 했을지 그 근거를 사용자가 알 수 있다면 모델을 조금 더 신뢰할 수 있을 것이다. 이러한 상황이 XAI 가 필요한 이유이다.

Map of Explainability Approaches



Transparaent(투명) vs Opaque(불투명)

- “transparent” 는 모델이 결과를 도출하는 과정에 대해 우리가 얼마나 이해가능할지를 의미한다. 그래서 transparent model 은 다소 단순한 알고리즘이 포함된다.
- 대조적으로, opaque model 은 이해가능하기 어려운 알고리즘을 포함한다. opaque model 은 모델의 판단에 대한 추가 설명이 필요하며, 이를 위해 사후 설명력 기법(post-hoc explainability technique) 이 필요하다.

Model-specific 기법

- 모델의 본질적인 구조를 이용하여 설명력을 제공하는, 특정 알고리즘에만 적용 가능한 기법
- ex. input image 에 대한 discriminative image region 을 localize 하는 CAM 기법
- ex. 트리 앙상블 모델에 사용되는 inTrees

> Interpreting Tree Ensembles with inTrees
>
> Tree ensembles such as random forests and boosted trees are accurate but difficult to understand, debug and deploy. In this work, we provide the inTrees (interpretable trees) framework that extracts, measures, prunes and selects rules from a tree ensemble, and calculates frequent variable interactions.
>
> https://arxiv.org/abs/1408.5456

Model-agnostic

- 일반적으로 사후 분석이 가능하며, 어떤 알고리즘에도 적용가능하다. 인기 있는 XAI 기법 상당수가 이 분류에 속한다.

- LIME :
  - Local Interpretable Model-agnostic Explanations
  - 특정 예측값 근처에서 지역적 해석력을 도출하는 기법이다.
  - 실제 적용 시에는, 설명력을 원하는 특정 데이터 포인트 x 주변에 샘플 데이터를 무작위로 생성한 후, π(근접성) 을 반영하기 위해 데이터가 x에 가까울 수록 높은 가중치를 적용한다. 그 다음, 샘플 데이터를 학습하여 x 주변 모델을 가장 잘 설명하는 local interpretable model 을 찾아 모델의 지역적 해석력을 도출한다.

- SHAP :
  - SHapley Additive exPlanation 은 Shapley value 를 이용하여 예측에 대해 각 feature 의 기여도를 계산하는 기법이다. 섀플리 값은 feature 의 모든 combination 에 대해 feature 의 기여도를 계산하여 그 값을 평균낸 값이다.

XAI 의 한계

XAI 에서 도출된 설명력을 과연 얼마나 믿을 수 있을까 하는 또다른 문제가 발생한다. 여러가지 XAI 기법을 시도했을 때 서로 다른 결고가 나온다면 어떤 결과를 따라야 하는가. 어떤 XAI 설명력 기법도 절대적인 지표가 되지 못하고, 이것은 설명력의 quality 를 파악하는 것이 매우 어려운 작업이기 때문이다.

설명력 성능 평가에 대한 연구는 다른 XAI 에 비해 상대적으로 덜 활발하고 아직까지는 설명력을 해석하는데 정량적인 기준대신 사용자의 주관적, 정성적인 기준이 들어가는 경우가 많다. 그렇기에 XAI 자체의 신뢰도를 높이고 XAI 간 정량화된 지표를 통한 비교 분석을 위한 더 많은 연구가 필요하다.