MIS 49Y APPLIED MACHINE LEARNING

HOMEWORK III (8 points)

Introduction

Explainable Artificial Intelligence (XAI) has been developed as a subfield of AI, focused on exposing complex AI and ML models to humans in a systematic and interpretable manner. XAI helps to understand, visualize and interprete machine learning models.

There are two tasks in the assignment (Grading (1st task: 4 points, 2nd task: 4 points)). Please note that using XAI techniques in your term project is highly encouraged. You will have also question(s) from XAI domain in your FINAL exam.

1. XAI Algorithms Presentation

Prepare a short report (or slides) for XAI. Your report template can be both in doc/docx (2 pages) or ppt/pptx (5-9 slides). Answer the following questions in your report (or slides). Don't forget to state your references.

a) List and explain at least 3 XAI algorithms.
b) Compare the listed XAI algorithms (properties, pros, cons, performance etc.)

2. XAI Algorithm Usage Demo via Jupyter Notebook

a) Build a Machine Learning model for any problem (and dataset) you are interested in.
b) Use one of the XAI algorithms to explain the model you built.
c) Support your solution with plot(s) and model evaluation metrics.

You may use any dataset(s) and any model you wish for the assignment. You may select your dataset from kaggle, UCI Machine Learning Repository or another publicly available datasets. You can find many others on the Web. Pick something that interests you.

Deliveries

Upload your report (in doc/ docx or ppt/ pptx format) , jupyter solution notebook(s) and dataset(s). Don't forget to put code descriptions (markdown or comments), mention about your references/sources in your notebook.

Please plan for a demo of up to 10 minutes.

1. Report (power point (max 9 slides) or word doc)
2. Jupyter Demo Notebook

Some resources to check **:**
(Feel free to check other and more recent resources.)

1. Explaining Explanations: An Overview of Interpretability of Machine Learning, Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, Lalana Kagal, 2019, https://arxiv.org/abs/1806.00069
2. Interpretable Machine Learning — A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/
3. SHAP: A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874
4. https://analyticsindiamag.com/8-explainable-ai-frameworks-driving-a-new-paradigm-for-transparency-in-ai/