Algorithmic Fairness, Accountability, and Ethics, Spring 2024, IT University of Copenhagen

Mandatory Assignment 1

Martin Aumüller, Anders Weile Larsen

- **Hand-out**: February 09, 2024
- **Hand-in**: March 7, 2024 at 23:59
- **What to hand-in:** A report as pdf summarizing the main findings, **max. 3 pages**, including plots. A jupyter notebook detailing the process. Upload the two files as a single zip file on learnIT.
- **Where to start:** You can find a template to get started in the assignment on learnIT.
- **Dataset:** US Census data from https://github.com/zykls/folktables. We use data of individuals from the state California in 2018, as detailed in the template.
  The template also details which attributes we use as feature vector.
  More details on the dataset can be found in the accompanying paper at https://arxiv.org/pdf/2108.04884.pdf.

Task 1 (Classifiers and fairness considerations)

1. Starting from the template, train two different classifiers on the training data: a white-box model using logistic regression, and a black-box model using a random forest. Consider feature engineering and scaling steps necessary for some of these classifiers and summarize the necessary changes in your report. For both models, report on the accuracy of the classifier on the test set.
2. For each classifier, measure statistical parity, equalized odds (both in terms of `T = 0` and `T = 1`), and equality of outcome (both in terms of `S = 0` and `S = 1`) (*Lecture 2*). Plot the results and discuss the differences that you observe.
3. Change the classification pipeline to (approximately) fulfill **one** of the fairness criteria by post-processing the results. How did the intervention influence the different fairness criteria, how did it change the accuracy of the classification?

Task 2 (Explaining white-box models)

1. Explain the trained logistic regression model. In particular, discuss which features in the model are deemed most relevant. Reflect on the interpretation. Does it fit your intuition about the prediction task?
2. Pick one data point in the test dataset. Find a counterfactual data point that contrasts the outcome of the inference on this data point (e.g., "had X had feature P >=, then it had been classified as ..."). Describe how you used the model explanation to find such a counterfactual.

Task 3 (Model-agnostic explanations)

1. Both for the white-box and the black-box classifier, use the `shap` module to explain predictions. Contrast the two models to each other: What are similarities, how do they differ?
2. For logistic regression, compare the model-agnostic explanation to your analysis in Task 2. How do the explanations differ?

Task 4 (Reflection)

Given the outcome of your study, which classifier is most suited for the prediction task under accuracy, explainability, and fairness considerations?

Checklist

To avoid surprises, please make sure that your hand-in covers the following parts:

Overall

- [ ] Assignment: Concise discussion of the results
- [ ] Plots: Labels and titles clear and readable?
- [ ] Code: can run the whole notebook? Modular, concise, and documented code?
- [ ] Page limit: At most 3 pages. Plots should still be readable!

Task 1

- [ ] Part 1: Discussion feature engineering and scaling steps
- [ ] Part 1: Correct implementation of one-hot-encoding (if used for model)
- [ ] Part 1: Build and train relevant models
- [ ] Part 2: Code to compute statistical parity, equalized odds and equalized outcome
- [ ] Part 2: Plot comparing the three metrics
- [ ] Part 3: Discussion on accuracy changes and how other measures were affected by the intervention

Task 2

- [ ] Part 1: Discussion on which features affect the prediction negatively / positively. Does it follow intuition?
- [ ] Part 2: Discussion on counterfactual

Task 3

- [ ] Part 1: Shap plot (and force plot if applicable). Make sure that you have plotted the correct values.
- [ ] Part 1: Is shap aggregated per feature?
- [ ] Part 2: Comparison to white-box model explanation from task 2

Task 4

- [ ] Discussion on classifier most suited for the prediction given all considerations