INFO-H420 Management of Data Science and Business Workflows

Project on Responsible Data Science

The goal of this project is to combine principles from responsible data science to study an ML pipeline. Specifically, you will study a classifier that predicts whether someone will have income “>50K” on the Adult data set. The project asks you to investigate several responsible data science aspects of the ML pipeline.

1. Classification

Preprocess the data, and binarize Age. Split the data into train, validation, test sets, and train a classifier; we will refer to it as the **classifier**. Measure the performance of the classifier on the test set.

2. Fairness

Assess the group fairness of the classifier, assuming the *protected* attributes are Age, Sex. Choose any fairness metric, and apply a technique to ensure the classifier is fair. We will refer to it as the **fair classifier**. Report the chosen fairness metric on the classifier and on the fair classifier.

3. Privacy

Assume that the attributes Age, Sex are *sensitive*. Compute a cross-tabulation showing how many people exist in value combinations of the two sensitive attributes.

Apply local differential privacy to the responses of the individuals about what is their Age and Sex, and create a private data set. You may want to explore various epsilon values.

Compute a cross-tabulation on the private data, and estimate how many people exist in value combinations of the two sensitive attributes. Quantify the errors in the estimation.

Split the private data in the same manner as in (1), and train a classifier; we will refer to it as the **private classifier**.

Measure the performance of the private classifier. Is there an impact on model performance due to privacy compared to the classifier?

4. Privacy and Fairness

Consider the same fairness metric and fairness mitigation method as in (2). Create a fair version of the private classifier; we will refer to is as **private+fair classifier**.

Assume, you’re an auditor that has access to the real sensitive values of Age and Sex. Using the real values of Age and Sex, measure the fairness of the private+fair classifier, and compare it to that of the fair classifier. Draw conclusions.

5. Explainability

Study the explainability of the private classifier. Identify instances where the model is wrong but highly confident, and explain them.

Assume you have access to the real sensitive values of Age and Sex. Investigate whether the noisy values for these attributes are responsible for the model being confident and wrong.

6. Explainability and LLMs

Select an explainability method and create a natural language interface to it. The idea is to take an explanation (e.g., feature-importance pairs, examples) and present it as text that a person can easily understand.

For this task, you may want to use small LLMs that can run locally on your computer, and LM Studio[1] might be a good option.

7. Free Exploration

Play around with the dataset and report any interesting (in terms of responsible data science) insights that you may find.

Instructions

The project contributes 30% to the overall grade (i.e., 6/20).

This project is to be made in **groups of six** persons. You are asked to form the groups via "Group Choice for Project" on the Université Virtuelle (UV). If you cannot find a partner, please post a request in the "Discussion Forum" on UV.

You are asked to submit a short **report** presenting your solution to the exercises, including justifications for the choices and assumptions made.

The report and any supporting files have to be uploaded to "Project" on UV by **December 10**, 2024.

[1] https://lmstudio.ai