Fairness in Machine Learning

Naive Bayes Classifier:

$$P(Y=y \mid X=x) = \frac{P(X=x \mid Y=y)\, P(Y=y)}{\sum_{y'}}$$

Both X and Y are random variables.

Logistic Regression

$$P(Y=y \mid x) = \text{Ber}(\sigma(w^T x))$$

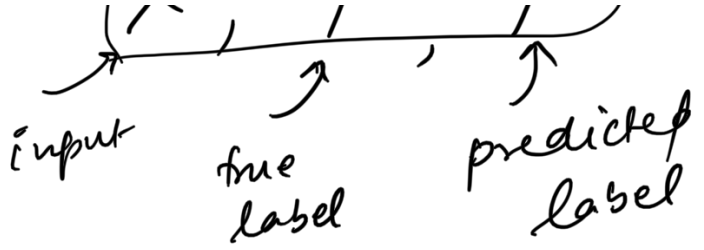
Only Y is a random variable

Generative classifiers $P(Y=y \mid X)$

vs.

Discriminative classifiers $P(y=y \mid x)$

$X \quad Y \quad \hat{Y}$



ignore, how to get $\hat{y}$ from x

Measuring performance:

x_1	y_1	$\hat{y}_1$
x_2	y_2	$\hat{y}_2$
⋮	⋮	⋮
x_N	y_N	$\hat{y}_N$

$x, y, \hat{y}$

$$\frac{\sum_{i=1}^{N} \mathbb{I}\left[y_i = \hat{y}_i\right]}{N} = \text{Accuracy}$$

$$\text{Accuracy} = P(y = \hat{y})$$

Accuracy is not always a good metric

X	Y	$\hat{Y}$
x_1	1	1
x_2	1	1
⋮	⋮	⋮
x_{90}	1	1
x_{91}	0	1
x_{92}	0	1
⋮	⋮	⋮
x_{100}	0	1

accuracy = 0.9

class imbalance

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

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

$Y \in \{0, 1\}$

give a loan - 1 - +ve

not give - 0 - -ve
a loan

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

$$P(Y = \hat{Y}) = \frac{7}{10} = 0.7$$

x_9	1	1
x_{10}	0	0

$$P(\hat{y}=1 \mid y=1) = \frac{3}{5}$$ True Positive Rate — recall +ve class

$$P(\hat{y}=0 \mid y=0) = \frac{4}{5}$$ True negative rate — recall -ve class

---

$$P(\hat{y}=0 \mid y=1) = \frac{2}{5}$$ False negative rate

$$P(\hat{y}=1 \mid y=0) = \frac{1}{5}$$ False positive rate

$$P(y=1 \mid \hat{y}=1) = \frac{3}{4}$$ Precision +ve class

Logistic regression

$$r = \sigma(w^T x) = \frac{1}{1+\exp(-w^T x)}$$

If $r \geq 0.5$ $\quad y = 1$

$\qquad\qquad\qquad y = 0$

$P(Y=1 \mid \_) = r$

SVM

$r = w^T x$

If $r \geq 0$ (t) $y = 1$

$y = 0$

$r(x) = \mathbb{E}[Y \mid X = x]$

R - random variable

How to construct an ROC curve.

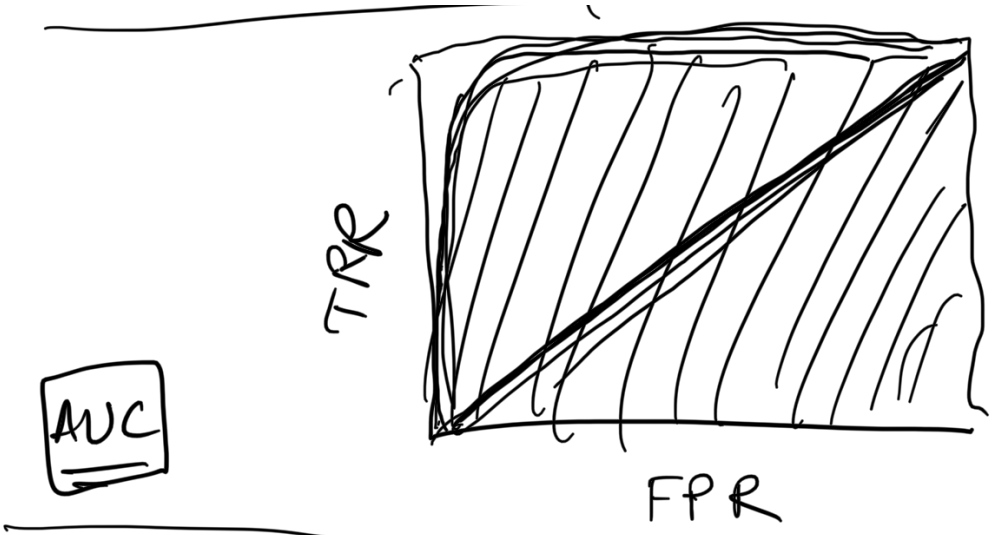
x	y	R	$\hat{Y}(t)$
x_1	1	0.7	
x_2	1	0.4	
x_3	0	0.3	
x_4	0	0.45	
x_5	0	0.6	
x_6	1	0.8	
x_7	1	0.9	
x_8	1	0.6	
x_9	0	0.7	
x_{10}	1	0.6	

$0 \leq t \leq 1$

Choose $t = 0$

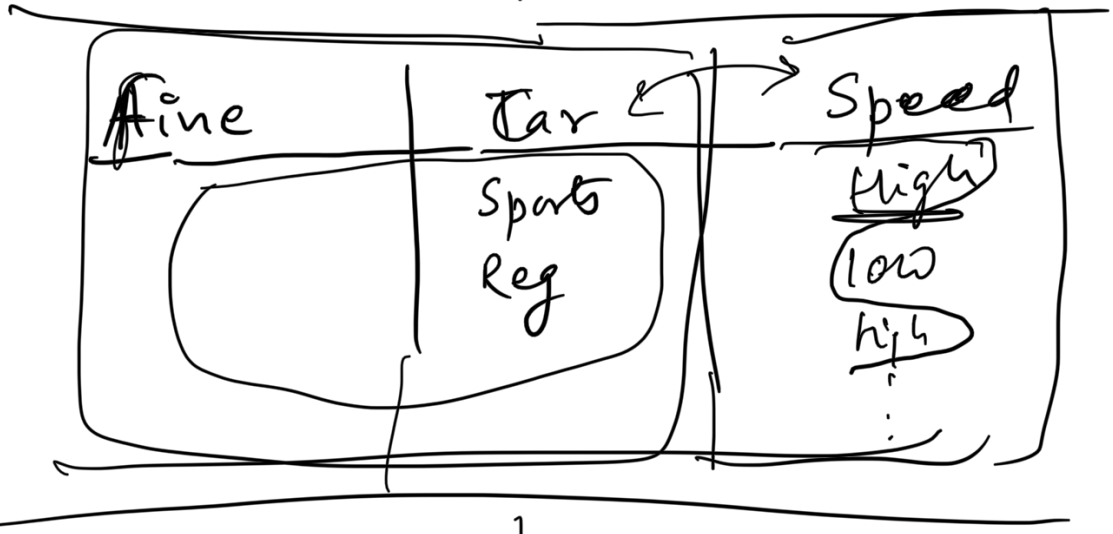
TPR, FPR

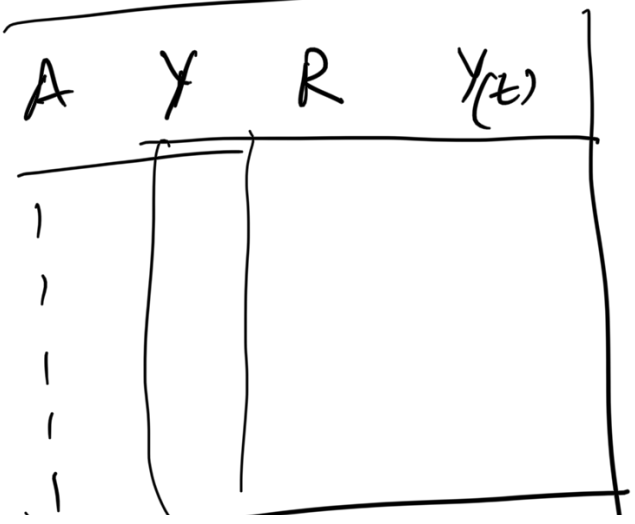
t	TPR	FPR
0		

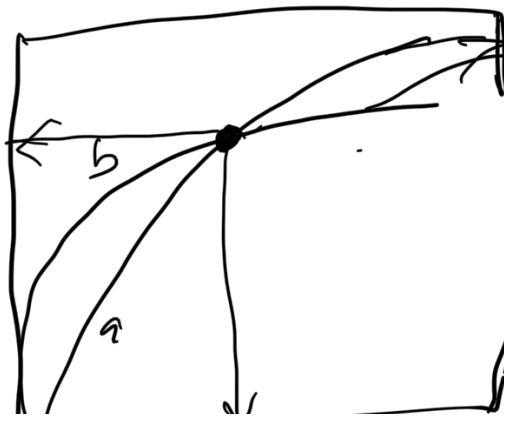


Sensitive Attribute A

$A \cap X = \phi$







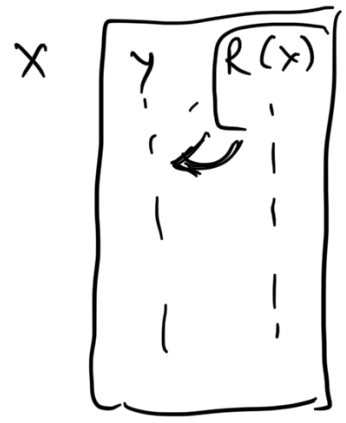
Calibration of scores

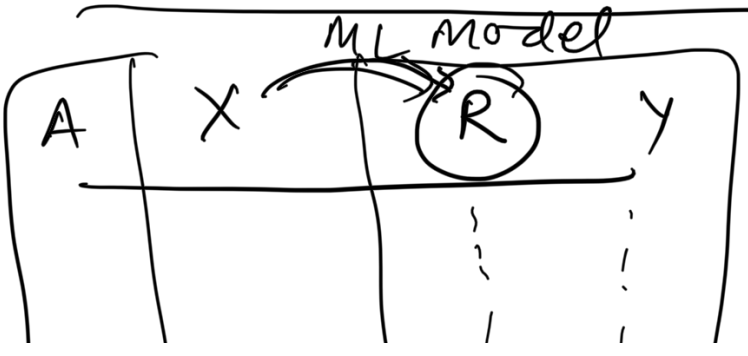
$R(x) \rightarrow$ score

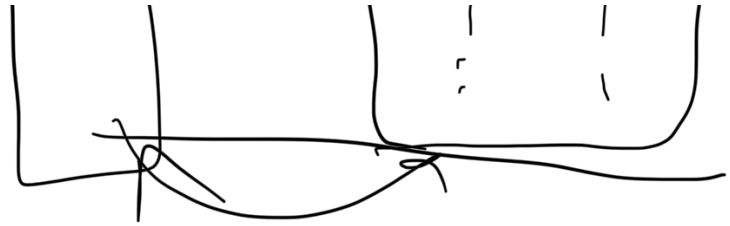
$$P(Y=1 \mid R=r) = r$$

$w^T x$

Platt Scaling







Strategy to convert $R \rightarrow \hat{y}$

~~tpr~~ $tp \rightarrow$ # applicants given a loan who repaid back

$fp \rightarrow$ # applicants given a loan who did not pay back

Profit - $\boxed{x\,tp - 6x\,fp}$

$$x\,(tp - 6\,fp)$$