Math 370 Lec 26 5/11/18

Logistic regression gives $\hat{p}$'s. How to validate?

"The" $\hat{p}$ is $f_{pr}(\vec{x})$ but you don't get to see it.

You only get y's which are $\in \{y=0,1\}$.

You need a scoring rule $S(\hat{p}, y)$.

A proper scoring rule is:

$\forall i \quad f_{pr}(\vec{x}_i) = \text{argmax}\{S(\hat{p}, y)\}$ i.e. maximal if you use the truth.

two proper rules:

- Log scoring rule:

let $s_i = y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i)$

$$S = \frac{1}{n}\sum_{i=1}^{n} s_i$$

- Brier score (1950)

let $s_i = -(y_i - \hat{p}_i)^2$, $S = \frac{1}{n}\sum s_i$

these scores are used as an "R^2", a means to evaluate models but don't have an interpretation like "% variance explained."

Using prob. classifiers to classify. Need a way to turn

$\hat{p} \rightarrow \hat{y}$. $\hat{p} = 90\% \Rightarrow \hat{y} = 1$, $\hat{p} = 2\% \Rightarrow \hat{y} = 0$

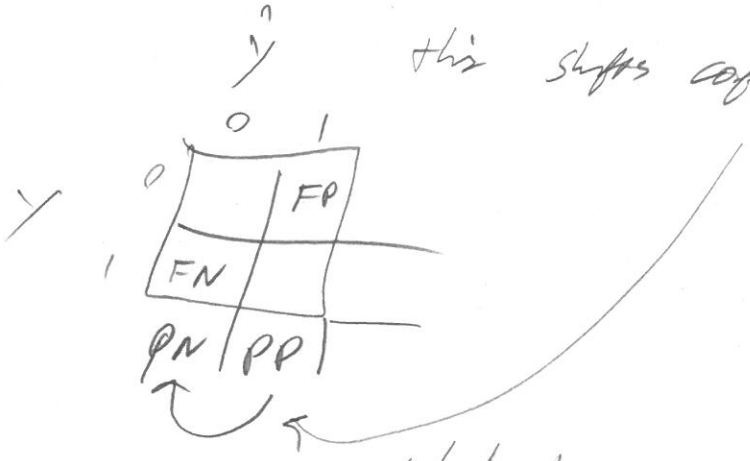
Naive:

$\hat{y} = \mathbb{1}_{\hat{p} \geq 50\%}$ HARD threshold at 50%.

What if ...

$\hat{y} = \mathbb{1}_{\hat{p} \geq 90\%}$? Makes it difficult to predict $\hat{y} = 1$.

You want this if $C_{FP} > C_{FN}$!



Generally ...

$\hat{y} = \mathbb{1}_{\hat{p} \geq \rho_{th}}$ ← tuning parameter based on asymmetric costs. Each value gives a different classifier. $\rho_{th} \in (0,1)$ not inclusive.

Which one to choose?

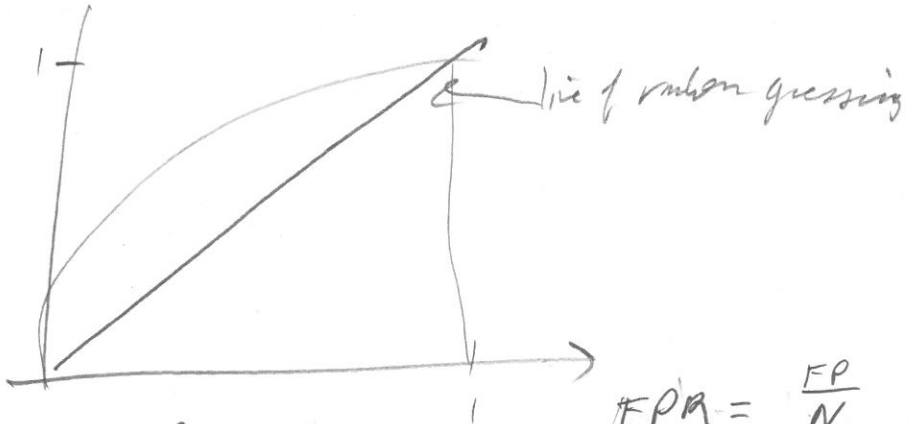
Look at all of them and choose!

$C_{FP} FP + C_{FN} FN$

p_{th}	TP	TN	FP	FN	precision	recall	FDR	FOR	other
0.01									
0.02									
0.03									
...									
0.98									
0.99									

all can be used

then the following curve is used



Recall

In null model...

why? 1 $\hat{y} = \mathbb{1}_{u \geq p_{th}}$ where $u \sim U(0,1)$

AUC = 50% of null prob. classifier

Ideal: AUC = 1

AUC > 1 ⇒ better than chance.

y \ $\hat{y}$	0	1	
0	$n(1-p_{th})(1-p_y)$	$n \cdot p_{th}(1-p_y)$	$n(1-p_y)$
1	$n(1-p_{th}) \cdot p_y$	$n \cdot p_{th} \cdot p_y$	$n p_y$
	$n(1-p_{th})$	$n p_{th}$	n

$$\text{recall} = \frac{n p_{th} p_y}{n p_y} = p_{th}$$

$$FPR = \frac{n p_{th}(1-p_y)}{n(1-p_y)} = p_{th}$$