Exploring the Unfairness of DP-SGD Across Settings

Frederik Noe, Rasmus Herskind and Anders Søgaard
University of Copenhagen

- We explore the dynamics of **fairness** and **privacy** across linear classification, deep learning, and dimensionality reduction
- We measure **fairness** as equal risk and the so-called p%-rule
- We evaluate the performance and fairness of linear classifiers, deep classifiers, and classifiers trained on reduced dimensions, for different levels of differential **privacy**
- We establish the **logarithmic** relation between fairness and differential privacy

Data: The Trustpilot Corpus

Implementations: diffprivlib and Opals