Algorithms with Statistical Support

Paulo Trigo Silva

Assumptions from the "Statistical Modeling"

Recall the main assumption of the1R method:

"a single attribute is enough to learn a significant rule"

The 1R approach is "opposite" to the **statistical modeling** assumption:

"all attributes are *equally important* and *independent* of one another"

Paulo Trigo Silva

... two essential assumptions

- All attributes are *equally important*
  - i.e., an estimation must always account for the influence of all attributes
- The attributes are statistically independent (given the class)
  - i.e., the value of an attribute is not influenced by the value of any another (even assuming that the attributes' class is known)

Those assumptions do not seem realistic!

We know that *rarely* all attributes have the same importance.

We know that *rarely* the attributes are independent among themselves.

But, they lead to a simple scheme that works surprisingly well in practice!

Paulo Trigo Silva

An example (classic)

Weather data and the decision about playing (`Play`) tennis!
(dataset represents “my” decision under different weather conditions)

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

... and now, what is my the expected decision?

... the weather conditions for "today" are:

Am I to play tennis?

Outlook	Temperature	Humidity	Windy	Play
sunny	cool	high	true	?

Search the most likely decision (*likelihood* evaluation) by means of a statistical analysis, starting by doing "synthesis table" with:

- **the number of times that the value of each attribute (i.e., each pair attribute-value) is associated with each value (*yes* or *no*) of *Play***

example

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot			high			false				
overcast			mild			normal			true				
rainy			cool										

complete this table

Paulo Trigo Silva

... the synthesis table (the counting)

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								

Now, represent in the format of "frequencies", or "observed probabilities":

- #(Attribute= '*vAtr*' | Play = '*vClass*') / #(Play = '*vClass*')
- #(Play = '*vClass*') / Σ_v #(Play = '*v*') ,
- *vClass* $\in$ {*yes*, *no*}, *v* $\in$ {*yes*, *no*}

example

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2/9	3/5	hot	?		high	?		false	?		9/14	
overcast			mild			normal			true			?	
rainy	?		cool										

Paulo Trigo Silva

… synthesis table (counting and observed probabilities)

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

... and now, given the following weather conditions:

Outlook	Temperature	Humidity	Windy	Play
sunny	cool	high	true	?

What is the *likelihood* of each value, *yes* and *no*, for the class *Play*?

Paulo Trigo Silva

… synthesis table (observed probabilities) and likelihood

Outlook	yes	no	Temperature	yes	no	Humidity	yes	no	Windy	yes	no	Play yes	Play no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

Outlook	Temperature	Humidity	Windy	Play
sunny	cool	high	true	?

likelihood **("verosimilhança") of *yes* = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053**

likelihood **("verosimilhança") of *no* = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206**

... given those conditions is more likely (≈ 4 times more) NOT to play tennis!

Paulo Trigo Silva

Bayes Rule

likelihood (“verosimilhança”) of *yes* = $2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$

likelihood (“verosimilhança”) of *no* = $3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$

probability of *yes* = 0.0053 / (0.0053 + 0.0206) = 20.5%

probability of *no* = 0.0206 / (0.0053 + 0.0206) = 79.5%

This example illustrates the application of the Bayes rule of conditional probabilities, which states that:

given an hypothesis *H* and an evidence *E* that bears on that hypothesis, then:

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]}$$

Paulo Trigo Silva

The hypothesis and the evidence

Given an hypothesis *H* and an evidence *E* that bears on that hypothesis, then:

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]}$$

Which is *H* (in the previous example)?

1. Play = *yes* and Play = *no*
2.

Outlook	Temperature	Humidity	Windy
sunny	cool	high	true

3. Pr(Play = *yes*)

Which is *E* (in the previous example)?

1. Play = *yes* and Play = *no*
2.

Outlook	Temperature	Humidity	Windy
sunny	cool	high	true

3. Pr(Play = *yes*)

... hypothesis (H) and evidence (E)

$$\underbrace{\text{Pr (Play=}\textit{yes}}_{H}\ |\ \underbrace{\text{Outlook=}\textit{sunny}\text{, Temperature=}\textit{cool}\text{, Humidity=}\textit{high}\text{, Windy=}\textit{true}\text{)}}_{E}$$

Pr (Play=*no* | Outlook=*sunny*, Temperature=*cool*, Humidity=*high*, Windy=*true*)

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]}$$

Outlook	Temperature	Humidity	Windy
sunny	cool	high	true

E

Example*: probability of class *yes* given *E

Pr (Play=*yes* | *E*) = ?
?

Paulo Trigo Silva

... hypothesis (H) and evidence (E) – example

$$\underbrace{\text{Pr (Play=}\textit{yes}}_{H}\ |\ \underbrace{\text{Outlook=}\textit{sunny}\text{, Temperature=}\textit{cool}\text{, Humidity=}\textit{high}\text{, Windy=}\textit{true}\text{)}}_{E}$$

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]}$$

Outlook	Temperature	Humidity	Windy
sunny	cool	high	true

$\} E$

Pr (Play=*yes* | *E*) = Pr(Outlook=*sunny* | Play=*yes*) ×
Pr(Temperature=*cool* | Play=*yes*) ×
Pr(Humidity=*high* | Play=*yes*) ×
Pr(Windy=*true* | Play=*yes*) ×
Pr(Play=*yes*)
/
Pr(*E*)
= 2/9 × 3/9 × 3/9 × 3/9 × 9/14
/
Pr(*E*)

Example*: probability of class *yes* given *E

the final normalization, i.e., Σ_iPr(i)=1, enables to discard Pr(*E*);

details next...

Paulo Trigo Silva

... in decision making, discard Pr(*E*)

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]}$$

Pr (*yes* | *E*) = p1 / Pr(*E*), where p1 = Pr(E | *yes*) × Pr(*yes*)

Pr (*no* | *E*) = p2 / Pr(*E*), where p2 = Pr(E | *no*) × Pr(*no*)

According to the law of probability Σ_iPr(i)=1; therefore,

Pr(*yes* | *E*) + Pr(*no* | *E*) = 1, or, equivalently,

p1 / Pr(*E*) + p2 / Pr(*E*) = 1 ⇔

(p1 + p2) / Pr(*E*) =1 ⇔

Pr(*E*) = p1 + p2

That is, the probability of the evidence to occur, i.e., Pr(*E*), only serves a "normalization" purpose, i.e., it guarantees that the summation is 1.

Pr(*E*) does not influence the relation among the probabilities of the hypothesis values because it is the same (constant) for all the values.

Thus, Pr(*E*) can be *discarded when using the Bayes rule to support a decision*.

Synthesis: the various components of the Bayes rule

- Probability of an event (hypothesis) *H* given the evidence *E*
  - *à-posteriori* probability relative to *à-priori* from evidence *E* given *H*

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]}$$

Thomas Bayes
1702 – 1761
England

- ***à-priori*** probability of the *H*
  - probability of the event *H* ***before*** the evidence's perception

$$\Pr[H]$$

- ***à-posteriori*** probability of the *H* (conditional on the *E*)
  - probability of the event *H* ***after*** (or "given that") the evidence's perception

$$\Pr[H|E]$$

Paulo Trigo Silva

Another important concept: “conditional independence”

- Consider three variables
  - A, B, C
- The conditional distribution of A given B and C
  - is written Pr(A | B, C)
- **If** Pr(A | B, C) **does not depend** on the value of B
  - we have Pr(A | B, C) = Pr(A | C)
- … in that case we say that
  - A is **conditionally independent** of B given C

Factorizing into marginals

- If the conditional distribution of A given B and C is independent of B
  – we have Pr(A | B, C) = Pr(A | C)

- The joint distribution of A and B conditioned on C
  – is written Pr(A, B | C)

- … it can be expressed in slightly different way

Pr(A, B | C) =

= Pr(A, B, C) / Pr(C) // product rule Pr(X,Y)=Pr(X|Y)Pr(Y), with X = A,B

= Pr(A | B, C) Pr(B, C) / Pr(C) // product rule with X = A

= Pr(A | B, C) Pr(B | C) Pr(C) / Pr(C) // product rule with X = B and Y = C

= Pr(A | B, C) Pr(B | C) // just "cut" Pr(C)

= **Pr(A | C) Pr(B | C)** // **conditional independence**

- i.e., the joint distribution factorizes into product of marginals
  – when the variables A and B are statistically independent given C

Paulo Trigo Silva

An illustrative example – using three binary variables

Let variables A, B, C represent colors:

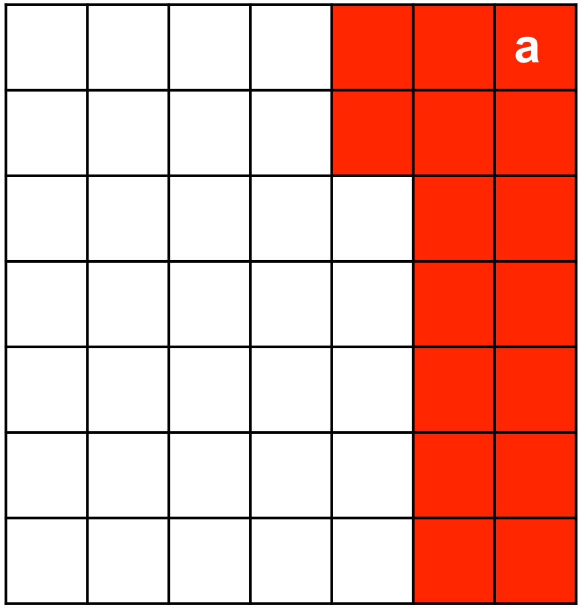
$A \in \{ a, \sim a \}$, where a is red and ~a is not red

$B \in \{ b, \sim b \}$, where b is blue and ~b is not blue

$C \in \{ c, \sim c \}$, where c is green and ~c is not green

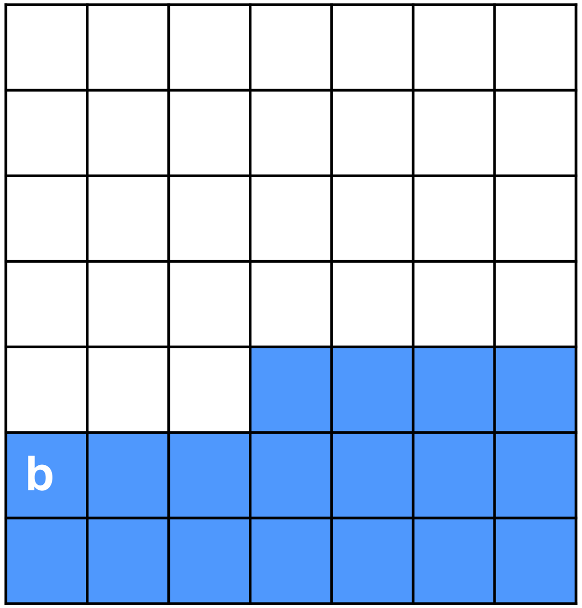
$P(A=a) = 16/49$

$P(A=\sim a) = 33/49$



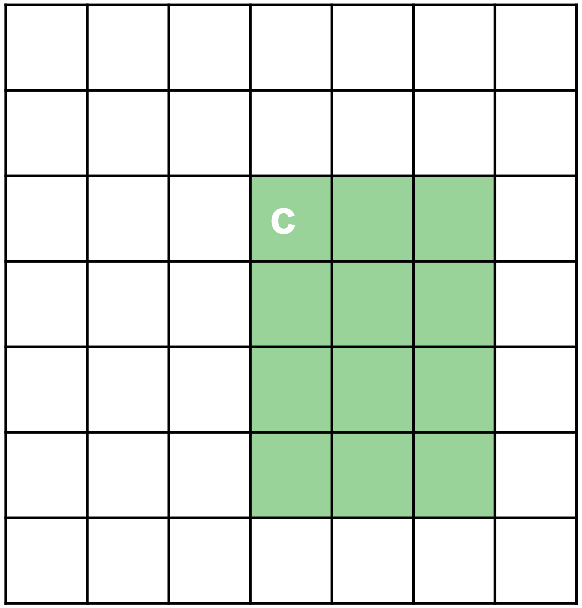
$P(B=b) = 18/49$

$P(B=\sim b) = 31/49$



$P(C=c) = 12/49$

$P(C=\sim c) = 37/49$

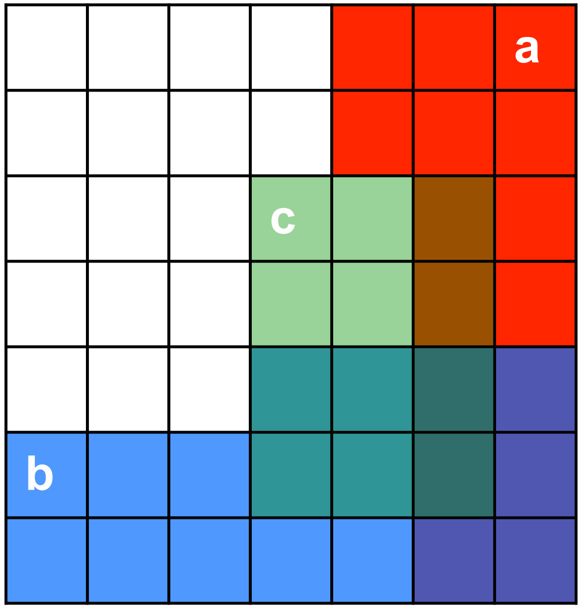


Paulo Trigo Silva

… mixing the three binary variables, and some notation

Let variables A, B, C represent colors:

$A \in \{ a, \sim a \}$
$B \in \{ b, \sim b \}$
$C \in \{ c, \sim c \}$



Some aspects of notation:

Let **X** be a random variable

if **X** is a discrete random variable it means that it may only take a countable number of distinct values, e.g., { 0, 1, 2 }

Pr(X) is the probability distribution of **X**, i.e., a list of probabilities associated with each of its possible values, e.g., P(X=0)=1/2, P(X=1)=1/4, P(X=2)=1/4

X=x is the event of **X** taking the value x, e.g., X=2

P(X=x) = p says that the event X=x occurs with the probability value p.

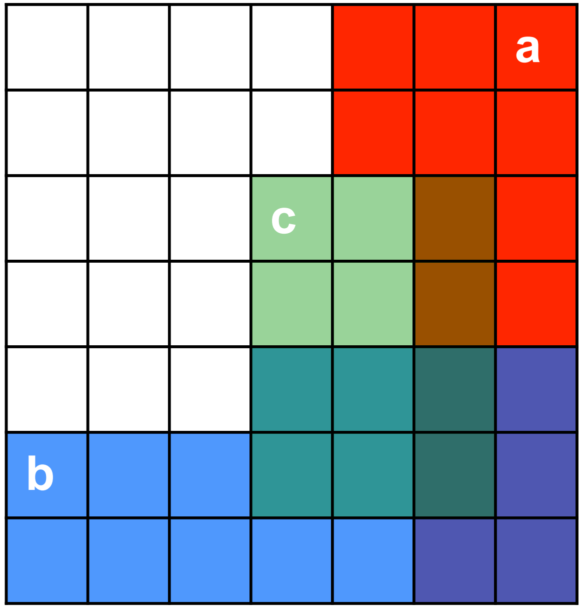
P(x) = p is a simplified way of writing P(X=x) = p

Paulo Trigo Silva

… illustrative example – marginal and joint distributions?

Let variables A, B, C represent colors:

A ∈ { a, ~a }
B ∈ { b, ~b }
C ∈ { c, ~c }



There are **26** marginal and joint probabilities in this problem!

Marginal probabilities (**6**):

Pr(A):	P(A=a) = ?	P(~a) = ?
Pr(B):	P(b) = ?	P(~b) = ?
Pr(C):	P(c) = ?	P(~c) = ?

Joint probabilities (**12** with 2 variables):

Pr(A,B):	P(A=a,B=b) = ?	P(a,~b) = ?
	P(~a,b) = ?	P(~a,~b) = ?
Pr(B,C):	P(b,c) = ?	P(b,~c) = ?
	P(~b,c) = ?	P(~b,~c) = ?
Pr(A,C):	P(a,c) = ?	P(a,~c) = ?
	P(~a,c) = ?	P(~a,~c) = ?

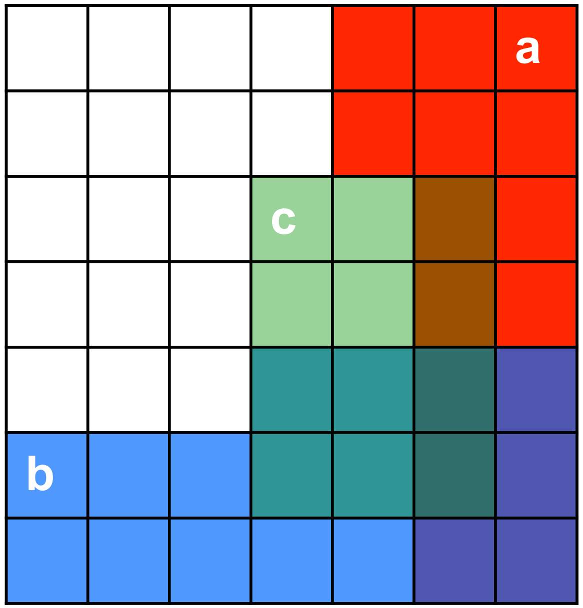
Joint probabilities (**8** with 3 variables):

Pr(A,B,C):	P(a,b,c) = ?	P(a,b,~c) = ?
	P(a,~b,c) = ?	P(a,~b,~c) = ?
	P(~a,b,c) = ?	P(~a,b,~c) = ?
	P(~a,~b,c) = ?	P(~a,~b,~c) = ?

… illustrative example – marginal and joint distributions

Let variables A, B, C represent colors:

$A \in \{ a, \sim a \}$
$B \in \{ b, \sim b \}$
$C \in \{ c, \sim c \}$



There are **26** marginal and joint probabilities in this problem!

Marginal probabilities (**6**):

Pr(A):	P(A=a) = 16/49	P(~a) = 33/49
Pr(B):	P(b) = 18/49	P(~b) = 31/49
Pr(C):	P(c) = 12/49	P(~c) = 37/49

Joint probabilities (**12** with 2 variables):

Pr(A,B):	P(A=a,B=b) = 6/49	P(a,~b) = 10/49
	P(~a,b) = 12/49	P(~a,~b) = 21/49
Pr(B,C):	P(b,c) = 6/49	P(b,~c) = 12/49
	P(~b,c) = 6/49	P(~b,~c) = 25/49
Pr(A,C):	P(a,c) = 4/49	P(a,~c) = 12/49
	P(~a,c) = 8/49	P(~a,~c) = 25/49

Joint probabilities (**8** with 3 variables):

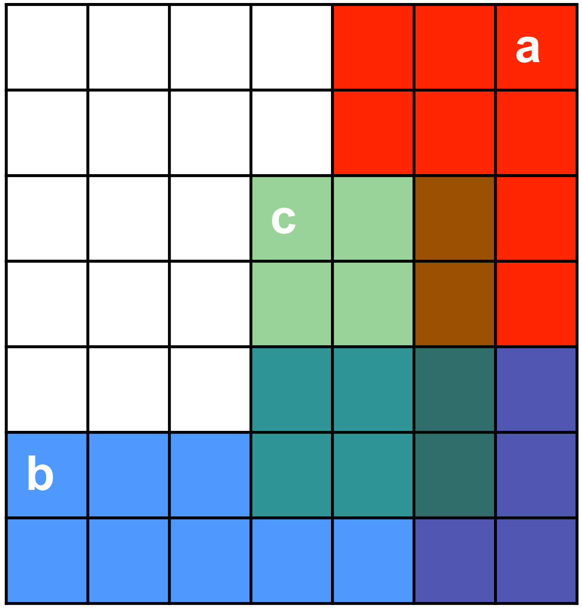
Pr(A,B,C):	P(a,b,c) = 2/49	P(a,b,~c) = 4/49
	P(a,~b,c) = 2/49	P(a,~b,~c) = 8/49
	P(~a,b,c) = 4/49	P(~a,b,~c) = 8/49
	P(~a,~b,c) = 4/49	P(~a,~b,~c) = 17/49

Paulo Trigo Silva

… illustrative example – single variable conditioned

Let variables A, B, C represent colors:

$A \in \{ a, \sim a \}$
$B \in \{ b, \sim b \}$
$C \in \{ c, \sim c \}$



There are **24** single variable conditional probabilities!

Single variables conditioned on single variables (**24**):

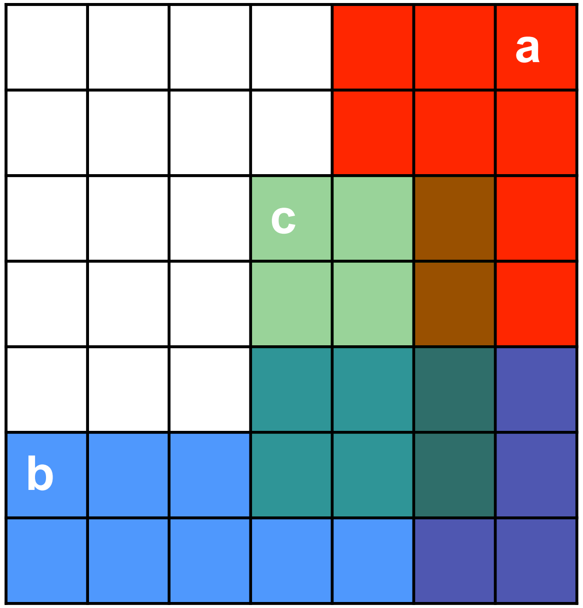
Pr(A|B):	P(a|b)=P(a,b)/P(b)=1/3	P(~a|b) = 2/3
	P(a|~b)=P(a,~b)/P(~b)=10/31	P(~a|~b)=21/31
Pr(A|C):	P(a|c)=P(a,c)/P(c)=1/3	P(~a|c) = 2/3
	P(a|~c)=P(a,~c)/P(~c)=12/37	P(~a|~c)=25/37
Pr(B|C):	P(b|c)=P(b,c)/P(c)=1/2	P(~b|c) = 1/2
	P(b|~c)=P(b,~c)/P(~c)=12/37	P(~b|~c)=25/37
Pr(B|A):	P(b|a)=P(b,a)/P(a)=?	P(~b|a) = ?
	P(b|~a)=P(b,~a)/P(~a)=?	P(~b|~a)=?
Pr(C|A):	P(c|a)=P(c,a)/P(a)=?	P(~c|a) = ?
	P(c|~a)=P(c,~a)/P(~a)=?	P(~c|~a)=?
Pr(C|B):	P(c|b)=P(c,b)/P(b)=?	P(~c|b) = ?
	P(c|~b)=P(c,~b)/P(~b)=?	P(~c|~b)=?

Paulo Trigo Silva

… example – conditioned (two and one variables)

Let variables A, B, C represent colors:

$A \in \{ a, \sim a \}$
$B \in \{ b, \sim b \}$
$C \in \{ c, \sim c \}$



24 two variable conditioned on one and **24** one variable conditioned on two!

Two variables conditioned on a single variables (**24**):

Pr(A,B|C): P(a,b|c) = P(a,b,c)/P(c)=1/6
P(a,b|~c) = P(a,b,~c)/P(~c) = 4/37
P(a,~b|c) = P(a,~b,c)/P(c) = 1/6
P(a,~b|~c) = P(a,~b,~c)/P(~c) = 8/37
P(~a,b|c) = P(~a,b,c)/P(c) = 1/3
P(~a,b|~c) = P(~a,b,~c)/P(~c) = 8/37
P(~a,~b|c) = P(~a,~b,c)/P(c) = 1/3
P(~a,~b|~c) = P(~a,~b,~c)/P(~c) = 17/37

Pr(A,C|B): ? (eight additional probabilities)

Pr(B,C|A): ? (eight additional probabilities)

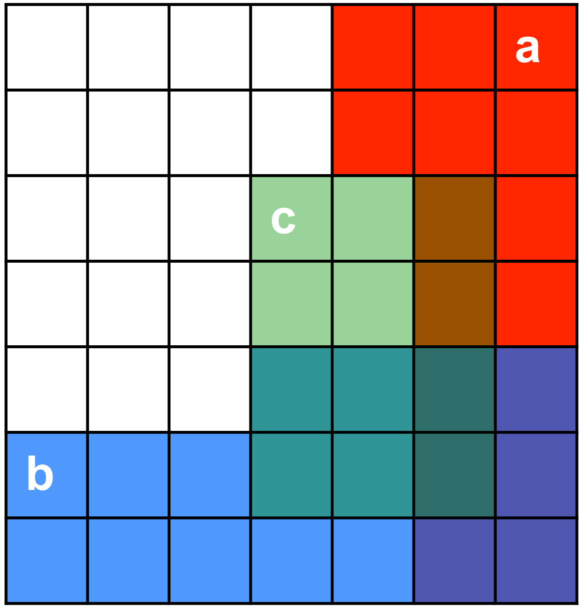
[similarly] one variable conditioned on two (**24**):

Pr(A|B,C): ? (eight additional probabilities)
Pr(B|A,C): ? (eight additional probabilities)
Pr(C|A,B): ? (eight additional probabilities)

… example – are there independent relations?

Let variables A, B, C represent colors:

A ∈ { a, ~a }
B ∈ { b, ~b }
C ∈ { c, ~c }



Independence:

Is A independent from B?
Is B independent from C?

Conditional Independence:

Is A independent from B given C?
Is C independent from A given B?

Paulo Trigo Silva

… example – are there independent relations? (cont.)

Independence:

Is A independent from B?

$P(a,b) \neq P(a)P(b)$

$P(a,\sim b) \neq P(a)P(\sim b)$

$P(\sim a,b) \neq P(\sim a)P(b)$

$P(\sim a,\sim b) \neq P(\sim a)P(\sim b)$

so A and B are not independent

(a single ≠ is sufficient)

Is B independent from C?

?

Conditional Independence:

Is A independent from B given C?

$P(a,b|c) = 1/6 = P(a|c)\,P(b|c)$

but

$P(a,b|\sim c) = 4/37 \neq P(a|\sim c)P(b|\sim c)$

so A and B are
not conditionally independent

Is C independent from A given B?

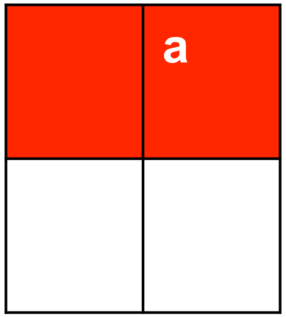
?

Paulo Trigo Silva

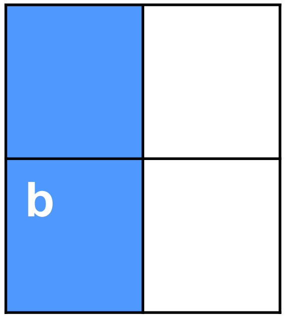
An illustrative example of two independent variables

Let variables A, B represent colors:
A ∈ { a, ~a }, where a is red and ~a is not red
B ∈ { b, ~b }, where b is blue and ~b is not blue

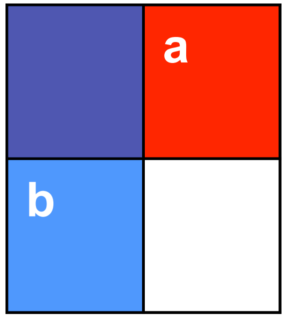
P(A=a) = 2/4 = 1/2
P(A=~a) = 2/4 = 1/2



P(B=b) = 2/4 = 1/2
P(B=~b) = 2/4 = 1/2



A and B are independent variables?



Paulo Trigo Silva

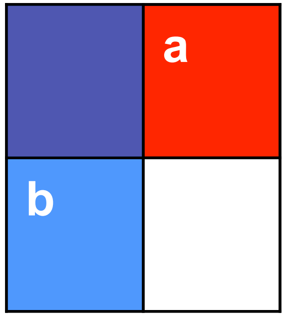
… example of two independent variables

Let variables A, B represent colors:

$A \in \{ a, \sim a \}$, where a is red and ~a is not red

$B \in \{ b, \sim b \}$, where b is blue and ~b is not blue

P(a,b)	= 1/4	P(a,~b)	= 1/4
P(~a,b)	= 1/4	P(~a,~b)	= 1/4



P(a,b)	= P(a)×P(b)	= 1/2×1/2 = 1/4	P(a,~b)	= P(a)×P(~b)	= 1/2×1/2 = 1/4
P(~a,b)	= P(~a)×P(b)	= 1/2×1/2 = 1/4	P(~a,~b)	= P(~a)×P(~b)	= 1/2×1/2 = 1/4

Pr(A, B) = Pr(A) × Pr(B)

therefore A and B are independent variables

Paulo Trigo Silva

… another example – two variables with different range

Let variables A, B represent colors:

A ∈ { **a, ~a** }, where a is red and ~a is not red

B ∈ { **b, bL, ~b** }, where b is blue, bL is blue light and ~b is not blue

draw a pattern (graphical representation) to illustrate A and B as independent variables

Paulo Trigo Silva

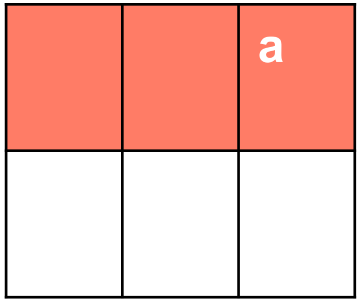
… example of two variables with different range

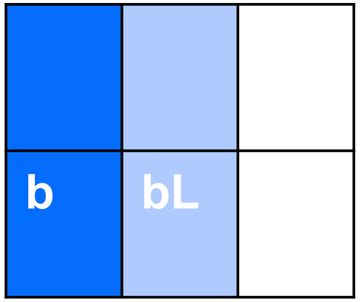
Let variables A, B represent colors:
A ∈ { a, ~a }, where a is red and ~a is not red
B ∈ { b, bL, ~b }, where b is blue, bL is blue light and ~b is not blue

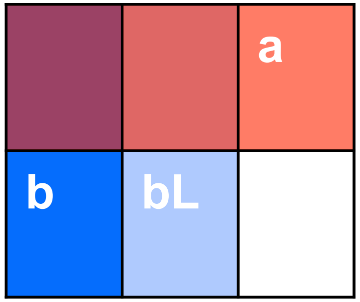
P(A=a) = 3/6 = 1/2
P(A=~a) = 3/6 = 1/2

P(B=b) = 2/6 = 1/3
P(B=bL) = 2/6 = 1/3
P(B=~b) = 2/4 = 1/3

A and B are independent variables?





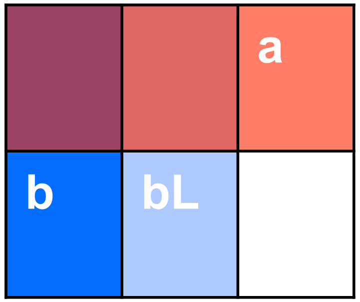


Paulo Trigo Silva

… two independent variables with different range

Let variables A, B represent colors:
A ∈ { a, ~a }, where a is red and ~a is not red
B ∈ { b, bL, ~b }, where b is blue, bL is blue light and ~b is not blue

P(a,b) = 1/6	P(a,bL) = 1/6	P(a,~b) = 1/6
P(~a,b) = 1/6	P(~a,bL)= 1/6	P(~a,~b)= 1/6



P(a,b) = P(a)×P(b) = 1/2×1/3 = 1/6	P(a, bL) = 1/2×1/3 = 1/6	P(a,~b) = 1/6
P(~a,b) = P(~a)×P(b) = 1/2×1/3 = 1/6	P(~a, bL) = 1/2×1/3 = 1/6	P(-a,~b) = 1/6

Pr(A, B) = Pr(A) × Pr(B)
therefore A and B are independent variables

Paulo Trigo Silva

… two independent variables each with three values

Let variables A, B represent colors:

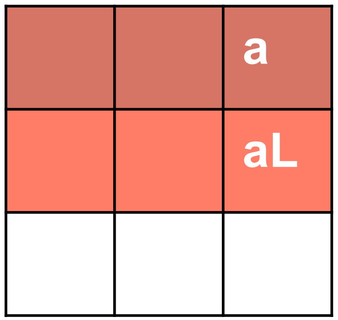
A ∈ { a, aL, ~a }, where a is red, aL is red light and ~a is not red

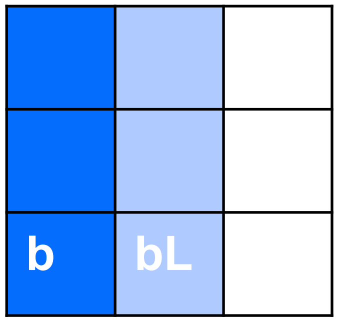
B ∈ { b, bL, ~b }, where b is blue, bL is blue light and ~b is not blue

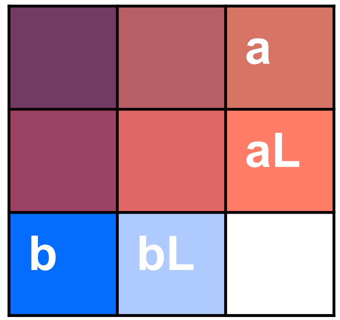
P(A=a) = 3/9 = 1/3
P(A=aL) = 3/9 = 1/3
P(A=~a) = 3/9 = 1/3

P(B=b) = 3/9 = 1/3
P(B=bL) = 3/9 = 1/3
P(B=~b) = 3/9 = 1/3

A and B are independent variables?







Similarly to previous examples we see Pr(A,B) is always 1/9 for all events, i.e.,

P(a,b)=1/9; P(a,bL)=1/9; P(a,~b)=1/9; P(aL,b)=1/9; P(aL,bL)=1/9; P(aL,~b)=1/9; P(~a,b)=1/9; P(~a,bL)=1/9; P(-a,~b)=1/9;

therefore, as Pr(A, B) = Pr(A) × P(B), the variables A and B are independent

Paulo Trigo Silva

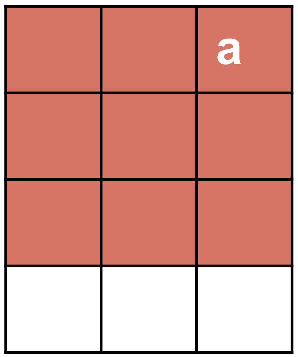
… two variables; now, values with different probabilities

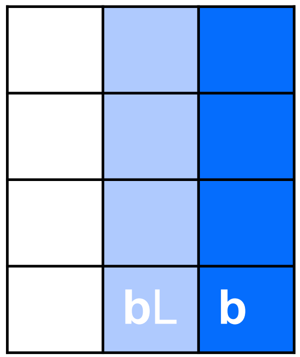
Let variables A, B represent colors:
$A \in \{ a, \sim a \}$, where a is red and ~a is not red
$B \in \{ b, bL, \sim b \}$, where b is blue, bL is blue light and ~b is not blue

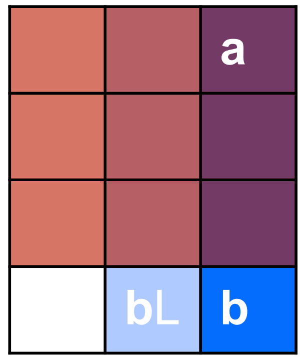
P(A=a) = 9/12 = **3/4**
P(A=~a) = 3/12 = **1/4**

P(B=b) = 4/12 = 1/3
P(B=bL) = 4/12 = 1/3
P(B=~b) = 4/12 = 1/3

A and B are independent variables?







variable A with different probability for each value

variable A has 2 values but its space of probabilities was “divided” in “4 portions”, so the whole space is 4 × 3 = 12 cells

Paulo Trigo Silva

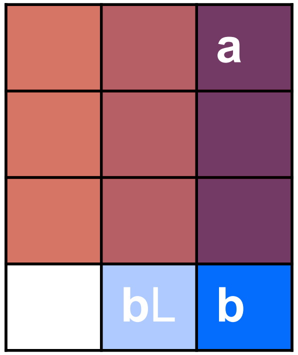
… two variables; values with different probabilities (cont.)

Let variables A, B represent colors:

$A \in \{ a, \sim a \}$, where a is red and ~a is not red

$B \in \{ b, bL, \sim b \}$, where b is blue, bL is blue light and ~b is not blue

$P(a,b) = 1/4$	$P(a,bL) = 1/4$	$P(a,\sim b) = 1/4$
$P(\sim a,b) = 1/12$	$P(\sim a,bL) = 1/12$	$P(\sim a,\sim b) = 1/12$



$P(a,b) = P(a) \times P(b) = 3/4 \times 1/3 = 1/4$	$P(a, bL) = 3/4 \times 1/3 = 1/4$	$P(a,\sim b) = 1/4$
$P(\sim a,b) = P(\sim a) \times P(b) = 1/4 \times 1/3 = 1/12$	$P(\sim a, bL) = 1/4 \times 1/3 = 1/12$	$P(-a,\sim b) = 1/12$

Pr(A, B) = Pr(A) × Pr(B)

therefore A and B are independent variables

Paulo Trigo Silva

… the “bottom line” idea on “independent variables”

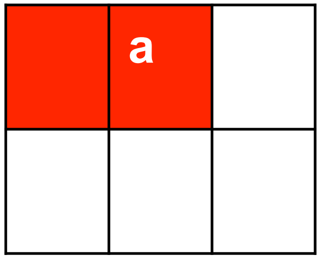
If A and B are independent variables
then
all joint events are likely to occur according to their proportionality relation

i.e.,
all combinations of the A and B values (joint events) occur the number of times delimited by the space that they occupy

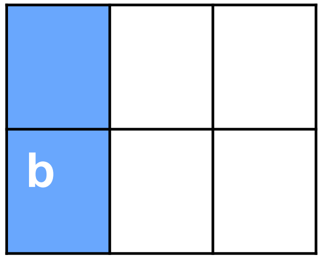
And now, A and B conditionally independent given C

Let variables A, B, C represent colors:
A ∈ { a, ~a }, where a is red and ~a is not red
B ∈ { b, ~b }, where b is blue and ~b is not blue
C ∈ { c, ~c }, where c is green and ~c is not green

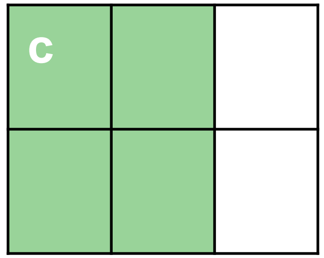
$P(A=a) = 2/6 = 1/3$
$P(A=\sim a) = 4/6 = 2/3$



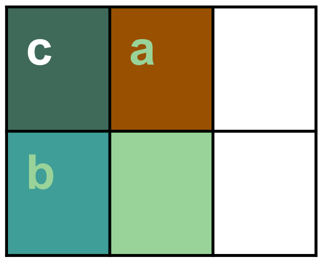
$P(B=b) = 2/6 = 1/3$
$P(B=\sim b) = 4/6 = 2/3$



$P(C=c) = 4/6 = 2/3$
$P(C=\sim c) = 2/6 = 1/3$



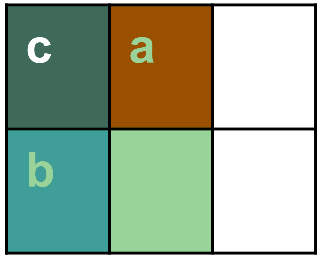
A and B are independent?
A and B are conditionally independent given C?



Paulo Trigo Silva

… A and B conditionally independent given C

A and B are independent?
A and B are conditionally independent given C?



A and B are independent?

P(a,b) = 1/6

P(a) = 2/6 = 1/3

P(b) = 2/6 = 1/3

P(a,b)=1/3 ≠ P(a)P(b)=1/9

so A and B are not independent

A and B conditionally independent given C?

P(a,b|c)=1/4	P(a|c)=1/2 P(b|c)=1/2
P(a,~b|c)=1/4	P(a|c)=1/2 P(~b|c)=1/2
P(~a,b|c)=1/4	P(~a|c)=1/2 P(b|c)=1/2
P(~a,~b|c)=1/4	P(~a|c)=1/2 P(~b|c)=1/2

P(~a,~b|~c)=1; P(a,b|~c)=0; … all zero in ~c

Pr(A,B|C) = P(A|C) × P(B|C)

so A and B conditionally independent given C

Notice that A and B follow the "pattern" independence (cf. previous example)
Now A and B follow that "pattern" in the "scope of C=c" (and do not occur in C=~c)

Paulo Trigo Silva

… the “bottom line” idea on “conditional independence”

If A and B are conditionally independent given C
then
all joint events of A and B are likely to occur according to their proportionality relation within the scope of each possible value for C

i.e.,
the variables A and B are independent within the “narrowed” space delimited by each value of C

using the “graphical intuition” the idea is to see if the “independence pattern” between A and B occurs within the scope of each value of C

Synthesis: important results on "conditional independence"

- If **A and B are conditionally independent given C**, then
  - we have $\mathbf{Pr(A \mid B, C) = Pr(A \mid C)}$

- … which can be expressed in a slightly different way

  $\mathbf{Pr(A, B \mid C)}$ = Pr(A, B, C) / Pr(C) = [Pr(A | B, C) Pr(B | C) Pr(C)] / Pr(C)

  = Pr(A | B, C) Pr(B | C) // product rule Pr(X,Y)=Pr(X|Y)Pr(Y)

  = $\mathbf{Pr(A \mid C) Pr(B \mid C)}$ // A and B **conditionally independent given C**

- If we have N variables, A_1, A_2, A_3, … A_n, that are all conditionally independent given C, the last expression can be generalized as

  $\mathbf{Pr(A_1, A_2, A_3, \ldots, A_n \mid C)} =$

  $= Pr(A_1 \mid A_2, A_3, \ldots A_n, C) Pr(A_2 \mid A_3, \ldots A_n, C) \ldots Pr(A_n \mid C)$

  $= Pr(A_1 \mid C) Pr(A_2 \mid C) Pr(A_3 \mid C) \ldots Pr(A_n \mid C)$

  $= \mathbf{\Pi_{i=1..n} Pr(A_i \mid C)}$

Now, back to the Bayes rule for "Decision Making"

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]}$$

Precisely the same formulation!

Only variables get a different name!

$$p\left(Decisão_i \mid x\right) = \frac{p\left(x \mid Decisão_i\right) p\left(Decisão_i\right)}{p\left(x\right)}$$

Algorithm for "Decision Making"

1. ***decision* = ∅, *max* = 0.0**
2. **for each *H* :**
   - ***a. p* ← Pr(*H* | *E*)**
   - **b. if *p* > *max*, then: *max* ← p, *decision* ← *H***
3. **return *decision***

The "decision" is to choose the hypothesis, *H*, with highest probability *à-posteriori*

Paulo Trigo Silva

... how to compute the likelihood, Pr(*E* | *H*) ?

... assumption – independence of the evidence E given H

$$\Pr[H|E] = \frac{\boxed{\Pr[E|H]}\Pr[H]}{\Pr[E]}$$

Assuming that the evidence, *E*, is composed of parts(i.e., attributes) that are independent given *H*, (conditional independence), so we have:

$$\Pr[E|H] = \Pr(E_1 \mid H) \times \Pr(E_2 \mid H) \ldots \Pr(E_n \mid H)$$

where E_i is the attribute *i* from the example to classify

Paulo Trigo Silva

... *à-posteriori*, Pr(H | E), and “Decision Making”

$$\Pr[H|E] = \frac{\boxed{\Pr[E|H]}\Pr[H]}{\Pr[E]}$$

*... and the assumption that each E_i component of the evidence are **conditionally independent** given H*

$$\Pr(H \mid E) = \frac{\overbrace{\Pr(E_1 \mid H) \times \Pr(E_2 \mid H) \ldots \Pr(E_n \mid H)}^{\Pr[E|H]} \times \Pr(H)}{\Pr(E)}$$

Recall that in decision making we can discard Pr(E), so:

$$\mathbf{Pr}(H \mid E) \propto \mathbf{Pr}(E_1 \mid H) \times \mathbf{Pr}(E_2 \mid H) \times \ldots \times \mathbf{Pr}(E_n \mid H) \times \mathbf{Pr}(H)$$

... to use in the “Decision Making” algorithm

this reads as: “it is proportional to”

Paulo Trigo Silva

Synthesis: Pr(H | E) in “Decision Making”

$$\Pr(H \mid E) \propto \Pr(E_1 \mid H) \times \Pr(E_2 \mid H) \times \ldots \times \Pr(E_n \mid H) \times \Pr(H)$$

Using an informal and intuitive style, we can write:

à-posteriori **∝ likelihood ×** ***à-priori***

or

à-posteriori **∝** ***à-priori*** **× likelihood**

Algorithm for “Decision Making”

1. ***decision*** **= ∅,** ***max*** **= 0.0**
2. **for each** ***H*** **:**
   a. $p \leftarrow \Pr(H) \times \Pr(E_1 \mid H) \times \Pr(E_2 \mid H) \times \ldots \times \Pr(E_n \mid H)$
   b. **if** ***p*** **>** ***max*****, then:** ***max*** **← p,** ***decision*** **←** ***H***
3. **return** ***decision***

Paulo Trigo Silva

The classification task (... Bayes rule provides support)

- To **classify** is to **decide** how to answer the question:
  - "In which class (group) is this example better framed (belonged)?"
  - i.e., learn a function $f: X \rightarrow \{ classe_1, classe_2, ..., classe_n \}$
  - ... each $x \in X$ is an example that we want to know the *class* it belongs
- The *input* data (dataset)
  - the set of classes to consider, e.g., $C = \{ c_1, c_2, ..., c_n \}$
  - pre-classified examples, e.g., $<x_1, c_3>$, $<x_2, c_4>$, $<x_3, c_3>$, $<x_4, c_1>$, ...
- The *output* (resulting model)
  - a function *f*, that is, a classification procedure,
  - ... that can be represented in several forms,
  - e.g., **<u>statistical model</u>**, decision tree, rule set, neural network, etc
  - ... *f* is used to associate a new object with its most likely class
  - ... by doing such an association we are classifying!

Paulo Trigo Silva

Classification task (global view of this scenario)

A set of pre-classified examples

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

concept to classify
or
class

class
or
class value

c_1 = yes,
c_2 = no

To which class does this example belong?

Outlook	Temperature	Humidity	Windy	Play
sunny	cool	high	true	?

Paulo Trigo Silva

The classification task – an intuitive scenario!

A set of pre-classified examples

the dataset would be formed by features computed from each image (e.g., the RGB of each pixel).

To which class does this example belong?

Paulo Trigo Silva

Classify – Bayes rule and the independence assumption

- "Bayes rule" context: "classify is to learn how to answer the question"
  - "What is the probability of each class given a certain instance?"

- In the "classification context" we have:
  - evidence E ≡ instance (or "example" when considering a training set)
  - hypothesis H ≡ the class of a concept (or the "class value" of a class)
  - ... prediction: "the instance belongs to the class with highest probability!"

- General case (*recall*: conditional independence assumption)

$$\boxed{\mathbf{Pr}(\, H \mid E \,) \propto \mathbf{Pr}(\, E_1 \mid H \,) \times \mathbf{Pr}(\, E_2 \mid H \,) \ldots \mathbf{Pr}(\, E_n \mid H \,) \times \mathbf{Pr}(\, H \,)}$$

- e.g., scenario of "weather conditions",

$$\Pr(\, \mathbf{yes} \mid E \,) \propto \Pr(\, E_1 \mid yes \,) \times \Pr(\, E_2 \mid yes \,) \times \Pr(\, E_3 \mid yes \,) \times \Pr(\, E_4 \mid yes \,) \times \mathbf{Pr}(\, \mathbf{yes} \,)$$
$$\Pr(\, \mathbf{no} \mid E \,) \propto \Pr(\, E_1 \mid no \,) \times \Pr(\, E_2 \mid no \,) \times \Pr(\, E_3 \mid no \,) \times \Pr(\, E_4 \mid no \,) \times \mathbf{Pr}(\, \mathbf{no} \,)$$

  - where E_i is the attribute i from the instance to classify

Paulo Trigo Silva

... independence assumption (of E_i given H) is Naïve

- "Evidence, E, is composed by independent parts (i.e., attributes)"
  - this is a naïve (simplistic, ingenuous) assumption
  - ... seems to be an ingenuous perspective of our surrounding reality!
- **Naïve Bayes** – designates the Bayes rule under the assumption that
  - $\mathbf{Pr(}\ \boldsymbol{E} \mid \boldsymbol{H}\ \mathbf{)} = \Pr(\ E_1, \ldots, E_n \mid H\ ) = \mathbf{Pr(}\ \boldsymbol{E_1} \mid \boldsymbol{H}\ \mathbf{)} \times \ldots \times \mathbf{Pr(}\ \boldsymbol{E_n} \mid \boldsymbol{H}\ \mathbf{)}$
  - i.e., assuming that all E_i are conditionally independent given H
- Despite its name, Naïve Bayes works very well in practical datasets!
  - mainly when the attributes are conditionally independent (given H)
- ... so, existence of **redundant attributes** skews the learning process
  - e.g., if we add a new attribute with the same values as *windy* the effect of *windy* would be multiplied; all its probabilities would be squared and increase its influence in the decision; if we add 10 such attributes, then the decisions would effectively be made on *windy* alone!
  - ... there are methods to **identify and eliminate redundant attributes**!

The problem of “zero-frequency”

Suppose the value of an attribute is not associated with any class,

i.e. the attribute, for that value, has zero-frequency,

... and, what happens to the probability *à-posteriori* ?

As an example, suppose we have,

P(Humidity=h*igh* | Play=*yes*) = 0

what happens to

P(Play=*yes* | *E*)

?

P(*yes* | *E*) would be:

- indefinite
- an infinite value (+ ∞)
- a zero value (0)
- an unity value (1)

What is the correct answer?

Paulo Trigo Silva

... zero-frequency ⇒ zero probability *à-posteriori* zero!

if P(Humidity=h*igh* | Play=*yes*) = 0

P(Play=*yes* | *E*) ∝ P(Outlook=*sunny* | Play=*yes*) ×
P(Temperature=*cool* | Play=*yes*) ×
P(Humidity=*high* | Play=*yes*) ×
P(Windy=*true* | Play=*yes*) ×
P(Play=*yes*)
= 2/9 × 3/9 × **0** × 3/9 × 9/14
= **0**

... probabilities that are zero hold a veto over the other ones;

and, we **loose all the information** about the remaining attributes.

A (standard) technique (named "Laplace estimator")**:**

add 1 to the count of each "value of the class attribute"

***Effect*: a (estimated) probability that is always > 0**

Paulo Trigo Silva

zero-frequency ⇒ probability modified by estimator

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

(2 **+ 1**) / (9 **+ 3**)
(4 **+ 1**) / (9 **+ 3**)
(3 **+ 1**) / (9 **+ 3**)

Σ = 12 / 12 = 1

...

(3 **+ 1**) / (9 **+ 2**)
(6 **+ 1**) / (9 **+ 2**)

Σ = 11 / 11 = 1

(4 **+ 1**) / (5 **+ 2**)
(1 **+ 1**) / (5 **+ 2**)

Σ = 7 / 7 = 1

... ...

Examples – add 1 to the count of each frequency
***Note*: add 1 to numerator; add to denominator the total amount of 1s**

Paulo Trigo Silva

... generalize the Laplace estimator

There is no particular reason for adding 1 to the counts.

It can be more adequate to add a constant value that is different from 1.

Example of *Outlook* for the class value *yes*

Outlook	
	yes
sunny	$\frac{(2 + \mu / 3)}{(9 + \mu)}$
overcast	$\frac{(4 + \mu / 3)}{(9 + \mu)}$
rainy	$\frac{(3 + \mu / 3)}{(9 + \mu)}$

Note that, in numerator, we have: $\mu \times 1/3$

i.e., the value of μ (e.g., 3) provides a weight that determines how influential the *a-priori* values of 1/3, 1/3 e 1/3 are for each of the three values *sunny*, *overcast* and *rainy* of attribute *Outlook*.

An increment of μ also increments the importance of the *à-priori* value (1/3 in this example) in the classification of a new instance; a decrement of μ also decrements such importance.

Paulo Trigo Silva

Frequencies weighted by the *à-priori* probabilities

There is no particular reason for dividing μ into equal parts in the numerators!

i.e.,, we generalize the approach of multiplying μ for 1/|range(Attribute)|.

Example of *Outlook* for the class value *yes*

Outlook	
	yes
sunny	$\dfrac{(2 + \mu \times \mathbf{p_1})}{(9 + \mu)}$
overcast	$\dfrac{(4 + \mu \times \mathbf{p_2})}{(9 + \mu)}$
rainy	$\dfrac{(3 + \mu \times \mathbf{p_3})}{(9 + \mu)}$

Note that, in the numerator, we now have:

$$\mu \times p_i$$

with the additional restriction: $p_1 + p_2 + p_3 = 1$

i.e., the three numbers $\mathbf{p_1}$, $\mathbf{p_2}$ e $\mathbf{p_3}$ now represent the ***à-priori*** probabilities of, respectively, the *sunny*, *overcast* and *rainy* of attribute *Outlook*.

Paulo Trigo Silva

(complete) Bayes formulation – (*à-priori* in all the terms)

In synthesis, the generalization of Laplace estimator gives:

$$\mathbf{Pr}(\, \boldsymbol{E_i = e_i \mid H = h_j} \,) = \frac{\#(\, E_i = e_i \,,\, H = h_j \,) + \mu \times p_i}{\#(\, H = h_j \,) + \mu} \,.$$

with the additional restriction: $\Sigma_i \, p_i = 1$

frequency perspective

Laplace estimator

à-priori probability of attribute *i*

Now we have a complete Bayes formulation with the à-priori probabilities, p_i e Pr(H), being present in all its terms (on the right side)

$$\mathbf{Pr}(\, H \mid E \,) \propto \mathbf{Pr}(\, E_1 \mid H \,) \times \mathbf{Pr}(\, E_2 \mid H \,) \times \ldots \times \mathbf{Pr}(\, E_n \mid H \,) \times \mathbf{Pr}(\, H \,)$$

i.e., in an informal and intuitive way we can write:

à-posteriori $\propto$ likelihood-&-**à-priori** $\times$ ***à-priori***

Paulo Trigo Silva

... the practice and the estimation of *à-priori* probabilities

- Advantage of the complete formulation (*à-priori* in all the terms)
  - completely rigorous
- Disadvantages of the complete formulation
  - it is not usually clear how the *à-priori* probabilities should be assigned
- In practice the *à-priori* probabilities may have small impact
  - provided that there are a reasonable number of training instances
  - … in that case the observed frequencies approximate the probability
- In practice, the usual approach is just to compute the frequencies
  - using the Laplace estimator
  - by initializing all counts to 1 (one) instead of initializing to 0 (zero)!

Missing Attributes

A "really nice" thing about the Bayes formulation is that dealing with missing values does not constitute an additional problem!

Outlook	yes	no	Temperature	yes	no	Humidity	yes	no	Windy	yes	no	Play yes	Play no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

Outlook	Temperature	Humidity	Windy	Play
omisso!	cool	high	true	?

***likelihood* of *yes* =** $\boxed{}$ **3/9 × 3/9 × 3/9 × 9/14 = 0.0238**

***likelihood* of *no* =** $\boxed{}$ **1/5 × 4/5 × 3/5 × 5/14 = 0.0343**

Paulo Trigo Silva

Missing Attribute (cont.)

Now (without the attribute)

likelihood of _yes_ = 3/9 × 3/9 × 3/9 × 9/14 = 0.0238

likelihood of _no_ = 1/5 × 4/5 × 3/5 × 5/14 = 0.0343

Previous Computation (with the attribute)

likelihood of *yes* = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

likelihood of *no* = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

Notice that the final numbers (individually) are much higher without the attribute (it is like if the attribute has a weight of 1)!

But, after normalization, the probability of *yes* and *no* is 41% and 59%

probability of _yes_ = 0.0238 / (0.0238 + 0.0343) = 41%

probability of _no_ = 0.0343 / (0.0238 + 0.0343) = 59%

Paulo Trigo Silva

Synthesis – missing attributes

- If a value is missing in a new instance (to classify)
  - then the likelihood is computed only with the values that really occur
  - ... the missing value(s) are simply not included in the computation
  - (such as in the previous example)
- If a value is missing in a training instance (data to analyze)
  - then, that value(s) is(are) simply not included in the frequency counts
  - ... i.e., the probability ratios are based on the number of values that actually occur rather than on the total number of instances

Paulo Trigo Silva

Attributes with Numeric Domain

The previous example has changed:

Temperature and *Humidity* now have a numeric domain!

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

... how to classify when attributes have numeric domain?

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

... today the weather conditions are as follows:

Am I going to play tennis?

Outlook	Temperature	Humidity	Windy	Play
sunny	66	90	true	?

Paulo Trigo Silva

For numeric attributes consider statistical measures

count the nominal attributes and list the vales of numeric attributes

Outlook			Temperature		Humidity		Windy			Play	
	yes	*no*	*yes*	*no*	*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	83	85	86	85	false	6	2	9	5
overcast	4	0	70	80	96	90	true	3	3		
rainy	3	2	68	65	80	70					
			64	72	65	95					
			69	71	70	91					
			75		80						
			75		70						
			72		90						
			81		75						

... and then,

calculate the <u>mean</u> and <u>standard deviation</u> of numeric attributes

mean

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

standard deviation

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2}$$

Paulo Trigo Silva

... mean and standard deviation of numeric attributes

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	*mean*	73	74.6	*mean*	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	*std. dev.*	6.2	7.9	*std. dev.*	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

... and then,

assume that the numeric values follow some probabilistic law …

Paulo Trigo Silva

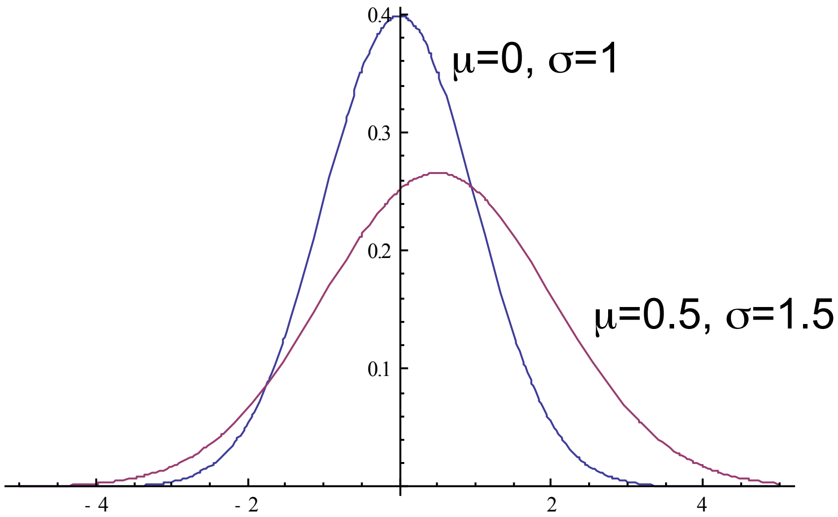
... numeric attributes follow a probabilistic law

Usually numeric values are handled assuming that they have a “normal” or “Gaussian” probability distribution with mean μ and standard deviation σ.

The probability density function for normal distribution is:

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}, -\infty < x < \infty, \sigma > 0$$

Carl Friedrich Gauss
1777 – 1855
Germany



Symmetric around the mean value.

When μ=0 e σ=1 it is called “centered and reduced” (or standard) Normal.

Paulo Trigo Silva

... example

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	*mean*	73	74.6	*mean*	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	*std. dev.*	6.2	7.9	*std. dev.*	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

What is the value of:

f(*temperature* = 66 | *Play* = 'yes')

?

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}, -\infty < x < \infty, \sigma > 0$$

Paulo Trigo Silva

... example – and classification of a "new day"

μ=73 e σ=6.2

f(*temperature* = *66* | *Play* = 'yes') =

$= 1 / (\mathbf{6.2} \times \mathrm{sqrt}(2 \times \pi)) \times e\hat{}(- (66 - \mathbf{73})^2 / (2 \times \mathbf{6.2}^2)) = \mathbf{0.0340}$

... and what is the value for:

f(*humidity* = 90 | *Play* = 'yes') ?

μ=79.1 e σ=10.2

f(*humidity* = *90* | *Play* = 'yes') =

$= 1 / (\mathbf{10.2} \times \mathrm{sqrt}(2 \times \pi)) \times e\hat{}(- (90 - \mathbf{79.1})^2 / (2 \times \mathbf{10.2}^2)) = \mathbf{0.0221}$

So, in this "new day" am I expected to play tennis?

Outlook	Temperature	Humidity	Windy	Play
sunny	66	90	true	?

... example – classification with numeric domain

likelihood of *yes* = 2/9 × **0.0340** × **0.0221** × 3/9 × 9/14 = 0.000036

likelihood of *no* = 3/5 × 0.0221 × 0.0381 × 3/5 × 5/14 = 0.000108

probability of *yes* **= 0.000036 / (0.000036 + 0.000108) = 25.0%**

probability of *no* **= 0.000108 / (0.000036 + 0.000108) = 75.0%**

These probability values are very close to the probabilities calculated earlier with nominal domains for the *temperature* and *humidity*.

i.e., the numeric values of *66* and *90* yield similar probabilities to, respectively, the nominal values of *cool* and *high* previously used.

Paulo Trigo Silva

The numeric domain & missing values

What to do with missing values in a numeric domain?

The missing values (in the training dataset) are simply not included in the calculation neither of the mean, μ, nor the standard deviation, σ.

The increase of missing values decreases the capability of properly expressing the distribution of the values of that attribute.

Paulo Trigo Silva

Naïve Bayes – final considerations about this technique

- Impressive results can be achieved
  - even when the the “independence assumption” is not guaranteed
- What is the justification for such good results?
  - the classification does not demand precise probability estimation values
- ... the classification only demands that the maximum probability
  - is assigned to the correct class!
- Nevertheless, adding too many redundant attributes (e.g., equal)
  - is problematic as it augments the attribute’s weight in classification

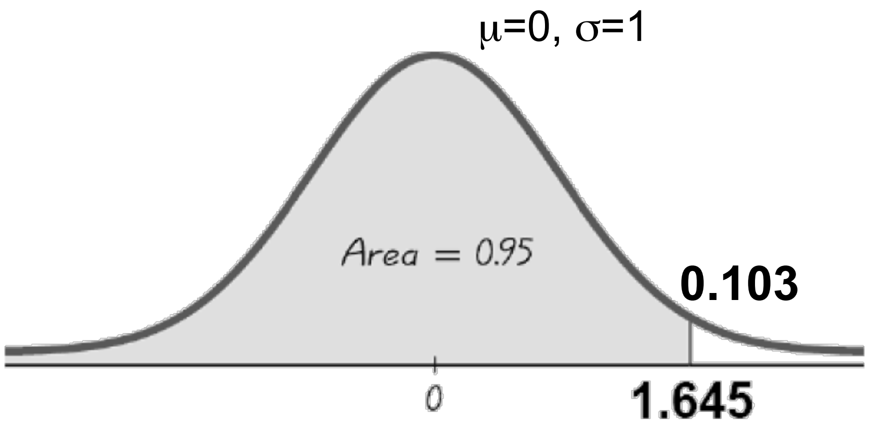
Additional – density function and cumulative function

The probability density function for the Normal distribution is given by:

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}, -\infty < x < \infty, \sigma > 0$$

The probability density function for Normal distribution has total area equal to 1.

$f(z, \mu, \sigma)$ describes a distributions and enables to characterize the cumulative distribution $\Pr(Z \leq z)$ as the area of function f until the point z.



$\Pr(Z \leq 1.645) = 0.95$

$= \int_{-\infty}^{1.645} f(x, 0, 1)$

$F(z) = \Pr(Z \leq z)$ is named the cumulative distribution function.

Paulo Trigo Silva

Additional – relation between density and probability

Notice that the probability of a continuous variable, Z, having exactly a certain value, z, is zero,

i.e., $\Pr(Z = z) = 0$

Therefore, the meaning of the density probability function $f(z)$ is that:

the probability of the value to occur in a "small" neighborhood, ε, of z, e.g., between $z - \varepsilon/2$ and $z + \varepsilon/2$,

is $\varepsilon f(z)$.

i.e.,

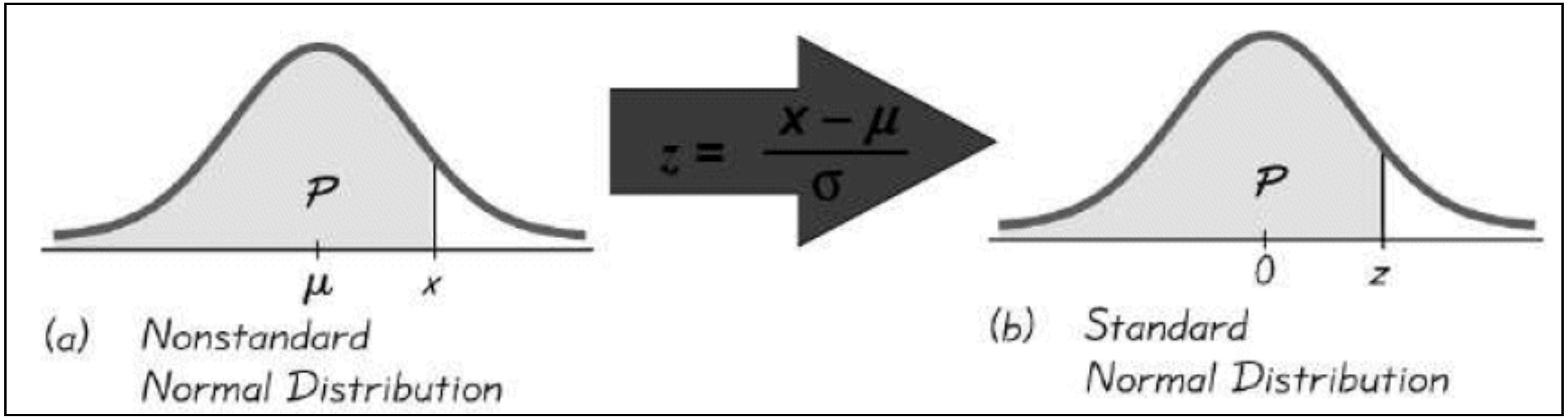
$$\Pr(z - \varepsilon/2 \leq Z \leq z + \varepsilon/2) \approx \varepsilon \times f(z)$$

In presented technique the ε does not need to be used because as it occurs in all likelihood values it is canceled when divided by summation of all likelihood values.

Additional – computation of the probabilities of the Normal

To compute the Pr($X \leq x$) one may resort to tables.

If X follows Normal distribution $N(\mu, \sigma)$ we need to transform the variable into the standard Normal $N(0, 1)$ because all tables has values for that distribution.



If X is $N(\mu, \sigma)$, then $Z = (X - \mu) / \sigma$ is $N(0, 1)$.

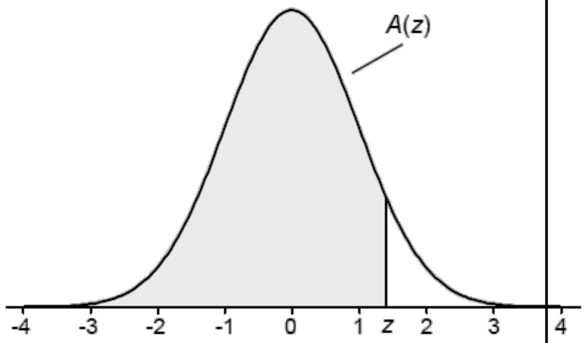
for example, if X is $N(5, 2)$ and we want to compute Pr($X \leq 7$):

$$\Pr(X \leq 7) = \Pr((X - 5) / 2 \leq (7 - 5) / 2) = \Pr(Z \leq 1) = 0.8413$$

Paulo Trigo Silva

... consult the table of the cumulative function $N(0, 1)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999							



$\Pr[\, Z \leq 1 \,] = 0.8413$