BIG DIVE

TECH. CUSTOM EDITION

A project by TOP-IX
designed for Intesa Sanpaolo

#DataScience

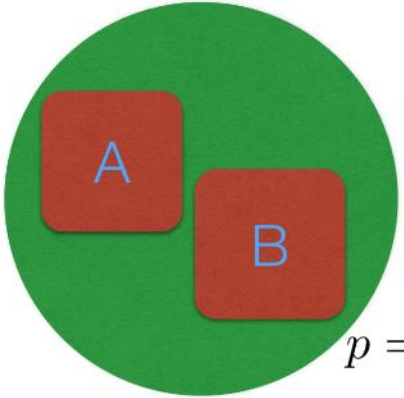
aizoon® AUSTRALIA EUROPE USA
TECHNOLOGY CONSULTING

Recap

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

Probability –basic concepts

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

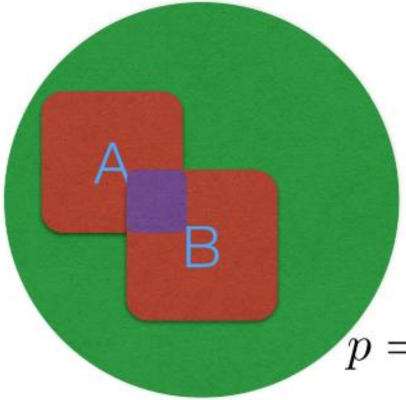


$$p = 1$$

$$P(A\ or\ B) = P(A) + P(B)$$

Probability –basic concepts

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



$$p = 1$$

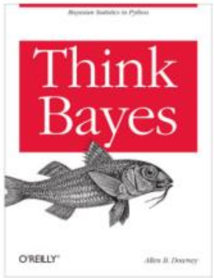
$$P(B|A) = \frac{P(A \; and \; B)}{P(A)}$$

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

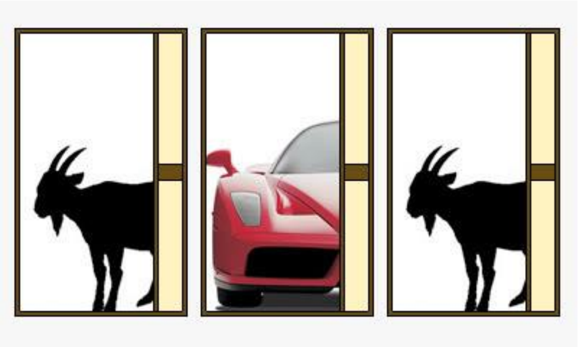
Bayes Theorem

$$P(A \; or \; B) = P(A) + P(B) - P(A \; and \; B)$$

$$P(A \; and \; B) = \text{overlap of A and B}$$



Monty Hall's problem



3 Doors, 1 car, 2 goats:

- $P(C1)=P(C2)=P(C3)=1/3$

Lets choose D1, and focus on Door 3.

Monty would open D3 with the following probabilities:

- $P(D3|C1)=1/2$
- $P(D3|C2)=1$
- $P(D3|C3)=0$

The Bayes part:

- $P(C1|D3)=P(D3|C1)P(C1)/P(D3)= (1/2*1/3)/1/2=1/3$
- $P(C2|D3)=P(D3|C2)P(C2)/P(D3)= (1*1/3)/1/2=2/3$

So, it's better to change door!

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:

$$P\left(disease|positive\ test\right) = \frac{P\left(positive\ test|disease\right) P\left(disease\right)}{P\left(positive\ test\right)}$$

Total probability

$$\begin{aligned} P\left(positive\ test\right) = & \ P\left(positive\ test|disease\right) P\left(disease\right) \\ & + P\left(positive\ test|no\ disease\right) P\left(no\ disease\right) \end{aligned}$$

Finally:

$$P\left(disease|positive\ test\right) \boxed{= 0.0098}$$

Screening test

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

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

Marginals

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

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

Screening test

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

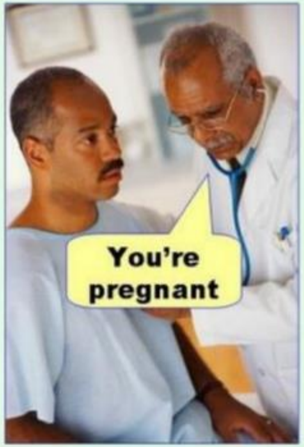
Marginals

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

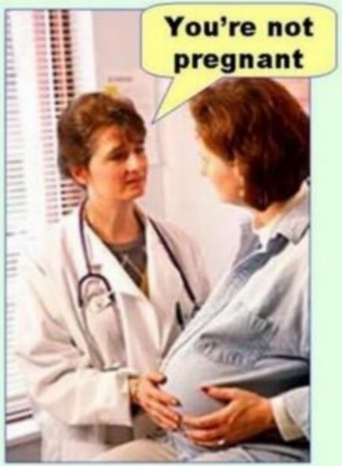
$$P\left(disease|positive\ test\right) = \frac{TP}{TP+FP} = 0.0098$$

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

Type I error
(false positive)



Type II error
(false negative)



Consider a second screening

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

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

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

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

Confusion Matrix

Test \ Feature	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

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

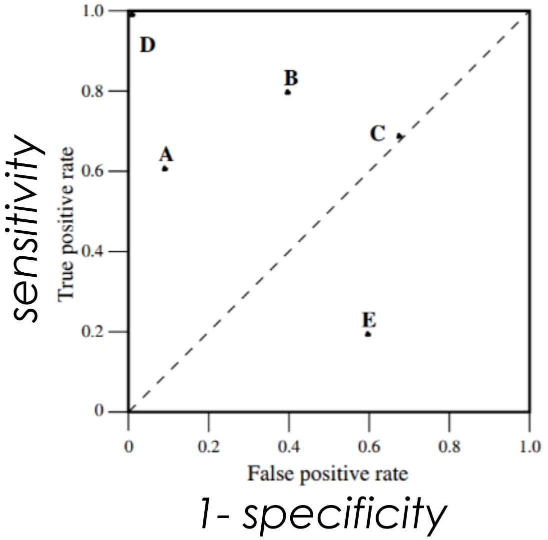
Confusion Matrix – ROC curve

Test \ Feature	positive	negative
positive	TP	FP
negative	FN	TN

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

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

Every dot is a model



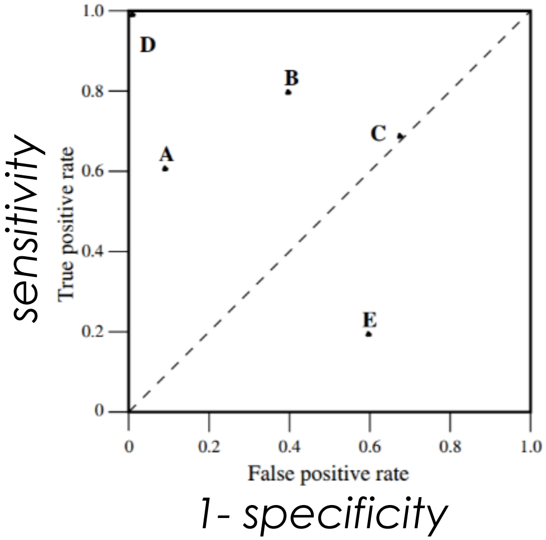
Confusion Matrix – ROC curve

Feature / Test	positive	negative
positive	TP	FP
negative	FN	TN

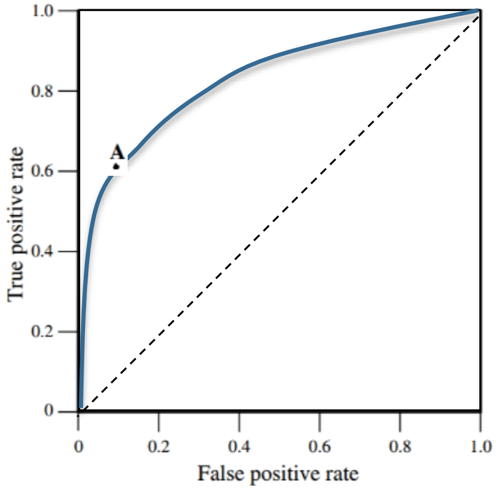
$$sensitivity = \frac{TP}{TP + FN}$$

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

Every dot is a model



Model A performance



Confusion Matrix – ROC curve

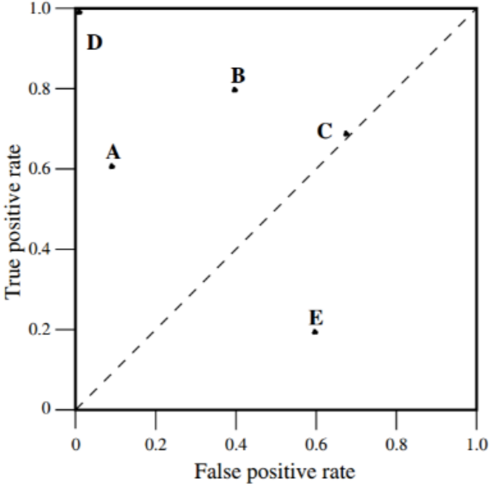
Test \ Feature	positive	negative
positive	TP	FP
negative	FN	TN

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

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

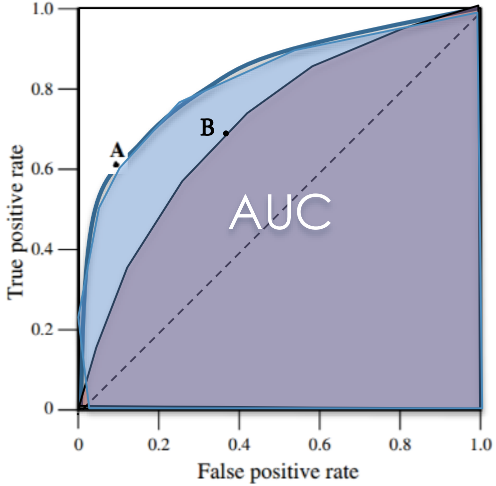
Every dot is a model

sensitivity



1- specificity

Model A performance



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

ELSEVIER

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Cosa NON abbiamo trattato

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

Q & A