Fairness Criteria

Exploring Fairness in Machine Learning

Mike Teodorescu
Assistant Professor of Information Systems, Boston College
Visiting Scholar, MIT

Potentially Sensitive Attributes in Machine Learning

Some countries have laws that protect specific groups of people from discrimination based on certain individual attributes (often referred to as ‘protected attributes’), such as:

- race;
- religion;
- national origin;
- gender;
- marital status;
- age;
- socioeconomic status.

Base Case: Fairness Through Unawareness

- One approach is fairness-through-unawareness (Kusner et al, 2017; Chen et al, 2019), which leaves out of the model protected social attributes such as gender, race, and other characteristics deemed sensitive

- The fairness-through-unawareness approach conceptually parallels the "color-blind" strategy for promoting cultural diversity at work (Apfelbaum et al, 2010).

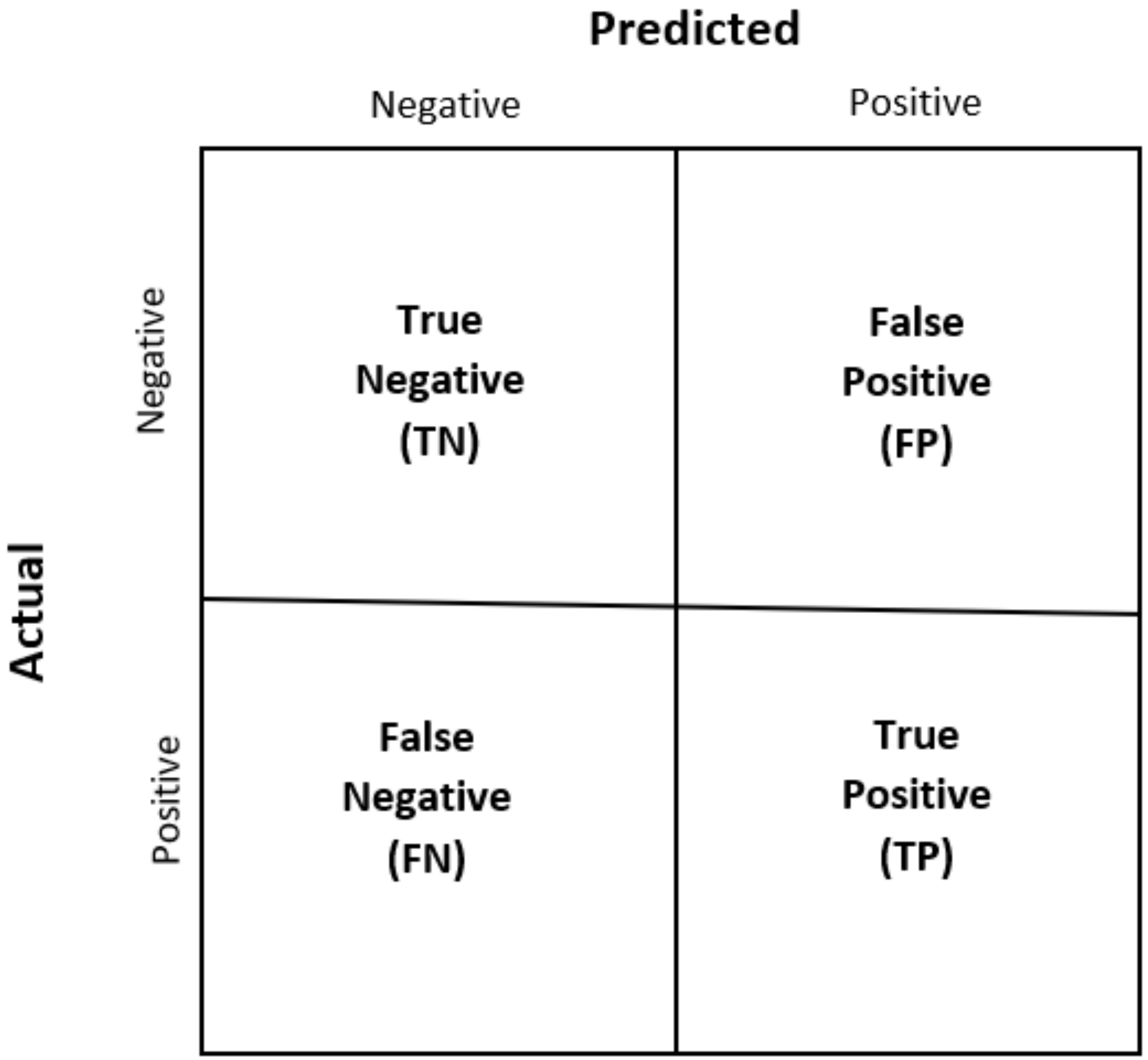
Confusion Matrix

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + FP + TN + FN}}$$

The four cells are:

- TP = true positive (Correctly classified as Positive)
- TN = true negative (Correctly classified as Negative)
- FP = false positive (Incorrectly classified as Positive)
- FN = false negative (Incorrectly classified as Negative**)**

Confusion Matrix



		Predicted	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Demographic Parity

- Demographic parity is the next step of the widely known remedies to unfairness in machine learning (Kusner et al, 2017) and is equivalent to independence of the outcome $\hat{Y}$ with respect to the protected attribute (A):

$$p(\hat{Y}|A = a) = p(\hat{Y}|A = a'), \qquad \hat{Y} \perp A,$$

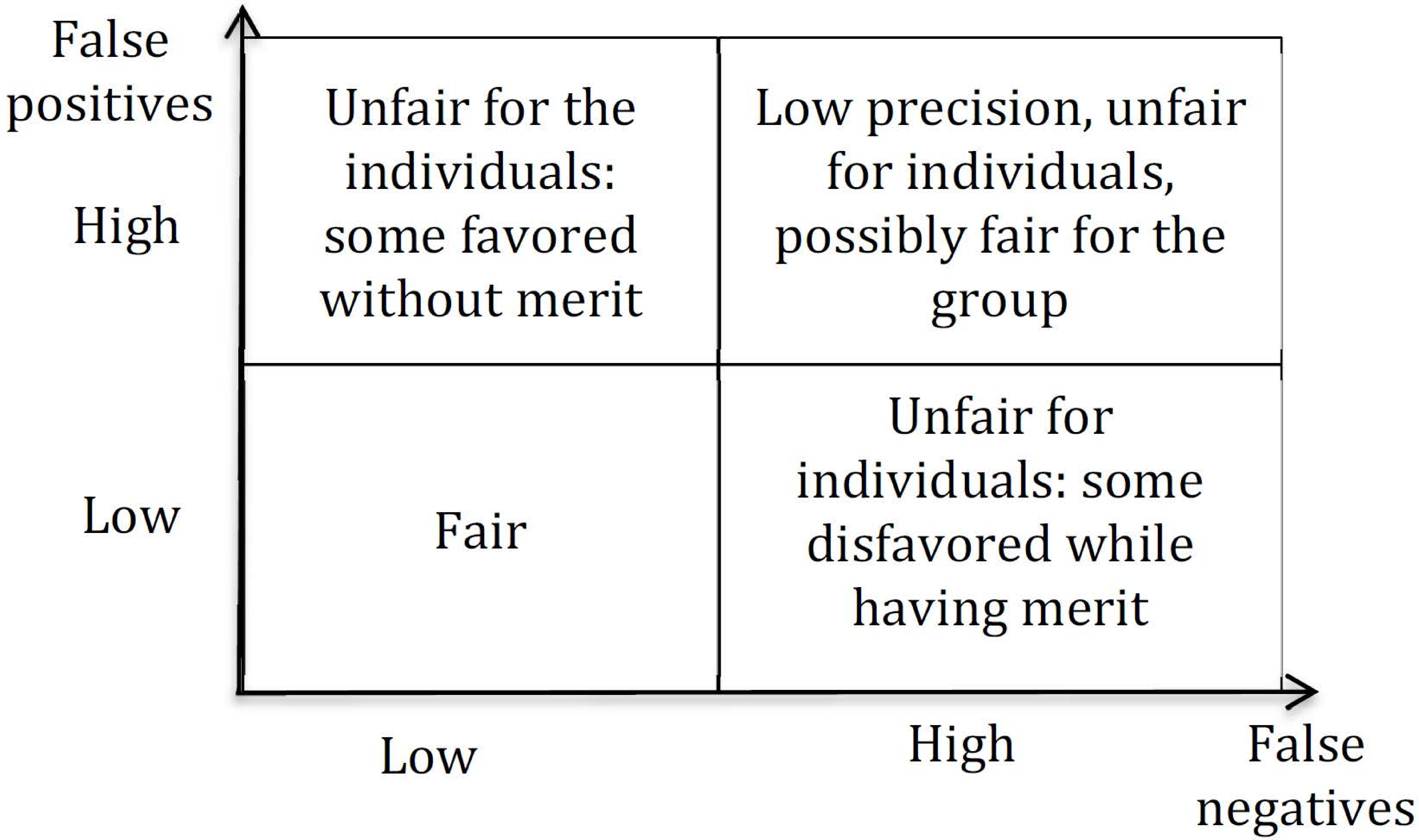
where $\hat{Y} \perp A$ denotes independence, and a and a' are any couple of values of the attribute.

- This definition expects the outcomes to be the same for groups, therefore the prediction independent of the protected attribute group membership.
- Example: probability of being hired is independent of gender

Demographic Parity

- Problems with demographic parity: what if we have people who are members of multiple protected groups?
- While enforcing group level fairness (say, same hiring rate for females and males), this can be unfair to the individual: it could force the algorithm to drop otherwise qualified individuals just to achieve independence of outcome with the attribute.
- Furthermore, there could be differences in qualifications across a non-protected attribute, say empathy, programming skills, communication skills, analytics, which would be washed off by forcing equality of probability of hire at the group level.

Fairness at the individual or group level



False positives	False negatives: Low	False negatives: High
High	Unfair for the individuals: some favored without merit	Low precision, unfair for individuals, possibly fair for the group
Low	Fair	Unfair for individuals: some disfavored while having merit

Equalized Odds

- Equalizing the odds = matching the True Positive Rate and False Positive Rate for different values of the protected attribute (Hardt et al, 2016)

$$\mathrm{p}(\hat{\mathrm{Y}}|\mathrm{A}=0,\mathrm{Y}=\mathrm{y})=\mathrm{p}(\hat{\mathrm{Y}}|\mathrm{A}=1,\mathrm{Y}=\mathrm{y}),\quad \mathrm{y}\in\{0,1\}$$

- This is hard to do but if achieved is one of the highest levels of algorithmic fairness

- If you'd like to learn more, see Hardt et al, 2016; Pleiss et al, 2017; Kilbertus et al, 2017.

Equalized Opportunity

- Equalized opportunity is concerned with treating fairly those who are determined to be worthy of acceptance (Y=1). It is not concerned with rejecting people fairly across protected groups. In other words, the false positive rates and the true positive rates do not both need to be equal across the protected categories. The equalized opportunity principle states the following condition for the probabilities: (Hardt et al., 2016)

$$\mathrm{p}\left(\hat{\mathrm{Y}} = 1 \middle| \mathrm{A} = 0, \mathrm{Y} = 1\right) = \mathrm{p}\left(\hat{\mathrm{Y}} = 1 \middle| \mathrm{A} = 1, \mathrm{Y} = 1\right).$$

- In a hiring example, this would be individuals deemed worthy of hiring by a human hiring officer, whereas $\hat{\mathrm{Y}}$ indicates those deemed worthy of hiring by the algorithm.

Review Questions

- What is “demographic parity”?
- What is “fairness through unawareness”?
- Is fairness at the group level always the best?
- What is the “confusion matrix”?
- What is the “equality of odds” criterion?

Acknowledgements

- This is joint work with Professors Lily Morse and Gerald Kane (Boston College) and Yazeed Awwad (Research Assistant, MIT D-Lab).

- The authors thank USAID-MIT Grant AID-OAA-A-12-00095 “Appropriate Use of Machine Learning in Developing Country Contexts” and the Carroll School of Management at Boston College for research funding. I thank Research Assistant Mariana Paredes (BC), Dr. Daniel Brown (HBS), Dr. Daniel Frey (MIT), Kendra Leith (MIT D-Lab), Nancy Adams (MIT D-Lab), Dr. Aubra Anthony (USAID), Dr. Shachee Doshi (USAID), Dr. Craig Jolley (USAID), Dr. Amy Paul (USAID), Dr. Rich Fletcher (MIT D-Lab), Dr. Maryam Najafian (MIT CSAIL), Amit Gandhi (MIT D-Lab), Lauren McKown (MIT CITE), and Dr. Sam Ransbotham (BC) for help with this work.

References

- Apfelbaum, E.P., Pauker, K., Sommers, S.R. and Ambady, N. (2010). In blind pursuit of racial equality?. In Psychological Science, 21(11), pp.1587-1592.

- Chen, J., Kallus, N., Mao, X., Svacha, G. and Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 339-348.

- Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems, pp. 3315-3323.

- Kilbertus, N., Carulla, M.R., Parascandolo, G., Hardt, M., Janzing, D. and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In Advances in Neural Information Processing Systems, pp. 656-666.

- Kusner, M.J., Loftus, J., Russell, C. and Silva, R. (2017). Counterfactual fairness. In Advances in Neural Information Processing Systems, pp. 4066-4076.

- Pleiss, G., et al. On Fairness and Calibration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), Advances in Neural Information Processing Systems 30 , pp. 5684–5693.

Thank you

Mike Teodorescu
Assistant Professor of Information Systems, Boston College
Visiting Scholar, MIT

hmteodor@mit.edu

MIT D-Lab USAID FROM THE AMERICAN PEOPLE CITE