Fairness in ML

$\hat{Y}$ ← prediction of the ML classifier.

$$X , Y , \hat{Y}$$

$$\boxed{\text{Accuracy} \equiv P(Y = \hat{Y})}$$

Test data

X_1	Y_1	$\hat{Y}_1$
X_2	Y_2	$\hat{Y}_2$
X_3	Y_3	$\hat{Y}_3$
$\vdots$	$\vdots$	$\vdots$
X_{10}	Y_{10}	$\hat{Y}_{10}$

$$\frac{\#(Y = \hat{Y})}{10}$$

$$P(\hat{Y} = 1 \mid Y = 1)$$

True Positive Rate.

X	Y	$\hat{Y}$
x_1	1	1
x_2	0	1
x_3	0	0
x_4	1	0
x_5	1	1
x_6	1	1
x_7	1	0
x_8	1	1
x_9	0	0
x_{10}	1	1

$$\text{Accuracy} \equiv P(\hat{y} = y) = \frac{7}{10} = \underline{\underline{0.7}}$$

$$\text{TPR} = P(\hat{y} = 1 \mid \underline{y=1}) = \frac{5}{7}$$

recall on +ve class

$$\text{FNR} = P(\hat{y} = 0 \mid y = 1) = \frac{2}{7}$$

$$\text{FPR} = P(\hat{y} = 1 \mid y = 0) = \frac{1}{3}$$

$$TNR = P(\hat{y}=0 \mid y=0) = \frac{2}{3}$$

recall on -ve class

$$P(y=1 \mid \hat{y}=1) = \frac{5}{6}$$

Precision (on +ve class)

$$P(y=0 \mid \hat{y}=0) = \frac{2}{4}$$

Precision on -ve class

---

f-measure for +ve class

$$= \frac{2}{\frac{1}{\text{recall}_{+ve}} + \frac{1}{\text{precision}_{+ve}}}$$

$$\mathbb{E}[y \mid \mathbb{X}=x]$$

$t=0.5$

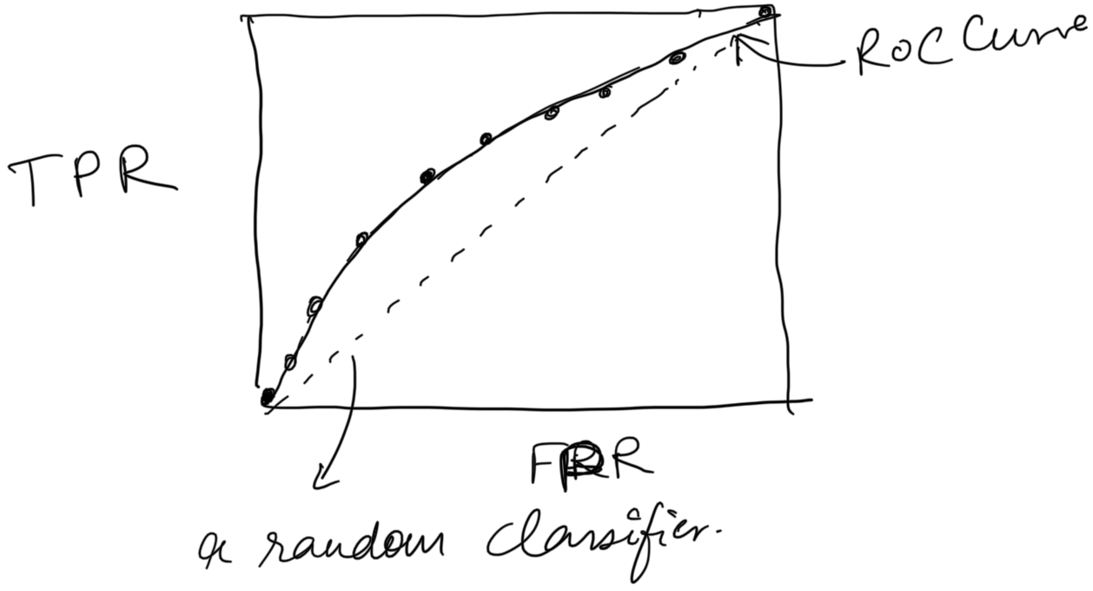
X	Y	R	$\hat{y}$
x_1	1	0.80	1
x_2	0	0.52	1

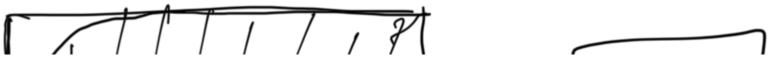
x_3	0	0.47	0
x_4	1	0.60	1
x_5	1	0.65	1
x_6	1	0.39	0
x_7	1	0.49	0
x_8	1	0.80	1
x_9	0	0.05	0
x_{10}	1	0.40	0

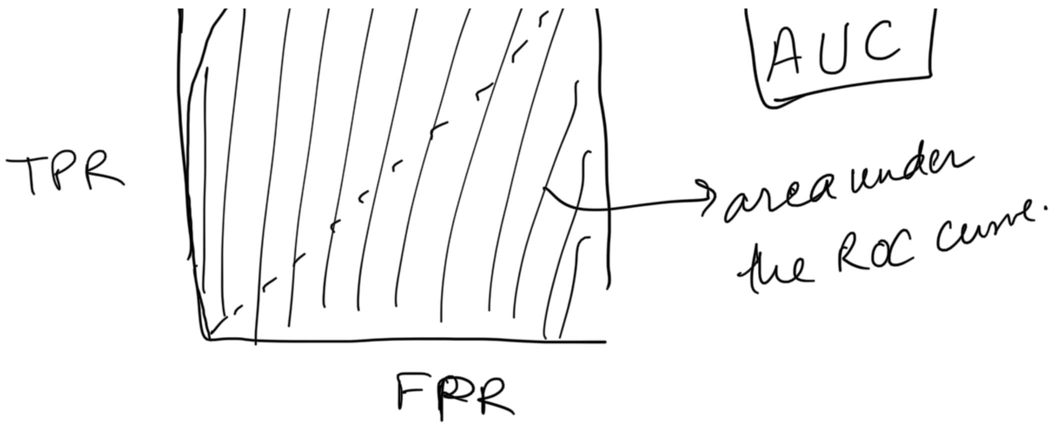
Receiver Operating Characteristic Curve (ROC)

$t=0$ FPR (+ve class) $P(\hat{Y}=1|y=0)$

TPR (+ve class) $P(\hat{Y}=1|y=1)$







Sensitive attribute A [A- binary]

$\perp\!\!\!\perp$ → independence

$A \perp\!\!\!\perp B \qquad P(A, B) = P(A)\,P(B)$

$A \perp\!\!\!\perp B \mid C \qquad P(A, B \mid C) = P(A \mid C)\,P(B \mid C)$

Independence

A classifier is fair if:

$$\underline{P(\hat{Y}=1 \mid A=a) = P(\hat{Y}=1 \mid A=b)}$$