Bias and Fairness in ML

Fairness

Fairness Definitions

Equality of Treatment

- Fairness through Unawareness
  - An algorithm is fair if protected attributes are not explicitly used in the decision-making process
- Counterfactual Fairness
  - An algorithm is fair if its output remains the same when the protected attribute is flipped to its counterfactual value

Equality of Outcomes

- Demographic Parity
  - Members of groups have an equal probability of being assigned to the positive class
- Conditional Statistical Parity
  - Demographic parity holds given a set of legitimate factors
- Fairness Through Awareness
  - An algorithm is fair if it gives similar predictions to similar individuals

Fairness Definitions

Equality of Performance/ Error

- Predictive Parity
  - Equalizing $FDR_g = \frac{FP_g}{FP_g + TP_g}$
- Sufficiency
  - Equalizing FDR_g and $FOR_g = \frac{FN_g}{FN_g + TN_g}$
- Equal Opportunity
  - Equalizing $FNR_g = \frac{FN_g}{FN_g + TP_g}$
- Equalized Odds
  - Equalizing FNR_g and $FPR_g = \frac{FP_g}{FP_g + TN_g}$
- Treatment Equality
  - Equalizing $\frac{FP_g}{FN_g}$
- Test Fairness
  - Considers complete score distribution across groups

Table: Confusion matrix

		Prediction 0	Prediction 1	
Reference	0	TN	FP	N'
	1	FN	TP	P'
		N	P	

→ Notions are in conflict with each other and with overall accuracy

Fairness Definitions

Equality of Performance/ Error

- Predictive Parity
  - Equalizing $FDR_g = \frac{FP_g}{FP_g + TP_g}$
- Sufficiency
  - Equalizing FDR_g and $FOR_g = \frac{FN_g}{FN_g + TN_g}$
- Equal Opportunity
  - Equalizing $FNR_g = \frac{FN_g}{FN_g + TP_g}$
- Equalized Odds
  - Equalizing FNR_g and $FPR_g = \frac{FP_g}{FP_g + TN_g}$
- Treatment Equality
  - Equalizing $\frac{FP_g}{FN_g}$
- Test Fairness
  - Considers complete score distribution across groups

Table: Confusion matrix

		Prediction 0	Prediction 1	
Reference	0	TN	FP	N'
	1	FN	TP	P'
		N	P	

$\rightarrow$ *Notions are in conflict with each other and with overall accuracy*

Fairness Definitions

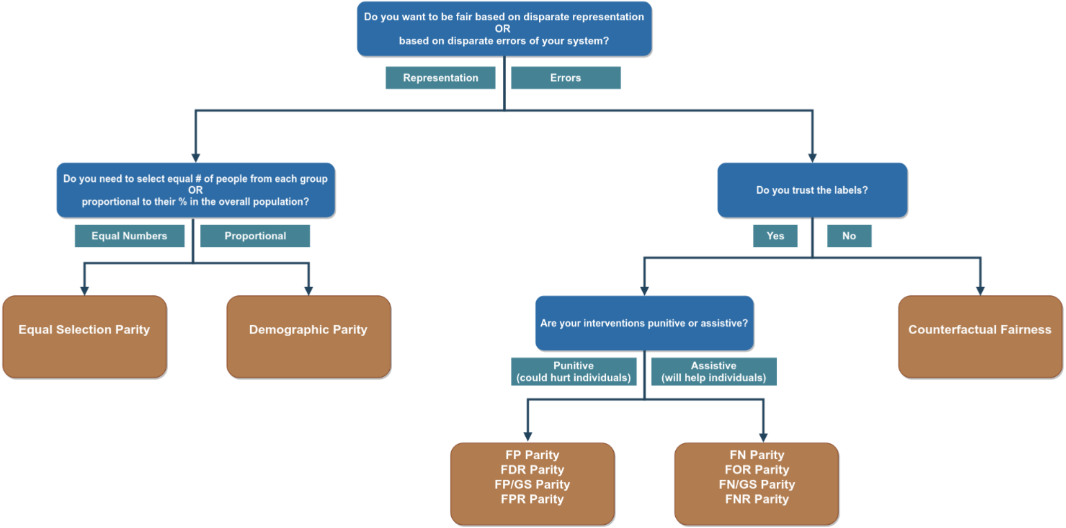
1. Individual Fairness: Give similar predictions to similar individuals
2. Group Fairness: Treat different groups equally
3. Subgroup Fairness: Extend group fairness to large collection of subgroups

Table: Categorizing Fairness Notions (Mehrabi et al. 2019)

	Group	Individual
Demographic parity	x	
Conditional statistical parity	x	
Equalized odds	x	
Equal opportunity	x	
Fairness through unawareness		x
Fairness through awareness		x
Counterfactual fairness		x

Fairness Definitions

Figure: Choosing Fairness Metrics (Saleiro et al. 2018)



Methods for Fair ML

Some Potential Solutions (Berk et al. 2017)

1. Pre-processing
   - Eliminating sources of unfairness in data before model training
     - Remove linear dependence between legitimate and protected predictors
     - Re-label some response values to make base rates comparable
     - Perturb class membership for protected attributes for some cases
2. In-processing
   - Making fairness adjustments as part of the model building process
     - Add fairness penalty to loss function
3. Post-processing
   - Adjust model output post-training to make it more fair
     - Randomly re-assign some predicted class labels

Software Resources

Resources for R

- PDP, ICE, ALE
  - Plot model surfaces: `plotmo`
  - Partial Dependence Plots: `pdp`
  - ICE plots: `ICEbox`
  - ALE plots: `ALEPlot`
- Interpretable Machine Learning in R: `iml`
- Descriptive mAchine Learning EXplanations: `DALEX`

References

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. https://arxiv.org/abs/1908.09635.

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/.

Rodolfa, K. T., Saleiro, P., Ghani, R. (2019). Bias and Fairness. In: Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J. (Eds.). Big Data and Social Science: A Practical Guide to Methods and Tools. https://coleridge-initiative.github.io/big-data-and-social-science/.

References

Apley, D. W. (2016). Visualizing the effects of predictor variables in black box supervised learning models. https://arxiv.org/abs/1612.08468.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A. (2017). Fairness in Criminal Justice Risk Assessments: The State of the Art. https://arxiv.org/abs/1703.09207

Fisher, A., Rudin, C., Dominici, F. (2018). Model Class Reliance: Variable importance measures for any machine learning model class, from the 'Rashomon' perspective. http://arxiv.org/abs/1801.01489.

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics, 29*(5), 1189–1232.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E. (2014). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. https://arxiv.org/abs/1309.6392.

Lum, K. and Isaac, W. (2016). To predict and serve? Significance 13, 14–19.

Murdoch, W., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. https://arxiv.org/abs/1901.04592.

Ribeiro, M. T., Singh, S., Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. https://arxiv.org/abs/1602.04938.

Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., Ghani, R. (2018). Aequitas: A Bias and Fairness Audit Toolkit. https://arxiv.org/abs/1811.05577