Fairness and Bias

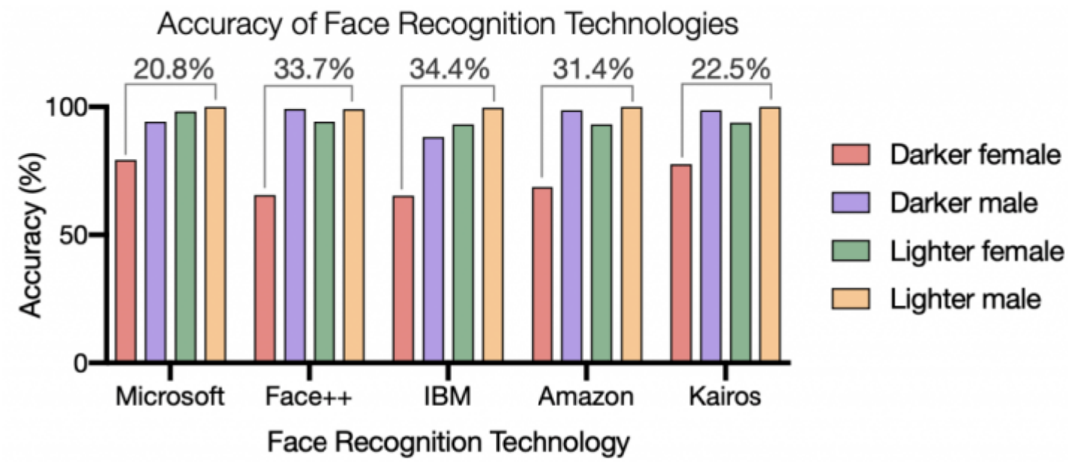
MIE223

Winter 2025

1 Introduction to Fairness and Bias

1.1 MIT Study on Racial/Gender Discrimination in Image Classifiers

- MIT Project known as Gender Shades
  - Examined facial recognition algorithms from Microsoft and IBM:
  - Did poorly on underrepresented groups (e.g. black females)



1.2 Data is Full of Biases

Representation and Collection Bias

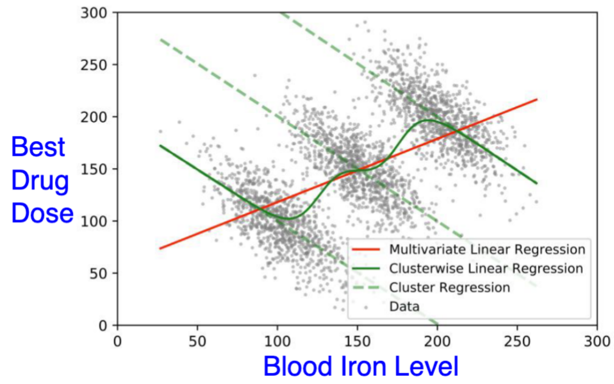
- The data is not representative of the population distribution it is intended to serve
  - Critical in medicine
- The data has insufficient representation of minority groups

Measurement and Historical Bias

- What you measure is misaligned with use case, or biased from historical data
  - E.g., using historical loan decisions to inform future decisions

Aggregation bias

- Effect for a group can be reversed when looking at data in aggregate



This data is multimodal. Gaussian means unimodal, so be careful applying unimodal tools to multimodal data. If you ignore modes, you get the red line rather than the green line.

1.3 Group and Individual Fairness

Much research on fairness aims to ensure "fair treatment" = parity
Group Fairness

- Achieve "parity" (equality) across protected and advantaged groups
- Groups usually determined by demographic attribute
  - e.g., Caucasian, non-Caucasian, Male, non-Male

Individual Fairness

- Similar individuals should have similar predictions / outcomes
- Requires measure of similarity between individuals
  - e.g., Euclidean or cosine similarity between demographic feature vectors

1.4 Group and Individual Fairness for Classification

Classification: predicting a discrete label, e.g., parole, no parole

- Group Fairness:
  - Fairness in Treatment: should not consider sensitive/protected attribute (e.g., gender, race, age)
    - But other attributes may be correlated with sensitive ones!
  - Fairness in Impact: classification outcome should be balanced across groups
    - E.g., false positive rate for each group should be the same
- Individual Fairness:
  - Enforce similar individuals to receive similar outputs (or distributions)
    - Requires a similarity function over sensitive attributes

1.5 On the chalkboard

Given your grades (vectorized) in previous courses, what letter grade will you get in this course? Your past history helps predict your future performance. You can regress directly to the grade, whereas classification goes to letter grades A,B,C,D,F Focusing on binary classification P, NP:

- True Positive (TP): predicted positive and actually positive
- True Negative (TN): predicted negative and actually negative
- False Positive (FP): predicted positive and actually negative
- False Negative (FN): predicted negative and actually positive

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN} \tag{3}$$

$$\text{False Discovery Rate (FDR)} = \frac{FP}{TP + FP} \tag{4}$$

$$\text{True Negative Rate (TNR)} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{False Negative Rate (FNR)} = \frac{FN}{TP + FN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

$$\text{F2 Score} = (1 + 2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + 2 \cdot \text{Recall}} \tag{10}$$

$$\text{F0.5 Score} = (1 + 0.5) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + 0.5 \cdot \text{Recall}} \tag{11}$$

Contingency table is a 2x2 table that summarizes the performance of a classification model Multiclass classification uses a confusion matrix

1.6 Achieving Parity

What types of parity in binary classification can we aim for?

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive,** Power	**False positive,** Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative,** Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= \frac{1}{\frac{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

What types of parity can we aim for in impact fairness?

- Demographic / statistical parity (predicted outcome only)
  - Different groups get same positive classification rates
    - ∗ E.g., same admission rate regardless of economic status
- Parity in errors (consider predicted and actual outcome)
  - Different groups have different classification rates
  - But we constrain the error rates (Accuracy, TPR, FPR, FDR, ...)

2 Can we have all desired Parities at once?

No

2.1 Incompatibility Between Fairness Metrics

Table 1: COMPAS Fairness Metrics

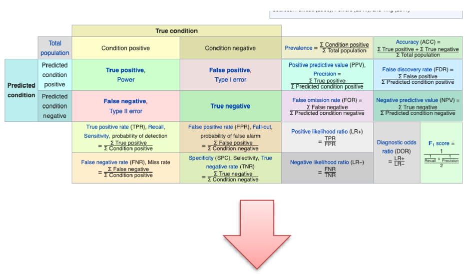
Metric	Caucasian	African American
False Positive Rate (FPR)	23%	45%
False Negative Rate (FNR)	48%	28%
False Discovery Rate (FDR)	41%	37%

ProPublica identified considerable disparities:

- The algorithm was twice as likely to incorrectly predict black individuals were at high risk of recidivating (FPR)
- It was also nearly twice as likely to incorrectly predict white individuals were at low risk of recidivating (FNR)

However, the creator of the algorithm pointed out that the algorithm is well- balanced across races on precision (equivalently, FDR), claiming this is the correct measure of fairness in this context.

- Who is right?
- Is the COMPAS algorithm biased?
- Can't the algorithm achieve both measures of fairness at the same time?



$$FPR = \frac{p}{1-p}\left(\frac{FDR}{1-FDR}\right)(1-FNR)$$

$$FPR = \frac{p}{1-p}\left(\frac{FDR}{1-FDR}\right)(1-FNR)$$

If prevalence is unequal across groups...

$$FPR = \frac{p}{1-p}\left(\frac{FDR}{1-FDR}\right)(1-FNR)$$

If prevalence is unequal across groups...

…and FDR (or precision) is equal across groups...

$$FPR = \frac{p}{1-p}\left(\frac{FDR}{1-FDR}\right)(1-FNR)$$

False Positive Rate
Among all actual 0's, fraction predicted to be 1

Prevalence
Fraction of actual 1's in population

False Discovery Rate
Among all predicted 1's, fraction that are actual 0's =(1 – precision)

False Negative Rate
Among all actual 1's, fraction predicted to be 0

$$FPR = \frac{p}{1-p}\left(\frac{FDR}{1-FDR}\right)(1-FNR)$$

If prevalence is unequal across groups...

…and FDR (or precision) is equal across groups...

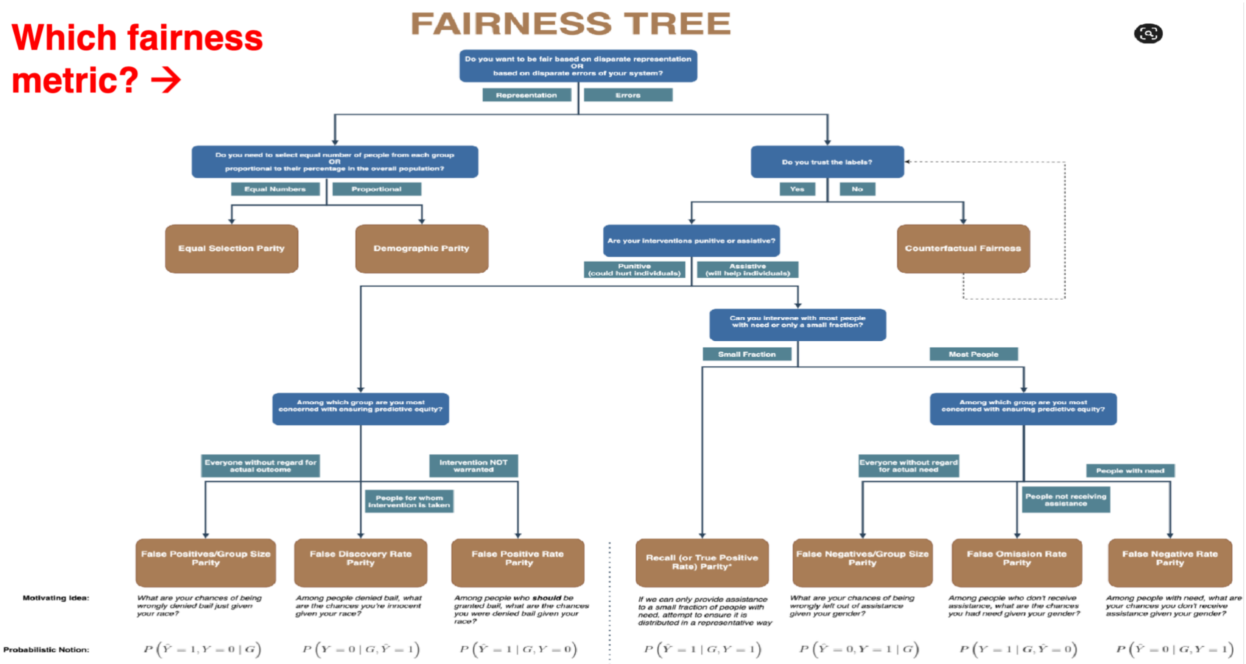
…then either FPR or FNR can be equal across groups, **but not both**

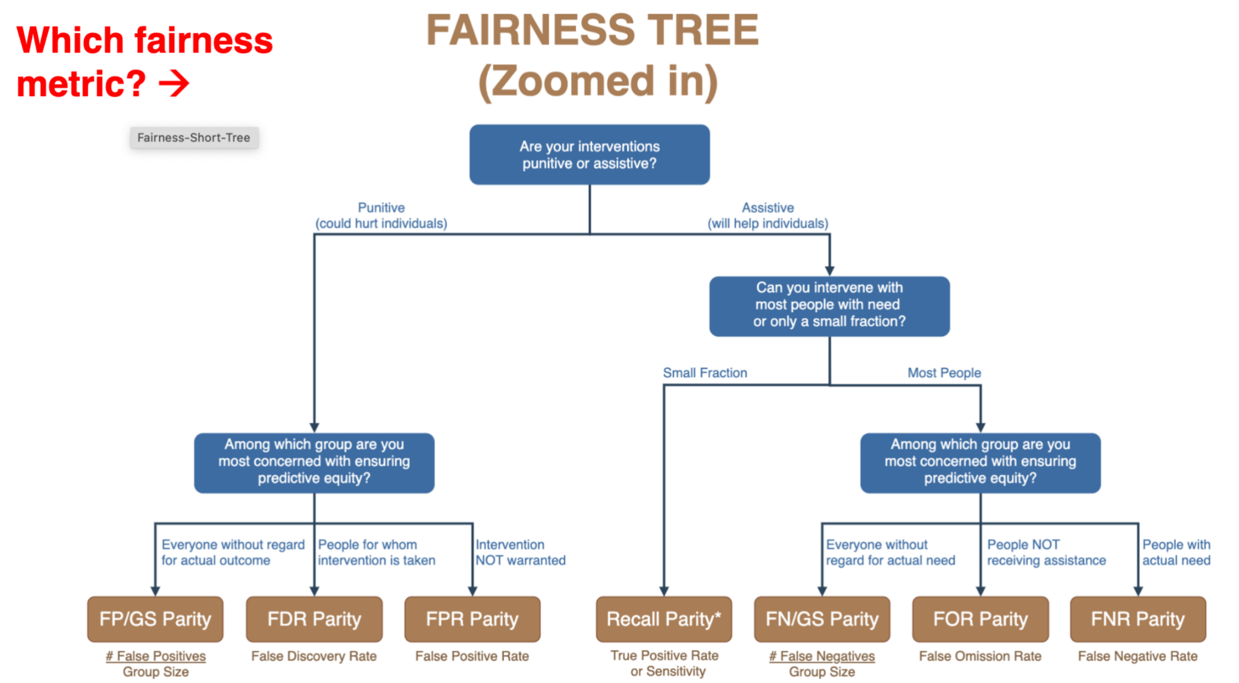
2.2 On the blackboard

$FPR_g = P_g/1 - P_g = FDR_g/(1 - FDR_g) * (1 - FNR_g)$
$FDR_g1 = FDR_g2$
$FNR_g1 = FNR_g2$
$FPR_g1 = FPR_g2$
You cannot have all 3 equal at once.
Not needed for exam

2.3 So... what does this mean?

Sadly, no... we can't have it all! But this also doesn't mean fairness is unachievable or we shouldn't bother. Instead, it means we have to carefully consider and choose the appropriate fairness metrics for our context





Not needed for exam

3 Working with Data Collected from Humans (e.g., Surveys)

Something you will likely do in your career Beware of many biases that can occur! (with potentially global impact: 2016 US election

3.1 Survey Methodology and Bias Prevention

- Survey Sampling and Bias:
- Question Design and Cognitive Biases: