**Fairness - Expert**

**Module 2 Technical interpretations of fairness/bias (3 lessons):**

This module will expand on the intermediate course on fairness. It will provide further tools for applying learnings on fairness/bias in the development of a technological solutions.

a. Mathematical formulations, quantification and metrics of bias and fairness and mitigation techniques for different of models

b. Performing a technical assessment of the fairness/bias quantification requirements of a problem (considering data used, model used, end-user requirements)

c. Overview of available tools for fairness/bias assessment and mitigation techniques

d. Communicating bias in the data/model to end-users, including metrics visualisation

Apply learnings to use cases from TRI's projects or products. The learner will be provided with datasets replicating samples of datasets used in TRI’s projects and predictive models and a Jupyter notebook to support data exploration and application of techniques to identify bias/fairness issues on both datasets and models performances.

**Syllabus material**

a. Mathematical formulations, quantification and metrics of bias and fairness and mitigation techniques for different of models

Detail mathematical definitions of specificity, sensitivity, precision, and recall etc. Link these to fairness metrics I.e.

Demographic parity

False Discovery rate

False Ommision rate

False Positive Rate parity

False Negative Rate Parity

Discuss incompatibility of some fairness metrics and the implications for selecting a fairness metric.

b. Performing a technical assessment of the fairness/bias quantification requirements of a problem (considering data used, model used, end-user requirements)

Examples of situations when different metrics of fairness might be used. Useful resource here is the ‘fairness tree’ from DSSG

Timeline

Description automatically generated

Fairness tree that helps decide which fairness metric might be most appropriate in various cases.

**Break out session:** Read through some project / use-case descriptions and try to ascertain which fairness metrics might be most appropriate, considering the perspective of both end users and demographics affected.

**Hands on example:** Use the case study from above about recruitment to a project. Load pseudo-data and walk through technical implementation of data cleaning, model training and evaluation. Build and train a collection of models and evaluate their performance, both overall and specifically for two subgroups.

c. Overview of available tools for fairness/bias assessment and mitigation techniques

Discuss bias assessment / auditing tools like SHAP and Aequitas.

And three mitigation strategies:

re-sampling – three substrategies here: Undersampling, Oversampling, Generating synthetic data

post-hoc disparity mitigation - selecting different thresholds different groups to try and achieve parity in some metric such as true positive rate

Regularisation – solve an optimisation problem in some fairness metric (Fairlearn)

d. Communicating bias in the data/model to end-users, including metrics visualisation

**Hands on example with notebook:** Load a pre-made model (or the previous model) and Use SHAP to identify important features of the data/model and comment on how this might lead to bias.

Refs:

Zafar, Muhammad Bilal, et al. "Fairness constraints: Mechanisms for fair classification." *Artificial intelligence and statistics*. PMLR, 2017.

Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." *Advances in neural information processing systems* 29 (2016).

Possible datasets:  
Credit card fraud - <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>