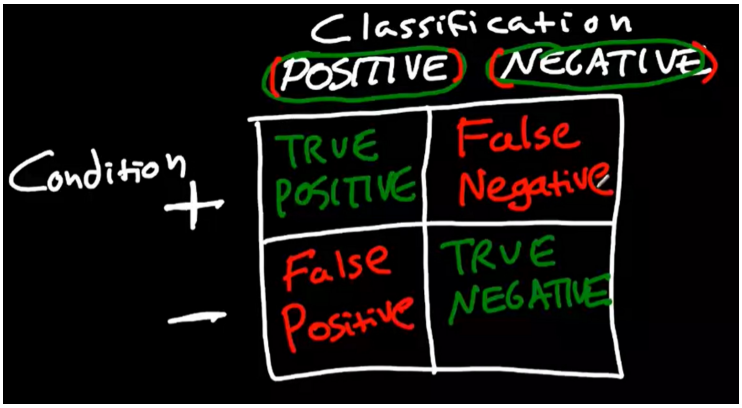
Week 2 Binary classification and confusion matrix

29 June 2020 18:02

Binary classification and confusion matrix

- binary outcomes and a binary classification -> confusion matrix



Confusion Matrix				
			Test Classification **Y**	
			"Positive"	"Negative"
			c	d
Condition **X**	"+"	a	e	f
	"-"	b	g	h

Individual Probabilities		**Name**
P("+")	a	Incidence of Condition "+"
p("-")	b	Incidence of "Condition "-"
p(Test POS)	c	Classification Incidence "POS"
p(Test NEG)	d	Classification Incidence "NEG"
p(Test POS, "+")	e	True Positives
p(Test NEG, "+")	f	False Negatives
p(Test "POS, "-")	g	False Positives
p(Test "NEG", "-")	h	True Negatives

Probability Distributions		**Name**
P(X)	p(a,b)	Marginal Probability of the Condition
P(Y)	p(c,d)	Marginal Probability of the Classification
p(X,Y)	p(e,f,g,h)	Joint Distribution of X and Y
P(X)p(Y)	p(ac,ad,bc,bd)	Product Distribution of X and Y
Conditional Probabilities		**Name**
p(Test POS | "+")	e/a	True Positive Rate
p(Test NEG | "+")	f/a	False Negative Rate
p(Test POS | "-")	g/b	False Positive Rate
p(Test NEG | "-")	h/b	True Negative Rate
p("+" | Test POS)	e/c	Positive Predictive Value (PPV)
p("-" | Test POS)	g/c	1- PPV
p("+" | Test NEG)	f/d	1- NPV
p("-" | Test NEG)	h/d	Negative Predictive Value (NPV)
Other implied rules		
a+b = 1	e+f = a	
c+d = 1	g+h = b	
e+f+g+h=1	e+g = c	
	f+h=d	

- ROC curve
  - x = false positive rate; y= true positive rate
    - FP rate: probability of FP conditional upon actually not sick
    - TP rate: probability of TP conditional upon actually sick
  - scatter plot with different threshold determined by cost of FN and FP
- Performance metrics for binary classification models: AUC(area under ROC curve)
  - between 1/2 - 1
    - 1/2: same as random
    - 1: perfect model