Session 8

Fairness

Recap

A binary classifier does not suffer from disparate impact if, given a sensitive variable z and a positive outcome $y = 1$,

$$P(\hat{y} = 1 | z = 0) = P(\hat{y} = 1 | z = 1)$$

Fairness

In the regression case, we can define

$$\text{fairness} = \left| \frac{1}{|Z_1|} \sum_{i \in Z_1} \hat{y}_i - \frac{1}{|Z_0|} \sum_{i \in Z_0} \hat{y}_i \right|$$

Scores for classification

In the classification case, we can approximate

$$\text{p-score} = \min\left(\frac{P(\hat{y}=1|z=1)}{P(\hat{y}=1|z=0)}, \frac{P(\hat{y}=1|z=0)}{P(\hat{y}=1|z=1)}\right),$$

or

$$\text{equal-op.} = \min\left(\frac{P(\hat{y}=1|z=1, y=1)}{P(\hat{y}=1|z=0, y=1)}, \frac{P(\hat{y}=1|z=0, y=1)}{P(\hat{y}=1|z=1, y=1)}\right)$$

Possible strategies

- Modify the data
- Modify the prediction process
- Modify model training

Modify the data

- Anonymize sensitive variable z
  - Probably not enough
- Drop sensitive variable z
  - *Might* not be enough
- Project out z covariance (analogous to PCA)
  - Makes model less *interpretable* / gain less insight

Modify the prediction process

- Most classification models can incorporate a decision threshold
    - Basically, use different thresholds dependent on z
- Needs knowledge of z at decision time

Modify model training

Training a Classifier

A classification algorithm with parameters θ is trained giving a convex loss function $L(\theta)$, such that we find the optimal parameters θ^*

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\ L(\theta)$$

Constrained fits

L is usually something that defines goodness-of-fit like cross-entropy, but can also contain terms that e.g. keep θ as small as feasible (ref. Lasso, Ridge regression). Why not try

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \; L(\theta)$$
$$\text{subject to} \quad |P(\hat{y}=1|z=0) - P(\hat{y}=1|z=1)| \leq \epsilon$$

But this is non-convex.

A better way

The way out is to constrain covariance of the outcome $\hat{y}$ and the sensitive variable z.

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \; L(\theta)$$
$$\text{subject to} \quad |\operatorname{Cov}\big(z, g_\theta(y,x)\big)| \leq c,$$

where

$$g_\theta(y,x) = \min\big(0, y\, d_\theta(x)\big),$$

and $d_\theta(x)$ is the signed distance from the classification boundary ($\theta^T x$ for linear models).