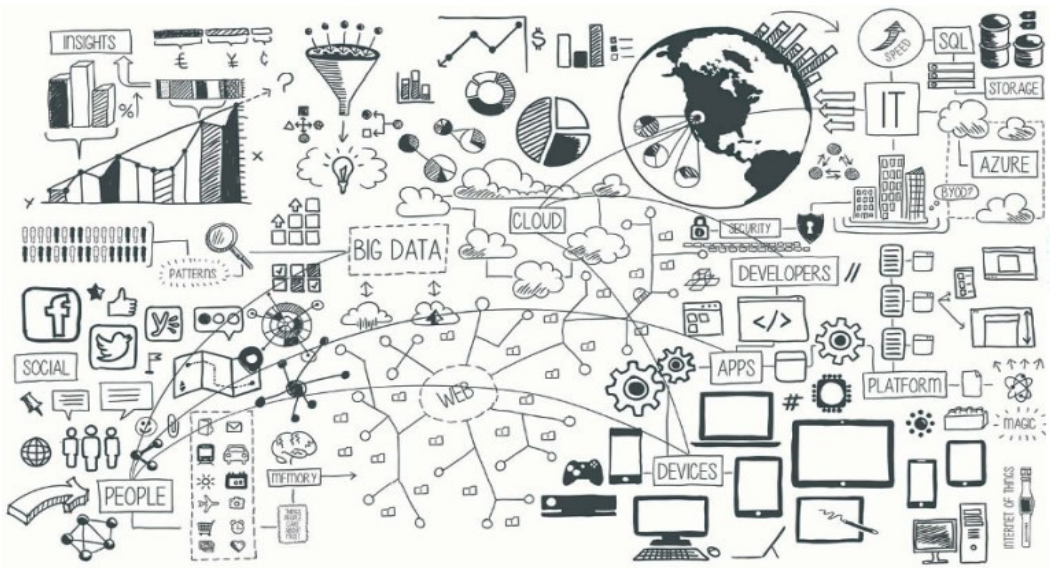
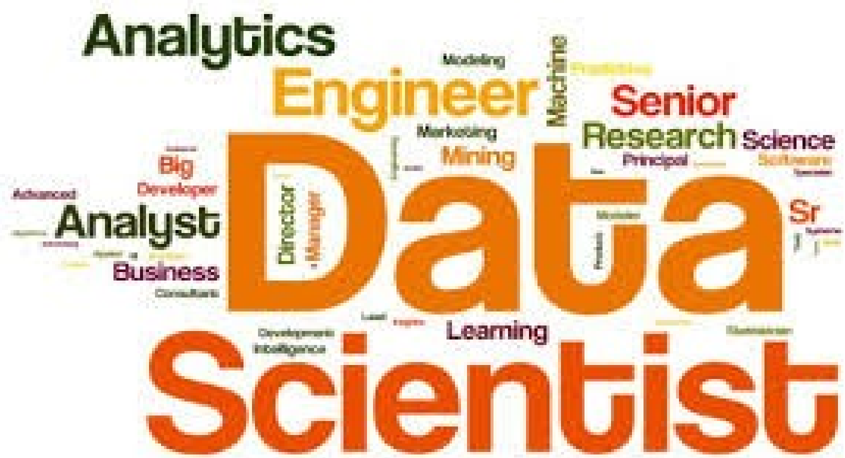
M1970 – Machine Learning II
Redes Probabilísticas Discretas (Inferencia)





Sixto Herrera (sixto.herrera@unican.es)

Grupo de Meteorología

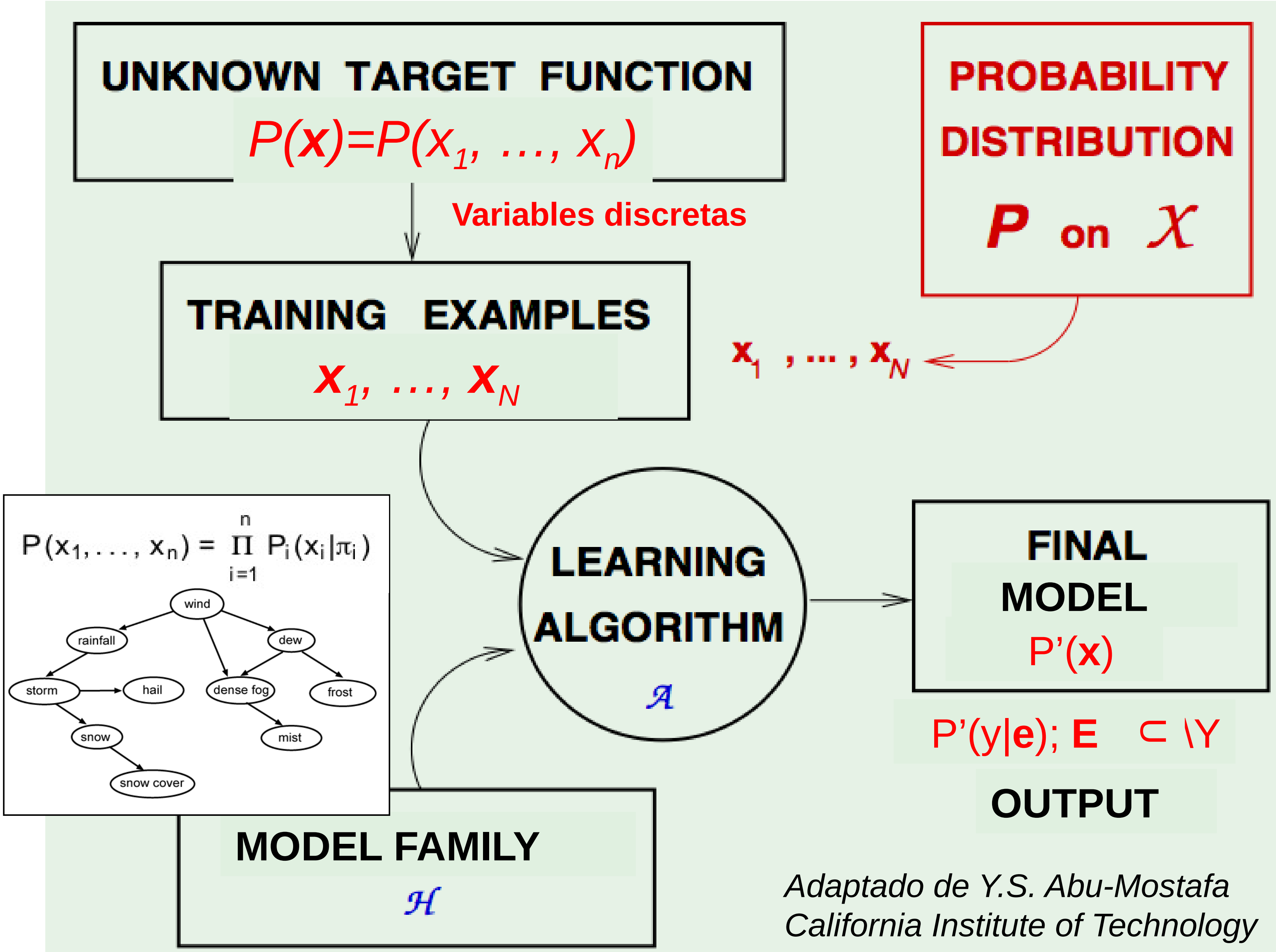
Univ. de Cantabria – CSIC
MACC / IFCA

M1970 – Machine Learning (L 15:30-17:30; X 15:30-17:30)

Feb	**28**	**L**	**Redes Probabilísticas Discretas (2h-T)**
Mar	**2**	**X**	**Redes Bayesianas: Creación e Inferencia (2h-L)**
	7	**L**	**Clasificacidores Bayesianos. Naive Bayes (2h-L)**
	9	**X**	**Redes Bayesianas: Aprendizaje Estructural (2h-T)**
	14	**L**	**Redes Bayesianas: Aprendizaje Paramétrico (2h-LT)**
	16	**X**	**Redes Bayesianas: Aprendizaje (2h-L)**
	21	**L**	**Evaluación (2h)**

NOTA: Las líneas de código de R en esta presentación se muestran sobre un fondo gris.

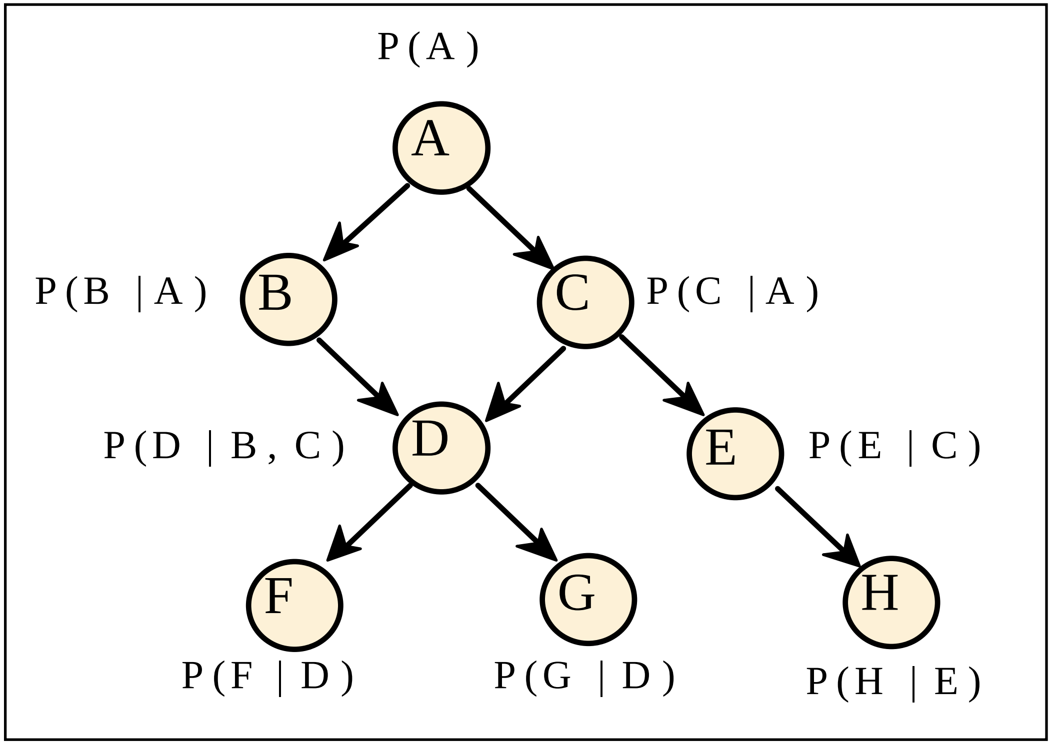
Graphical Probabilistic Models



Adaptado de Y.S. Abu-Mostafa
California Institute of Technology

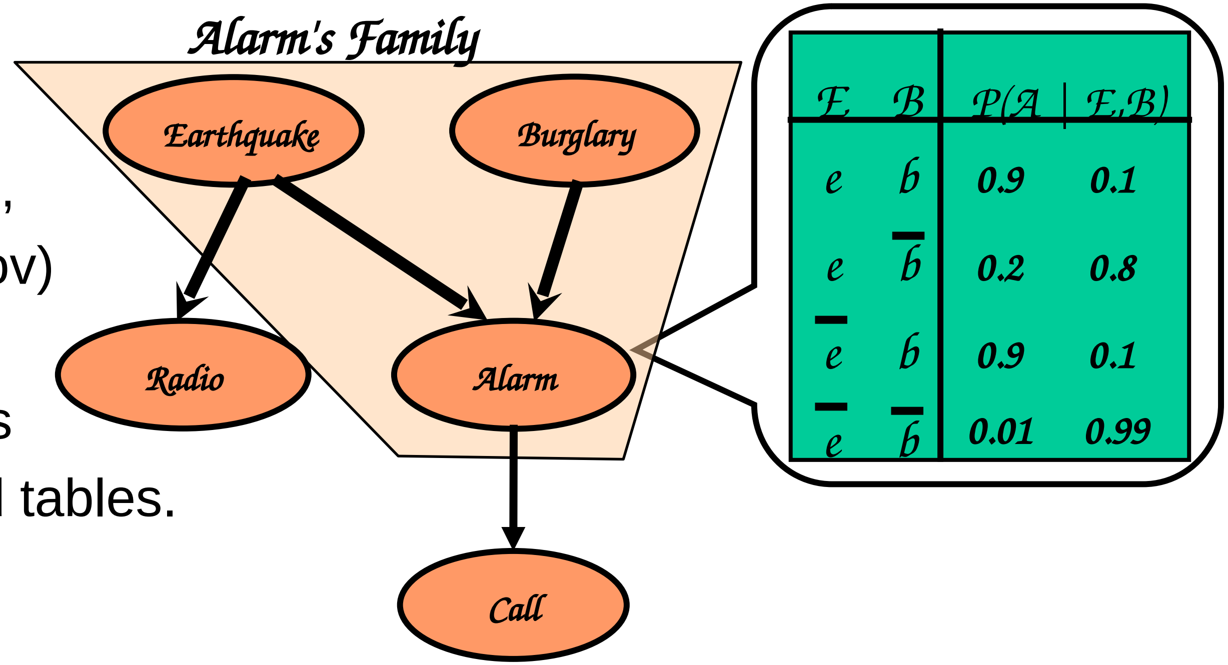
Directed graphs lead to a probabilistic model directly obtained from the graph, defining the factorization of the joint probability function as product of conditional probabilities of each node x_i given his parents π_i.

$$P(X)=\prod_{i=1}^{n} P(X_i|\pi_i)$$



$$P(A,B,C,D,E,F,G,H)=P(A)P(B|A)P(C|A)P(D|B,C)x...$$
$$xP(E|C)P(F|D)P(G|D)P(H|E)$$

Bayesian Networks obtain a compact representation of the joint probability function through the conditional independences.



Structure:

Acyclic Directed Graph (DAG),

or non-directed graphs (Markov)

- Nodes – variables
- Links – direct dependences

Parameters: Probabilities and tables.

Once the Bayesian Network (**DAG**+**CPT**) is defined, some questions grow. In particular, given a **new evidence**

- Which is the probability of an event? ← **CPT-Inference**

- There are new (in)dependences between variables? ← **DAG-Inference**

...

Bayesian Networks obtain a compact representation of the joint probability function through the conditional independences.

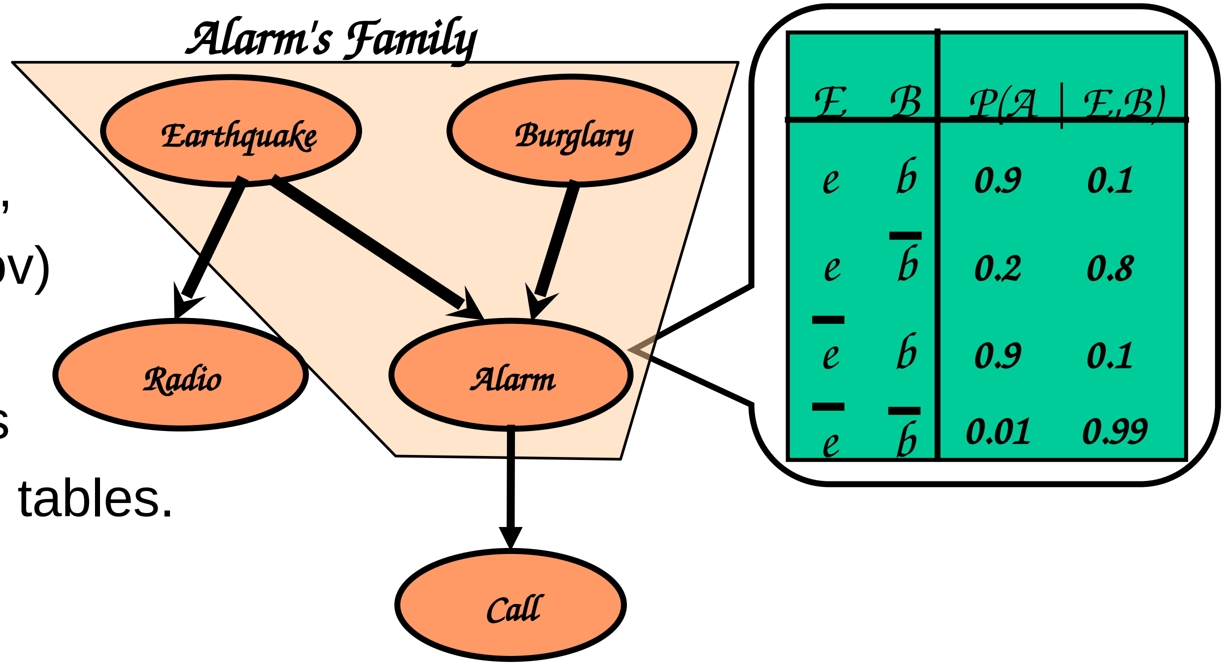
Structure:

Acyclic Directed Graph (DAG),

or non-directed graphs (Markov)

- Nodes – variables
- Links – direct dependences

Parameters: Probabilities and tables.

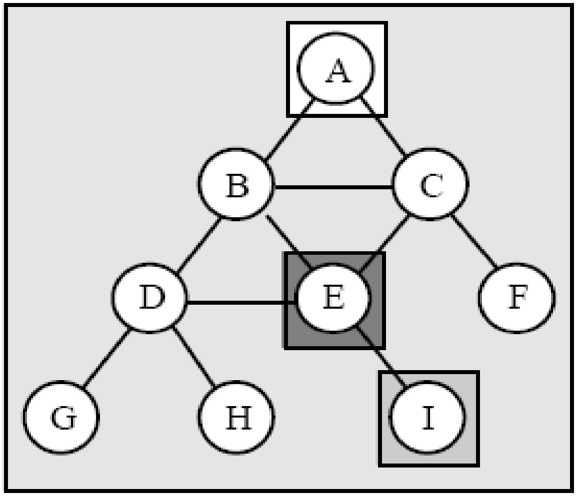


Once the Bayesian Network (**DAG**+**CPT**) is defined, some questions grow. In particular, given a **new evidence**

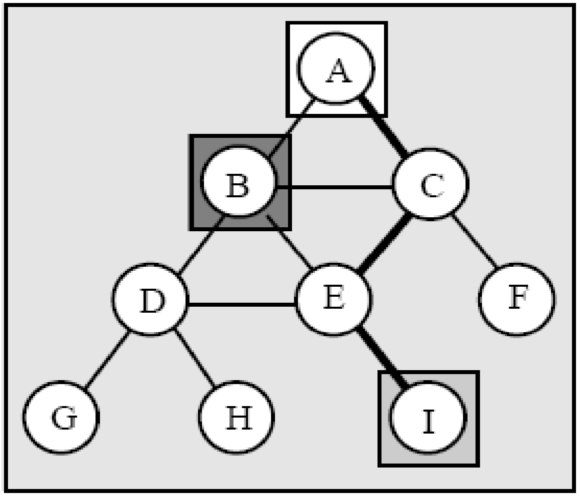
- Which is the probability of an event? ← **CPT-Inference**

- There are new (in)dependences between variables? ← **DAG-Inference** → **02/03**

...



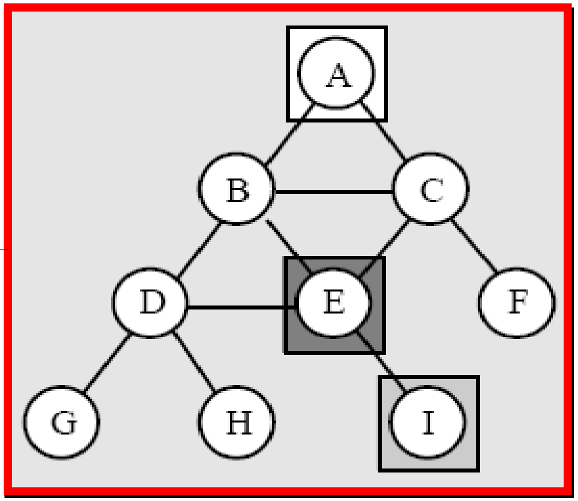
(a) $I(A, I \mid E)$



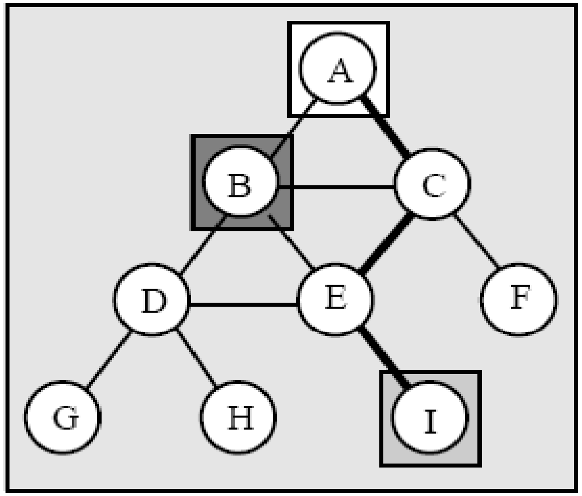
(b) $D(A, I \mid B)$

Links of the graph reflect **dependences** between the linked variables.

Non directed graphs define the conditional dependence through the **d-separation** concept.



(a) $I(A, I \mid E)$

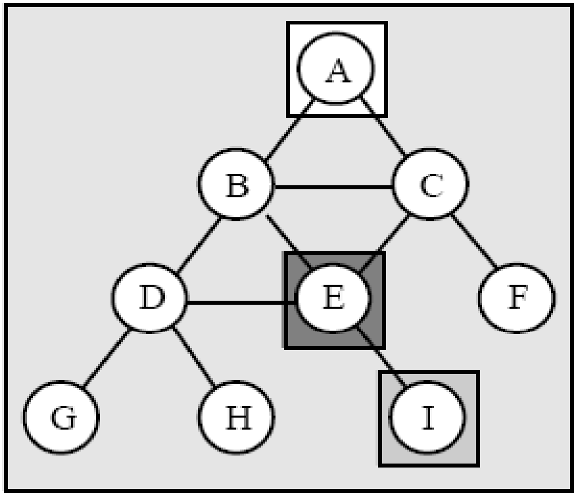


(b) $D(A, I \mid B)$

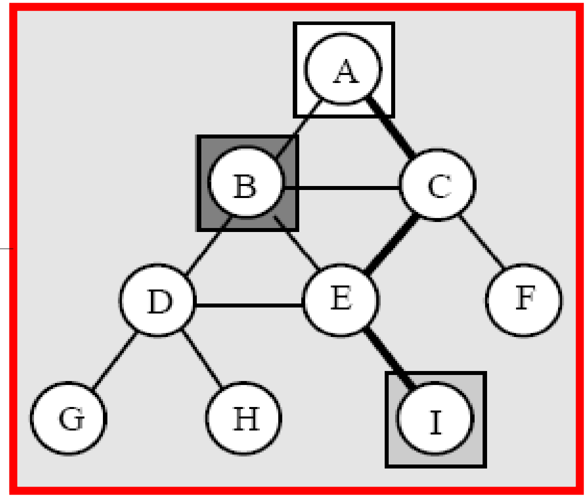
Links of the graph reflect **dependences** between the linked variables.

Non directed graphs define the conditional dependence through the **d-separation** concept.

There is not a path linking A and I not passing for E.
Thus A and I are dependent but conditional independent given E.



(a) $I(A, I \mid E)$

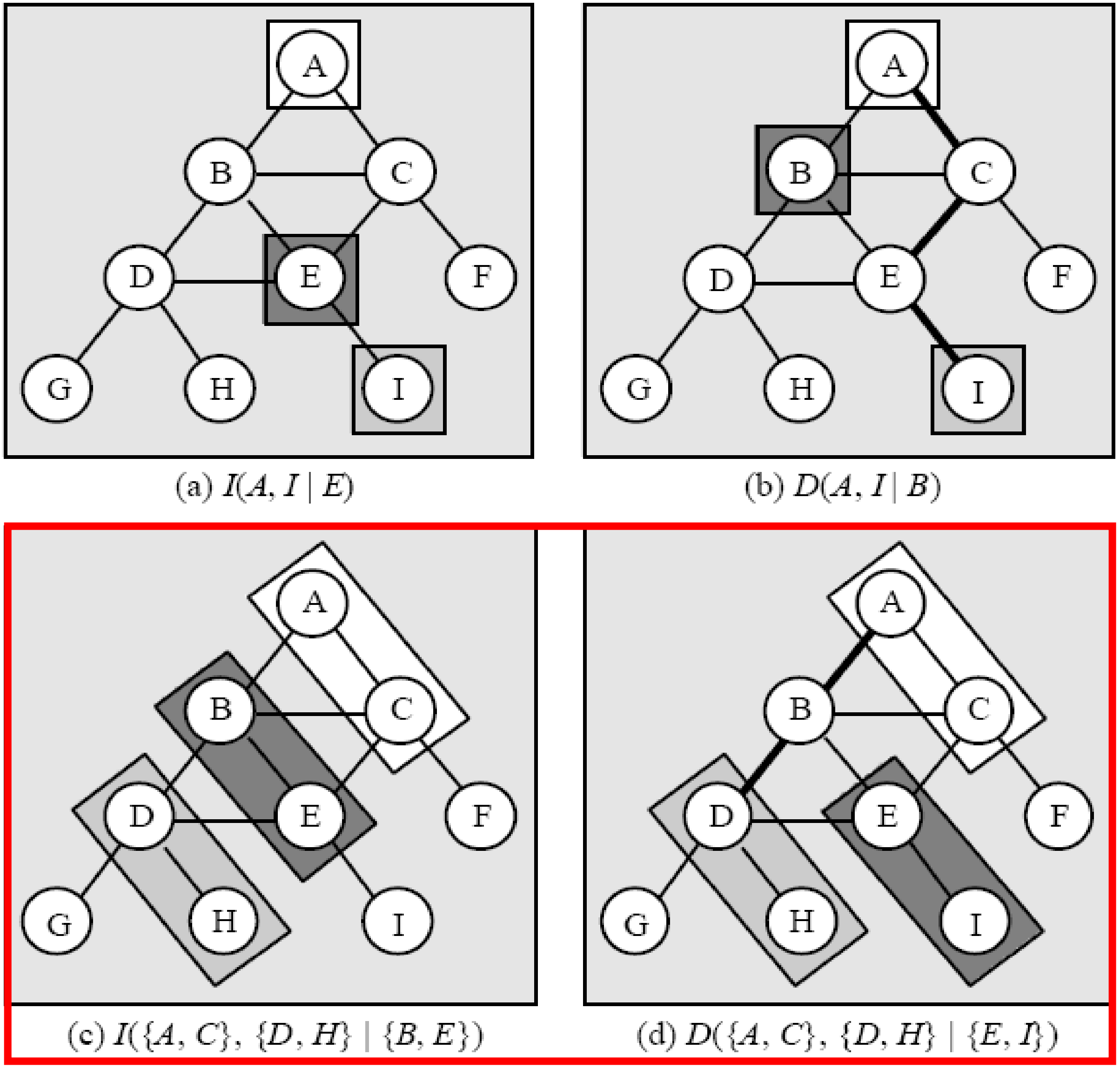


(b) $D(A, I \mid B)$

Links of the graph reflect **dependences** between the linked variables.

Non directed graphs define the conditional dependence through the **d-separation** concept.

There is a path linking A and I not passing for B (A->C->E->I).
Thus A and I are dependent given B and B doesn't d-separate A and I.



(a) $I(A, I \mid E)$

(b) $D(A, I \mid B)$

(c) $I(\{A, C\}, \{D, H\} \mid \{B, E\})$

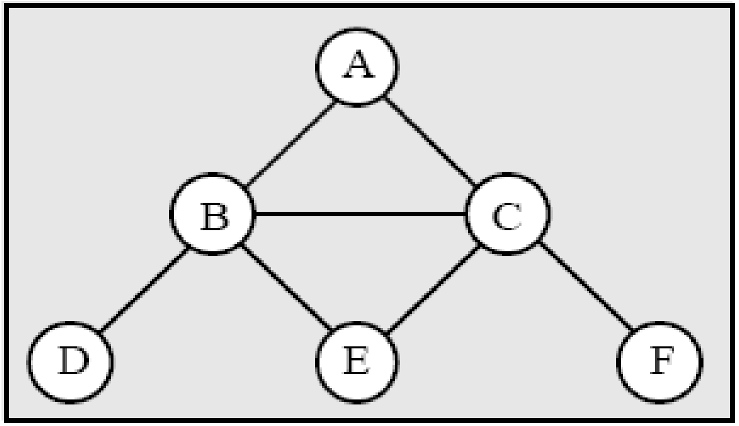
(d) $D(\{A, C\}, \{D, H\} \mid \{E, I\})$

Links of the graph reflect **dependences** between the linked variables.

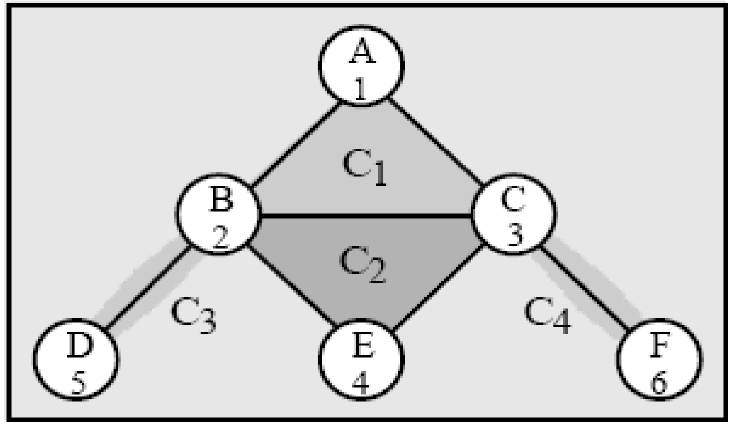
Non directed graphs define the conditional dependence through the **d-separation** concept.

D-separation is extended to set of variables.

Non-directed graphs define a graphical probabilistic model family based on the **cliques** of the graph and the factorization of the joint probability function given by them.



(a)



(b)

$$C_1 = \{A, B, C\}, \quad C_2 = \{B, C, E\},$$
$$C_3 = \{B, D\}, \quad C_4 = \{C, F\}.$$

i	Clique C_i	Separator S_i	Residual R_i
1	A, B, C	ϕ	A, B, C
2	B, C, E	B, C	E
3	B, D	B	D
4	C, F	C	F

$$\begin{aligned} p(a,b,c,d,e,f) &= \psi_1(c_1)\psi_2(c_2)\psi_3(c_3)\psi_4(c_4) \\ &= \psi_1(a,b,c)\psi_2(b,c,e)\psi_3(b,d)\psi_4(c,f). \end{aligned}$$

$$p(a,b,c,d,e,f) = \prod_{i=1}^{4} p(r_i|s_i) = p(a,b,c)p(e|b,c)p(d|b)p(f|c).$$

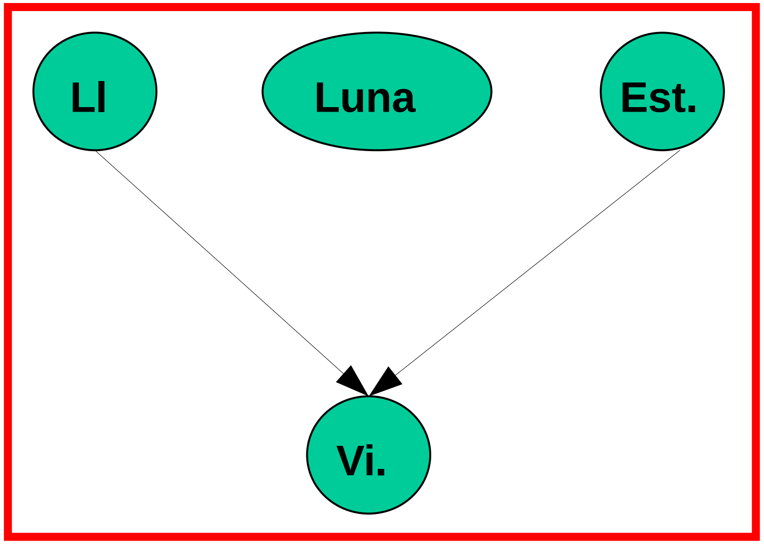
$$P_Z(Y|X)=P(Y|X,Z)=P(Y|Z)=P_Z(Y)\Rightarrow I(X,Y|Z)$$

	Anual		Invierno		Primavera		Verano		Otoño	
	S	Ll	S	Ll	S	Ll	S	Ll	S	Ll
NE	1014	516	190	99	287	166	360	162	177	89
SE	64	57	24	18	6	4	1	9	33	26
SW	225	661	98	223	18	119	15	71	94	248
NW	288	825	49	150	95	277	108	251	36	147
Total	1591	2059	361	490	406	566	484	493	340	510

P (Ll / Primavera) = 0.576
P (Ll / Invierno) = 0.582

Direct independence variables → Involve only two variables

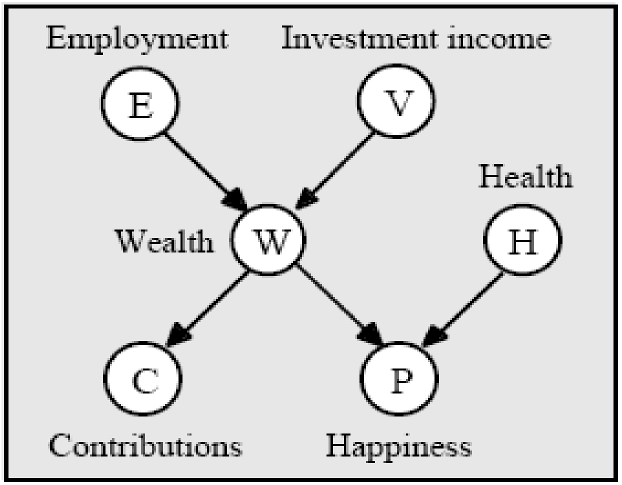
P (Ll) = 0.564

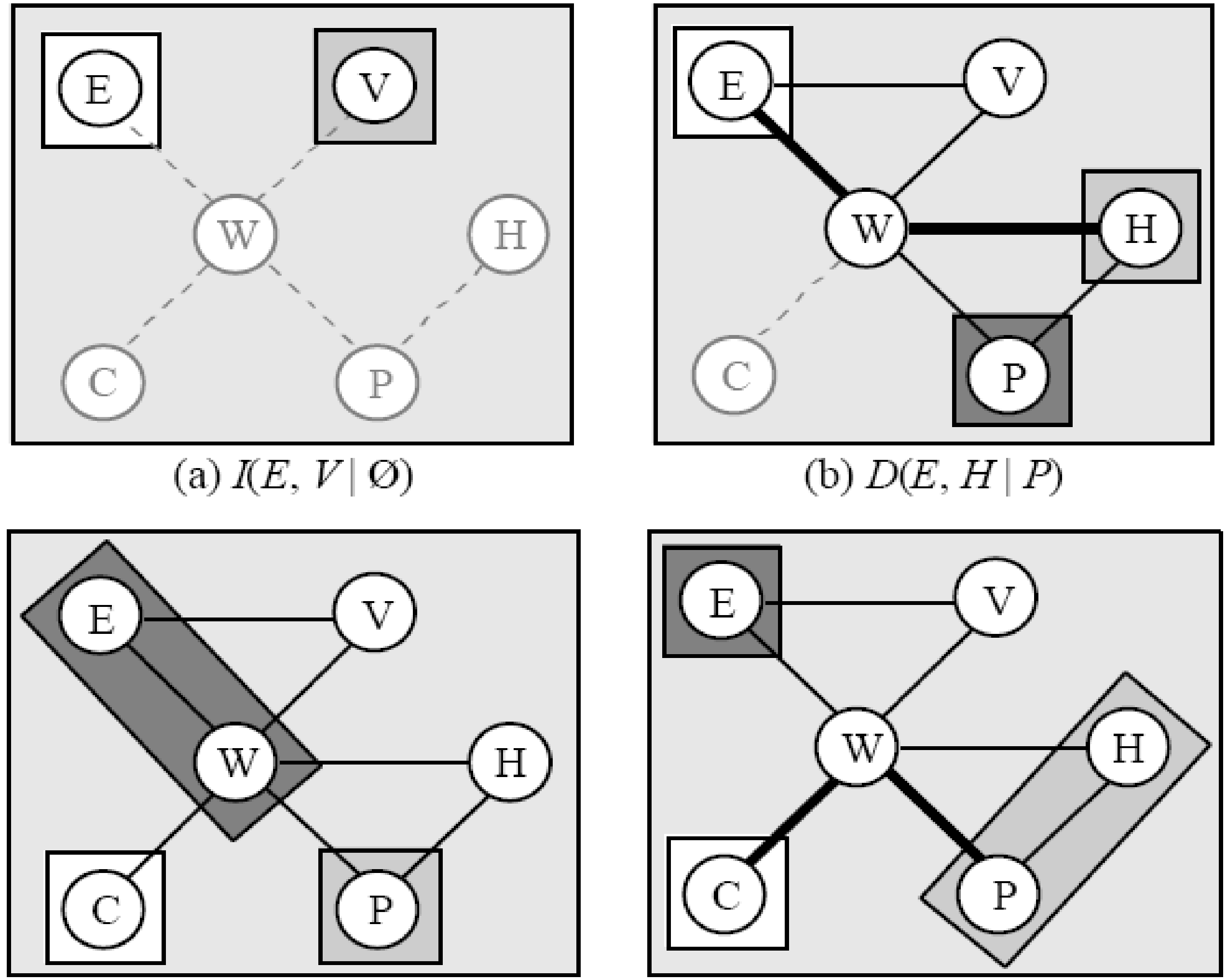


Non-directed graphs are not able to represent this kind of dependence!!!

Conditional dependence between rainfall and season, given the wind

D-separation concept for directed graphs enrich the representativity of the model → **Moral graph.**





(a) $I(E, V \mid \varnothing)$

(b) $D(E, H \mid P)$

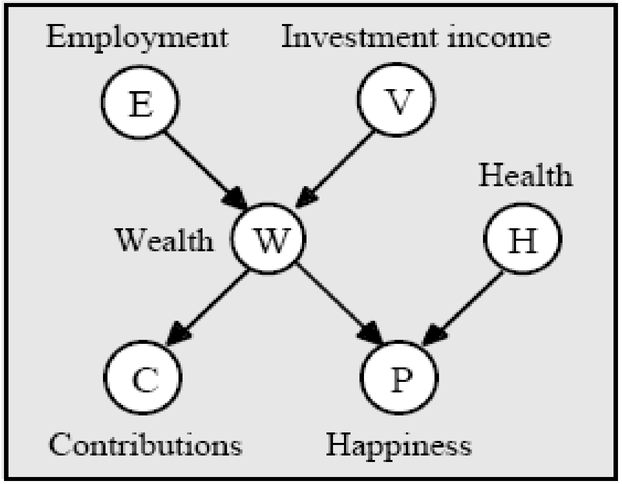
(c) $I(C, P \mid \{E, W\})$

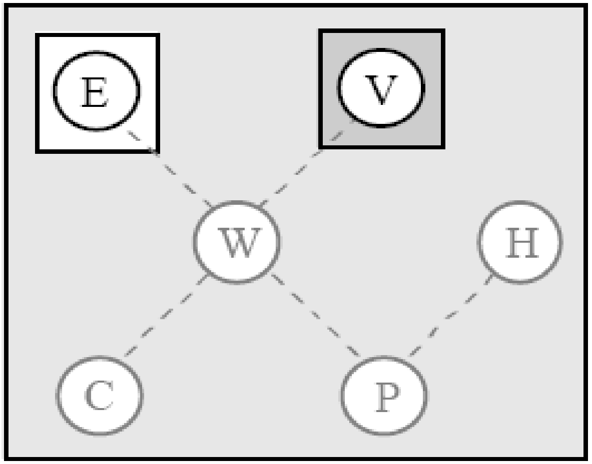
(d) $D(C, \{H, P\} \mid E)$

Links between variables imply probabilistic dependence **NOT CAUSALITY !!!!!!**

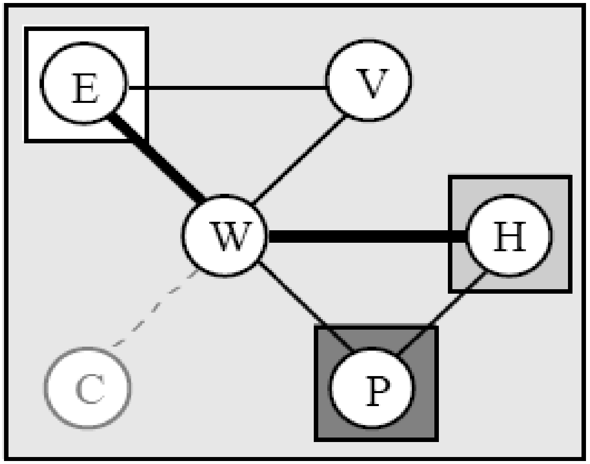
Causal Networks (not seen)

D-separation concept for directed graphs enrich the representativity of the model → **Moral graph.**





(a) $I(E, V \mid \emptyset)$



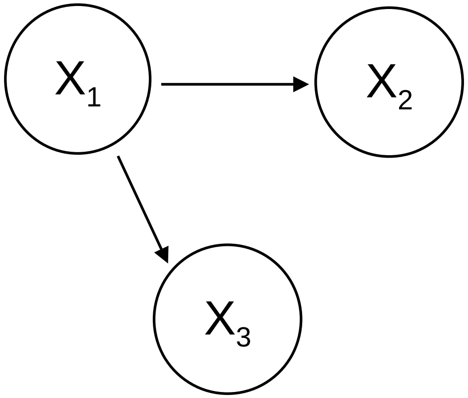
(b) $D(E, H \mid P)$

```
## Load bnlearn:
library(bnlearn)
## Defining an empty graph:
dag<-empty.graph(nodes=c("E","V","W","H","C","P"))
class(dag)
print(dag)
plot(dag)
## Adding link between nodes:
dag<-set.arc(dag,from="E",to="W")
dag<-set.arc(dag,from="V",to="W")
## Complete and plot the graph:
## Evaluate the separation included in the previous slide (See ? dsep and ?path):
```

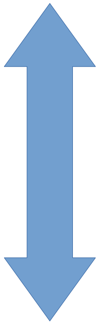
D-separation concept for directed graphs enrich the representativity of the model → **Moral graph.**

Two directed graph are **equivalents** when they lead to the same probabilistic model:

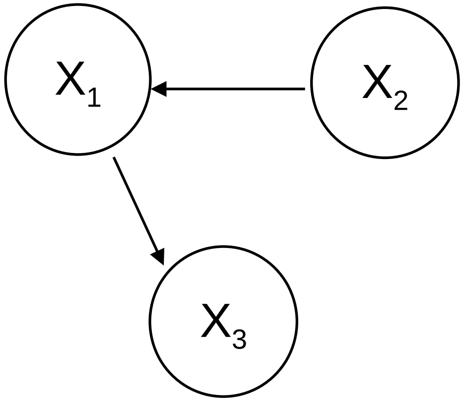
$$P(X_1,X_2,X_3)=P(X_1)P(X_2|X_1)P(X_3|X_1)=P(X_1,X_2)P(X_3|X_1)$$



Equivalents



$$P(X_1,X_2,X_3)=P(X_2)P(X_1|X_2)P(X_3|X_1)=P(X_1,X_2)P(X_3|X_1)$$



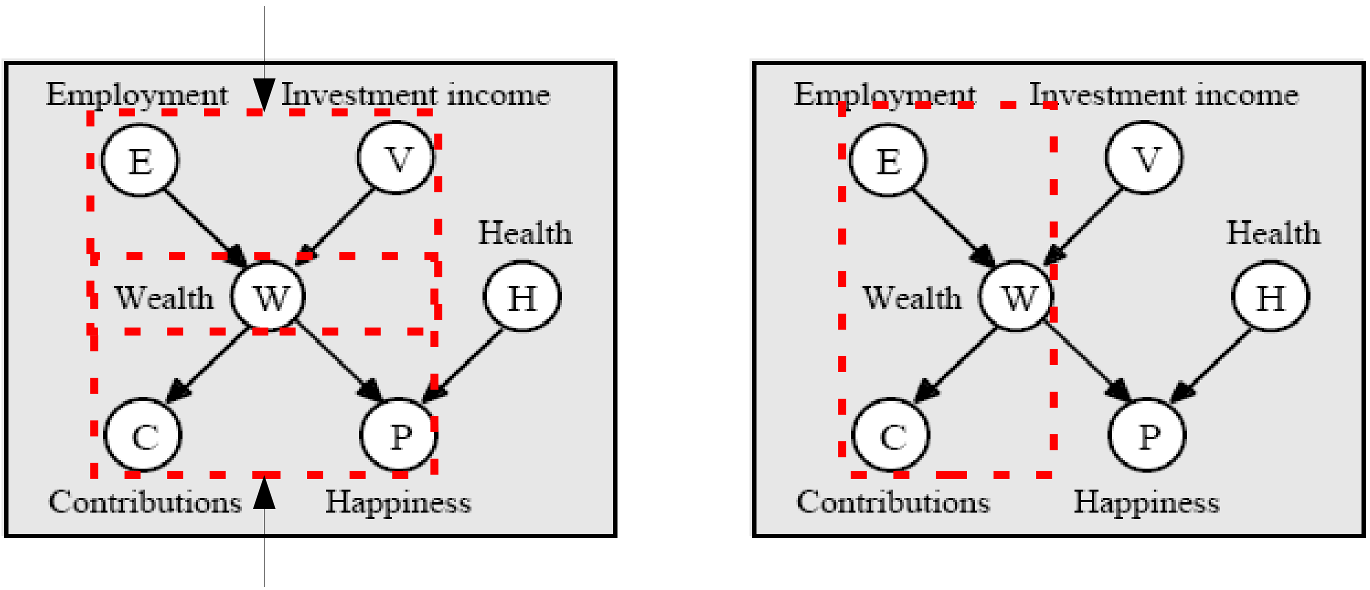
D-separation concept for directed graphs enrich the representativity of the model → **Moral graph.**

Two directed graph are **equivalents** when they lead to the same probabilistic model.

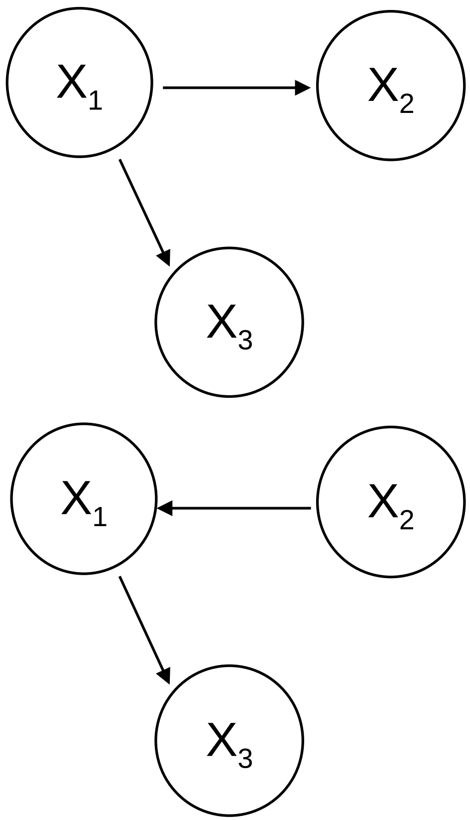
This occurs when the **subyacent non-directed graph** is the same and include the same **V-structures**.

$$P(X_1,X_2,X_3)=P(X_1,X_2)P(X_3|X_1)$$

Common effect



Common cause **Indirect evidential/causal effect**



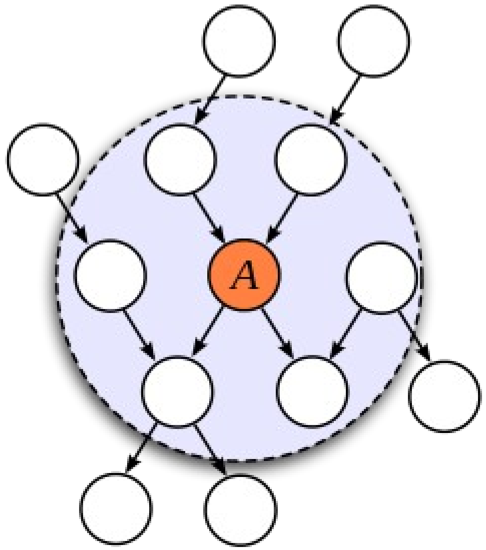
D-separation concept for directed graphs enrich the representativity of the model → **Moral graph.**

Two directed graph are **equivalents** when they lead to the same probabilistic model.

This occurs when the **subyacent non-directed graph** is the same and include the same **V-structures**.

The **Skeleton** of the graph is the undirected graph underlying.

The **Markov Blanket** of a node ***A*** is the set of nodes that completely separates ***A*** from the rest of the graph. In particular, it includes the parents and childrens of the node ***A***, and those children's other parents.



Source: Image from https://en.wikipedia.org/wiki/Markov_blanket

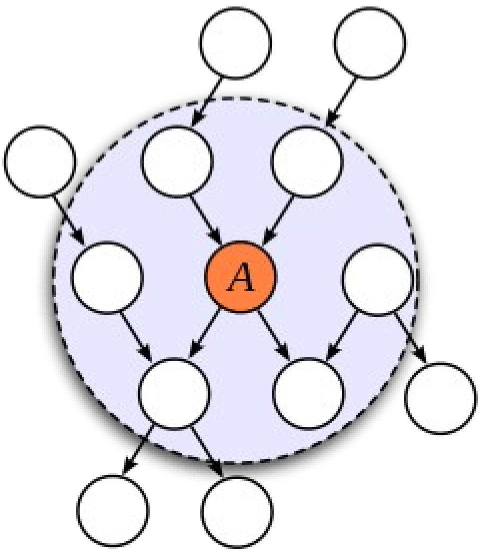
D-separation concept for directed graphs enrich the representativity of the model → **Moral graph.**

Two directed graph are **equivalents** when they lead to the same probabilistic model.

This occurs when the **subyacent non-directed graph** is the same and include the same **V-structures**.

The **Skeleton** of the graph is the undirected graph underlying.

The **Markov Blanket** of a node ***A*** is the set of nodes that completely separates ***A*** from the rest of the graph. In particular, it includes the parents and childrens of the node ***A***, and those children's other parents.



The **Markov Blanket** of is the set of nodes that includes all the knowledge needed to do inference on the node ***A***, from estimation to hypothesis testing to prediction.

Source: Image from https://en.wikipedia.org/wiki/Markov_blanket

Bayesian Networks obtain a compact representation of the joint probability function through the conditional independences.

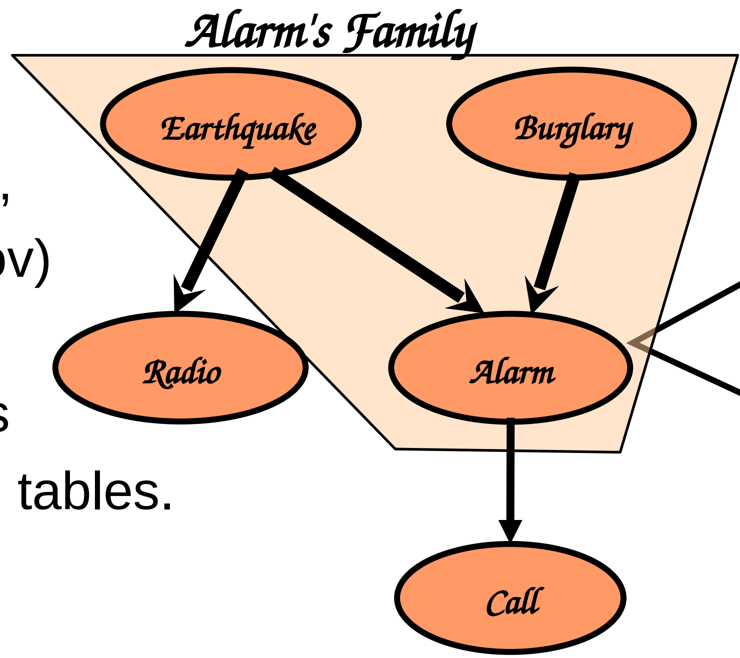
Structure:

Acyclic Directed Graph (DAG),

or non-directed graphs (Markov)

- Nodes – variables
- Links – direct dependences

Parameters: Probabilities and tables.



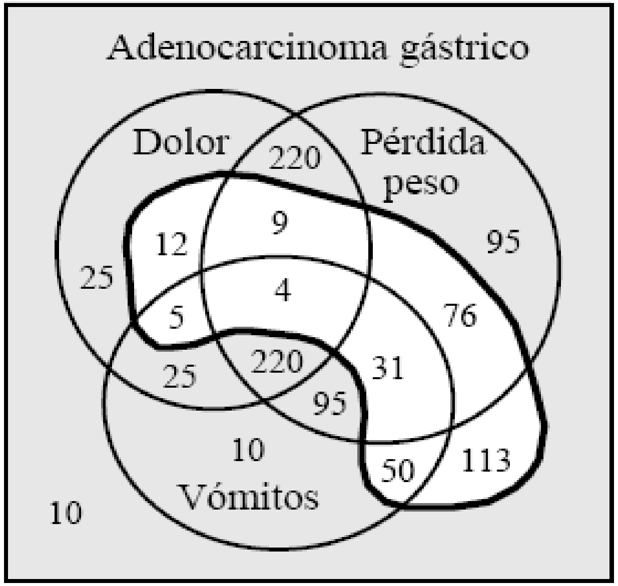
E	B	$P(A \mid E,B)$	
e	b	0.9	0.1
e	$\bar{b}$	0.2	0.8
$\bar{e}$	b	0.9	0.1
$\bar{e}$	$\bar{b}$	0.01	0.99

Once the Bayesian Network (**DAG**+**CPT**) is defined, some questions grow. In particular, given a **new evidence**

- Which is the probability of an event? ← **CPT-Inference**

- There are new (in)dependences between variables? ← **DAG-Inference**

...



Initial Probabilities:

Gray → Adenocarcinoma

White → Not Adenocarcinoma

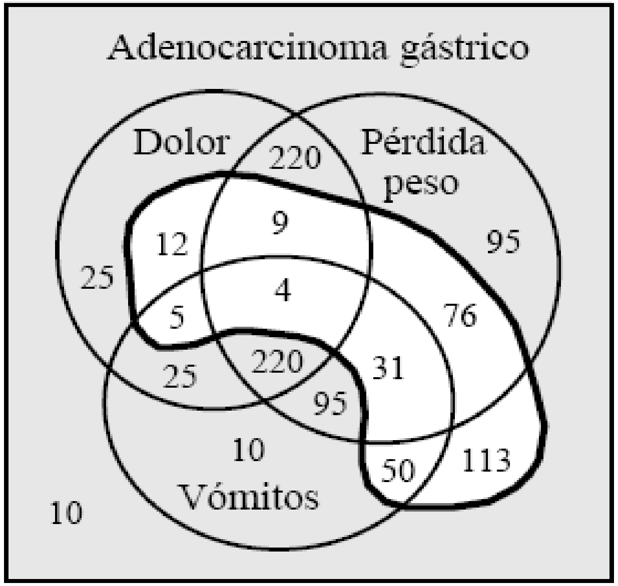
$$P(g)=\frac{700}{700+300}=\frac{700}{1000}=0.7$$

$$P(\neg g)=1-P(g)=1-0.7=0.3$$

Could we predict the probability of a disease based on the symptoms?

Bayes' Theorem (Predictands vs. Predictors), Factorization, etc.

$$\{X_1,\ldots,X_n : X_1\cup\ldots\cup X_n=M \wedge X_i\cap X_j=\emptyset\ \forall i\neq j\}\Rightarrow P(X_i|B)=\frac{P(B|X_i)P(X_i)}{\sum_{j=1}^{n}P(B|X_j)P(X_j)}$$



Initial Probabilities:

Gray $\rightarrow$ Adenocarcinoma

White $\rightarrow$ Not Adenocarcinoma

$$P(g)=\frac{700}{700+300}=\frac{700}{1000}=\boxed{0.7}$$

$$P(\neg g)=1-P(g)=1-0.7=0.3$$

Patient has suffered threw up:

$$\{V=v\}\Rightarrow P(g|v)=\frac{P(g)P(v|g)}{P(g)P(v|g)+P(\neg g)P(v|\neg g)}=\frac{0.7*0.5}{0.7*0.5+0.3*0.3}=\boxed{0.795}$$

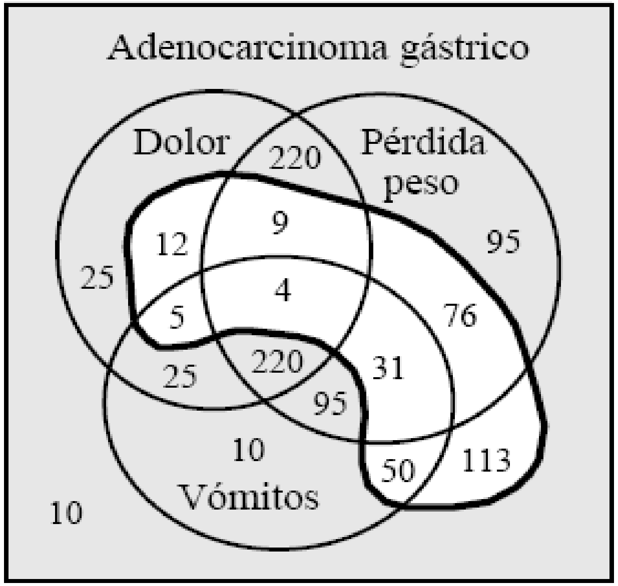
Initial Probabilities:

Gray $\rightarrow$ Adenocarcinoma

White $\rightarrow$ Not Adenocarcinoma

$$P(g)=\frac{700}{700+300}=\frac{700}{1000}=\boxed{0.7}$$

$$P(\neg g)=1-P(g)=1-0.7=0.3$$



Patient has suffered threw up:

$$\{V=v\}\Rightarrow P(g|v)=\frac{P(g)P(v|g)}{P(g)P(v|g)+P(\neg g)P(v|\neg g)}=\frac{0.7*0.5}{0.7*0.5+0.3*0.3}=\boxed{0.795}$$

Patient has suffered of weight loss and threw up:

$$\{P=p\wedge V=v\}\Rightarrow P(g|v,p)=\frac{P(g)P(v,p|g)}{P(g)P(v,p|g)+P(\neg g)P(v,p|\neg g)}=\frac{0.7*0.45}{0.7*0.45+0.3*0.12}=\boxed{0.9}$$

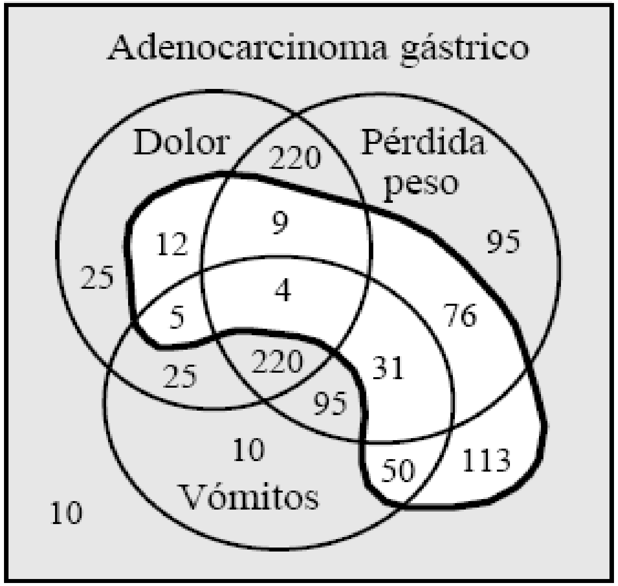
Initial Probabilities:

Gray → Adenocarcinoma

White → Not Adenocarcinoma

$$P(g)=\frac{700}{700+300}=\frac{700}{1000}=\boxed{0.7}$$

$$P(\neg g)=1-P(g)=1-0.7=0.3$$



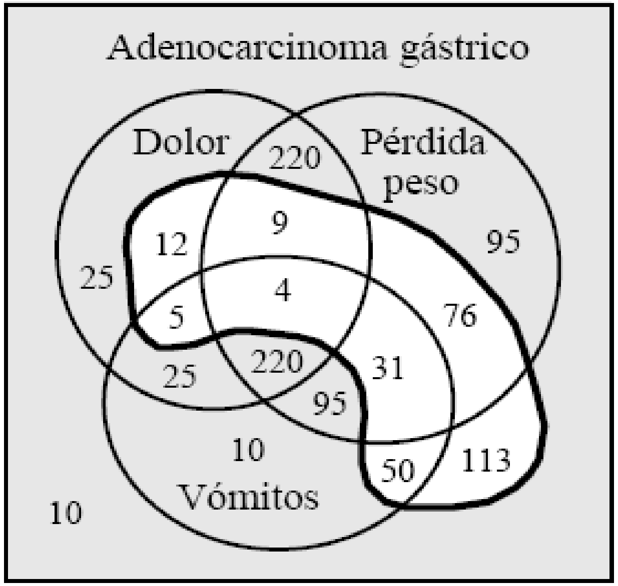
Patient has suffered threw up:

$$\{V=v\}\Rightarrow P(g|v)=\frac{P(g)P(v|g)}{P(g)P(v|g)+P(\neg g)P(v|\neg g)}=\frac{0.7*0.5}{0.7*0.5+0.3*0.3}=\boxed{0.795}$$

Patient has suffered of weight loss and threw up:

$$\{P=p\wedge V=v\}\Rightarrow P(g|v,p)=\frac{P(g)P(v,p|g)}{P(g)P(v,p|g)+P(\neg g)P(v,p|\neg g)}=\frac{0.7*0.45}{0.7*0.45+0.3*0.12}=\boxed{0.9}$$

Once the graph has been obtained, how change the probabilities when an evidence is given? Have we any method to estimate it efficiently?



Initial Probabilities:

Gray → Adenocarcinoma

White → Not Adenocarcinoma

$$P(g)=\frac{700}{700+300}=\frac{700}{1000}=\boxed{0.7}$$

$$P(\neg g)=1-P(g)=1-0.7=0.3$$

Significant changes in the probabilities reflect the dependence between predictand and predictors.

↓

Predictability

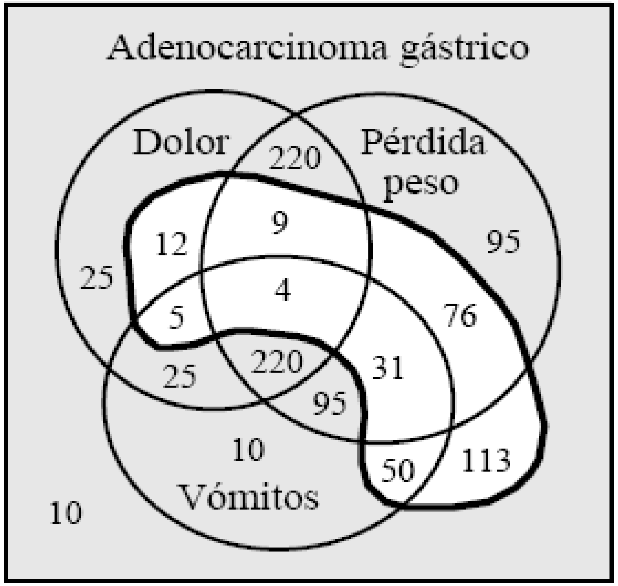
Could we predict the probability of a disease based on the symptoms?

Patient has suffered threw up:

$$\{V=v\}\Rightarrow P(g|v)=\frac{P(g)P(v|g)}{P(g)P(v|g)+P(\neg g)P(v|\neg g)}=\frac{0.7*0.5}{0.7*0.5+0.3*0.3}=\boxed{0.795}$$

Patient has suffered of weight loss and threw up:

$$\{P=p\wedge V=v\}\Rightarrow P(g|v,p)=\frac{P(g)P(v,p|g)}{P(g)P(v,p|g)+P(\neg g)P(v,p|\neg g)}=\frac{0.7*0.45}{0.7*0.45+0.3*0.12}=\boxed{0.9}$$



Initial Probabilities:

Gray → Adenocarcinoma

White → Not Adenocarcinoma

$$P(g)=\frac{700}{700+300}=\frac{700}{1000}=\boxed{0.7}$$

$$P(\neg g)=1-P(g)=1-0.7=0.3$$

Significant changes in the probabilities reflect the dependence between predictand and predictors.

↓

Predictability

Hypothesis Testing to Compare Two Population Proportions

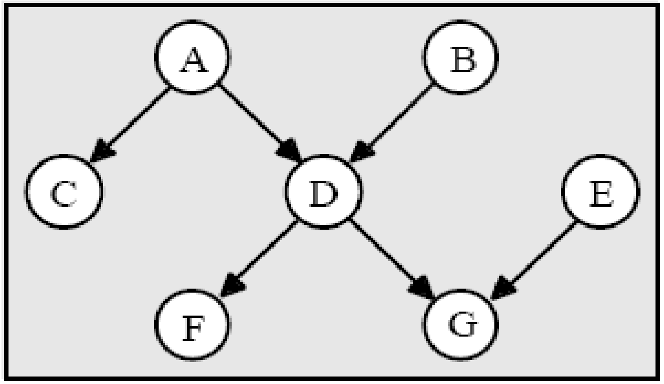
$$Z=\frac{p_1-p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$p=\frac{x_1+x_2}{n_1+n_2}$$

$$If\ Z>N_{(0,1)}^{-1}(\alpha)\Rightarrow p_1\neq p_2$$

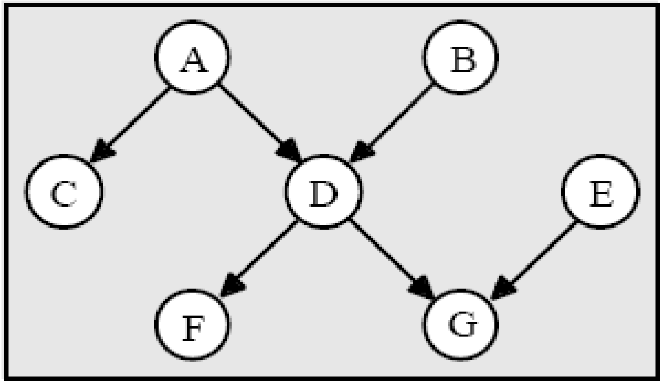
$$\boxed{p(x_i|e)} \quad E \subset X \quad X_i = e_i \quad X_i \in E$$

$$p(x) = p(a)p(b)p(c|a)p(d|a,b)p(e)p(f|d)p(g|d,e),$$



$$\boxed{p(x_i|e)} \quad E \subset X \quad X_i = e_i \quad X_i \in E$$

$$p(x) = p(a)p(b)p(c|a)p(d|a,b)p(e)p(f|d)p(g|d,e),$$



$$p(d) = \sum_{x \setminus d} p(x) = \sum_{a,b,c,e,f,g} p(a,b,c,d,e,f,g).$$

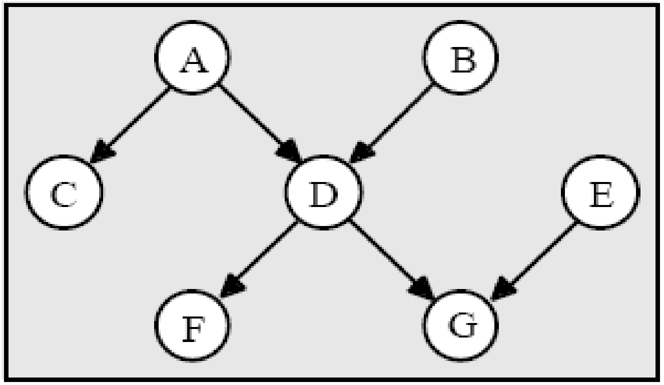
$$p(d) = \sum_{a,b,c,e,f,g} p(a)p(b)p(c|a)p(d|a,b)p(e)p(f|d)p(g|d,e)$$

$$= \left(\sum_{a,b,c} p(a)p(b)p(c|a)p(d|a,b)\right)\left(\sum_{e,f,g} p(e)p(g|d,e)p(f|d)\right),$$

$$\sum_a \left[p(a) \sum_c \left[p(c|a) \sum_b p(b)p(d|a,b)\right]\right] \sum_e \left[p(e) \sum_f \left[p(f|d) \sum_g p(g|d,e)\right]\right]$$

$$\boxed{p(x_i|e)} \quad E \subset X \quad X_i = e_i \quad X_i \in E$$

$$p(x) = p(a)p(b)p(c|a)p(d|a,b)p(e)p(f|d)p(g|d,e),$$



$$p(d) = \sum_{a,b,c,e,f,g} p(a)p(b)p(c|a)p(d|a,b)p(e)p(f|d)p(g|d,e)$$

$$= \left(\sum_{a,b,c} p(a)p(b)p(c|a)p(d|a,b)\right)\left(\sum_{e,f,g} p(e)p(g|d,e)p(f|d)\right),$$

$$\sum_a \left[p(a)\sum_c \left[p(c|a)\sum_b p(b)p(d|a,b)\right]\right]\sum_e \left[p(e)\sum_f \left[p(f|d)\sum_g p(g|d,e)\right]\right]$$

Moralized non-directed graph is obtained and efficient graphs algorithms are applied to obtain the new probabilities. → Exact Inference

Exact inference suffers when the graph is dense (hyper-conected) or there are many variables in the model, losing most of their efficiency and making more adequate the use of aproximated algorithms based on simulation.
Herer we include a brief description of the general approach used by this algorithms:

Input: Real probability function P(X) and distribution considered for the simulation h(X) (e.g. uniform), sample size N and a subset Y⊂ X.
Output: Approximated value for P(y) for y in Y.

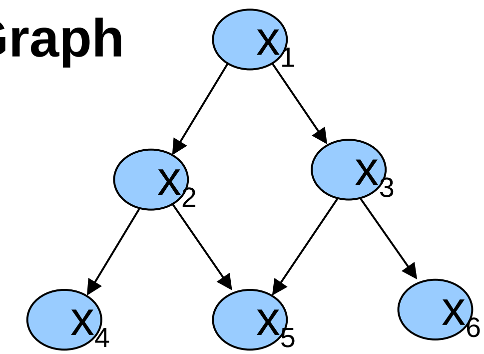
1. For $j=1 .. N$
   - Generate $x^j = (x^j_1,...,x^j_n)$ from $h(x)$.
   - Estimate $s(x^j) = p(x^j)/ h(x^j)$.

2. For each ***y***, estimate $P(y) \approx \Sigma_y\, s(x^j) / \Sigma_j\, s(x^j)$

1. For j=1 .. N
   - Generate $x^j = (x^j_1,\ldots,x^j_n)$ from h(x).
   - Estimate $s(x^j) = p(x^j)/\, h(x^j)$.

2. For each ***y***, estimate $P(y) \approx \Sigma_y\, s(x^j) \,/\, \Sigma_j\, s(x^j)$

Graph



Joint Probability Function

$$P(X_1,\ldots,X_6)=P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2)P(X_5|X_2,X_3)P(X_6|X_3)$$

x_1	$p(x_1)$
0	0.3
1	0.7

x_1	x_2	$p(x_2\|x_1)$
0	0	0.4
0	1	0.6
1	0	0.1
1	1	0.9

x_1	x_3	$p(x_3\|x_1)$
0	0	0.2
0	1	0.8
1	0	0.5
1	1	0.5

x_2	x_4	$p(x_4\|x_2)$
0	0	0.3
0	1	0.7
1	0	0.2
1	1	0.8

x_3	x_6	$p(x_6\|x_3)$
0	0	0.1
0	1	0.9
1	0	0.4
1	1	0.6

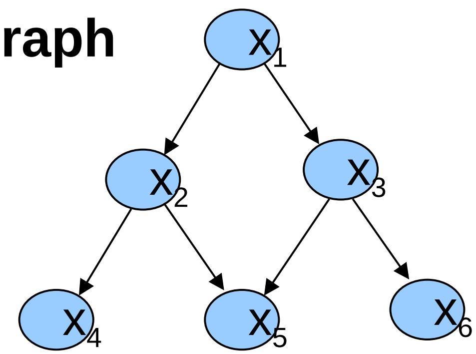
x_2	x_3	x_5	$p(x_5\|x_2,x_3)$
0	0	0	0.4
0	0	1	0.6
0	1	0	0.5
0	1	1	0.5
1	0	0	0.7
1	0	1	0.3
1	1	0	0.2
1	1	1	0.8

For example, for the event (0,1,1,1,0,0) this is the probability:
$p(0,1,1,1,0,0) = p(x_1=0)p(x_2=1|x_1=0)p(x_3=1|x_1=0)p(x_4=1|x_2=1)$
$p(x_5=0|x_2=1,x_3=1)p(x_6=0|x_3=1) = 0.3 \times 0.6 \times 0.8 \times 0.8 \times 0.2 \times 0.4 = 0.009216$

1. For j=1 .. N
   - Generate $x^j = (x^j_1,\ldots,x^j_n)$ from h(x).
   - Estimate $s(x^j) = p(x^j)/ h(x^j)$.

2. For each ***y***, estimate $P(y) \approx \Sigma_y\, s(x^j) / \Sigma_j\, s(x^j)$

Graph



Joint Probability Function

$$P(X_1,\ldots,X_6)=P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2)P(X_5|X_2,X_3)P(X_6|X_3)$$

x_1	$p(x_1)$
0	0.3
1	0.7

x_1	x_2	$p(x_2\|x_1)$
0	0	0.4
0	1	0.6
1	0	0.1
1	1	0.9

x_1	x_3	$p(x_3\|x_1)$
0	0	0.2
0	1	0.8
1	0	0.5
1	1	0.5

x_2	x_4	$p(x_4\|x_2)$
0	0	0.3
0	1	0.7
1	0	0.2
1	1	0.8

x_3	x_6	$p(x_6\|x_3)$
0	0	0.1
0	1	0.9
1	0	0.4
1	1	0.6

x_2	x_3	x_5	$p(x_5\|x_2,x_3)$
0	0	0	0.4
0	0	1	0.6
0	1	0	0.5
0	1	1	0.5
1	0	0	0.7
1	0	1	0.3
1	1	0	0.2
1	1	1	0.8

Six binary variables → 2^6=64 posibilities → Suppose ***h*** uniform → h(x)=1/64

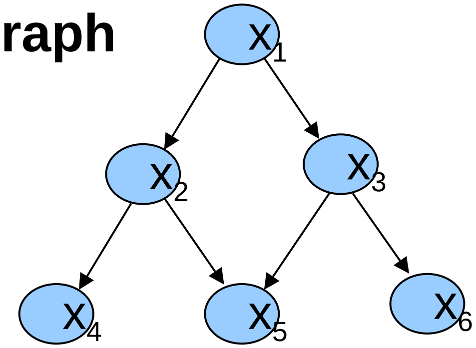
Step 1

Realization x^j	$p(x^j)$	$h(x^j)$	$s(x^j)= p(x^j)/h(x^j)$
x^1=(0,1,1,1,0,0)	0.0092	1/64	0.5898
x^2=(1,1,0,1,1,0)	0.0076	1/64	0.4838
x^3=(0,0,1,0,0,1)	0.0086	1/64	0.5529
x^4=(1,0,0,1,1,0)	0.0015	1/64	0.0941
x^5=(1,0,0,0,1,1)	0.0057	1/64	0.3629

1. For j=1 .. N
   - Generate $x^j = (x^j_1,...,x^j_n)$ from $h(x)$.
   - Estimate $s(x^j) = p(x^j)/ h(x^j)$.

2. For each ***y***, estimate $P(y) \approx \Sigma_y\, s(x^j) / \Sigma_j\, s(x^j)$

Graph



Joint Probability Function

$$P(X_1,\dots,X_6)=P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2)P(X_5|X_2,X_3)P(X_6|X_3)$$

x_1	$p(x_1)$
0	0.3
1	0.7

x_1	x_2	$p(x_2\|x_1)$
0	0	0.4
0	1	0.6
1	0	0.1
1	1	0.9

x_1	x_3	$p(x_3\|x_1)$
0	0	0.2
0	1	0.8
1	0	0.5
1	1	0.5

x_2	x_4	$p(x_4\|x_2)$
0	0	0.3
0	1	0.7
1	0	0.2
1	1	0.8

x_3	x_6	$p(x_6\|x_3)$
0	0	0.1
0	1	0.9
1	0	0.4
1	1	0.6

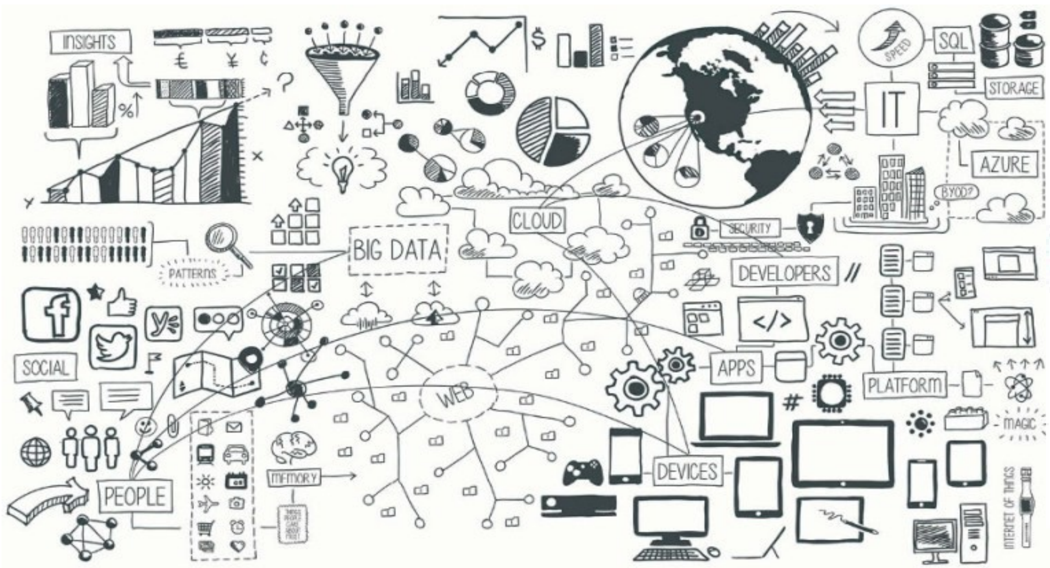
Step 2

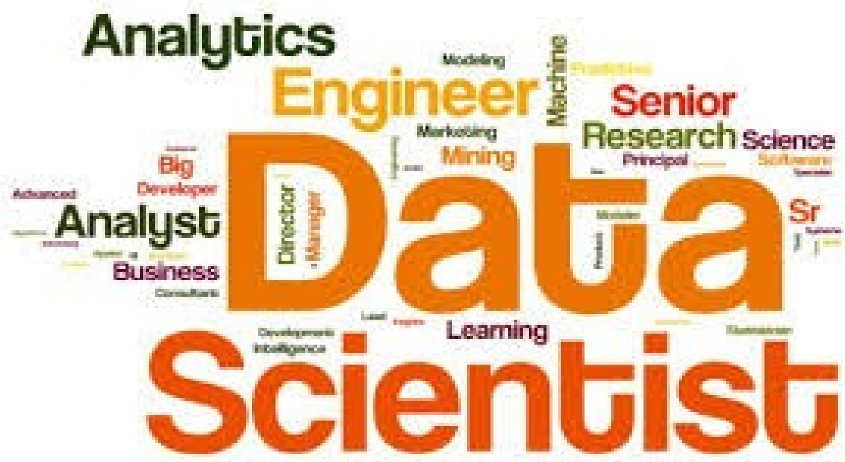
Realization x^j	$p(x^j)$	$h(x^j)$	$s(x^j)= p(x^j)/h(x^j)$
$x^1=(0,1,1,1,0,0)$	0.0092	1/64	0.5898
$x^2=(1,1,0,1,1,0)$	0.0076	1/64	0.4838
$x^3=(0,0,1,0,0,1)$	0.0086	1/64	0.5529
$x^4=(1,0,0,1,1,0)$	0.0015	1/64	0.0941
$x^5=(1,0,0,0,1,1)$	0.0057	1/64	0.3629

Poor estimation due to the number of simulations (5)

$p(X_1=0)\approx[s(x^1)+s(x^3)]/\Sigma_j s(x^j)=[0.5898+0.5529]/2.0835=$ **0.5485**

M1970 – Machine Learning II
Redes Probabilísticas Discretas
(Clasificadores Bayesianos)





Sixto Herrera (sixto.herrera@unican.es)

Grupo de Meteorología

Univ. de Cantabria – CSIC
MACC / IFCA

$$P:\ \mathbf{M} \longrightarrow [0,1]$$
$$A \longmapsto a$$

	Anual		Invierno		Primavera		Verano		Otoño	
	S	Ll	S	Ll	S	Ll	S	Ll	S	Ll
NE	1014	516	190	99	287	166	360	162	177	89
SE	64	57	24	18	6	4	1	9	33	26
SW	225	661	98	223	18	119	15	71	94	248
NW	288	825	49	150	95	277	108	251	36	147
Total	1591	2059	361	490	406	566	484	493	340	510

$$P(X) \in [0,1], X \subseteq M$$
$$P(\emptyset) = 0 \wedge P(M) = 1$$
$$P(X_1 \cup X_2) = P(X_1) + P(X_2) - P(X_1 \cap X_2)$$
$$P(X_1 \cap X_2) = P(X_1) * P(X_2) \Rightarrow X_1 \wedge X_2 \, independent$$

```
## States of the variables:
estados.Wind < - c("NE","SE","SW","NW")
estados.Season < - c("Anual","Invierno","Primavera","Verano","Otono")
estados.Precip < - c("Seco","Lluvioso")

## Table of Absolute frequencies:
table.freq < - array(c(1014, 64, 225, 288, 190, 24, 98, 49, 287, 6, 18, 95, 360, 1, 15, 108,
                    177, 33, 94, 36, 516, 57, 661, 825, 99, 18, 223, 150,
                      166, 4, 119, 277, 162, 9, 71, 251, 89, 26, 248, 147), dim = c(4,5,2),
               dimnames = list(W=estados.Wind, S=estados.Season, P = estados.Precip))
```

$$P: \mathbf{M} \longrightarrow [0,1]$$
$$A \longrightarrow a$$

	Anual		Invierno		Primavera		Verano		Otoño	
	S	Ll	S	Ll	S	Ll	S	Ll	S	Ll
NE	1014	516	190	99	287	166	360	162	177	89
SE	64	57	24	18	6	4	1	9	33	26
SW	225	661	98	223	18	119	15	71	94	248
NW	288	825	49	150	95	277	108	251	36	147
Total	1591	2059	361	490	406	566	484	493	340	510

$$P(X) \in [0,1], X \subseteq M$$
$$P(\emptyset) = 0 \wedge P(M) = 1$$
$$P(X_1 \cup X_2) = P(X_1) + P(X_2) - P(X_1 \cap X_2)$$
$$P(X_1 \cap X_2) = P(X_1) * P(X_2) \Rightarrow X_1 \wedge X_2 \textit{ independent}$$

$$X = \{Inv, \mathrm{Ll}, NW\} \Rightarrow P(X) = \frac{freq(Inv, \mathrm{Ll}, NW)}{N} = \frac{150}{3650} = 0.041$$

```
## States of the variables:
estados.Wind < - c("NE","SE","SW","NW")
estados.Season < - c("Anual","Invierno","Primavera","Verano","Otono")
estados.Precip < - c("Seco","Lluvioso")

## Table of Absolute frequencies:
table.freq < - array(c(1014, 64, 225, 288, 190, 24, 98, 49, 287, 6, 18, 95, 360, 1, 15, 108,
                       177, 33, 94, 36, 516, 57, 661, 825, 99, 18, 223, 150,
                        166, 4, 119, 277, 162, 9, 71, 251, 89, 26, 248, 147), dim = c(4,5,2),
                  dimnames = list(W=estados.Wind, S=estados.Season, P = estados.Precip))
## Obtain the probability:
table.freq["NW","Invierno","Lluvioso"]/sum(table.freq[,"Anual",])
```

P: **M** $\longrightarrow$ [0,1]

A $\longrightarrow$ a

	Anual		Invierno		Primavera		Verano		Otoño	
	S	Ll	S	Ll	S	Ll	S	Ll	S	Ll
NE	1014	516	190	99	287	166	360	162	177	89
SE	64	57	24	18	6	4	1	9	33	26
SW	225	661	98	223	18	119	15	71	94	248
NW	288	825	49	150	95	277	108	251	36	147
Total	1591	2059	361	490	406	566	484	493	340	510

$$P(X)\in[0,1], X\subseteq M$$

$$P(\emptyset)=0 \wedge P(M)=1$$

$$P(X_1\cup X_2)=P(X_1)+P(X_2)-P(X_1\cap X_2)$$

$$P(X_1\cap X_2)=P(X_1)*P(X_2)\Rightarrow X_1\wedge X_2 \, independent$$

$$X=\{Inv,\mathrm{Ll},NW\}\Rightarrow P(X)=\frac{freq(Inv,\mathrm{Ll},NW)}{N}=\frac{150}{3650}=0.041$$

$$\boxed{X=\{Inv\}}\Rightarrow P(X)=\frac{freq(Inv)}{N}=\frac{\sum_{p\in Pr}\sum_{v\in Vi} freq(Inv,p,v)}{N}=0.233$$

```
## Obtain the probability:
sum(table.freq[,"Invierno",])/sum(table.freq[,"Anual",])
```

P: **M** → [0,1]

A → a

$$P(X) \in [0,1], X \subseteq M$$

$$P(\emptyset) = 0 \wedge P(M) = 1$$

$$P(X_1 \cup X_2) = P(X_1) + P(X_2) - P(X_1 \cap X_2)$$

$$P(X_1 \cap X_2) = P(X_1) * P(X_2) \Rightarrow X_1 \wedge X_2 \, independent$$

	Anual		Invierno		Primavera		Verano		Otoño	
	S	Ll	S	Ll	S	Ll	S	Ll	S	Ll
NE	1014	516	190	99	287	166	360	162	177	89
SE	64	57	24	18	6	4	1	9	33	26
SW	225	661	98	223	18	119	15	71	94	248
NW	288	825	49	150	95	277	108	251	36	147
Total	1591	2059	361	490	406	566	484	493	340	510

$$X = \{Inv, \mathrm{Ll}, NW\} \Rightarrow P(X) = \frac{freq(Inv, \mathrm{Ll}, NW)}{N} = \frac{150}{3650} = 0.041$$

$$X = \{Inv\} \Rightarrow P(X) = \frac{freq(Inv)}{N} = \frac{\sum_{p \in Pr} \sum_{v \in Vi} freq(Inv, p, v)}{N} = 0.233$$

$$X = \{NW\} \Rightarrow P(Y|X) = \frac{P(Y, X)}{P(X)}$$

$$Y = \{Inv\} \Rightarrow P(Y|X) = \frac{freq(Y, X)}{freq(X)} = \frac{199}{1113} = 0.179$$

P: **M** $\longrightarrow$ [0,1]

A $\longrightarrow$ a

	Anual		Invierno		Primavera		Verano		Otoño	
	S	Ll	S	Ll	S	Ll	S	Ll	S	Ll
NE	1014	516	190	99	287	166	360	162	177	89
SE	64	57	24	18	6	4	1	9	33	26
SW	225	661	98	223	18	119	15	71	94	248
NW	288	825	49	150	95	277	108	251	36	147
Total	1591	2059	361	490	406	566	484	493	340	510

$$P(X)\in[0,1], X\subseteq M$$

$$P(\emptyset)=0 \wedge P(M)=1$$

$$P(X_1\cup X_2)=P(X_1)+P(X_2)-P(X_1\cap X_2)$$

$$P(X_1\cap X_2)=P(X_1)*P(X_2)\Rightarrow X_1\wedge X_2 \, independent$$

$$X=\{Inv,\mathrm{Ll},NW\}\Rightarrow P(X)=\frac{freq(Inv,\mathrm{Ll},NW)}{N}=\frac{150}{3650}=0.041$$

$$X=\{Inv\}\Rightarrow P(X)=\frac{freq(Inv)}{N}=\frac{\sum_{p\in Pr}\sum_{v\in Vi} freq(Inv,p,v)}{N}=0.233$$

$$\boxed{X=\{NW\}\Rightarrow P(Y|X)=\frac{P(Y,X)}{P(X)}}$$

$$Y=\{Inv\}\Rightarrow P(Y|X)=\frac{freq(Y,X)}{freq(X)}=\frac{199}{1113}=0.179$$

```
## New probability-space:
cond.table.freq <- table.freq["NW",,]
print(cond.table.freq)
```

$$P: \mathbf{M} \longrightarrow [0,1]$$
$$A \longrightarrow a$$

	Anual		Invierno		Primavera		Verano		Otoño	
	S	Ll	S	Ll	S	Ll	S	Ll	S	Ll
NE	1014	516	190	99	287	166	360	162	177	89
SE	64	57	24	18	6	4	1	9	33	26
SW	225	661	98	223	18	119	15	71	94	248
NW	288	825	49	150	95	277	108	251	36	147
Total	1591	2059	361	490	406	566	484	493	340	510

$$P(X) \in [0,1], X \subseteq M$$
$$P(\emptyset)=0 \wedge P(M)=1$$
$$P(X_1 \cup X_2)=P(X_1)+P(X_2)-P(X_1 \cap X_2)$$
$$P(X_1 \cap X_2)=P(X_1)*P(X_2) \Rightarrow X_1 \wedge X_2 \, independent$$

$$X=\{Inv, \mathrm{Ll}, NW\} \Rightarrow P(X)=\frac{freq(Inv, \mathrm{Ll}, NW)}{N}=\frac{150}{3650}=0.041$$

$$X=\{Inv\} \Rightarrow P(X)=\frac{freq(Inv)}{N}=\frac{\sum_{p \in Pr} \sum_{v \in Vi} freq(Inv, p, v)}{N}=0.233$$

$$\boxed{X=\{NW\} \Rightarrow P(Y|X)=\frac{P(Y,X)}{P(X)}}$$

$$Y=\{Inv\} \Rightarrow P(Y|X)=\frac{freq(Y,X)}{freq(X)}=\frac{199}{1113}=0.179$$

```
## Obtain the probability:
sum(cond.table.freq["Invierno",])/sum(cond.table.freq["Anual",])
```

$$P: \mathbf{M} \longrightarrow [0,1]$$
$$A \longmapsto a$$

	Anual		Invierno		Primavera		Verano		Otoño	
	S	Ll	S	Ll	S	Ll	S	Ll	S	Ll
NE	1014	516	190	99	287	166	360	162	177	89
SE	64	57	24	18	6	4	1	9	33	26
SW	225	661	98	223	18	119	15	71	94	248
NW	288	825	49	150	95	277	108	251	36	147
Total	1591	2059	361	490	406	566	484	493	340	510

$$P(X)\in[0,1], X\subseteq M$$
$$P(\emptyset)=0 \wedge P(M)=1$$
$$P(X_1\cup X_2)=P(X_1)+P(X_2)-P(X_1\cap X_2)$$
$$P(X_1\cap X_2)=P(X_1)*P(X_2)\Rightarrow X_1\wedge X_2 \textit{ independent} \Leftarrow P(X_1|X_2)=P(X_1)\wedge P(X_2|X_1)=P(X_2)$$

Bayes' Theorem (Predictands vs. Predictors), Factorization, etc.

$$\{X_1,\dots,X_n: X_1\cup\dots\cup X_n=M\wedge X_i\cap X_j=\emptyset\ \forall i\neq j\}\Rightarrow P(X_i|B)=\frac{P(B|X_i)P(X_i)}{P(B)}$$

$$X=\{NW\}\Rightarrow P(Y|X)=\frac{P(Y,X)}{P(X)}$$

$$Y=\{Inv\}\Rightarrow P(Y|X)=\frac{freq(Y,X)}{freq(X)}=\frac{199}{1113}=0.179$$

$$P: \mathbf{M} \longrightarrow [0,1]$$
$$A \longmapsto a$$

	Anual		Invierno		Primavera		Verano		Otoño	
	S	Ll	S	Ll	S	Ll	S	Ll	S	Ll
NE	1014	516	190	99	287	166	360	162	177	89
SE	64	57	24	18	6	4	1	9	33	26
SW	225	661	98	223	18	119	15	71	94	248
NW	288	825	49	150	95	277	108	251	36	147
Total	1591	2059	361	490	406	566	484	493	340	510

$$P(X) \in [0,1], X \subseteq M$$
$$P(\emptyset)=0 \wedge P(M)=1$$
$$P(X_1 \cup X_2)=P(X_1)+P(X_2)-P(X_1 \cap X_2)$$
$$P(X_1 \cap X_2)=P(X_1) * P(X_2) \Rightarrow X_1 \wedge X_2 \, independent \Leftarrow P(X_1|X_2)=P(X_1) \wedge P(X_2|X_1)=P(X_2)$$

Bayes' Theorem (Predictands vs. Predictors), Factorization, etc.

Probability "***a priori***"

$$\{X_1,\dots,X_n : X_1 \cup \dots \cup X_n = M \wedge X_i \cap X_j = \emptyset \; \forall i \neq j\} \Rightarrow \boxed{P(X_i|B)} = \frac{P(B|X_i)\boxed{P(X_i)}}{P(B)}$$

Probability "***a posteriori***"

$$X=\{NW\} \Rightarrow P(Y|X)=\frac{P(Y,X)}{P(X)}$$

$$Y=\{Inv\} \Rightarrow P(Y|X)=\frac{freq(Y,X)}{freq(X)}=\frac{199}{1113}=0.179$$

$$P: \mathbf{M} \longrightarrow [0,1]$$
$$A \longmapsto a$$

	Anual		Invierno		Primavera		Verano		Otoño	
	S	Ll	S	Ll	S	Ll	S	Ll	S	Ll
NE	1014	516	190	99	287	166	360	162	177	89
SE	64	57	24	18	6	4	1	9	33	26
SW	225	661	98	223	18	119	15	71	94	248
NW	288	825	49	150	95	277	108	251	36	147
Total	1591	2059	361	490	406	566	484	493	340	510

$$P(X)\in[0,1], X\subseteq M$$
$$P(\emptyset)=0 \wedge P(M)=1$$
$$P(X_1\cup X_2)=P(X_1)+P(X_2)-P(X_1\cap X_2)$$
$$P(X_1\cap X_2)=P(X_1)*P(X_2)\Rightarrow X_1\wedge X_2 \, independent \Leftarrow P(X_1|X_2)=P(X_1)\wedge P(X_2|X_1)=P(X_2)$$

[Bayes' Theorem (Predictands vs. Predictors), Factorization, etc.]

Verosimilitud

$$\{X_1,\dots,X_n: X_1\cup\dots\cup X_n=M\wedge X_i\cap X_j=\emptyset\ \forall i\neq j\}\Rightarrow P(X_i|B)=\frac{\boxed{P(B|X_i)}P(X_i)}{P(B)}$$

$$X=\{NW\}\Rightarrow P(Y|X)=\frac{P(Y,X)}{P(X)}$$

$$Y=\{Inv\}\Rightarrow P(Y|X)=\frac{freq(Y,X)}{freq(X)}=\frac{199}{1113}=0.179$$

$$P: \mathbf{M} \longrightarrow [0,1]$$
$$A \longmapsto a$$

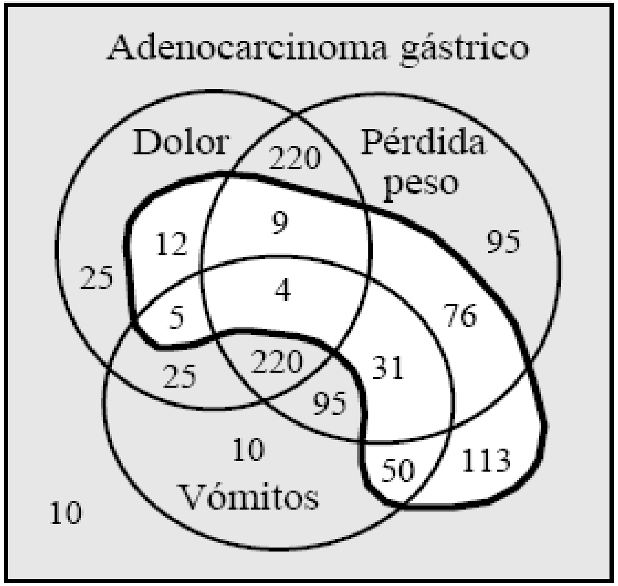
	Anual		Invierno		Primavera		Verano		Otoño	
	S	Ll	S	Ll	S	Ll	S	Ll	S	Ll
NE	1014	516	190	99	287	166	360	162	177	89
SE	64	57	24	18	6	4	1	9	33	26
SW	225	661	98	223	18	119	15	71	94	248
NW	288	825	49	150	95	277	108	251	36	147
Total	1591	2059	361	490	406	566	484	493	340	510

$$P(X) \in [0,1], X \subseteq M$$
$$P(\emptyset) = 0 \wedge P(M) = 1$$
$$P(X_1 \cup X_2) = P(X_1) + P(X_2) - P(X_1 \cap X_2)$$
$$P(X_1 \cap X_2) = P(X_1) * P(X_2) \Rightarrow X_1 \wedge X_2 \, independent \Leftarrow P(X_1 | X_2) = P(X_1) \wedge P(X_2 | X_1) = P(X_2)$$

Bayes' Theorem (Predictands vs. Predictors), Factorization, etc.

$$\{X_1, \ldots, X_n : X_1 \cup \ldots \cup X_n = M \wedge X_i \cap X_j = \emptyset \; \forall i \neq j\} \Rightarrow P(X_i | B) = \frac{P(B | X_i) P(X_i)}{\boxed{\sum_{j=1}^{n} P(B | X_j) P(X_j)}}$$

$$X = \{NW\} \Rightarrow P(Y | X) = \frac{P(Y, X)}{P(X)} \qquad \Rightarrow \qquad Y = \{Inv\} \Rightarrow P(Y | X) = \frac{freq(Y, X)}{freq(X)} = \frac{199}{1113} = 0.179$$



Initial Probabilities:

Gray → Adenocarcinoma

White → Not Adenocarcinoma

$$P(g)=\frac{700}{700+300}=\frac{700}{1000}=\boxed{0.7}$$

$$P(\neg g)=1-P(g)=1-0.7=0.3$$

Significant changes in the probabilities reflect the dependence between predictand and predictors.

↓

Predictability

Could we predict the probability of a disease based on the symptoms?

Patient has suffered threw up:

$$\{V=v\}\Rightarrow P(g|v)=\frac{P(g)P(v|g)}{P(g)P(v|g)+P(\neg g)P(v|\neg g)}=\frac{0.7*0.5}{0.7*0.5+0.3*0.3}=\boxed{0.795}$$

Patient has suffered of weight loss and threw up:

$$\{P=p\wedge V=v\}\Rightarrow P(g|v,p)=\frac{P(g)P(v,p|g)}{P(g)P(v,p|g)+P(\neg g)P(v,p|\neg g)}=\frac{0.7*0.45}{0.7*0.45+0.3*0.12}=\boxed{0.9}$$

$$C \in \{c_1, \ldots, c_m\}$$ Target variable with ***m*** states/classes

$$X = \{X_1, \ldots, X_n\}$$ Predictors in a ***n-dimensional*** space

Bayes' Theorem (Predictands vs. Predictors)

$$P(C|\{X_1,\ldots,X_n\}) = \frac{P(\{X_1,\ldots,X_n\}|C)\,P(C)}{P(\{X_1,\ldots,X_n\})}$$

Bayesian Classifier

$$Arg_C[Max(P(C|\{X_1,\ldots,X_n\}))]$$

$$C \in \{c_1, \ldots, c_m\}$$

Target variable with ***m*** states/classes

$$X = \{X_1, \ldots, X_n\}$$

Predictors in a ***n-dimensional*** space

Bayes' Theorem (Predictands vs. Predictors)

$$P(C|\{X_1,\ldots,X_n\}) = \frac{P(\{X_1,\ldots,X_n\}|C)P(C)}{P(\{X_1,\ldots,X_n\})}$$

Bayesian Classifier

$$Arg_C[Max(P(C|\{X_1,\ldots,X_n\}))]$$

$P(\{X_1,\ldots,X_n\})$ **Constant**

$$Arg_C[Max(P(\{X_1,\ldots,X_n\}|C)P(C))]$$

$$C \in \{c_1, \ldots, c_m\}$$

Target variable with ***m*** states/classes

$$X = \{X_1, \ldots, X_n\}$$

Predictors in a ***n-dimensional*** space

Bayes' Theorem (Predictands vs. Predictors)

$$P(C|\{X_1,\ldots,X_n\}) = \frac{P(\{X_1,\ldots,X_n\}|C)\,P(C)}{P(\{X_1,\ldots,X_n\})}$$

Bayesian Classifier

$$Arg_C\left[Max\left(P(C|\{X_1,\ldots,X_n\})\right)\right]$$

$$P(\{X_1,\ldots,X_n\})$$ ← **Constant**

Probability "***a priori***"

$$Arg_C\left[Max\left(P(\{X_1,\ldots,X_n\}|C)\boxed{P(C)}\right)\right]$$

$$C \in \{c_1, \ldots, c_m\}$$ Target variable with ***m*** states/classes

$$X = \{X_1, \ldots, X_n\}$$ Predictors in a ***n-dimensional*** space

Bayes' Theorem (Predictands vs. Predictors)

$$P(C|\{X_1, \ldots, X_n\}) = \frac{P(\{X_1, \ldots, X_n\}|C)\,P(C)}{P(\{X_1, \ldots, X_n\})}$$

Bayesian Classifier

$$Arg_C[Max(P(C|\{X_1, \ldots, X_n\}))]$$

$P(\{X_1, \ldots, X_n\})$ ← **Constant**

Probability "***a priori***"

$$Arg_C[Max(\boxed{P(\{X_1, \ldots, X_n\}|C)}\,P(C))]$$

Verisimilitude

$$C \in \{c_1, \ldots, c_m\}$$

Target variable with ***m*** states/classes

$$X = \{X_1, \ldots, X_n\}$$

Predictors in a ***n-dimensional*** space

[Bayes' Theorem (Predictands vs. Predictors)]

$$P(C|\{X_1,\ldots,X_n\}) = \frac{P(\{X_1,\ldots,X_n\}|C)\,P(C)}{P(\{X_1,\ldots,X_n\})}$$

Bayesian Classifier

$$Arg_C[Max(P(C|\{X_1,\ldots,X_n\}))]$$

m	n		parámetros
3	10	$\simeq$	$8 \cdot 10^3$
5	20	$\simeq$	$33 \cdot 10^6$
10	50	$\simeq$	$11 \cdot 10^{17}$

$P(\{X_1,\ldots,X_n\})$ **Constant**

Probability "***a priori***"

$$Arg_C[Max(\boxed{P(\{X_1,\ldots,X_n\}|C)}\,P(C))]$$

Verisimilitude

$$C \in \{c_1, \ldots, c_m\}$$ Target variable with ***m*** states/classes

$$X = \{X_1, \ldots, X_n\}$$ Predictors in a ***n-dimensional*** space

[Bayes' Theorem (Predictands vs. Predictors)](#)

$$P(C|\{X_1,\ldots,X_n\}) = \frac{P(\{X_1,\ldots,X_n\}|C)P(C)}{P(\{X_1,\ldots,X_n\})}$$

[+](#) $P(X_i|\{X_j, C\}) = P(X_i|C) \; \forall \; j \neq i$

Naive Bayesian Classifier

Exclusive states/classes
Predictors conditionally independent given the state.

$$C \in \{c_1, \ldots, c_m\}$$ Target variable with ***m*** states/classes

$$X = \{X_1, \ldots, X_n\}$$ Predictors in a ***n-dimensional*** space

[Bayes' Theorem (Predictands vs. Predictors)]

$$P(C|\{X_1,\ldots,X_n\}) = \frac{P(\{X_1,\ldots,X_n\}|C)\,P(C)}{P(\{X_1,\ldots,X_n\})}$$

\+ $P(X_i|\{X_j, C\}) = P(X_i|C)\ \forall\ j \neq i$

Naive Bayesian Classifier

Exclusive states/classes
Predictors conditionally independent given the state.

$$Arg_C[Max(P(\{X_1,\ldots,X_n\}|C)\,P(C))] = Arg_C[Max(P(X_1|C)\ldots P(X_n|C)\,P(C))]$$

Ideas: https://sw23993.wordpress.com/2017/02/17/naive-bayes-classification-in-r-part-2/

$$C \in \{c_1, \ldots, c_m\}$$ Target variable with ***m*** states/classes

$$X = \{X_1, \ldots, X_n\}$$ Predictors in a ***n-dimensional*** space

Bayes' Theorem (Predictands vs. Predictors)

$$P(C|\{X_1,\ldots,X_n\}) = \frac{P(\{X_1,\ldots,X_n\}|C)\,P(C)}{P(\{X_1,\ldots,X_n\})}$$

\+ $P(X_i|\{X_j, C\}) = P(X_i|C) \; \forall\, j \neq i$

Naive Bayesian Classifier

Exclusive states/classes
Predictors conditionally independent given the state.

$$Arg_C[Max(P(\{X_1,\ldots,X_n\}|C)\,P(C))] = Arg_C[Max(P(X_1|C)\ldots P(X_n|C)\,P(C))]$$

Bayesian Classifier

m	n		parámetros
3	10	$\simeq$	$8 \cdot 10^3$
5	20	$\simeq$	$33 \cdot 10^6$
10	50	$\simeq$	$11 \cdot 10^{17}$

Naive Bayesian Classifier

m	n	parámetros
3	10	32
5	20	104
10	50	509

$$C \in \{c_1, \ldots, c_m\}$$ Target variable with ***m*** states/classes

$$X = \{X_1, \ldots, X_n\}$$ Predictors in a ***n-dimensional*** space

Bayes' Theorem (Predictands vs. Predictors)

$$P(C|\{X_1,\ldots,X_n\}) = \frac{P(\{X_1,\ldots,X_n\}|C)P(C)}{P(\{X_1,\ldots,X_n\})}$$

$+$ $P(X_i|\{X_j,C\}) = P(X_i|C)\,\forall\, j \neq i$

Naive Bayesian Classifier

$$Arg_C[Max(P(\{X_1,\ldots,X_n\}|C)P(C))] = Arg_C[Max(P(X_1|C)\ldots P(X_n|C)P(C))]$$

How should be the graph for a Naive Bayesian Classifier?

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

- Define the corresponding graph

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

- Define the corresponding graph

- Define the Bayesian Network (graph + probabilities)

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

- Define the corresponding graph.

- Define the Bayesian Network (graph + probabilities).

- Could we play golf today? Use the formula and the Bayesian Network

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

- Define the corresponding graph.
- Define the Bayesian Network (graph + probabilities).
- Could we play golf today? Use the formula and the Bayesian Network
- Which is the accuracy of the classifier?

Pros:

It is easy and fast to predict class of test data set. It also perform well in multi class prediction

When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.

It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Cons:

If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction.

Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

$$C \in \{c_1, \ldots, c_m\}$$ Target variable with ***m*** states/classes

$$X = \{X_1, \ldots, X_n\}$$ Predictors in a ***n-dimensional*** space

Bayes' Theorem (Predictands vs. Predictors)

$$P(C|\{X_1,\ldots,X_n\}) = \frac{P(\{X_1,\ldots,X_n\}|C)\,P(C)}{P(\{X_1,\ldots,X_n\})}$$

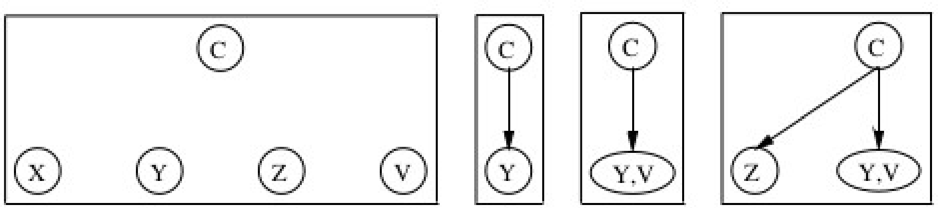
+ $$P(X_i|\{X_j, C\}) = P(X_i|C)\ \forall\ j \neq i$$

Naive Bayesian Classifier

Exclusive states/classes
Predictors conditionally independent given the state.

Very restrictive hypothesis

Semi-Naive Bayesian Classifier



(a) (b) (c) (d)

$$C \in \{c_1, \dots, c_m\}$$ Target variable with ***m*** states/classes

$$X = \{X_1, \dots, X_n\}$$ Predictors in a ***n-dimensional*** space

Bayes' Theorem (Predictands vs. Predictors)

$$P(C|\{X_1,\dots,X_n\}) = \frac{P(\{X_1,\dots,X_n\}|C)\,P(C)}{P(\{X_1,\dots,X_n\})}$$

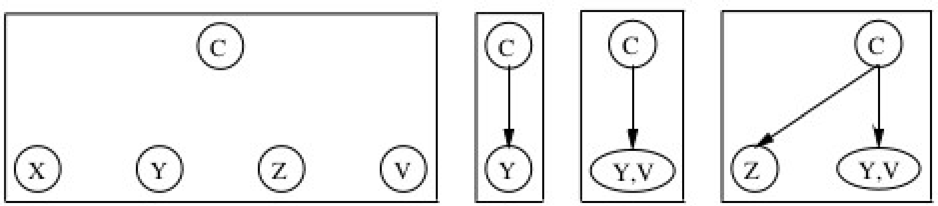
+ $P(X_i|\{X_j, C\}) = P(X_i|C)\ \forall\ j \neq i$

Naive Bayesian Classifier

Exclusive states/classes
Predictors conditionally independent given the state.

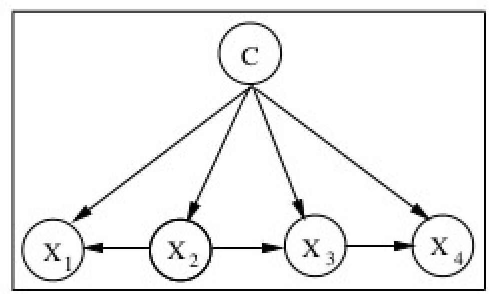
Very restrictive hypothesis

Semi-Naive Bayesian Classifier



(a) (b) (c) (d)

Tree Augmented-Naive (TAN)



$$C \in \{c_1, \ldots, c_m\}$$ Target variable with ***m*** states/classes

$$X = \{X_1, \ldots, X_n\}$$ Predictors in a ***n-dimensional*** space

[Bayes' Theorem (Predictands vs. Predictors)]

$$P(C|\{X_1,\ldots,X_n\}) = \frac{P(\{X_1,\ldots,X_n\}|C)\,P(C)}{P(\{X_1,\ldots,X_n\})}$$

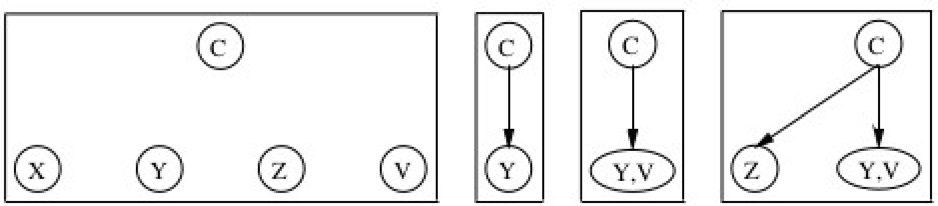
[+] $P(X_i|\{X_j, C\}) = P(X_i|C)\ \forall\ j \neq i$

Naive Bayesian Classifier

Exclusive states/classes
Predictors conditionally independent given the state.

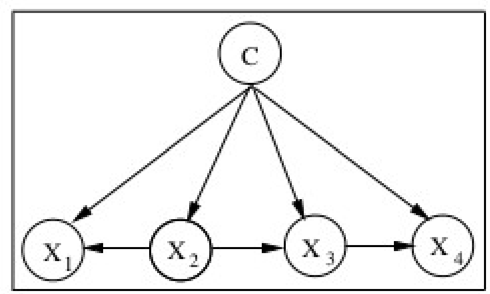
Very restrictive hypothesis

Semi-Naive Bayesian Classifier



Structural Improvement

Tree Augmented-Naive (TAN)



Extensions

$$C \in \{c_1, \dots, c_m\}$$ Target variable with ***m*** states/classes

$$X = \{X_1, \dots, X_n\}$$ Predictors in a ***n-dimensional*** space

[Bayes' Theorem (Predictands vs. Predictors)]

$$P(C|\{X_1,\dots,X_n\}) = \frac{P(\{X_1,\dots,X_n\}|C)\,P(C)}{P(\{X_1,\dots,X_n\})}$$

\+ $P(X_i|\{X_j, C\}) = P(X_i|C)\ \forall\ j \neq i$

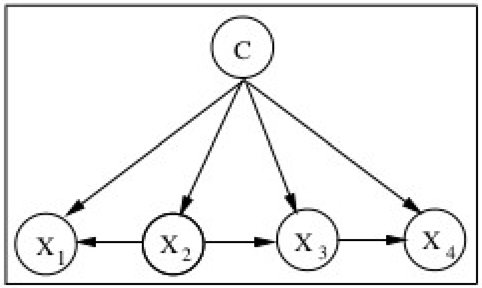
Naive Bayesian Classifier

Exclusive states/classes
Predictors conditionally independent given the state.

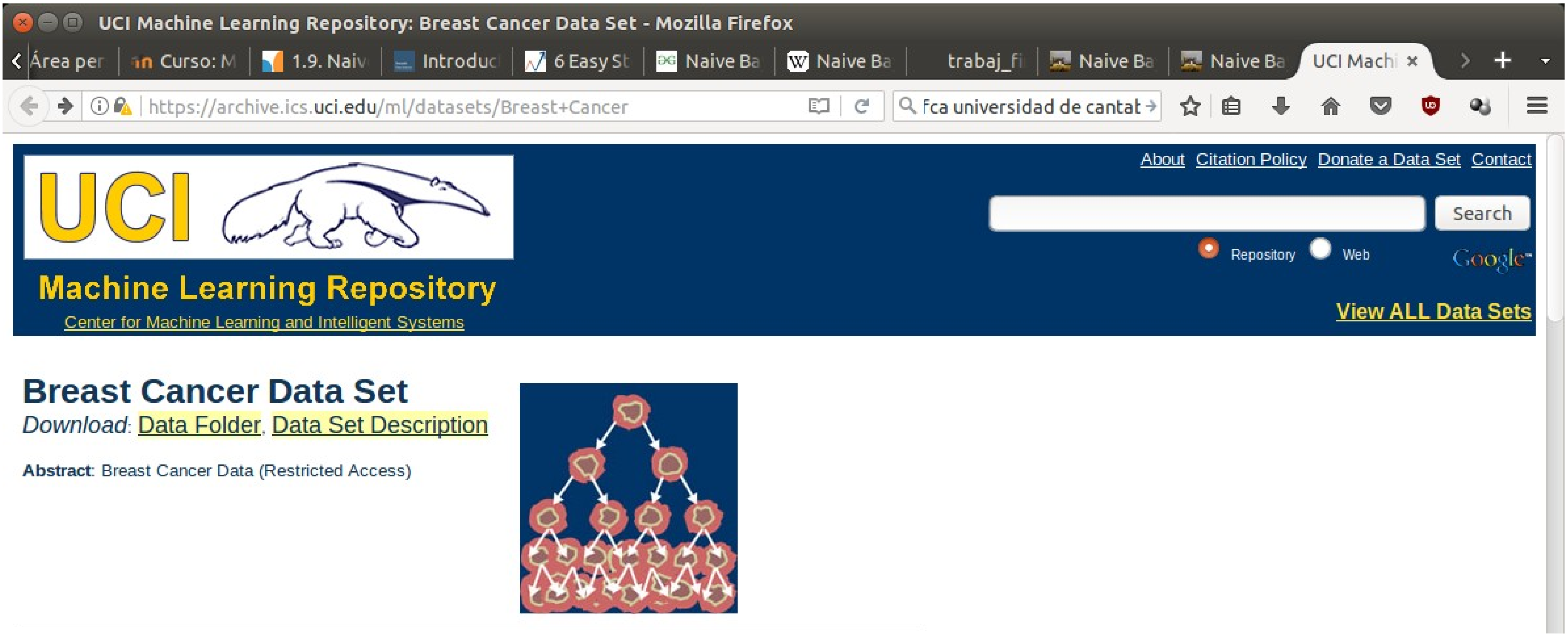
Very restrictive hypothesis

Particular case of Bayesian Networks

Tree Augmented-Naive (TAN)



Extensions



The Naive Bayesian Classifier is included in the R-package ***e1071*** (see function ***naiveBayes***).

An example with the Breast Cancer data set could be found here:
https://sw23993.wordpress.com/2017/02/17/naive-bayes-classification-in-r-part-2/

The data set could be download from the UCI repository:
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

Exercise (~1h)