

BIG DIVE

TECH. CUSTOM EDITION

A project by **TOP-IX**
designed for **Intesa Sanpaolo**

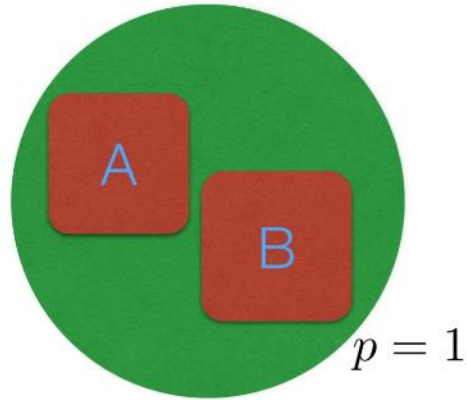


#DataScience



- Data-analytic thinking
- Data quality
- Descriptive statistics
- Correlation & causation
- Bias
- Regression
- Comparison between groups

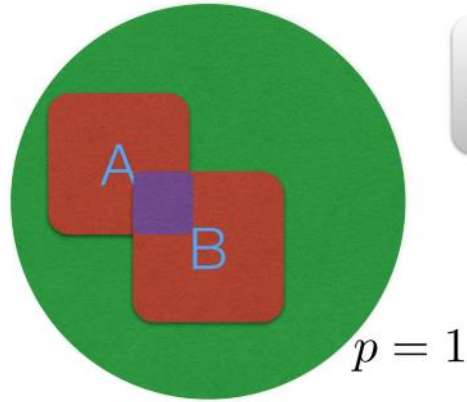
$$P(A) = \text{Area of } A$$



$$P(A \text{ or } B) = P(A) + P(B)$$

Probability –basic concepts

$$P(A) = \text{Area of } A$$

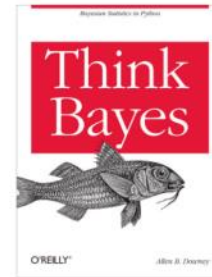


$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$
$$P(A \text{ and } B) = \text{overlap of } A \text{ and } B$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Bayes Theorem



Screening test



Your doctor thinks you might have a rare disease that affects **1 person in 10,000**. A test that is **99%** accurate comes out **positive**. What's the probability of you having the disease?

Bayes Theorem:
$$P(\text{disease}|\text{positive test}) = \frac{P(\text{positive test}|\text{disease}) P(\text{disease})}{P(\text{positive test})}$$

Total probability
$$P(\text{positive test}) = P(\text{positive test}|\text{disease}) P(\text{disease}) + P(\text{positive test}|\text{no disease}) P(\text{no disease})$$

Finally:
$$P(\text{disease}|\text{positive test}) = 0.0098$$

Screening test



Consider a population of 1,000,000 individuals. The numbers we should expect in the **contingency matrix** are:

Marginals
↓

	disease	no disease	
positive	99	9,999	10,098
negative	1	989,901	989,902
Marginals →	100	999,900	1,000,000

$$P(\text{disease}|\text{positive test}) = \frac{TP}{TP + FP} = 0.0098$$

$$P(\text{no disease}|\text{negative test}) = \frac{TN}{TN + FN} = 0.99999$$

Screening test

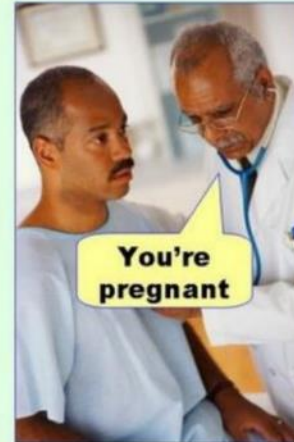
Consider a population of 1,000,000 individuals. The numbers we should expect in the **contingency matrix** are:

	disease	no disease	Marginals
positive	99	9,999	10,098
negative	1	989,901	989,902
Marginals	100	999,900	1,000,000

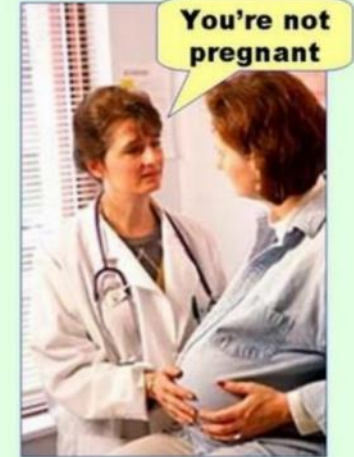
$$P(\text{disease}|\text{positive test}) = \frac{TP}{TP + FP} = 0.0098$$

$$P(\text{no disease}|\text{negative test}) = \frac{TN}{TN + FN} = 0.99999$$

Type I error
(false positive)



Type II error
(false negative)



Consider a second screening



Bayes Theorem still looks the same: $P(\text{disease}|\text{positive test}) = \frac{P(\text{positive test}|\text{disease}) P(\text{disease})}{P(\text{positive test})}$

but now the probability that we have the disease has been updated: $P^{\dagger}(\text{disease}) = 0.0098$

So this time we find: $P^{\dagger}(\text{disease}|\text{positive test}) = 0.4949$

Each test is providing **new evidence**, and Bayes theorem is simply telling us how to use it to **update our beliefs**. 

Confusion Matrix

Feature Test	positive	negative
positive	TP	FP
negative	FN	TN

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$specificity = \frac{TN}{FP + TN}$$

$$precision = \frac{TP}{TP + FP}$$

$$sensitivity = \frac{TP}{TP + FN}$$

harmonic mean

$$F1 = \frac{2TP}{2TP + FP + FN}$$

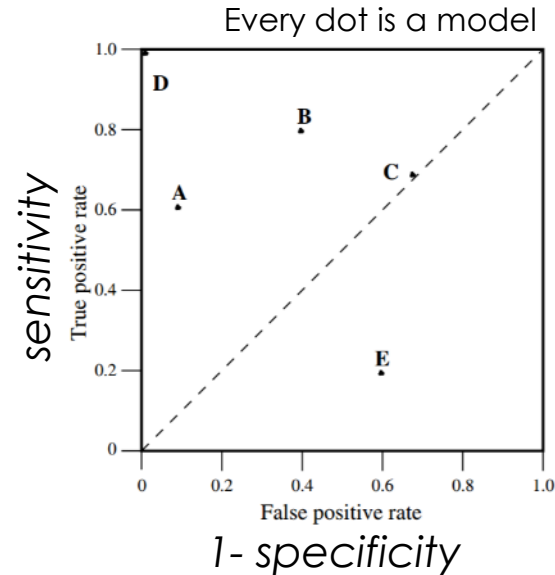
Confusion Matrix – ROC curve

Feature Test	positive	negative
positive	TP	FP
negative	FN	TN



$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specificity} = \frac{TN}{FP + TN}$$

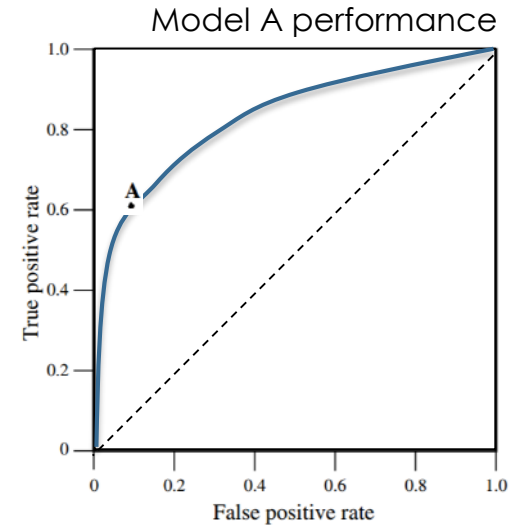
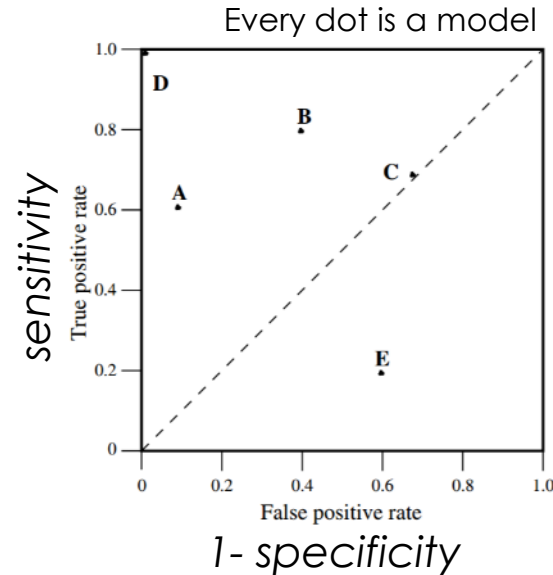


Confusion Matrix – ROC curve

Feature Test \	positive	negative
positive	TP	FP
negative	FN	TN

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specificity} = \frac{TN}{FP + TN}$$

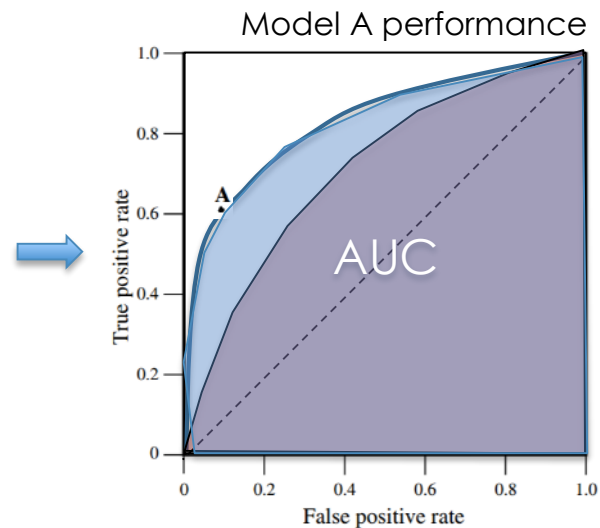
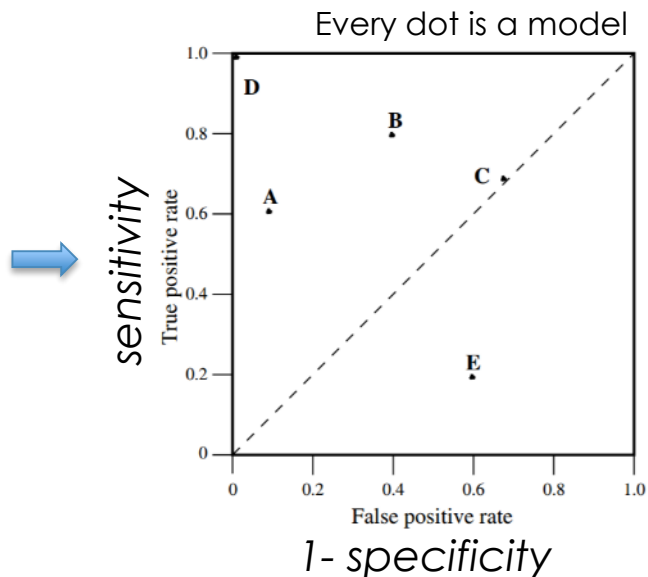



Confusion Matrix – ROC curve

Feature Test \	positive	negative
positive	TP	FP
negative	FN	TN

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specificity} = \frac{TN}{FP + TN}$$





An introduction to ROC analysis

Tom Fawcett

Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306, USA

Available online 19 December 2005

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Q & A



- Regressioni (per predizioni) → Modulo Machine Learning
- Riduzione dimensionalità (PCA, ICA...)
- Analisi di serie storiche
- Analisi di sopravvivenza
- Network Bayesiani
-