Laboratoire d'Informatique et d'Automatique pour les Systèmes

Proposition of Master 2 Subject

Title: On Machine Learning Models Interpretability and Explainability

Laboratory: LIAS/ENSMA

Encadrant(s): Prof. Allel HADJALI (allel.hadjali@ensma.fr)
Dr. Seif-Eddine Benkabou (seif.eddine.benkabou@univ-poitiers.fr)
Dr. …….

Key Words: Models of ML, Explanation, SHAP value, Sensor data

Subject description

A key component of an artificially intelligent system is the ability to explain the decisions, recommendations, predictions or actions made by it and the process through which they are made. Despite the high accuracy in their predictions/recommendations, Machine Learning (ML) models are not able to explain their results, they are considered as "black box" models. This nature of ML-models limit their adoption and practical applicability in many real world domains and affect the human trust in them. Starting from the rule "Better Interpretability Leads to Better Adoption", the issue related to explanation and interpretation in ML is considered as one of the current hot topics in Data Science field.

Explainable AI (XAI) refers to the tools and techniques that can be used to make any black-box ML to be understood by human experts. There are many such tools available in the market such as LIME, SHAP, ELI5, Interpretml, etc. For instance, the SHapley Additive exPlanations (SHAP) methodology is recently introduced to explain and interpret any ML prediction. The idea is to show how much has each feature value contributed to the value predicted.

The objective of this work is twofold:

- First, provide a comprehensive and complete survey about approaches dedicated to ML models explanation. Then, propose a categorisation of such approaches w.r.t. to some criteria conveniently chosen.
- From this categorisation, identify the family of tools that are more appropriate to explain the prediction/recommendation in the Multisensor Data context.

Bibliographie

- Erik Štrumbelj and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions". In: Knowledge and information systems 41.3 (2014), pp. 647–665.
- Lundberg, Scott et al. – "Consistent individualized feature attribution for tree ensembles", 2019. (https://arxiv.org/pdf/1802.03888.pdf)

- Scott M. Lundberg, Su-In Lee, "A Unified Approach to Interpreting Model Predictions", NIPS 2017: 4765-4774
- Rich Caruana, Scott Lundberg, Marco Túlio Ribeiro, Harsha Nori, Samuel Jenkins, "Intelligible and Explainable Machine Learning: Best Practices and Practical Challenges", KDD 2020: 3511-3521
- Goodman, Bryce, and Seth Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation", *AI magazine* 38.3 (2017): 50-57, aaai.org.