XAI 설명 가능한 인공지능 3장 리뷰

작성일시	@2022년 3월 2일 오전 9:39
최종 편집일시	@2022년 3월 2일 오전 10:41
회의 유형	XAI책
작성자	
참가자	
속성	

3.1 머신러닝 이해

머신러닝이란, 인간이 직접 논리를 구축하는 것이 아닌 학습 방식을 먼저 입력한 후 기계가 스스로 logic을 만들어가게 제작하는 과정

이 과정의 결과물로 의사 결정을 할 수 있게 된 머신러닝 산출물 = 모델(or 머신러닝 모델)

인간이 이해하기에 너무 많고 복잡한 매개변수를 가질 때, 머신러닝의 모델의 의사 결정을 인간이 이해할 수 없을 때의 모델을 블랙박스(Black box)라고 부른다.

3.2 블랙박스 들여다 보기

XAI - 머신러닝 모델의 블랙박스 성향을 인간이 이해할 수 있는 수준까지 분해하는 기술

그림 3.1을 참고해 환자의 당뇨병 유무를 진단해보자

3.1은 인공지능 모델이 그린 당뇨병 진단 의사 결정 트리(Decision Tree)다.

트리 모형을 따라가면서 0.101999916이라는 마지막 데이터(leaf data)를 얻었다.

이 수치는 logistic function으로 (logistic function = $e^{(X_i)}/1 + e^{(X_i)}$)

로지스틱 공식에 따라, 모델이 환자가 당뇨병에 걸렸다고 진단할 확률은 52.55%이며, 이는 전문가보다는 근거가 부족하나 비전문가의 근거 없는 예상보다는 해석적이라는 것을 알 수

있다.

XAI의 기본 기법의 하나인 피처 중요도(Feature Importance)는 모델이 의사 결정을 수행하는 과정에서 어떤 Feature들이 가장 크게 기여했는지를 측정한다. ex) ‘사람의 신체가 성장하는 요인’

- 이를 통해 모델이 어떤 데이터를 비중 있게 다루는지 들여다볼 수 있다.

3.3 시각화와 XAI의 차이 이해하기

XAI $\neq$ Visualization(시각화)

즉 XAI가 대개의 경우 시각화 기법에 의존하고 있지만 머신러닝 모델을 시각화했다고 해서 모두 XAI로 볼 수 없다. → 핵심은 해석 가능성

- 해석 가능성이란 해당 모델을 신뢰해야하거나 하지 않는 이유, 모델의 특정 모델 결정 근거, 어떤 결과가 예상되는지 판단하는 과정

데이터 및 모델 설명 기법으로 다음과 같다.

1. 대리 분석 (Surrogate Analysis)
2. 부분 의존성 플롯 (Partial Dependence Plots, PDPs)
3. 유사도 분석 (Similarity Measure)
4. 피처 중요도 (Feature Importance)

시각화 (Visualization) 과 도해법 (Graphical Method)은 모델을 분석한 이후에 추가로 수행하는 후처리 방식이다.