Fairness

Recitation 11/11/2022

Dataset

- Our review of Fairness concepts will be done in the credit rating context
- We will use the German Credit Rating Dataset

Fairness Concepts Review

1. **Anti Classification:**
   aka *Fairness through Blindness*, ignores sensitive attributes when making a decision.
   Operationalized as:

$$\forall x.\ f(x[p \leftarrow 0]) = f(x[p \leftarrow 1])$$

Is Anti Classification always good?
What about proxies?

Fairness Concepts Review

2. **Group Fairness:**
   aka *Independence*, states that the prediction should be independent of the sensitive attribute

$$P[Y' = 1|A = a] = P[Y' = 1|A = b]$$

What if the label and protected attribute are correlated?

Fairness Concepts Review

3. Separation:

aka *Equalized Odds*, states that the prediction should be independent of the sensitive attribute conditional on the target variable

$$P[Y' = 1 \mid Y = 0, A = a] = P[Y' = 1 \mid Y = 0, A = b]$$
$$P[Y' = 0 \mid Y = 1, A = a] = P[Y' = 0 \mid Y = 1, A = b]$$

I.e, all groups have the same false positive/negative rates

Exercise

- Like the American FICO scores, German citizens have Schufa scores
- Schufa scores are used to inform financial decisions in contexts like insurance and rentals
- The Schufa scoring system is owned by a private company and the algorithm is not public
- This makes it difficult to discern whether Schufa may (inadvertently) make unfair decisions against certain groups of people

Exercise

- There have been attempts at unearthing the inner workings of the system and identifying potential bias (most notable the OpenSCHUFA project)
- Today we will train a model on the Schufa score dataset and evaluate it's fairness using anti-classification considering gender to be a protected attribute
- Make a copy of this notebook