Exercise 12

Group Fairness

Reliable and Trustworthy Artificial Intelligence
ETH Zurich

Problem 1 (Group Fairness via Post-processing)**.** Consider the post-processing method described in the lecture (slide 6). In this task we will explore this method on a toy example. Consider the following dataset of 10 points from distribution $\mathcal{X}$, represented as tuples (x, s, y), denoting respectively a 1D feature vector, binary sensitive group membership, and the target label:

$$D = \{(0.1, 0, 0), (0.2, 0, 0), (0.3, 0, 0), (0.8, 0, 1), (0.9, 0, 0),\\ (0.1, 1, 0), (0.3, 1, 1), (0.4, 1, 0), (0.5, 1, 0), (0.7, 1, 1)\}.$$

Further, assume our binary classifier g is the identity function.

1. For standard thresholds of $t_0 = 0.5$ and $t_1 = 0.5$ (no post-processing), estimate the accuracy of g using D?

2. Estimate the fairness of g using D, i.e., calculate the values of demographic parity distance, equalized odds distance, and equal opportunity distance? Definitions of these fairness constraints are given in the first fairness lecture. To interpret them as a distance follow the example of DP-distance given in the group fairness lecture (for equalized odds, the distances for two cases should be averaged).

3. Assume we keep $t_0 = 0.5$ fixed and want to change t_1 as a way to apply post-processing to make g more fair. For what value of t_1 is g the most fair with respect to three distances? To answer this question, fill in the code in the provided `task1.py` to calculate the relevant metrics, and plot the dependence of accuracy and fairness metrics on t_1.

Problem 2 (Bounding Unfairness with the Optimal Adversary)**.** Prove the key inequality used by FNF and FARE to upper bound the DP-distance of downstream classifiers:

$$\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) \leq 2 \cdot BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^\star) - 1,$$

where we use the notation from the lecture. Namely, $\mathcal{Z}_0$ and $\mathcal{Z}_1$ are the conditional distributions of z for $s = 0$ and $s = 1$, respectively. The DP-distance is defined as:

$$\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) = \left| \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_0} g(z) - \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_1} g(z) \right|,$$

and the balanced accuracy of the adversary as:

$$BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h) = \frac{1}{2} \left(\mathop{\mathbb{E}}_{z\sim\mathcal{Z}_0} (1 - h(z)) + \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_1} h(z) \right).$$

Problem 3 (FNF with Categorical Features)**.** Real-world datasets often contain categorical data, in which case the optimal bijections f_0 and f_1 can be directly computed, instead of using normalizing flows. Consider discrete samples x coming from a probability distribution $q(x)$ where each component takes a value from a finite set $\{1, 2, \ldots, d_i\}$. This implies a finite $\mathcal{X} = \{x_1, \ldots, x_m\}$. As before, our goal is to find bijections $f_0 : \mathcal{X} \to \mathcal{Z}$ and $f_1 : \mathcal{X} \to \mathcal{Z}$ that minimize the statistical distance of the latent distributions, i.e., minimize the adversary advantage. For simplicity, you can assume $\mathcal{Z} = \{1, \ldots, m\}$.

1. Describe the procedure used to construct the optimal f_0 and f_1.
2. Provide a proof that this procedure minimizes the balanced accuracy of the optimal adversary.
3. Are such f_0 and f_1 always the best choice in practice? If not, why?