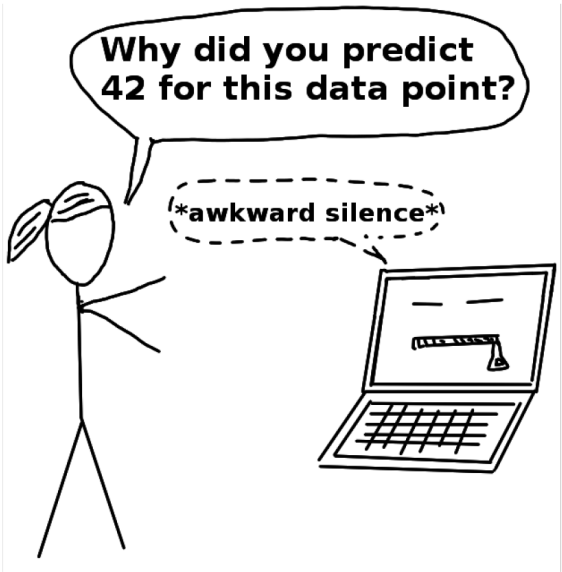
Chapter 1 Introduction

Interpretable Macbine Learning 이란?

- **refers to methods and models that make the behavior andpredictions of machine learning systems understandable to humans**
- Below is Black Box Model, not interpretable



- Interpretability is
  - the degree to which a human can understand the cause of a decision
  - the degree to which a human can consistently predict the model's result

Interpretablilty 참고 자료

Demo on MNIST (https://lrpserver.hhi.fraunhofer.de/handwriting-classification)

Demo on NLP (https://lrpserver.hhi.fraunhofer.de/text-classification)

google distill (https://distill.pub/2018/building-blocks/)

http://interpretable-ml.org/cvpr2018tutorial/ (http://interpretable-ml.org/cvpr2018tutorial/)

Chapter 2 Interpretability

Interpretable Macbine Learning 이 왜 중요한가?

knowing the 'why' can help you learn more about the problem, ex) why model fail

- need for interpretability arise from an incompleteness in problem formalization

Unexpected events make human curious, and needs an explanation of that

- ex) AI 기반 대출 거부, AI 기반 면접 탈락 => 설명 요구
- ex) 뜻밖의 추천 => 설명 요구

very-high risk 에서는 safty measure가 필요

- ex) 최근 비행기 사고는 자동 항법 조정 장치의 오류 때문. 설명 여구

Bias finding

- ex) 특정 소수 민족 집단에 대한 많은 대출 거부 => 설명 필요

socal acceptance

- 사람은 AI를 의인화하려고 할 것이고, 이 때 사람과의 interaction처럼 설명 필요

model debugging and auditting

- An interpretation for an erroneous prediction helps to understand the cause of the error.

기타 관련 사항

- Fairness = bias
- Privacy
- Robustness
- Causality

Interpretable Macbine Learning 이 중요하지 않은 경우는?

- When the model has no significant impact
- when the problem is well studied

IML 관련 용어집

Intrinsic or post-hoc

- intrinsic interpretability
  - interpretability is achieved by restricting the complexity of the machine learning model
  - ex) Linear Model, Decision Tree
- Post-hoc interpretability
  - application of interpretation methods after model training
  - ex) permutation feature importance

Result of interpretation method

Feature summary statistic

- single number per feature, ex) feature importance

Feature summary visualization

- ex) heatmap of feature importance
- ex) partial dependency plot
- ex) feature dector visualization of CNN

Datapoint

- ex) To explain the prediction of a data instance, the method finds a similar data point bychanging some of the features for which the predicted outcome changes in a relevant way
- ex) the identification of prototypes of predicted classes.

intrinsically interpretable model

- approximate black-box model with an interpretable model
- ex) LIME : DNN 기반 모델을 Linear한 interpretable model로 설명

Model-specific or model-agnostic

- Model-specific interpretation tools are limited to specific model classes.
  - ex) interpretation of weight is limited to simple linear model
  - ex) GRAD-CAM : limited to convolutional layer
- Model-agnostic can be used to any ML model, and post-hoc

Local or global

Scope of Interpretability

Algorithm Transparency

- How does the algorithm create model?
- ex) CNN in image may learn edge detectors...
- ex) DNN are less transparent...

Global, Holistic(총체적) Model Interpretability

- You comprehend the entire model at once
- holistic view of features
- very difficult in practice

Global Model Interpretability on a Modular Level

- How do parts of the model affect predictions?
- ex) BatchNorm이 미치는 영향 파악

Local Interpretability for a Single Prediction (이 책의 주요 관심사)

- Why did the model make a certain prediction for an instance?

Local Interpretability for a Group of prediction

- ex) 특정 subset of data에 대한 높은 오분류 원인 분석시 필요할 듯

Evaluation of Interpretability

- Application level evaluation (real task)
  - Make it product, and be evaluated by End User (Domain Expert)
  - ex) 의료 쪽 이상 탐지 시스템의 설명을 전문가인 의사가 듣고 평가
- Human level evaluation (simple task)
  - 쉬운 설명을 비전문가가 평가
- Function level evaluation (proxy task)
  - 사실상 자동화된 평가
  - ex) Shorter Decision Tree => better explanation

Properties of Explanations

Robnik-Sikonja and Bohanec, 2018

Properties of Explanation Methods

Expressive Power

- 설명은 일종의 언어이다.
- IF-Then rule, decision tree, weighted sum, NLP 등등으로 (이해 가능할 만한) 표현력이 있어야 한다.

Translucency (투명성)

- how much the explanation method relies on looking into ML model
- ex) linear model은 완전 투명하다.
- ex) Black box model은 입력 대비 출력 변화 양상 기반 설명이므로, 완전 불투명하다.

Portability (이식성)

- range of ML model with which explanation method can be used
- 불투명한 설명 방법은 오히려 높은 이식성을 가진다.
- ex) GradCAM은 CNN에서는 잘 되는데 RNN에는 적용 불가라서 이식성이 빵이다.

Algorithm Complexity

- 설명 방법의 계산 복잡도.
- inference는 빠르게 되는데, 설명 생성이 너무 느리면 곤란

Properties of Indivisual Explanations

Fidelity

- How well does the explanation approximate the prediction of the black box model

- ex) 기존 빚이 많고 무직이이라서 대출을 승인했습니다. => zero fidelity

Consistency

- 같은 데이터셋, 같은 테스크라면 모델이 달라도 같은 prediction에 대해서 비슷한 설명을 해야 일관된 설명이다.
- 반면 모델이 너무 상이하면(ex. SVM, Linear Model), 서로 다른 feature를 쓰므로 설명도 다를 것이 당연하다. => Rashomon Effect
- 그러나 비슷한 모델(ex. vgg16, vgg8)이라면 설명은 유사해야..

Stability

- How similar ar the explanations for similiar instances?

Comprehensibility

- How wee do human understand explanations?
- 코끼리 다리 만지기 식으로 정확히 정의하기는 쉽지는 않지만... 가장 중요한..

Certainty

- reflection of model uncertainty
- 모델의 판단 자체가 불확실하다면, 어떤 식으로 이것이 설명에도 반영되어야 한다.

Degree of Importance

- 설명의 각 부분부분에도 경중이 있으므로.. 이것이 잘 분간되게.. 강약중간약 설명..

Novelty

- 설명 대상 데이터 instance가 novelty하다면..
- model 결과도 uncertain하고,.. 설명도 사실성 무용지물..

Human-friendly Explanations

Miller 2017

Contrastive Explanation (Lipton 1990)

- conterfactual explanation
- How would the prediction have been if input X had been different?
- ex) 이러저러해서 대출 거절했어요 => 대출을 받으신 다른 분 사례에 견주어 봐서, 이런저런 점이 보강되면 대출 승인 될 수 있어요
- ex) 임상실험에서 실험군, 대조군
- 머신러닝에서는 reference instance을 정하고, 이에 견주어 설명해야 한다.

Explanation are selected

- 긴 설명(모든 원인을 다 나열하는 것은)은 거부감이 든다.
- 핵심 원인 몇몇을 찍어서 설명

Explanation are social

- 설명은 사람에게 하는 소통이다.
- 머신러닝에서는 설명의 대상(explainee)이 누구인지에 따라 설명이 달라져야 한다.
  - ex) 자율 주행차의 사고 유발 설명 => 차 소유자, 교통 당국, 법정

Explanation focus on the abnormal

- 사람은 독특한 원인 기반 설명을 선호한다.
- ex) 저 집은 커서 비싸 => 평범한 설명
- ex) 저 집은 마이클 조단이 살던 대라서 비싸 => 독특한 원인 기반 설명

Explanation are truthful

- 실제 현실에 부합해야 한다.
- selective한 설명을 하다보면 단순화를 시도하다보니 truthful한 설명이 되지 않을수도 있다.

Consisten with prior belief of explainee

- 사람의 직관과 선입견에 너무 벗어나는 설명이면 곤란

In [ ]: