Interpretable Machine Learning

Introduction, Motivation and History

Learning goals

- Why do we need interpretability?
- What have been the developments until now?

WHY INTERPRETABILITY?

- Machine learning (ML) has a huge potential to aid the decision-making process in various scientific and business applications due to its predictive power.
- ML models usually are intransparent black boxes (or rather greyish boxes), e.g., bagged or boosted trees, RBF SVM, deep neural networks.
- The lack of explanation hurts trust and creates a barrier for the adoption of ML, especially in critical areas where decisions can affect human life (e.g., medicine).
- As a result, a lot of disciplines where trust in the model is critical still rely on traditional statistical models, e.g., GLMs, with less predictive performance.

BRIEF HISTORY OF INTERPRETABILITY

- Linear regression models date back as far as Gauss (1777 - 1855), Legendre (1752 - 1833), and Quetelet (1796 - 1874).
- Rule-based ML, which covers decision rules and decision trees, has been an active research area since the middle of the 20th century.
- Active research in the interpretation of modern ML models began in the 2000s, which reused a lot of concepts from sensitivity analysis (SA). Publications in SA date back as far as the 1940s.
- The built-in feature importance measure of random forests was one of the first important milestones.
- The deep learning hype in the 2010s resulted in a lot of publications related to explainable AI (XAI).
- IML as an independent field of research took off around 2015 with novel techniques such as LIME.

WHEN DO WE NEED INTERPRETABILITY?

- To **Justify** (and increase trust in models): investigate if and why biased, unexpected or discriminatory predictions were made.
- To **Control**: debug models, identify and correct vulnerabilities and flaws.
- To **Improve**: understanding why a prediction was made makes it easier to improve the model.
- To **Discover**: learn new facts, gather information and gain insights.

Figure: Reasons for IML.

Doshi-Velez, F., and Kim, B. (2017) Adadi, Amina, and Mohammed Berrada (2018)