



Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI



이상용 / 2020-02-24



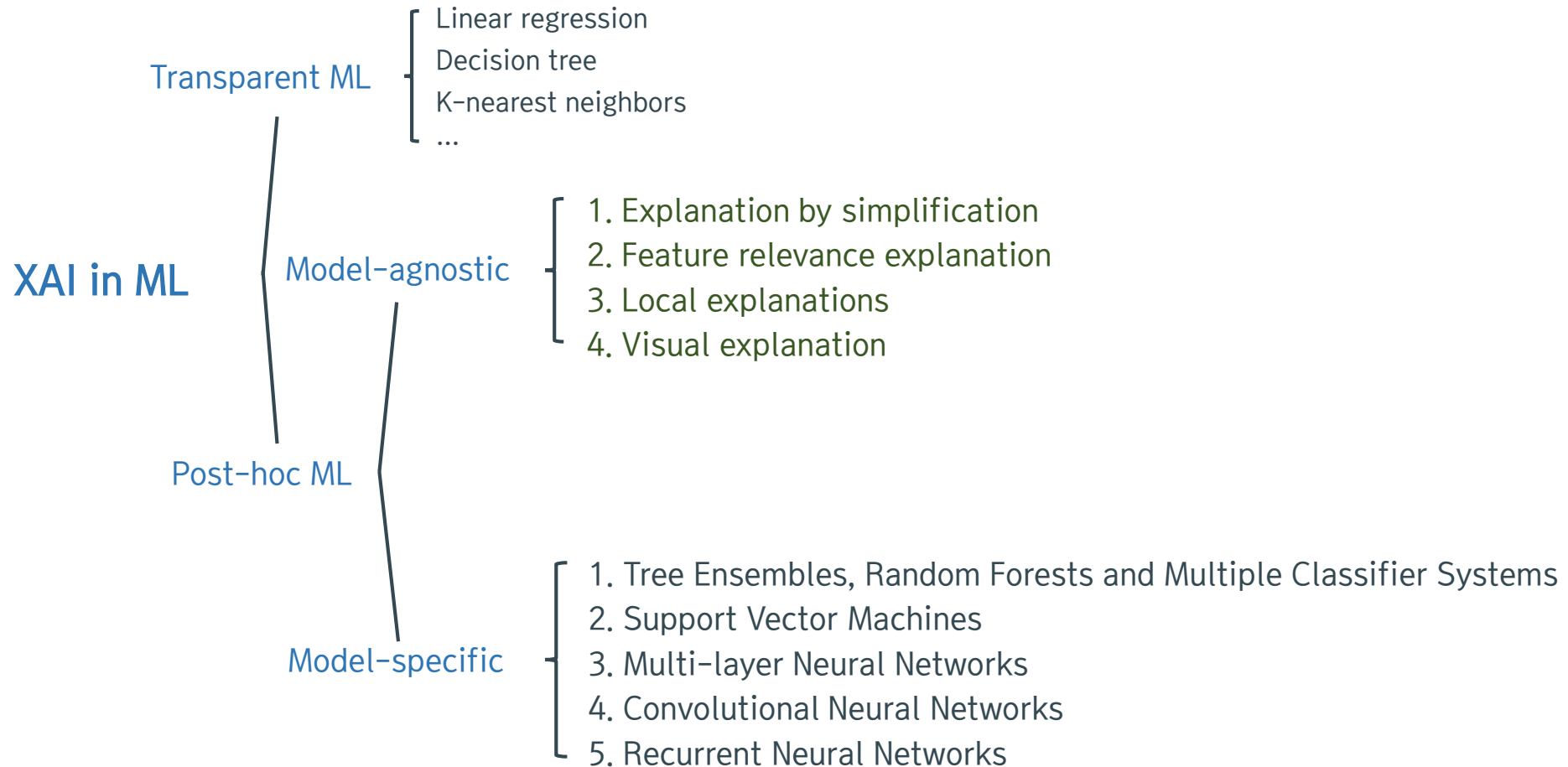
Computational Data Science LAB



CONTENTS

1. Post-hoc Explainability Techniques for Machine Learning Models: Taxonomy, Shallow Models and Deep Learning
- 
- 

02 | Post-hoc Explainability Techniques for Machine Learning Models: Taxonomy, Shallow Models and Deep Learning



02 | Post-hoc Explainability Techniques for Machine Learning Models: Taxonomy, Shallow Models and Deep Learning

- Explanation by simplification
 - ✓ Model agnostic post-hoc 방법들 중 가장 광범위하고, 이 범주에 속하는 거의 모든 방법들은 rule extraction technique에 기반함
 - ✓ *Local explanations* 또한 모델의 특정 부분만 나타내기 때문에 이 범주에 속함
 - ✓ LIME, G-Rex, ...
- Feature relevance explanation
 - ✓ 모델의 출력된 prediction에 대해 각 feature가 가지는 influence, relevance, importance를 측정하여 순위를 매기거나 모델의 기능을 설명하는 것을 목표로 함
 - ✓ SA (Sensitivity Analysis), LRP, SHAP (SHapley Additive exPlanations), ...

02 | Post-hoc Explainability Techniques for Machine Learning Models: Taxonomy, Shallow Models and Deep Learning

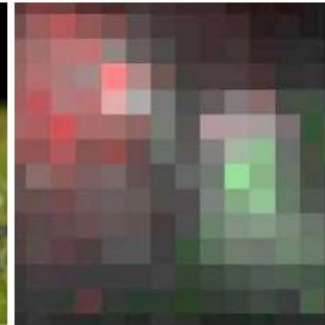
- Visual explanation
 - ✓ Visual explanation은 모델에 구매 받지 않는 설명을 달성하기 위한 수단으로 Post-hoc Explainability 방법들 중 덜 흔함
 - ✓ 이러한 방법의 설계는 모델의 내부 구조에 상관없이 어떤 모델에도 원활하게 적용될 수 있도록 보장해야 하므로 불투명한 모델의 input과 output 만으로 시각화를 하는 것은 복잡함
 - ✓ Global SA (GSA), ICE (Individual Conditional Expectation), ...

02 | Post-hoc Explainability Techniques for Machine Learning Models: Taxonomy, Shallow Models and Deep Learning

- Explainability in Deep Learning
 - ✓ Multi-layer Neural Networks
 - DeConvNet, Guided BackProp, LRP, ...
 - ✓ Convolutional Neural Networks
 - LRP, CAM, LIME, ...
 - ✓ Recurrent Neural Networks



(a) Heatmap [168]



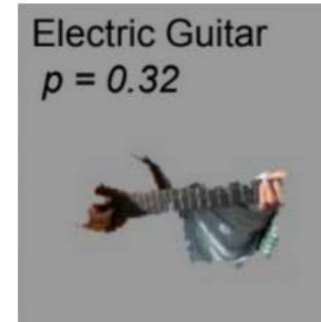
(b) Attribution [293]



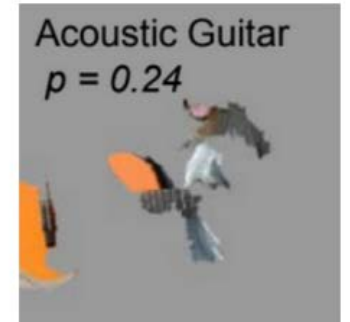
(c) Grad-CAM [292]



(a) Original image



(b) Explaining *electric guitar*



(c) Explaining *acoustic guitar*

Q&A

감사합니다.

02 | Visual explanation

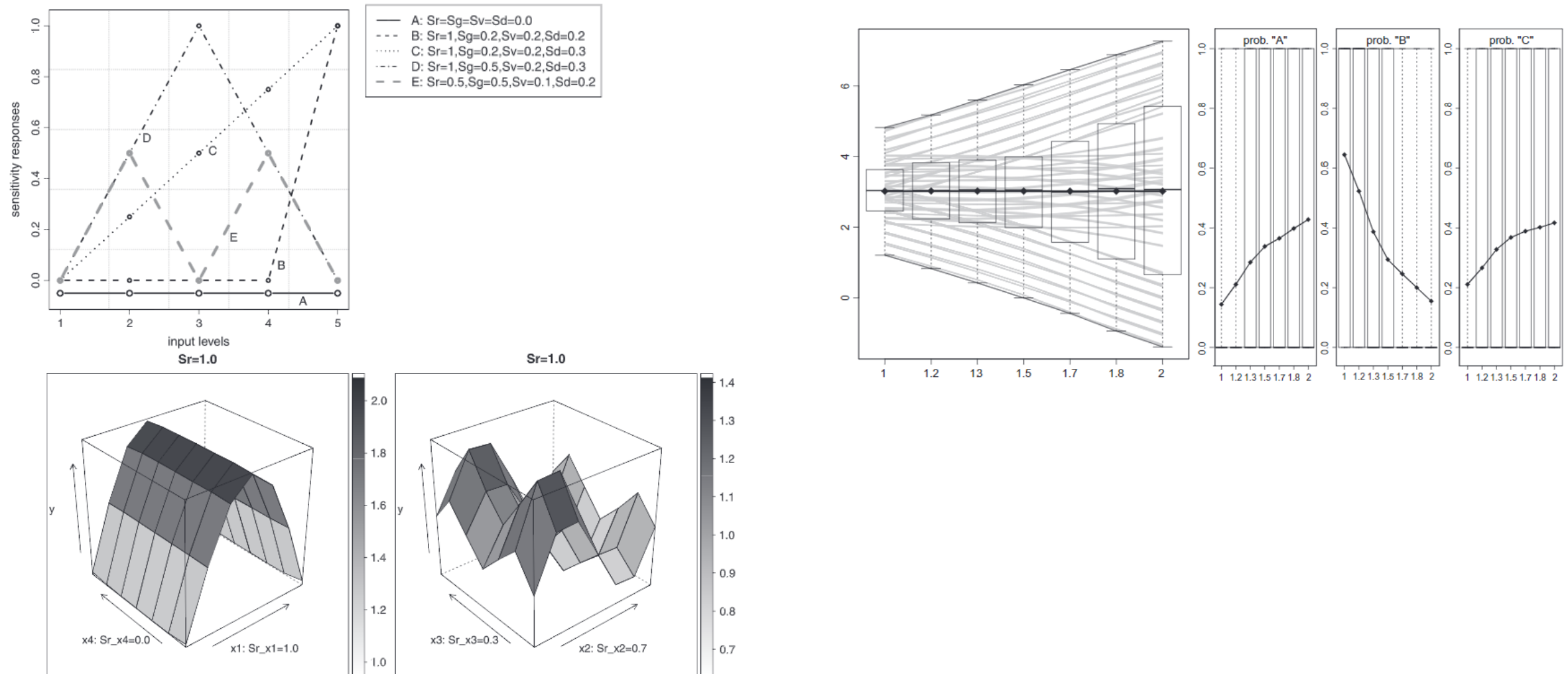


Fig. 2. Variable effect characteristic surface for the psin task and pairs (x_1, x_4) (left) and (x_2, x_3) (right).