Explainability

Sini Suresh
M1 – DATA AI
IP Paris

11/01/2021

"But Why?" Understanding Explainable Artificial Intelligence

- CV rejection in a few seconds of submission
- Black-box algorithms
  - Ethics and Trust
- Explainable AI (XAI)
- Challenges:
  - Opaqueness – Weighing/ Interpretable models/ Discard Deep-NN
  - Exploiting human strength
  - Human centeredness

"But Why?" Understanding Explainable Artificial Intelligence

- Ethical concerns and lack of trust in these technologies will continue to limit their adoption of AI
- XAI will be one piece of this solution.
- Combine computer science, social science, and human- computer interaction
- Explanatory systems that interact naturally with non-experts is a necessity

Explaining Explanations: An Overview of Interpretability of Machine Learning

- Explainable models are interpretable by default, but the reverse is not always true.
- The goal of *interpretability* is to describe the internals of a system in a way that is understandable to humans.
- The goal of *completeness* is to describe the operation of a system in an accurate way.
- Challenge of XAI - Completeness and Interpretablity

Explaining Explanations: An Overview of Interpretability of Machine Learning

- Ethical concerns:
  - When is it unethical to manipulate an explanation to better persuade users?
  - How do we balance our concerns for transparency and ethics with our desire for interpretability?
- Deep Networks – *Processing, Representations, Explanation-Producing Systems*
- Related Work – Interpretability
  - Taxonomy - application-grounded, human-grounded, and functionally grounded

Explaining Explanations: An Overview of Interpretability of Machine Learning

- Taxonomy

Processing	Representation	Explanation Producing
Proxy Methods	Role of layers	Scripted conversations
Decision Trees	Role of neurons	Attention-based
Salience mapping	Role of vectors	Disentangled rep.
Automatic-rule extraction		Human evaluation

TABLE I
THE CLASSIFICATIONS OF TOP LEVEL METHODS INTO OUR TAXONOMY.

Report Structure

- Introduction
- Fairness & Bias
  - Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges
  - What does it mean to ‘solve’ the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems
  - Fairness Through Awareness
  - On the Apparent Conflict Between Individual and Group Fairness
- Explainability
  - "But Why?" Understanding Explainable Artificial Intelligence
  - Explaining Explanations: An Overview of Interpretability of Machine Learning
  - A Survey on Explainable Artificial Intelligence (XAI): towards Medical XAI
  - Explanation in Artificial Intelligence: Insights from the Social Sciences

Thank you!