Foryciarz et al 2021 - Supplement A

1 Regularization approaches for equalized odds at a threshold

Let $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^N \sim P(X, Y, A)$ be a dataset, where X designates patient-level features, Y is a binary indicator for the occurrence of an outcome, and A is a discrete attribute that stratifies the population into K disjoint groups. We learn a model $f_\theta(x)$ to estimate $\mathbb{E}[Y \mid X = x] = P(Y = 1 \mid X = x)$ using empirical risk minimization:

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell(y, f_\theta(x)), \tag{1}$$

where ℓ is the cross-entropy loss and the expectation with respect to the dataset $\mathcal{D}$ indicates an empirical average over the dataset.

To construct a training objective that penalizes violation of equalized odds, we add a regularizer M_{EqOdd} to the training objective, with the degree of regularization controlled by a non-negative scalar λ:

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell(y, f_\theta(x)) + \lambda M_{\text{EqOdd}}. \tag{2}$$

We derive a differentiable regularizer for this setting by first noting that the true positive rate at a threshold t can be defined as

$$TPR := \mathbb{E}_{x\sim\mathcal{D}|Y=1}\, h(f_\theta(x) - t), \tag{3}$$

where $\mathbb{E}_{(x,y)\sim\mathcal{D}|Y=1}$ indicates the expectation over the subset of the population for which $Y = 1$, and $h(z) = \mathbb{1}\{z > 0\}$ is the step function. Similarly, the false positive rate is given by

$$FPR := \mathbb{E}_{x\sim\mathcal{D}|Y=0}\, h(f_\theta(x) - t). \tag{4}$$

In order to achieve equal sensitivity (true positive rate) across groups and equal specificity (true negative rate) across groups, it is sufficient for the model to achieve a true positive rate and false positive rate for each group equal to the true positive rate and false positive rate of the population overall. The direct incorporation of the true positive rate and false positive rate, as formulated above, does not provide a useful signal for stochastic gradient descent when incorporated into a regularizer because the step function $h(z)$ has a derivative of zero everywhere that the derivative is defined. To address this issue, we replace the step function $h(z)$ with the sigmoid function $\sigma(z) = \frac{1}{1+\exp(-z)}$, following Cotter et al. [1]. Furthermore, in practice, we compare the log of the scores to the log of the threshold. Terms representing the difference between the relaxed rates for a group A_k with the population are given by:

$$M_{TPR}^{k,t} = \mathbb{E}_{x\sim D|A=A_k,Y=1}\Big[\sigma(\log f_\theta(x) - \log(t))\Big] - \mathbb{E}_{x\sim D|Y=1}\Big[\sigma(\log f_\theta(x) - \log(t))\Big] \tag{5}$$

and

$$M_{FPR}^{k,t} = \mathbb{E}_{x\sim D|A=A_k,Y=0}\Big[\sigma(\log f_\theta(x) - \log(t))\Big] - \mathbb{E}_{x\sim D|Y=0}\Big[\sigma(\log f_\theta(x) - \log(t))\Big]. \quad (6)$$

To construct the regularizer, we combine these terms into a single non-negative regularizer over all groups and thresholds:

$$M_{\text{EqOdd}} = \sum_{t_i\in\{t_1,t_2\}} \sum_{k=1}^{K} \Big[(M_{TPR}^{k,t_i})^2 + (M_{FPR}^{k,t_i})^2\Big]. \quad (7)$$

To adjust for censoring, we use an inverse probability of censoring (IPCW) approach, and replace the empirical averages in equations 1, 2, 5, and 6 with weighted ones that incorporate the IPCW weights.

References

[1] Andrew Cotter, Heinrich Jiang, Maya R Gupta, Serena Wang, Taman Narayan, Seungil You, Karthik Sridharan, Maya R Gupta, Seungil You, and Karthik Sridharan. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *Journal of Machine Learning Research*, 20(172):1–59, sep 2019. URL `http://arxiv.org/abs/1809.04198`.