Introduction aux

Réseaux bayésiens

Antoine Cornuéjols
(antoine@lri.fr)

I.I.E.
&
L.R.I., Université d'Orsay

Plan général

2. R. Incertain : Approches intensionnelles

- **Mesure de certitude assignée à des *ensembles de mondes***
  - ➥ **Connecteurs = combinaison ensembliste des ensembles de mondes**

    Exemple : $\mathbf{p(A \wedge B) = p(A \cap B)}$

    MAIS non déterminable à partir de p(A) et p(B) seulement

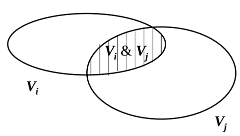
$$\mathbf{A} \xrightarrow{m} \mathbf{B} \;\equiv\; \mathrm{p}(A|B) = m$$

2- Approches intensionnelles

- **Probabilités**
  - **On suppose que le monde peut être décrit par une collection de <u>variables aléatoires</u> $V_1, V_2, \ldots V_k$**
    - **Chaque variable V_i prend une valeur dans un domaine de définition (booléen, réel, …) que nous noterons v_i**
    - **On étudiera la probabilité qu'une variable V_i puisse prendre la valeur v_i : $p(V_i{=}v_i)$**
    - **Probabilité jointe : $p(V_1{=}v_1, V_2{=}v_2, \ldots, V_k{=}v_k)$**
    - **Fonction de probabilité jointe : $p(V_1{=}v_1, V_2{=}v_2, \ldots, V_k{=}v_k)$ æ R**
    - **$\Sigma\, p(V_1, V_2, \ldots, V_k) = 1$**
  - **Exemple :**
    - **p(*face*) = 1/2**
    - **p(*face, pile, pile, pile, face*) = 1/32**

2- Probabilités conditionnelles

$$p(V_i \mid V_j) \;=\; \frac{p(V_i, V_j)}{p(V_j)}$$



$$p(\mathrm{B,L,M,G}) \;=\; p(\mathrm{B \mid L,G,M}).p(\mathrm{L \mid G,M}).p(\mathrm{G \mid M}).p(\mathrm{G})$$

Règle de Bayes (de révision des probabilités conditionnelles) **:**

$$\boxed{p(V_i \mid V_j) \;=\; \frac{p(V_j \mid V_i).p(V_i)}{p(V_j)}}$$

2- Approche intensionnelle directe : le calcul des probabilités

- **Probabilité conditionnelle :**

$$P(A|B) \;=\; \frac{P(A \cap B)}{P(B)} \qquad\qquad P(A|B) \;=\; \frac{P(B|A).P(A)}{P(B)}$$

$$P(H|E) \;=\; \frac{P(E|H).P(H)}{P(E|H).\,p(H) \;+\; P(E|\neg H).P(\neg H)}$$

$$P(H_i|E) \;=\; \frac{P(E|H_i).P(H_i)}{\sum_{k=1}^{m} P(E|H_k).\,p(H_k)}$$

$$\boxed{P(H_i|E_1E_2\ldots E_n) \;=\; \frac{P(E_1E_2\ldots E_n|H_i).P(H_i)}{\sum_{k=1}^{m} P(E_1E_2\ldots E_n|H_k).\,p(H_k)}}$$

2- Exploitation des indépendances conditionnelles

- **Une variable *V* est conditionnellement indépendante d'un ensemble de variables V_i étant donné un ensemble de variables V_j, si :**

$$p(V \mid V_i, V_j) \; = \; p(V \mid V_j) \quad \Leftrightarrow \quad I(V, V_i \mid V_j)$$

<u>Exemple</u> : $p(\mathrm{M} \mid \mathrm{B}, \mathrm{G}) \; = \; p(\mathrm{M} \mid \mathrm{B})$

- **Des variables sont *mutuellement conditionnellement indépendantes* étant donné un ensemble *V* si :**

$$p(V_1, V_2, \ldots, V_k \mid V) \; = \; \prod_{i=1}^{k} p(V_i \mid V_{i-1}, \ldots, V_1, V) \; = \; \prod_{i=1}^{k} p(V_i \mid V)$$

➥ **Réseaux bayésiens**

3- Réseaux bayésiens : plan

➥ **Introduction**

- **Les réseaux bayésiens**
  - **Le versant qualitatif**
  - **Le versant quantitatif**
- **Inférences dans les réseaux bayésiens**
- **Applications**
- **Apprentissage des réseaux bayésiens**

Réseaux bayésiens : plan

1. **Introduction**

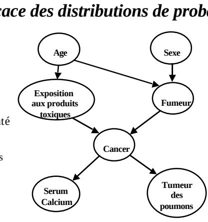
➥ **Les réseaux bayésiens**
  - **Principe**
  - **Le versant qualitatif**
  - **Le versant quantitatif**

2. **La circulation de l'information dans les réseaux bayésiens**
3. **Inférences dans les réseaux bayésiens**
4. **Applications**
5. **Apprentissage des réseaux bayésiens**

3.1- Les réseaux bayésiens : principe

Une représentation efficace des distributions de probabilités

- Un **versant qualitatif** : expression des indépendances : causalité dans un **graphe *G***
  - Graphe acyclique orienté (DAG)
    - Nœuds : variables aléatoires d'intérêt (exhaustives et mutuellement exclusives)
    - Arcs : influence (causale) directe



p(F | A,S) = 0.9
p(F | A,S) = 0.05
p(F | ¬A,S) = 0.0
p(F | ¬A,S) = 0.0

- Un **versant quantitatif** : Une distribution de probabilités ***P***. Modèle probabiliste local.
  - Ensemble de distributions de probabilités conditionnelles
  - Une par nœud et par parent.

Réseaux bayésiens : plan

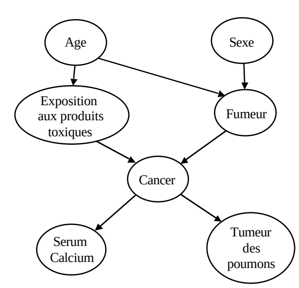
1. **Introduction**

➥ **Les réseaux bayésiens**
  - **Le versant qualitatif**
  - **Le versant quantitatif**

2. **La circulation de l'information dans les réseaux bayésiens**
3. **Inférences dans les réseaux bayésiens**
4. **Applications**
5. **Apprentissage des réseaux bayésiens**

3.1- Les réseaux bayésiens : structure

- **Un réseau bayésien :**



3.1- Les réseaux bayésiens : structure

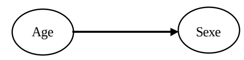
- ***Age*** et ***Sexe*** sont indépendants

$P(A,S) = P(A).P(S)$

$P(A|S) = P(A)$ car $A \perp S$
$P(S|A) = P(S)$ car $A \perp S$

$P(A,S) = P(S|A).P(A) = P(S).P(A)$
$P(A,S) = P(A|S).P(S) = P(A).P(S)$

3.1- Les réseaux bayésiens : structure

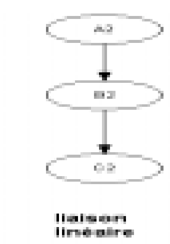


- Dans certaines espèces, ***Age*** et ***Sexe*** sont **<u>dépendants</u>**

$$P(A,S) = P(A) \,.\, P(S|A)$$

$$P(S = s) \;=\; \sum_{\hat{a}ge} P(S \mid \hat{a}ge) \;.\; P(\hat{a}ge)$$

3.1- Liaison linéaire



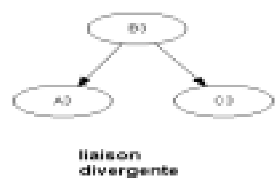
< Il pleut

< La chaussée est humide

<La chaussée est glissante

Une fois que je sais que « la chaussée est humide », le fait que « la chaussée est glissante » est indépendant de « il pleut »

3.1- Liaison divergente

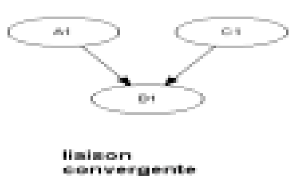


- **A3 = «Le patient renifle»**
- **B3 = «Le patient a un rhume»**
- **C3 = «Le patient tousse»**

A3 et C3 ne sont dépendants que si B3 est <u>inconnu</u>

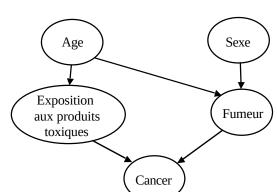
3.1- Liaison convergente

- **A1 = «Automne»**
- **B1 = «Perte de feuilles»**
- **C1 = «Maladie»**



A1 et C1 ne sont dépendants que si B1 est <u>connu</u>

3.1- Les réseaux bayésiens : structure



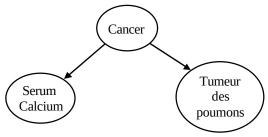
Cancer est indépendant de ***Age*** et de ***Sexe*** étant donné ***Fumeur***

$\boxed{P(C \mid A,S,F) = P(C \mid F)}$ car : $C \perp A, C \perp S$

3.1- Les réseaux bayésiens : structure

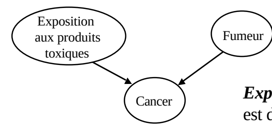
- ***Serum Calcium*** et ***Tumeur des poumons*** sont dépendants
- ***Serum Calcium*** est indépendant de ***Tumeur des poumons*** étant donné ***Cancer***

$$\boxed{\mathrm{P}(T \mid SC,C) = \mathrm{P}(T \mid C)}$$



3.1- Indépendance conditionnelle : explaining away

Exposition aux produits toxiques et ***Fumeur*** sont indépendants : $E \perp F$



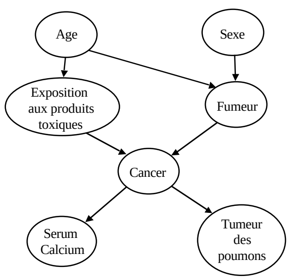
Exposition aux produits toxiques est dépendant de ***Fumeur*** étant donné ***Cancer***

$$\mathrm{P}(E = lourde \mid C = maligne) \; > \; \mathrm{P}(E = lourde \mid C = maligne, F = gros)$$

(*Explaining away*)

3.1- Les réseaux bayésiens : structure

- **Tout ensemble :**



$$\mathrm{P}(A,S,E,F,C,T,SC) =$$
$$\mathrm{P}(A)\,.\,\mathrm{P}(S)\,.$$
$$\mathrm{P}(E \mid A)\,.\,\mathrm{P}(F \mid A,S)\,.$$
$$\mathrm{P}(C \mid E,F)\,.$$
$$\mathrm{P}(SC \mid C)\,.\,\mathrm{P}(T \mid C)$$

3.1- Théorème de factorisation : condition de Markov

- Si le graphe G est une représentation des indépendances conditionnelles de la distribution P :

$$\boxed{\mathrm{P}(X_1, X_2, \ldots, X_n) \;=\; \prod_{i=1}^{n} \mathrm{P}(X_i \mid \boldsymbol{Pa_i})}$$

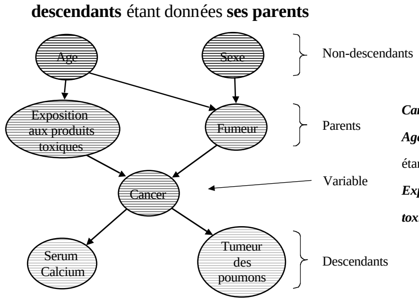
avec : $\boldsymbol{Pa_i}$ = Parents(X_i)

au lieu de :

$$\mathrm{P}(X_1,X_2,\ldots,X_n) = \mathrm{P}(X_1|X_2,\ldots,X_n).\,\mathrm{P}(X_2|X_3,\ldots,X_n)\ldots\mathrm{P}(X_{n-1}|X_n)$$

3.1- Indépendance conditionnelle

- Une variable (nœud) est **conditionnellement indépendante** de ses **non-descendants** étant données **ses parents**

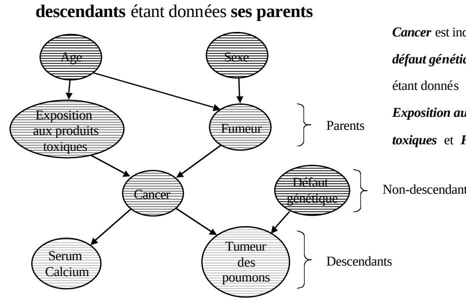


Cancer est indépendant de ***Age*** et ***Sexe*** étant donnés ***Exposition aux produits toxiques*** et ***Fumeur***

3.1- Indépendance conditionnelle

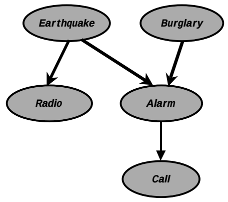
- Une variable (nœud) est **conditionnellement indépendante** de ses **non-descendants** étant données **ses parents**

Cancer est indépendant de ***défaut génétique*** étant donnés ***Exposition aux produits toxiques*** et ***Fumeur***



3.1- Exemple

- **Dans cet exemple:**
  - ***Ind(E; B)***
  - ***Ind(B; E, R)***
  - ***Ind(R; A, B, C | E)***
  - ***Ind(A; R | B,E)***
  - ***Ind(C; B, E, R | A)***



3.1- Réseaux bayésiens : résumé

- **Un réseau bayésien spécifie une distribution de probabilités grâce à deux composantes :**
  - **Un (Graphe Acyclique Orienté) DAG *G***
  - **Une collection de distributions de probabilités conditionnelles $P(X_i|Pa_i)$**
- **La distribution jointe *P* est définie par la factorisation**

$$P(X_1,\ldots,X_n) = \prod_i P(X_i \mid Pa_i)$$

- **Contrainte additionnelle : *G* est une carte d'indépendance (minimale) de *P***
- **Note : un DAG *G* est une *carte d'indépendance* d'une distribution *P* si toutes les conditions de Markov impliquées par *G* sont vérifiées par *P*.**

3.1- Conséquences

- ***P* peut s'écrire en termes de probabilités conditionnelles "locales"**

Si *G* est clairsemé,

- **C'est-à-dire, $|Pa(X_i)| < k$,**

⇒ chaque probabilité conditionnelle peut se spécifier de manière compacte

- **e.g. pour des variables binaires, cela requièrt $O(2^k)$ paramètres.**

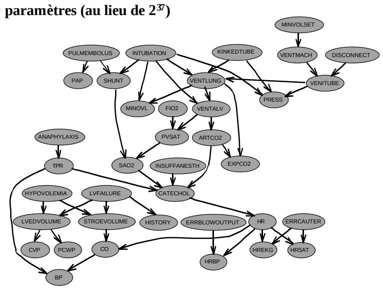
⇒ la représentation de *P* est compacte

- **linéaire dans le nombre de varaibles (vs. exponentielle)**

3.1- Exemple de réseau bayésien

Le réseau "ICU alarm"

37 variables, 509 paramètres (au lieu de 2^{37})



XXX D-séparation

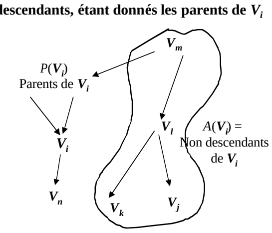
- **Rapport entre G en tant que I-map et indép. cond. [Tirgul 2-3,tr-12]**
- **Le processus d-sep qui contrôle l'indépendance [Tirgul 2-3,tr-20-…]**
- **Le pb du choix de la structure : ordre sur les nœuds [Tirgul 2-3,tr-29] et [Bishop]**

3.1- Pour résumer :

- Un ***réseau bayésien*** est un **graphe orienté acyclique (DAG)**
  - Chaque **nœud** représente une variable aléatoire
  - Chaque **arc** représente une dépendance causale
- **Chaque nœud V_i dans le graphe est <u>*conditionnellement indépendant*</u> de tout ensemble de nœuds qui ne sont pas ses descendants, étant donnés les parents de V_i**

On a : $\boldsymbol{I(V_i, A(V_i) \mid P(V_i))}$

cad : $\mathbf{p}(V_i \mid A(V_i), P(V_i)) = \mathbf{p}(V_i \mid P(V_i))$



3.1- Pour résumer :

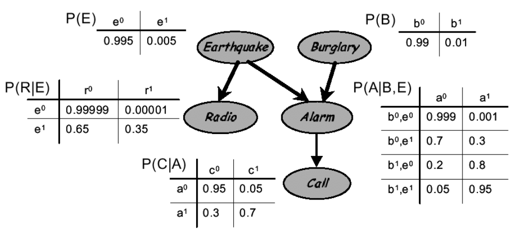
- **Soient V_1, V_2, …, V_k les nœuds d'un réseau bayésien**
- **Étant données les hypothèses d'indépendance conditionnelles impliquées par le réseau, la probabilité jointe de tous les nœuds peut s'écrire :**

$$p(V_1, V_2, ..., V_k) \;=\; \prod_{i=1}^{k} p(V_i \mid P(V_i))$$

➥ Pour calculer la valeur d'une probabilité jointe, il est nécessaire de connaître les fonctions de probabilité de chaque nœud dans le réseau <u>conditionnées uniquement sur les parents</u>

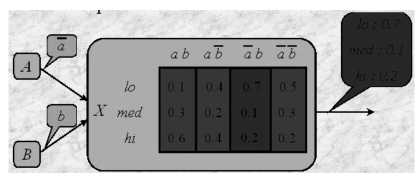
Réseaux bayésiens : plan

3.1- Tables de probabilités dans les noeuds



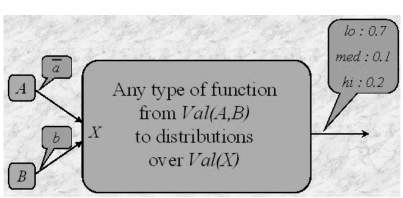
3.1- Fonctions réalisées dans les noeuds

- **Un nœud réalise une fonction de distribution conditionnelle**
  - ❑ **Les entrées sont les valeurs des parents**
  - ❑ **La sortie est une distribution sur ses valeurs**



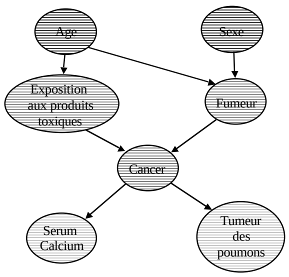
3.1- Fonctions réalisées dans les noeuds

- **N'importe quelle fonction est possible**



Réseaux bayésiens : plan

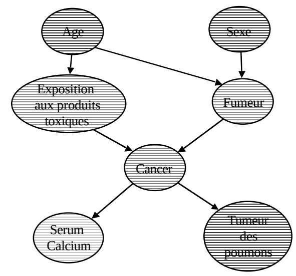
3.1- Inférence prédictive



Quelle est la probabilité que les hommes seniors aient un cancer malin ?

P(C= malin | Age > 60, sexe = mâle) = ?

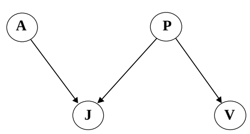
3.1- Inférence prédictive



Quelle est la probabilité qu'un homme senior avec un fort taux de Serum Calcium ait un cancer malin ?

P(C= malin | Age > 60, sexe = mâle, Serum Calcium = fort) = ?

3.1- Les réseaux bayésiens et la causalité



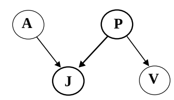
A : J'ai oublié de débrancher mon arroseur automatique

P : Il a plu cette nuit

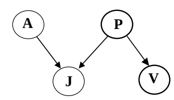
J : L'herbe de mon jardin est humide

V : L'herbe du jardin de mon voisin est humide

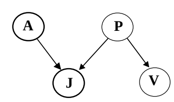
3.1- Les réseaux bayésiens et la causalité



S'il a plu cette nuit, l'herbe de mon jardin est humide

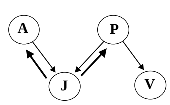


S'il a plu cette nuit, l'herbe du jardin de mon voisin est humide

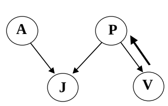


Si j'ai oublié de débrancher mon arroseur automatique, l'herbe de mon jardin est humide

3.1- Réseaux bayésiens et circulation de l'information



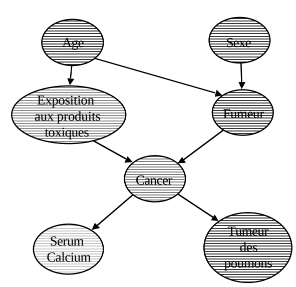
La connaissance de *J* renforce la croyance dans l'une des deux causes *A* ou *P*



La connaissance de *V* augmente la croyance dans la cause *P*.

La cause *A* devient moins plausible.

3.1- Inférence intercausale *(« explaining away »)*



Si Tumeur des poumons est constaté, les probabilités de Fumeur et de Exposition aux produits toxiques montent.

Si l'on apprend alors que gros fumeur, la probabilité de Exposition aux produits toxiques redescend.

3.1- Inférences de base

- Calculer P(Q=q | $\boldsymbol{E}$=$\boldsymbol{e}$)
  - ❑ Q est la variable de requête
  - ❑ $\boldsymbol{E}$ est l'ensemble des nœuds *faits*

$$P(q \mid \boldsymbol{e}) = \frac{P(q,\boldsymbol{e})}{P(\boldsymbol{e})}$$

  - ❑ $X_1, \ldots, X_n$ sont les variables de réseau exceptées Q et $\boldsymbol{E}$

$$\boxed{P(q,\boldsymbol{e}) = \sum_{x_1,\ldots,x_n} P(q,\boldsymbol{e},x_1,\ldots,x_n)}$$

(Marginalisation)

3.1- Les réseaux bayésiens : inférences de bases

$S \in \{no, light, heavy\}$ Smoking $\rightarrow$ Cancer

$C \in \{none, benign, malignant\}$

P(S=no)	0.80
P(S=light)	0.15
P(S=heavy)	0.05

Smoking=	no	light	heavy
P(C=none)	0.96	0.88	0.60
P(C=benign)	0.03	0.08	0.25
P(C=malig)	0.01	0.04	0.15

3.1- Règle du produit

- P(C,F) = P(C | F) . P(F)

$F \rightarrow C$

C⇒ F⇓	*Aucun*	*Bénin*	*Malin*
Non	0.768	0.024	0.008
Léger	0.732	0.012	0.006
Gros	0.035	0.010	0.005

3.1- Les réseaux bayésiens : inférences de base

- **Marginalisation**

S⇓ C⇒	none	benign	malig	total
no	0.768	0.024	0.008	.80
light	0.132	0.012	0.006	.15
heavy	0.035	0.010	0.005	.05
total	0.935	0.046	0.019	

P(Smoke)

P(Cancer)

3.1- Les réseaux bayésiens : inférence de base

- **Utilisation de la règle de Bayes**

$$P(S \mid C) = \frac{P(C \mid S)P(S)}{P(C)} = \frac{P(C,S)}{P(C)}$$

S⇓ C⇒	none	benign	malig
no	0.768/.935	0.024/.046	0.008/.019
light	0.132/.935	0.012/.046	0.006/.019
heavy	0.030/.935	0.015/.046	0.005/.019

Cancer=	none	benign	malignant
P(S=no)	0.821	0.522	0.421
P(S=light)	0.141	0.261	0.316
P(S=heavy)	0.037	0.217	0.263

3.1- Les réseaux bayésiens : inférence de chaîne

$A \rightarrow B \rightarrow C$

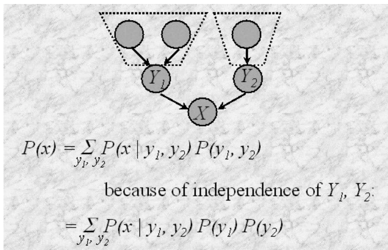
$$P(b) = \sum_a P(a, b) = \sum_a P(b \mid a)\, P(a)$$

$$P(c) = \sum_b P(c \mid b)\, \overbrace{P(b)}$$

$$P(c) = \sum_{b,a} P(a, b, c) = \sum_{b,a} P(c \mid b)\, P(b \mid a)\, P(a)$$

$$= \sum_b P(c \mid b) \underbrace{\sum_a P(b \mid a)\, P(a)}_{P(b)}$$

3.1- Les réseaux bayésiens : *inférence dans les arbres*



XXX- les types de réseaux

- **chaînes**
- **polyarbres**
- **à boucles**

[Naïm et Becker]

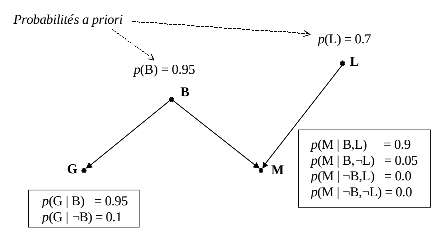
3.1- Probabilités : exemple

- **Soit un robot dont l'état est caractérisé par 4 variables booléennes** (***B*** : batterie ok, ***M*** : le bras bouge, ***L*** : le bloc peut être soulevé, ***G*** : la jauge indique que la batterie est chargée)
- **Il est possible de définir le robot par une *table de probabilités jointes***

(*B, M, L, G*)	Probabilité jointe
(Vrai, Vrai, Vrai, Vrai)	0,5686
(Vrai, Vrai, Vrai, Faux)	0,0299
(Vrai, Vrai, Faux, Vrai)	0,0135
(Vrai, Vrai, Faux, Faux)	0,0007
….	

Probabilité marginale : $p(\mathrm{B} = \mathrm{Vrai}) \;=\; \sum_{B=\mathrm{Vrai}} p(\mathrm{B,M,L,G})$

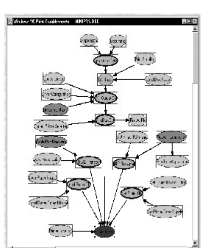
3.1- Exemple de réseau bayésien



$$p(\mathrm{G,B,M,L}) = p(\mathrm{G} \mid \mathrm{B}).p(\mathrm{M} \mid \mathrm{B,L}).p(\mathrm{B}).p(\mathrm{L})$$

3.1- Réseaux bayésiens : quelle économie ?

- **Indépendance causale :** de 2^n à n+1 paramètres
- **Estimation assymétrique :** même économie en pratique
- **Economie typique :**
  - **De 145 on passe à 55 paramètres pour un petit réseau**
  - **De 133 931 430 on passe à 8254 pour le réseau CPCS !!**



3.1- Raisonnement dans les réseaux bayésiens

- Les **types de calculs les plus communs**
  - **Mise à jour des probabilités** $P(x \mid e)$
  - **Validité d'une hypothèse** $P(A \mid e)$
  - **Explication la plus probable** $\arg\max_{v \in V \setminus E} P(V = v \mid e)$
  - **K-èmes explications les plus probables**
  - **Maximum a posteriori** $\arg\max_{v \in W \subset V \setminus E} P(W = w \mid e)$
  - **Analyse de sensibilité** $impact(e,h) \;=\; \dfrac{P(h \mid e)}{P(h)}$
  - **Gain d'information**

$$\begin{aligned} G(W) &= I(H) - I_W(H) \\ &= -\sum_H P(H).\log_2 P(H) \;-\; P(W)\sum_H P(H \mid W).\log_2 P(H \mid W) \end{aligned}$$

3.1- Inférences dans les réseaux bayésiens

- ***Inférence causale*** **(descendante)**
  - **Exemple :** $p(\text{M} \mid \text{L}) = p(\text{M,B} \mid \text{L}) + p(\text{M},\neg\text{B} \mid \text{L})$

    $p(\text{M} \mid \text{L}) = p(\text{M} \mid \text{B,L})\, p(\text{B} \mid \text{L}) + p(\text{M} \mid \neg\text{B,L})\, p(\neg\text{B} \mid \text{L})$

    or : $p(\text{B} \mid \text{L}) = p(\text{B})$ et : $p(\neg\text{B} \mid \text{L}) = p(\neg\text{B})$

    $p(\text{M} \mid \text{L}) = p(\text{M} \mid \text{B,L})\, p(\text{B}) + p(\text{M} \mid \neg\text{B,L})\, p(\neg\text{B})$

    d'où : $p(\text{M} \mid \text{L}) = 0.855$

  - **Principe général :**
    - **Ré-écrire la probabilité conditionnelle désirée du nœud *V* étant donné les faits en termes de probabilité jointe de *V* et de tous ses parents (**qui ne sont pas des faits**), étant donnés les faits**
    - **Ré-exprimer cette probabilité jointe sous forme de la probabilité de *V* conditionnée sur tous ses parents**

3.1- Inférences dans les réseaux bayésiens

- ***Inférence de diagnostic*** **(ascendante)**
  - **Exemple :** $p(\neg\text{L} \mid \neg\text{M}) = \dfrac{p(\neg\text{M} \mid \neg\text{L}).p(\neg\text{L})}{p(\neg\text{M})}$

$$\begin{aligned} p(\neg\text{M} \mid \neg\text{L}) &= p(\neg\text{M,B} \mid \neg\text{L}) + p(\neg\text{M},\neg\text{B} \mid \neg\text{L}) \\ &= p(\neg\text{M} \mid \text{B},\neg\text{L}).p(\text{B} \mid \neg\text{L}) + p(\neg\text{M} \mid \neg\text{B},\neg\text{L}).p(\neg\text{B} \mid \neg\text{L}) \\ &= p(\neg\text{M} \mid \text{B},\neg\text{L}).p(\text{B}) + p(\neg\text{M} \mid \neg\text{B},\neg\text{L}).p(\neg\text{B}) \\ &= 0.95 \times 0.95 + 1 \times 0.05 = 0.9525 \end{aligned}$$

$$p(\neg\text{L} \mid \neg\text{M}) + p(\text{L} \mid \neg\text{M}) = \frac{0.9525 \times 0.3}{p(\neg\text{M})} + \frac{p(\neg\text{M} \mid \text{L}) \times p(\text{L})}{p(\neg\text{M})}$$

$$= \frac{1}{p(\neg\text{M})}\{(0.9525 \times 0.3) + (0.0595 \times 0.7)\} = 1 \quad \Longrightarrow \quad p(\neg\text{M}) = 0.3274$$

$$\boxed{p(\neg\text{L} \mid \neg\text{M}) = \frac{0.28575}{0.3274} = 0.8728}$$

3.1- Inférences dans les réseaux bayésiens

- ***Inférence de disculpation*** **("explaining away")**
  - **Exemple :**
    - soit ¬M (*le bras ne bouge pas*). On peut alors calculer p(¬L).
    - Que devient p(¬L) quand on apprend ¬B (*la batterie est déchargée*) ?

$$\begin{aligned} p(\neg\text{L} \mid \neg\text{B},\neg\text{M}) &= \frac{p(\neg\text{M},\neg\text{B} \mid \neg\text{L})\ p(\neg\text{L})}{p(\neg\text{M},\neg\text{B})} && (\textit{Règle de Bayes}) \\ &= \frac{p(\neg\text{M} \mid \neg\text{B},\neg\text{L})\ p(\neg\text{B} \mid \neg\text{L})\ p(\neg\text{L})}{p(\neg\text{M},\neg\text{B})} && (\textit{Déf. des prob. cond.}) \\ &= \frac{p(\neg\text{M} \mid \neg\text{B},\neg\text{L})\ p(\neg\text{B})\ p(\neg\text{L})}{p(\neg\text{M},\neg\text{B})} && (\textit{d'après le réseau}) \\ &= \frac{1 \times 0.05 \times 0.3}{p(\neg\text{M},\neg\text{B})} \end{aligned}$$

$$p(\neg\text{M},\neg\text{B}) = 0.05 \quad \Rightarrow \quad \boxed{p(\neg\text{L} \mid \neg\text{B},\neg\text{M}) = \frac{0.015}{0.05} = 0.3}$$

3.1- Inférences plus générales

- En général, on veut pouvoir calculer la distribution de probabilités d'une ou plusieurs variables, connaissant la valeur de certaines autres variables

➥ ***Comment faire ?***

- Il est intéressant de savoir comment s'influencent les variables, c'est-à-dire ***comment circule l'information***

➠ Le concept de ***d-séparation*** [Pearl, 1988] permet de calculer ces influences **directement sur le graphe G**.

Réseaux bayésiens : plan

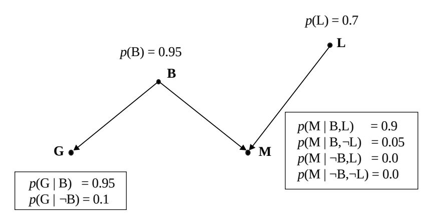
1. **Introduction**
2. **La circulation de l'information dans les réseaux bayésiens**
   - ❑ **La d-séparation**
3. **Inférences dans les réseaux bayésiens**
4. **Applications**
5. **Apprentissage des réseaux bayésiens**

3.2- Principe de la méthode de d-séparation

Idée : exploiter des indépendances conditionnelles impliquées par le réseau autres que celles impliquant seulement les parents d'un nœud

Exemple : $p(\text{M}|\text{G,B}) = p(\text{M}|\text{B})$

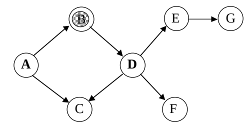
(M est conditionnellement indépendant de G sachant B).



3.2- La d-séparation : intuition

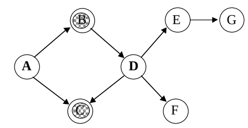
- **La d-séparation permet de préciser dans quelles conditions une information peut être traitée localement sans perturber l'ensemble du graphe**
- Le fait que ***X et Y sont d-séparés par Z*** signifie que **Z bloque le passage de l'information entre X et Y** dans le cas où Z est la seule information connue dans le graphe.
- La d-séparation est une propriété purement graphique qui a des conséquences sémantiques sur la causalité

3.2- La d-séparation : intuition



A est d-séparé de D par B

Comme B est la seule information connue dans ce graphe, une connaissance sur A ne modifiera pas la connaissance sur D : le circuit d'information de A à D est bloqué par B.



A est toujours d-séparé de D par B

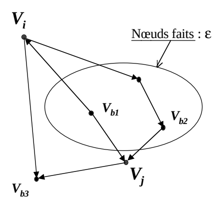
Cependant, comme C est connu maintenant, un chemin de circulation de l'information est ouvert de A à D.

A *n'est pas* d-séparé de D par B et C

3.2- Types d'indépendances conditionnelles

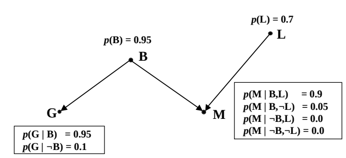
- **Deux nœuds V_i et V_j sont conditionnellement indépendants** étant donné un ensemble de nœuds ε **(c.a.d. I(V_i,V_j| ε)) <u>si</u> pour tout chemin non orienté entre V_i et V_j, il existe un nœud V_b, sur ce chemin vérifiant l'une des trois propriétés :**

1. V_b ε et : Vb æ
2. V_b ε et : æ Vb æ
3. Ni V_b, ni aucun de ses descendants ε, et : æ Vb



3.2- Types d'indépendances conditionnelles : *exemples*

- **I(G,L|B) (*règle 1*) :** B bloque le (seul) chemin entre G et L, étant donné B.
- **I(G,L|B) (*règle 3*) :** M bloque ce chemin, étant donné B, car M n'appartient pas à l'ensemble des nœuds faits.
- **I(G,L) et I(B,L) (*règle 3*) :** M bloque le chemin entre G et L, et entre B et L, étant donné l'ensemble vide d'éléments faits (M n'est pas membre de cet ensemble vide).



3.2- La d-séparation

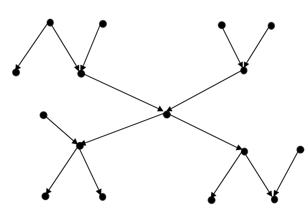
- **La d-séparation peut être étendue aux ensembles de nœuds.**

  Deux ensembles de nœuds V_i et V_j sont d-séparés par ε <u>si</u> tous les chemins entre tous les nœuds de V_i et tous les nœuds de V_j sont bloqués étant donné ε.

- Même en tenant compte de la d-séparation, **le calcul d'inférence probabiliste est en général NP-difficile.**
- **Il existe une classe importante de réseaux pour lesquels le calcul est réalisable : *les polyarbres*.**

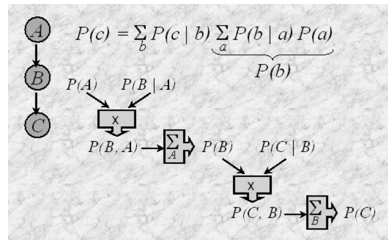
3.2- La d-séparation : le cas des polyarbres

- ***Définition* :** Un ***polyarbre*** est un DAG (graphe acyclique orienté) pour lequel il n'y a qu'un chemin entre toute paire de nœuds.



Réseaux bayésiens : plan

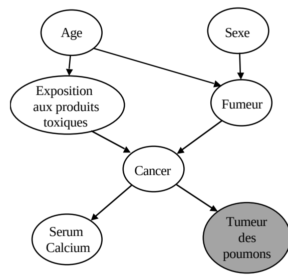
3.3- Élimination de variables

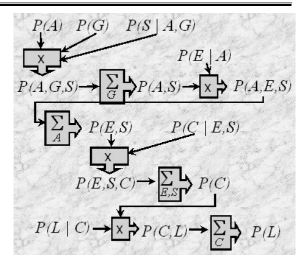


3.3- Inférence comme élimination de variable

- Un ***facteur*** sur *X* est une fonction de *val(X)* sur [0,1] :
  - Un CPT est un facteur
  - Une distribution jointe est aussi un facteur
- Inférence dans un réseau bayésien
  - Les facteurs sont multipliés pour en donner de nouveaux
  - Les variables dans les facteurs sont éliminées
- Une variable peut être éliminée dès que tous les facteurs qui la mentionnent ont été multipliées

3.3- Élimination de variable avec des boucles





La complexité est exponentielle dans la taille des facteurs

3.3- Inférences exactes avec élimination de variables

- **Pour ceux qui sont intéressés : voir** `Tirgul-4.pdf`

  (Tiré des transparents de Tirgul dans le cours CS-228 de Stanford)

- Voir aussi les transparents de `Bishop-2.pdf`

  (Tiré des transparents de Bishop dans un cours à une école d'été)

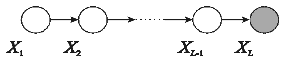
3.3- L'inférence exacte : principe d'élimination des variables

- Grouper les variables cachées *H* en H_1 et H_2 pour lesquelles nous voulons marginaliser sur H_2 pour trouver la distribution de probabilité a posteriori sur H_1
- Le problème d'inférence le plus général implique l'évaluation de

$$P(H_1|V) = \sum_{H_2} P(H_1, H_2|V)$$

- Pour une unité à *M* états discrets, il y a $M^{|H_2|}$ termes dans la sommation où $|H_2|$ est le nombre de nœuds cachés
- Peut facilement devenir impraticable
- Peut-on exploiter la structure d'indépendance conditionnelle (les arcs manquants) pour trouver des algorithmes plus efficaces ?

3.3- Exemple



$$P(X_1, \ldots, X_L) = P(X_1)P(X_2|X_1)\ldots P(X_L|X_{L-1})$$

- **But : trouver** $P(X_1|X_L)$
- **L'évaluation directe donne :**

$$P(X_1|X_L) = \frac{\sum_{X_2} \cdots \sum_{X_{L-1}} P(X_1, \ldots, X_L)}{\sum_{X_1} \sum_{X_2} \cdots \sum_{X_{L-1}} P(X_1, \ldots, X_L)}$$

où, pour des variables à M valeurs, le dénominateur implique de faire des sommations sur M^{L-1} termes (exponentiel en la longueur de la chaîne)

3.3- Exemple (suite)

- **En utilisant la structure d'indépendance conditionnelle, il est possible de ré-ordonner les sommations dans le dénominateur :**

$$\sum_{X_{L-1}} P(X_L|X_{L-1}) \cdots \sum_{X_2} P(X_3|X_2) \sum_{X_1} P(X_2|X_1)P(X_1)$$

qui implique $\sim LM^2$ sommations (linéaire en la longueur de la chaîne) – et similairement pour le numérateur

- **Peut être vu comme un algorithme de transmission de messages**

3.3- La propagation de croyances

- **Extension aux structures générales de graphes**
- **Implique la transmission d'un message dans chaque direction à travers chaque arc**
- **Solution exacte en temps linéaire par rapport à la taille du graphe**

3.3- Limites

- Calcul des probabilités conditionnelles **impraticable en général pour les réseaux assez grands**
  - **Complexité** dans le pire cas **<u>exponentielle dans le nombre de variables</u>**
  - **Et problème des boucles dans le réseau**

➥ ***Il faut utiliser des méthodes de simplification des calculs*** (en simplifiant les graphes)

- ➢ **Méthode de D-séparation**
- ➢ **Méthode de regroupement** ("arbres de jonction")
- ➢ **Méthodes d'approximation dans les calculs** (e.g. Monte Carlo)

Réseaux bayésiens : plan

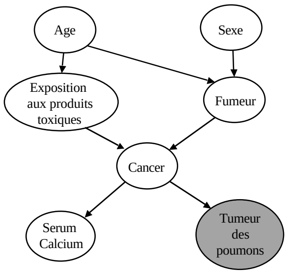
1. **Introduction**
2. **La circulation de l'information dans les réseaux bayésiens**
3. **Inférences dans les réseaux bayésiens** ⬅
   1. **Les grandes approches**
   2. **Élimination des variables**
   3. **Méthodes de regroupement (arbres de jonction)** ⬅
   4. **Méthodes d'approximation par échantillonnage**
4. **Applications**
5. **Apprentissage des réseaux bayésiens**

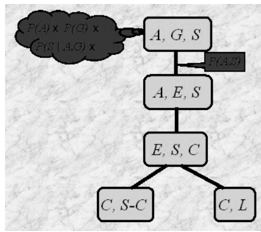
3.3- L'algorithme de l'arbre de jonction

- **Un algorithme efficace pour un graphe général**
- **S'applique aussi bien aux graphes orientés que non orientés**
- **Compile le graphe original en une structure d'arbre puis réalise la transmission de messages sur cet arbre**

3.3- Arbres de jonction

- Un ***arbre de jonction*** est une factorisation partiellement précompilée



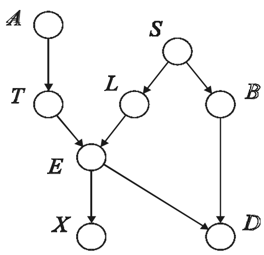


3.3- Résumé de l'algorithme d'arbre de jonction (JT)

- **Étapes clés :**
  1. **Moraliser**
  2. **Absorber les faits mesurés**
  3. **Trianguler**
  4. **Construire l'arbre de jonction**
  5. **Transmettre les messages pour réaliser la cohérence**

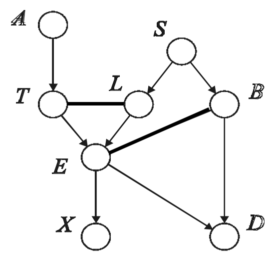
3.3- Exemple de l'algorithme de JT

- **Graphe orienté original**



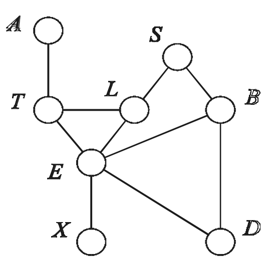
3.3- Exemple de l'algorithme de JT (suite)

- **Moralisation**



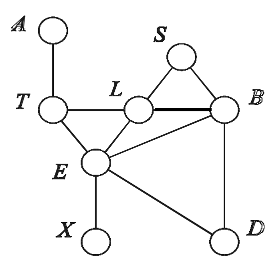
3.3- Exemple de l'algorithme de JT (suite)

- **Graphe non orienté**



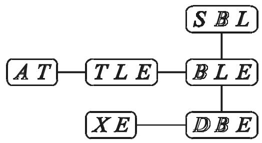
3.3- Exemple de l'algorithme de JT (suite)

- **Triangulation**



3.3- Exemple de l'algorithme de JT (suite)

- **Arbre de jonction**



3.3- Limitations de l'inférence exacte

- Le **coût computationnel** de l'algorithme de l'arbre de jonction est déterminé par la **taille de la plus grande crique**
- Pour les graphes densément connectés, **l'inférence exacte peut être impraticable**
- Il existe **trois approches pour réaliser des inférences approchées**
  - **Faire comme si le graphe était un arbre : "loopy belief propagation"**
  - **Markov chain Monte Carlo (e.g. échantillonnage de Gibbs)**
  - **Inférence variationnelle**

Réseaux bayésiens : plan

1. **Introduction**
2. **La circulation de l'information dans les réseaux bayésiens**
3. **Inférences dans les réseaux bayésiens**
   1. **Les grandes approches**
   2. **Élimination des variables**
   3. **Méthodes de regroupement (arbres de jonction)**
   4. **Méthodes d'approximation par échantillonnage**

4 **Applications**

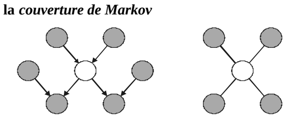
5 **Apprentissage des réseaux bayésiens**

3.3- MCMC (Markov Chain Monte Carlo)

- ***L'Échantillonnage de Gibbs*** **échantillonne chaque variable à son tour en utilisant sa distribution conditionnelle conditionnée sur les autres variables**

$$P(X_i|X_{\{j\neq i\}})$$

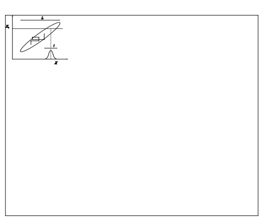
- **Non limité aux distributions exponentielles conjuguées**
- **Pour les modèles graphiques, ces distributions conditionnelles ne dépendent que des variables de la *couverture de Markov***



- **Implémentation : BUGS (Spiegelhalter *et al.*)**

3.3- MCMC (suite)

- **Le problème avec l'échantillonnage de Gibbs est que les points successifs sont hautement corrélés**



- Dans cet exemple, il faut de l'ordre de $(L/l)^2$ étapes pour engendrer des échantillons indépendants (marche aléatoire)

3.3- L'inférence variationnelle

- **Méthode de plus en plus utilisée**
- **Une sorte d'adaptation de l'algorithme EM (Expectation-Maximization)**
- **Voir les transparents de** `Bishop-2.pdf`

3.3- Les difficultés et limites actuelles de ces méthodes

- Encore beaucoup de travaux théoriques et pratiques pour permettre une mise en œuvre simple et efficace des réseaux bayésiens
  - Méthodes de simplification
  - Extension aux variables continues (seulement avec des distributions Gaussiennes)
- En particulier, beaucoup de travaux sur **l'apprentissage des réseaux bayésiens**
  - **Apprentissage des paramètres quand la structure est connue**
  - **Apprentissage de la structure**

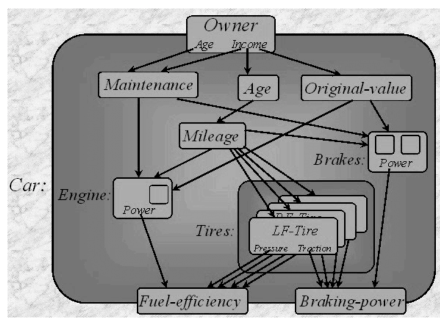
Réseaux bayésiens : plan

1. **Introduction**
2. **La circulation de l'information dans les réseaux bayésiens**
3. **Inférences dans les réseaux bayésiens**
4. **Applications**
5. **Apprentissage des réseaux bayésiens**

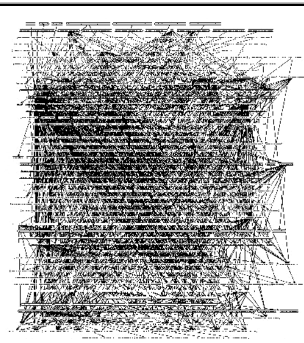
3.4- Domaines d'application

- **Diagnostic médical** (ex : PATHFINDER)
- **Systèmes de contrôle** (Hugin : sous-marin autonome)
- **Aide à la décision en temps réel** (NASA : application VISTA) (25000 données en temps réel)
- **Fusion de données (multi-capteurs)**
- **Détection de fraude sur cartes bancaires**
- **Projet LUMIÈRE de Microsoft :** sélection des données et de l'interface en fonction de l'utilisateur et du contexte
- **Il existe des outils permettant la définition et l'utilisation des réseaux bayésiens**

3.4- Modèle d'un véhicule



3.4- Exemple : le réseau CPCS

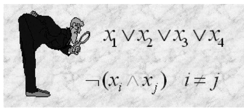


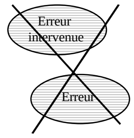
3.5- La construction de réseaux bayésiens

- Quelles **variables ?**
- Quelle **structure ?**
- Quels **nombres ?**

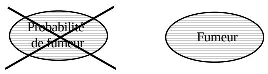
3.5- Les variables

- **Collectivement exhaustives et mutuellement exclusives**





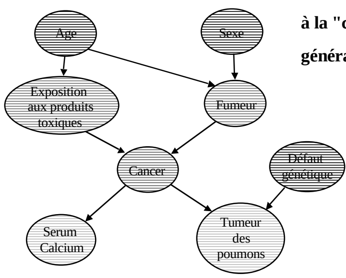
- **Des valeurs et non des probabilités**



3.5- Les variables

- **Test de clarté : valeur connaissable en principe**
  - **Temps : {Ensoleillé, nuageux, pluvieux, neige}**
  - **Carburant : euros / litres**
  - **Température : {<100°C, ≥ 100°C}**
  - **L'utilisateur a besoin d'aide sur Unix : {oui, non}**
  - **Personnalité de l'utilisateur : {rapide, lent}**

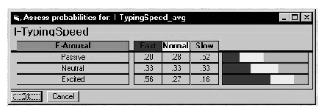
3.5- La structure



La structuration correspondant à la "causalité" est généralement favorable

3.5- Les nombres

- **Généralement la deuxième décimale ne change rien**
- **Ce sont les probabilités relatives qui comptent**
- **Donc :**
  - **0 et 1**
  - **Ordres de magnitudes : 10^{-9} vs. 10^{-6}**
  - **Analyse de sensibilité**

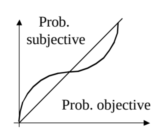


3.6- Bilan : avantages et limites des réseaux bayésiens

- **Avantages :**
  - La **représentation des connaissances par liens entre causes et effets** est **souvent plus naturelle** que la représentation par règles de production
  - **La représentation des connaissances est assez lisible** (par opposition aux réseaux de neurones par exemple, ou même aux arbres de décision)
  - **Les types d'inférences réalisables** à partir de la même représentation **sont très variés** (diagnostic, effets, analyse de sensibilité, …)
  - **Des méthodes d'apprentissage existent**
- **Limites :**
  - **L'utilisation des probabilités et leur donnée par l'expert sont problématiques**
  - **Problème des variables continues**
  - **Complexité des algorithmes**

4. Biais dans le jugement humain

- **Précision limitée (échelle 7 ± 2 échelons)** (jugement absolu sur un stimulus unidimensionnel)
- **Calibration**



- **Trop conservatif**
  - Moins si petit échantillon, intéressement, fortes prob. a priori
- **Mais ne tiennent pas compte des probabilités a priori**
- **Corrélations positives OK, sauf si contre préjugés**