

SEMANTICS DERIVED...

(How do we remove the bias?)

CERTIFYING AND REMOVING DISPARATE IMPACT

CERTIFY: How do we detect "potential for bias"?

REMOVE: How do we create a classifier less prone to discrimination?

3 NB APPROACHES FOR DISCRIMINATION-FREE CLASSIFICATION

1 NB:

2. NB:

FAIRNESS: DEFINITIONS

Y : True Label

\hat{Y} : Predicted Label

A : protected attribute

"TYPICAL" CONFUSION MATRIX

	$Y=0$	$Y=1$
$\hat{Y}=0$	TN	FN
$\hat{Y}=1$	FP	TP

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

$$\text{TPR} = \frac{TP}{FN + TP}$$

Let's use better notation:

	$Y=0$	$Y=1$
$\hat{Y}=0$	NN	NP
$\hat{Y}=1$	PN	PP

(Prediction, then truth)

Let's also say that:

$$*N = NN + PN$$

$$P* = PN + PP$$

$$** = *N + *P$$

$$\text{Accuracy: } \frac{NN + PP}{**}$$

$$\text{TPR: } PP / P* \quad \text{FPR: } PN / *N$$

$$\text{PPV: } PP / P*$$

$$\text{prevalence: } \frac{*P}{**}$$

"FAIRNESS-AWARE" CONFUSION MATRIX

A=0	Y=0	Y=1	A=1	Y=0	Y=1
$\hat{Y}=0$	UNN	UNP	$\hat{Y}=0$	PNN	PNP
$\hat{Y}=1$	UPN	OPP	$\hat{Y}=1$	PPN	PPP
"unprivileged"			"privileged"		

DEMOGRAPHIC PARITY:

$$\frac{UP*}{U**} - \frac{PP*}{P**} = 0$$

DISPARATE MISTREATMENT:

$$\frac{UNN + OPP}{U**} - \frac{PNN + PPP}{P**} = 0 \quad (\text{MISCLASS.})$$

Similar for FPR, FNR, etc.

$$OPP/UPP - PPP/PPP = 0$$

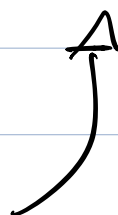
EQUALIZED ODDS:

$$\frac{UPN}{U*N} - \frac{PPN}{P*N} = 0$$

$$\frac{OPP}{U*P} - \frac{PPP}{P*P} = 0$$

" $\hat{Y} \perp A \mid Y$ "

EQUAL OPPORTUNITY: just



WHAT'S THE PROBLEM W. DEMOGRAPHIC PARITY?

$$\frac{U * P}{U * \pi} - \frac{P * P}{P * \pi} = 0$$

WHAT'S THE PROBLEM W. EQUALITY OF OPP.?

$$\text{Calibration: } P(Y=1 | \hat{Y}=1, A=0) = P(Y=1 | \hat{Y}=1, A=1)$$

WHAT'S THE PROBLEM W. DISPARATE MISTREATMENT?

- Impossibility theorem:

If prevalence (p) is different between two groups, you can't have calibration and equal FPR + FNR rates.

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