Binary Classification

Stat 432

Dalpiaz

(x_i, y_i) FROM → VALIDATION → EVAL → MODEL FIT TO ESTIMATION

OR

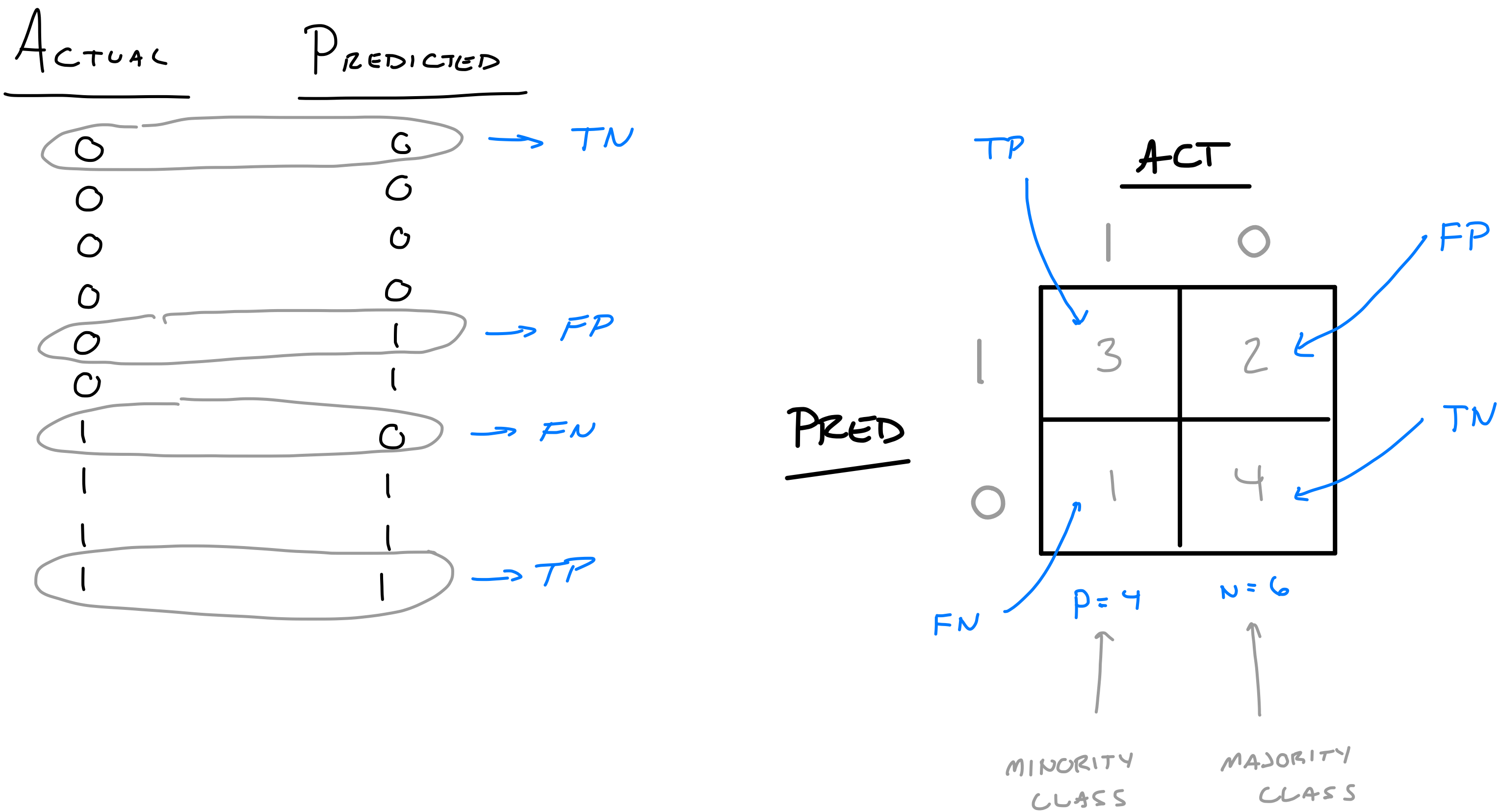
TEST → REPORT → MODEL FIT TO TRAIN

y_i → ACTUAL	$C(x_i)$ → PREDICTED	
0	0	
0	0	
0	0	
0	0	
0	1	→ "POSITIVE"
0	1	
1	0	→ "NEGATIVE"
1	1	
1	1	
1	1	

CONFUSION MATRIX

PRED \ ACT	1	0
1	III	II
0	I	IIII

POSITIONS OF ACT / PRED AND 0/1 COULD CHANGE!!!



ACTUAL
PREDICTED
0 0
0 0
0 0
0 0
0 1
0 1
1 0
1 1
1 1
1 1
TN
FP
FN
TP
TP
ACT
1
0
FP
1
3
2
PRED
TN
0
1
4
FN
P = 4
N = 6
MINORITY CLASS
MAJORITY CLASS

		ACT	
		1	0
PRED	1	3	2
	0	1	4

TP FP FN TN

$P = 4$ $N = 6$

$$\text{Prev} = \frac{P}{P+N} = \frac{4}{4+6} = 0.4$$

$$\text{Sens} = \frac{TP}{P} = \frac{3}{4} = 0.75$$

$$\text{Spec} = \frac{TN}{N} = \frac{4}{6} = 0.66\overline{6}$$

$$\text{PPV} = \frac{TP}{TP+FP} = \frac{3}{3+2} = 0.6$$

$$\text{FDR} = \frac{FP}{FP+TP} = \frac{2}{2+3} = 0.4$$

1 − misclass

$$* \text{Acc} = \frac{TP+TN}{P+N} = \frac{3+4}{4+6} = 0.7$$

And many more... (F_1, MCC, ...)

See book / wikipedia

No information Rate → PROPORTION OF MAJORITY CLASS

$$NIR = \max\left\{ \underbrace{\frac{P}{P+N}}_{\text{PREV}} , \underbrace{\frac{N}{P+N}}_{1-\text{PREV}} \right\}$$

IF ACC > NIR → REASONABLE CLASSIFIER

IF ACC < NIR → USELESS CLASSIFIER

$0.7 > 0.6$ ✓

$Y = 0$ OR $Y = 1$

$$p(x) \triangleq P[Y = 1 \mid X = x]$$

$$1 - p(x) = P[Y = 0 \mid X = x]$$

$P[Y=1 \mid X=x] \geq P[Y=0 \mid X=x]$

$$C^B(x) = \begin{cases} 1 & p(x) \geq 0.5 \\ 0 & p(x) < 0.5 \end{cases}$$

ESTIMATE $p(x)$ USING

- KNN
- TREE
- LOGISTIC
- NAIVE BAYES

$\hat{p}(x)$

"CUTOFF"

SET $0 \le \alpha \le 1$

$$C_\alpha(x) = \begin{cases} 1 & \hat{p}(x) \ge \alpha \\ 0 & \hat{p}(x) < \alpha \end{cases}$$

ASSUMED $\alpha = 0.5$

AS $\alpha \uparrow$ HARDER TO CLASSIFY AS $Y=1$

$$\text{CLASSIFIER} = f(\hat{p}(x), \alpha)$$

y_i	$\hat{p}(x_i)$	$C_{0.0}(x_i)$	$C_{0.25}(x_i)$	$C_{0.5}(x_i)$	$C_{0.75}(x_i)$	$C_{1.0}(x_i)$
0	0.1	1	0	0	0	0
0	0.1	1	0	0	0	0
0	0.2	1	0	0	0	0
0	0.3	1	1	0	0	0
0	0.6	1	1	1	0	0
0	0.7	1	1	1	0	0
1	0.4	1	1	0	0	0
1	0.7	1	1	1	0	0
1	0.8	1	1	1	1	0
1	0.9	1	1	1	1	0

TP/P =	SENS =	1.00	1.00	0.75	0.50	0.00
TN/N =	SPEC =	0.00	0.50	0.66	1.00	1.00
	ACC =	0.40	0.70	0.70	0.80	0.60

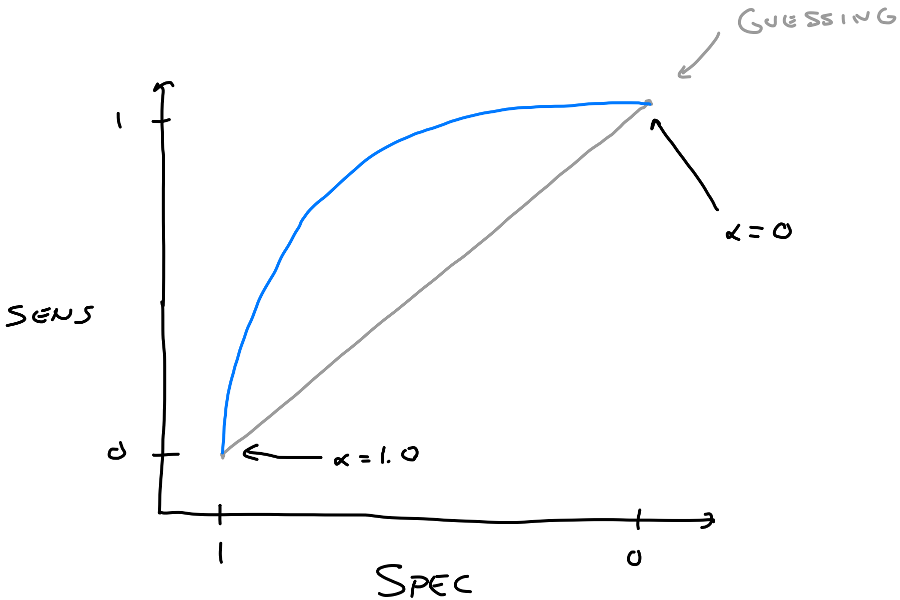
EXPECTED TO BE BEST

Evaluate $\hat{p}(x)$ instead of $C(x)$? ROC Curve!

Inputs

y

$\hat{p}(x)$



AUC

- Bigger = Better
- 1 = Perfect
- 0.50 = "Worst"