

XAI_04. 참고자료들


🕒 생성일	@2022년 8월 31일 오후 10:09
🏷️ 유형	머신러닝/딥러닝
👤 작성자	

Model-agnostic 기법

'해석할 수 있는 기계학습/5. 모델 불특정성 방법' 카테고리의 글 목록

'해석할 수 있는 기계학습/5. 모델 불특정성 방법' 카테고리의 글 목록

🔗 <https://eair.tistory.com/category/%ED%95%B4%EC%84%9D%ED%95%A0%20%EC%88%98%20%EC%9E%88%EB%8A%94%20%EA%B8%B0%EA%B3%84%ED%95%99%EC%8A%B5/5.%20%EB%AA%A8%EB%8D%B8%20%EB%B6%88%ED%8A%B9%EC%A0%95%EC%84%B1%20%EB%B0%A9%EB%B2%95>



Model-specific (주로 computer vision 분야)

'Deep learning study/Explainable AI, 설명가능한 AI' 카테고리의 글 목록

포항에서 대학원생활 하고 있습니다. 질문 및 논의 정말 환영합니다! <https://www.linkedin.com/in/wonju-seo-5a9922132/>

🔗 <https://wewinserv.tistory.com/category/Deep%20learning%20study/Explainable%20AI%2C%20%EC%84%A4%EB%AA%85%EA%B0%80%EB%8A%A5%ED%95%9C%20AI>




Model agnostic - LIME (지역 대체모델) 에 대해서

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

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model.

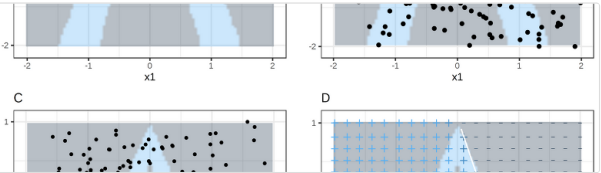
🌐 <https://arxiv.org/abs/1602.04938>



[해석할 수 있는 기계학습(5-7)] 지역 대체모델(LIME)

지역 대체모델은 블랙박스 기계 학습 모델의 개별 예측을 설명하는 데 사용되는 해석할 수 있는 모델입니다. 지역적 해석 가능한 모델 불특정성 설명(Local interpretable model-agnostic explanations, LIME) 은 저자들이 지역 대체모델의 구체적 구현을 제안하는 논문입니다. 대체모델은 기본 블랙박스 모델의 예측에 근사하게 학습됩니다. LIME은 전역 대체모 델을 양성하는 대신 지역 대체모델을 학습시켜 개별 예측을 설명하는 데 주력합니다.

🔗 <https://eair.tistory.com/26?category=883307>



LIME 은 'Local interpretable model-agnostic explanations (지역적 해석 가능한 모델 불특정성 설명)' 의 약자이다. LIME 은 전역 대체 모델을 양성하는 대신 지역 대체모델을 학습시켜 개별 예측을 설명하는데 주력한다.

LIME 은 데이터의 변화를 기계학습 모델에 줄 때 예측값에 어떤 일이 일어나는지 테스트한다. LIME 은 순열 샘플과 블랙박스 모델의 해당 예측으로 구성된 새로운 데이터 집합을 생성한다. 이 새 데이터 집합을 이용 해 해석가능한 모델을 학습하고, 이 모델은 샘플링된 객체와 관심있는 객체의 근접성에 의해 가중치가 정해진다. 학습된 모델은 기계 학습 모델 예측의 좋은 근사값이어야 하지만, 좋은 전역 근사값이 될 필요는 없다. 단지 local에 집중하기만 하면 된다.

수학적으로 지역 대체 모델은 다음과 같이 표현된다.

$$\text{explanation}(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

- 인스턴스 x 에 대한 설명 모델은 손실함수 L 을 최소화하는 모델 g 라고 할 수 있다.
- 원본 모델 f 의 예측과 설명이 얼마나 가까운지를 측정한다.
- 모델 복잡성 $\Omega(g)$ 은 낮게 유지됨을 보장한다.
- G 는 가능한 설명의 집합체이다.
- 근접도 측정치 π_x 는 인스턴스 x 의 주변이 얼마나 넓은지를 고려하기 위해 정의한다. 실제로 LIME 은 손실 부분만 최적화한다.

지역 대체모델을 학습시키는 다음과 같은 방법이 있습니다.

- 블랙박스 예측값에 대한 설명을 확인하고자 하는 관심 인스턴스를 선택한다.
- 데이터셋에 작은 변화를 준 다음 새로운 데이터 포인트에 대한 블랙박스 예측값은 얻는다.
- 관심 인스턴스에 대한 근접도에 따라 새로운 가중치를 측정한다.
- 변동 사항이 있는 데이터셋에 가중치가 적용된 해석할 수 있는 모델을 학습시킨다.
- 지역 모델을 해석하여 예측에 대해 설명한다.

텍스트에 대한 LIME

텍스트에 대한 LIME 은 원본 텍스트에서 시작하여, 새로 생성된 텍스트는 원본 텍스트에서 임의로 단어를 삭제하여 생성된다.

간단한 예시로, 모델이 어떤 단어를 기준으로 스팸 메일 혹은 스팸 댓글을 분류하는지 알고자 한다고 가정해보자.

스팸 클래스에 대한 양성, 음성 원본 데이터셋은 다음과 같다.

	CONTENT	CLASS
267	PSY is a good guy	0
173	For Christmas Song visit my channel! ;)	1

LIME 방법을 적용하면, 로컬 모델에 사용되는 데이터셋에 변형을 만들어낸다.

	For	Christmas	Song	visit	my	channel!	;)	prob	weight
2	1	0	1	1	0	0	1	0.17	0.57
3	0	1	1	1	1	0	1	0.17	0.71
4	1	0	0	1	1	1	1	0.99	0.71
5	1	0	1	1	1	1	1	0.99	0.86
6	0	1	1	1	0	0	1	0.17	0.57

prob 열은 각 문장 변형에 대한 스팸 예측 확률을 나타낸다. weight 열은 우너래 문장에 대한 변동의 근접성을 1에서 제거된 단어의 비율을 뺀 값으로 계산한다. 물론 각 단어 당 부여되는 가중치값은 1/n 으로 동일하다.

case	label_prob	feature	feature_weight
1	0.1701170	good	0.000000
1	0.1701170	a	0.000000
1	0.1701170	is	0.000000
2	0.9939024	channel!	6.180747
2	0.9939024	For	0.000000
2	0.9939024	;)	0.000000

위의 이미지는 LIME 알고리즘에 의해 확인된 추정 지역 가중치이며, channel! 이란 단어가 문장에 있을 때 스팸일 가능성이 높게 측정된다.

정리

지역 대체모델은 해석할 수 있는 모델을 훈련시키고 해석할 수 있는 문맥과 경험으로부터 이익을 얻는다. 인간 친화적인 설명을 통해 모델에 대한 어느 정도의 신뢰성을 확보할 수 있다. 다만, 예측을 완전히 설명해야 할 수 있는 적합성 시나리오에는 LIME 을 사용하는 것이 항상 합리적일 수는 없다.

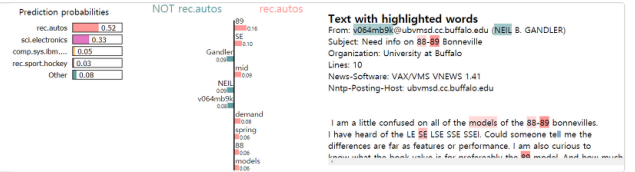
LIME 의 가장 큰 장점은 해석할 수 있는 모델이 블랙박스 예측에 얼마나 근접한지를 수치화(fidelity)해서 데이터 인스턴스의 인접 지역에서 블랙박스 예측을 설명하는데 있어 해석할 수 있는 모델이 얼마나 신뢰 가능한지 알 수 있다.

텍스트 문서 분류와 이미지 분류에 LIME 적용

인공지능이 궁금하다고 ? 들어와봐

본 글은 투윅스 블로그를 참고하여 작성한 것임을 미리 말씀드립니다. 오늘은 파이썬에 구현된 LIME 라이브러리를 직접 실행해보도록 하겠습니다. 크게 텍스트와 이미지 데이터에 대해 LIME을 돌려보고 그 결과를 해석하려고 합니다. 먼저, 사이킷런에서 제공하는 20가지의 카테고리를 포함하는 뉴스 기사를 대상으로 LIME for Text 과정을 소개합니다.

🌐 <https://moondol-ai.tistory.com/397>



구현파일

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/52132f34-e980-4235-adf3-0837ae74c850/LIME.ipynb>

Brain-Tumor-Classification-with-Efficient-Net-and-Grad-CAM-Visualization/Notebook.ipynb at main · baotramduong/Brain-Tumor-Classification-with-Efficient-Net-and-Grad-CAM-Visualization

Brain Tumor Classification with Efficient Net Convolutional Neural Network (CNNs) - Brain-Tumor-Classification-with-Efficient-Net-and-Grad-CAM-Visualization/Notebook.ipynb at main · baotramduong/Br...

<https://github.com/baotramduong/Brain-Tumor-Classification-with-Efficient-Net-and-Grad-CAM-Visualization/blob/main/Notebook.ipynb>

Explainable AI: Brain Tumor Classification with EfficientNet and Gradient-Weighted Class Activation...

A brain tumor is a growth of abnormal cells that have formed in the brain. Brain and other nervous system cancer is the 10th leading cause of death for men and women.

<https://medium.com/mlearning-ai/explainable-ai-brain-tumor-classification-with-efficientnet-and-gradient-weighted-class-activation-24c57ae6175d>

