02.02.2018

CLASSIFICATION

Bayes classifier: $P(Y = k \mid X = x)$

$k = 1, \dots, K$

covariates
explanatory variables

training data → fit method to produce
x, Y

$\hat{P}(Y = k \mid X = x)$

- LDA
- logistic regression
- ⋮

test data

Test data: true class and $\hat{P}(Y = k \mid X = x)$
x, Y

how to use this to classify?

confusion matrix ← choose k with $\hat{P}(Y = k \mid X = x)$ max over $1, \dots, K$

true \ pred	1	2	…	K
1	o			
2		o		
⋮			o	
K				o

misclassification rate

to evaluate the goodness of our method

Special case of $K = 2$ →

÷ nondisease $Y = 0$

+ disease $Y = 1$

		predict		total
		$-$	$+$	
true	$-$	TN	FP	N
	$+$	FN	TP	P
total		N^*	P^*	

truth

predictions

sensitivity $= \frac{TP}{P}$

specificity $= \frac{TN}{N}$

With $k=2$ then the max. class will have $\hat{P}(Y=k \mid X=x) > 0.5$

Bayes class. rule

1) Let $p(x) = \hat{P}(Y=1 \mid X=x)$ then we classify as disease (1) if $p(x) > 0.5$.

→ example (Afr. Heart disease)

sens = 0.625

spec = 0.778

2) But – is the cost of misclassification the same for both mistakes?

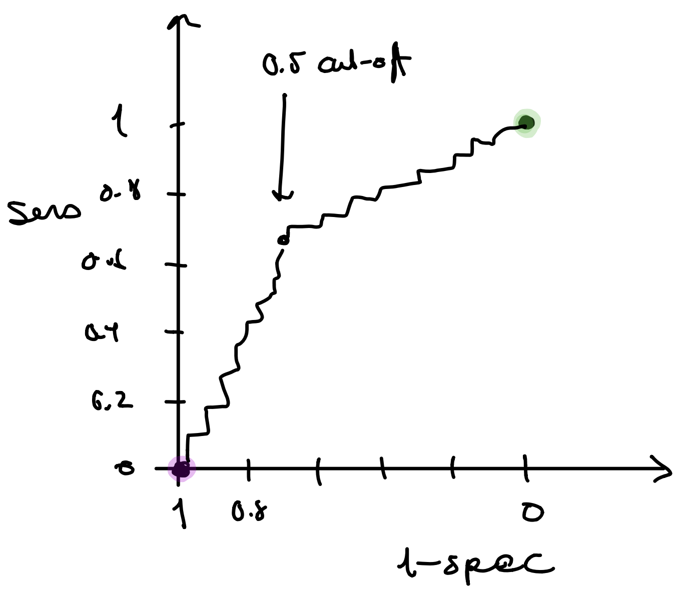
say + if $-$ Type I, false positive

say $-$ if + Type II, false negative

→ want to investigate different cut-offs on p(x) for classification.

p(x) > 0.1 → sens , spec
> 0.2 → ⋮ ⋮
⋮

⇒ plot (sens, 1-spec) = ROC curve



p(x) ≥ 0 ⇒ all Y=1

$sens = \frac{TP}{P} = \frac{P}{P} = 1$

$spec = \frac{TN}{N} = \frac{0}{N} = 0$

p(x) ≥ 1 all Y=0

$sens = \frac{TP}{P} = \frac{0}{P} = 0$

$spec = \frac{TN}{N} = \frac{N}{N} = 1$

Good if sens & spec both high

⇒ curve hug upper left corner

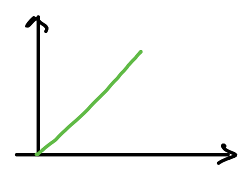
AUC = area under (ROC) curve

What if we just do "random guessing"?

* if 0.1 is used as cut-off, for each individual we draw $bin(1, 0.1) \rightarrow \frac{TP}{P} \approx 0.1$, $\frac{TN}{N} \approx 0.9$
* 0.2 $bin(1, 0.2) \rightarrow \frac{TP}{P} \approx 0.2$, $\frac{TN}{N} \approx 0.8$

This is similar to assign uniformly drawn p's $[0,1]$ to each observation

$\Rightarrow$ this will give a roc curve with AUC = 0.5



This is often used for comparison. An AUC ≈ 0.5 is thus <u>not</u> good.