Introduction aux processus stochastiques
Leçon 2

Louis Wehenkel

Département EEI
Université de Liège

Montefiore - Liège - 13/2/2007

Find slides: http://montefiore.ulg.ac.be/∼lwh/ProcStoch/

Objectifs de cette leçon

- *Expliquer et illustrer la notion d'indépendance (conditionnelle) de variables aléatoires, dans le cas discret.*
- *Montrer comment cette notion peut être exploitée pour modéliser une expérience aléatoire.*
- *Expliquer ce que c'est que l'inférence probabiliste.*
- *Montrer comment exploiter les propriétés d'indépendance et d'indépendance conditionnelle pour l'inférence.*
- *Introduire une représentation graphique des modèles probabilistes et des relations d'indépendance conditionnelle.*

Notion d'indépendance conditionnelle

Soient $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ trois v.a. discrètes.

- On dit que $\mathcal{X}$ est indépendante de $\mathcal{Y}$ *(noté par $\mathcal{X} \perp \mathcal{Y}$)* si et seulement si $\forall i, j : P(X_i, Y_j) = P(X_i)P(Y_j)$.
- On dit que $\mathcal{X}$ est indépendante de $\mathcal{Y}$ conditionnellement à $\mathcal{Z}$ *(noté par $\mathcal{X} \perp \mathcal{Y}|\mathcal{Z}$)* si et seulement si $\forall i, j, k : P(X_i, Y_j|Z_k) = P(X_i|Z_k)P(Y_j|Z_k)$.

Remarque:

- Nous utilisons la notation $P(\mathcal{X}, \mathcal{Y}) = P(\mathcal{X})P(\mathcal{Y})$ pour résumer la condition $\forall i, j : P(X_i, Y_j) = P(X_i)P(Y_j)$,
- et la notation $P(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) = P(\mathcal{X}|\mathcal{Z})P(\mathcal{Y}|\mathcal{Z})$ pour résumer la condition $\forall i, j, k : P(X_i, Y_j|Z_k) = P(X_i|Z_k)P(Y_j|Z_k)$.

Double pile-ou-face

On lance en même temps une pièce d'1 Euro et une de 2 Euros. L'ensemble des résultats possibles est l'ensemble des trajectoires (position, vitesse, orientation . . .) dans $\mathbf{R}^3$ *de chaque pièce, soit une fonction du temps à valeurs dans* $\mathbf{R}^{24}$*.*

Nous définissons trois variables aléatoires binaires à partir de cette expérience:

- $\mathcal{H}_1$: elle vaut 1 si la pièce d'1 Euro présente son côté pile vers le haut (une fois au repos), 0 sinon.
- $\mathcal{H}_2$: elle vaut 1 si la pièce de 2 Euros présente son côté pile vers le haut, 0 sinon.
- $\mathcal{S}$: elle vaut 1 si les deux pièces tombent du même côté, 0 sinon.

Modélisation

Hypothèses physiques:

- La connaissance de la face sur laquelle est tombée une des deux pièces ne nous apprend rien sur la face sur laquelle l'autre est tombée: $P(\mathcal{H}_1, \mathcal{H}_2) = P(\mathcal{H}_1)P(\mathcal{H}_2)$.
- La première est légèrement biaisée: $P(\mathcal{H}_1 = 1) = 0.6$.
- La seconde est fortement biaisée: $P(\mathcal{H}_2 = 1) = 0.7$.

Construction de la distribution de probabilité conjointe

En général, on a $P(\mathcal{H}_1, \mathcal{H}_2, \mathcal{S}) = P(\mathcal{H}_1)P(\mathcal{H}_2|\mathcal{H}_1)P(\mathcal{S}|\mathcal{H}_1, \mathcal{H}_2)$.
Ici nous avons : $P(\mathcal{H}_1, \mathcal{H}_2, \mathcal{S}) = P(\mathcal{H}_1)P(\mathcal{H}_2)P(\mathcal{S}|\mathcal{H}_1, \mathcal{H}_2)$, et
$P(\mathcal{S} = 1|\mathcal{H}_1, \mathcal{H}_2) = 1$ si $\mathcal{H}_1 = \mathcal{H}_2$ et $P(\mathcal{S} = 1|\mathcal{H}_1, \mathcal{H}_2) = 0$ sinon

$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{S}$	$P(\mathcal{H}_1, \mathcal{H}_2, \mathcal{S})$
1	1	1	$0.6 \times 0.7 \times 1 = 0.42$
1	1	0	$0.6 \times 0.7 \times 0 = 0.00$
1	0	1	$0.6 \times 0.3 \times 0 = 0.00$
1	0	0	$0.6 \times 0.3 \times 1 = 0.18$
0	1	1	$0.4 \times 0.7 \times 0 = 0.00$
0	1	0	$0.4 \times 0.7 \times 1 = 0.28$
0	0	1	$0.4 \times 0.3 \times 1 = 0.12$
0	0	0	$0.4 \times 0.3 \times 0 = 0.00$

Inférences

A partir de la table qui précède et en oubliant la nature du problème, répondez aux questions suivantes :

- Quelle est la probabilité pour que $\mathcal{H}_1 = 1$ et $\mathcal{H}_2 = 0$?
- Quelle est la probabilité pour que $\mathcal{S} = 1$?
- Sachant que $\mathcal{S} = 1$, quelle est la probabilité que $\mathcal{H}_1 = 1$?
- Sachant que $\mathcal{S} = 0$, quelle est la probabilité que $\mathcal{H}_2 = 1$?
- Calculer la distribution conditionnelle $P(\mathcal{S}|\mathcal{H}_1 = 1)$.
- Calculer la distribution conditionnelle $P(\mathcal{H}_1|\mathcal{S} = 1)$.
- Est-ce que $\mathcal{H}_i \perp \mathcal{S}$?
- Est-ce que $\mathcal{H}_1 \perp \mathcal{H}_2|\mathcal{S}$?
- Est-ce que $\mathcal{H}_1 \perp \mathcal{S}|\mathcal{H}_2$?

Et si les deux pièces sont équilibrées ?

L'ours, le mal-voyant, le mal-entendant et l'enfant

Pierre est mal-voyant et se promène dans la forêt; il y rencontre un ours. Au retour il rencontre Tom qui est mal-entendant, et lui dit qu'il a rencontré un ours d'une certaine couleur. Tom rentre chez lui, et dit à son fils qu'il y avait un ours d'une certaine couleur dans la forêt.

Nous définissons trois variables aléatoires à partir de cette histoire:

- $\mathcal{O}$: la vraie couleur (brune ou noire) de l'ours.
- $\mathcal{P}$: la couleur (brune ou noire) que Pierre précise à Tom.
- $\mathcal{T}$: la couleur (brune ou noire) que Tom mentionne à son fils.

Modélisation

Hypothèses physiques explicites:

- Il y a 50% d'ours bruns dans la forêt: $P(\mathcal{O} = B) = 0.5$.
- Pierre se trompe de couleur 1 fois sur 10:
  $P(\mathcal{P} = B|\mathcal{O} = B) = 0.9$ et $P(\mathcal{P} = N|\mathcal{O} = N) = 0.9$
- De même, Tom entend mal 1 fois sur 10:
  $P(\mathcal{T} = B|\mathcal{P} = B) = 0.9$ et $P(\mathcal{T} = N|\mathcal{P} = N) = 0.9$

Hypothèses implicites:

- Sachant la couleur que Pierre précise à Tom, celle que Tom mentionne à son fils ne dépend pas de la vraie couleur de l'ours: $P(\mathcal{T}|\mathcal{O}, \mathcal{P}) = P(\mathcal{T}|\mathcal{P})$.

Construction de la distribution de probabilité conjointe

En général on a $P(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3) = P(\mathcal{X}_1)P(\mathcal{X}_2|\mathcal{X}_1)P(\mathcal{X}_3|\mathcal{X}_1, \mathcal{X}_2)$.
Ici, nous avons $P(\mathcal{O}, \mathcal{P}, \mathcal{T}) = P(\mathcal{O})P(\mathcal{P}|\mathcal{O})P(\mathcal{T}|\mathcal{P}, [\mathcal{O}])$.

$\mathcal{O}$	$\mathcal{P}$	$\mathcal{T}$	$P(\mathcal{O}, \mathcal{P}, \mathcal{T})$
B	B	B	$0.5 \times 0.9 \times 0.9 = 0.405$
B	B	N	$0.5 \times 0.9 \times 0.1 = 0.045$
B	N	B	$0.5 \times 0.1 \times 0.1 = 0.005$
B	N	N	$0.5 \times 0.1 \times 0.9 = 0.045$
N	B	B	$0.5 \times 0.1 \times 0.9 = 0.045$
N	B	N	$0.5 \times 0.1 \times 0.1 = 0.005$
N	N	B	$0.5 \times 0.9 \times 0.1 = 0.045$
N	N	N	$0.5 \times 0.9 \times 0.9 = 0.405$

Inférences

A partir de la table qui précède et en oubliant la nature du problème, répondez aux questions suivantes :

- Quelle est la probabilité pour que $\mathcal{P} = B$?
- Quelle est la probabilité pour que $\mathcal{T} = B$?
- Sachant que $\mathcal{O} = B$ quelle est la probabilité que $\mathcal{P} = B$?
- Sachant que $\mathcal{O} = B$ quelle est la probabilité que $\mathcal{T} = B$?
- Sachant que $\mathcal{T} = B$ quelle est la probabilité que $\mathcal{O} = B$?
- Sachant que $\mathcal{T} = N$ quelle est la probabilité que $\mathcal{O} = B$?
- Sachant que $\mathcal{P} = B$ quelle est la probabilité que $\mathcal{O} = B$?

Et en faisant appel aux propriétés physiques du problème, comment feriez-vous pour répondre à ces questions ?

La promenade des ours

Pierre surveille l'entrée de la clairière pour observer les ours qui passent. Les ours ont tendance à venir en famille, en file indienne. Les familles sont de taille variable. Pierre note la couleur de chaque ours qui passe. Sachant que Pierre a noté 'B, B' pour les deux premiers, quelle est la probabilité pour que le troisième soit brun?

Nous définissons quatre variables aléatoires:

- $\mathcal{O}_1$: la couleur du premier ours qui entre dans la clairière.
- $\mathcal{O}_2$: la couleur du second ours qui entre dans la clairière.
- $\mathcal{P}_1$: la couleur du premier ours notée par Pierre.
- $\mathcal{P}_2$: la couleur du second ours notée par Pierre.
- $\mathcal{O}_3$: la couleur du troisème ours.

Modélisation

Hypothèses physiques explicites:

- Il y a 50% d'ours bruns dans la forêt: $P(\mathcal{O}_1 = B) = 0.5$.
- Pierre se trompe de couleur 1 fois sur 10:
  $P(\mathcal{P}_i = B|\mathcal{O}_i = B) = 0.9$ et $P(\mathcal{P}_i = N|\mathcal{O}_i = N) = 0.9$.
- Quelle que soit la couleur des ours qui viennent de passer, il y a une chance sur 5 pour que le suivant ne soit pas de la même couleur que le dernier arrivé:
  $P(\mathcal{O}_2 = B|\mathcal{O}_1 = B) = P(\mathcal{O}_2 = N|\mathcal{O}_1 = N) = 0.8$.

Hypothèses implicites:

- Pierre se trompe de couleur de manière indépendante de ses erreurs passées et de la couleur des ours qu'il a vu passer auparavant: $P(\mathcal{P}_2|\mathcal{O}_2, \mathcal{O}_1, \mathcal{P}_1) = P(\mathcal{P}_2|\mathcal{O}_2)$.
- Les ours ne savent pas ce que Pierre note sur son cahier.

Construction de la distribution de probabilité conjointe

Nous avons: $P(\mathcal{O}_1, \mathcal{P}_1, \mathcal{O}_2, \mathcal{P}_2, \mathcal{O}_3) =$
$P(\mathcal{O}_1)P(\mathcal{P}_1|\mathcal{O}_1)P(\mathcal{O}_2|\mathcal{O}_1[\mathcal{P}_1])P(\mathcal{P}_2|\mathcal{O}_2[\mathcal{O}_1, \mathcal{P}_1])P(\mathcal{O}_3|\mathcal{O}_2[\mathcal{O}_1, \mathcal{P}_1, \mathcal{P}_2])$

$\mathcal{O}_1$	$\mathcal{P}_1$	$\mathcal{O}_2$	$\mathcal{P}_2$	$\mathcal{O}_3$	$P(\mathcal{O}_1, \mathcal{P}_1, \mathcal{O}_2, \mathcal{P}_2, \mathcal{O}_3)$
B	B	B	B	B	$0.5 \times 0.9 \times 0.8 \times 0.9 \times 0.8 = \ldots$
B	B	B	B	N	$0.5 \times 0.9 \times 0.8 \times 0.9 \times 0.2 = \ldots$
B	B	B	N	B	$0.5 \times 0.9 \times 0.8 \times 0.1 \times 0.8 = \ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	N	N	N	B	$0.5 \times 0.9 \times 0.8 \times 0.9 \times 0.2 = \ldots$
N	N	N	N	N	$0.5 \times 0.9 \times 0.8 \times 0.9 \times 0.8 = \ldots$

Il faut $(2^5 = 32) \times 4 = 128$ multiplications pour remplir la table.

Après le $270^{\text{ème}}$ ours: 5.1e+83 multiplications et une table de 1.9e+81 lignes.
Mais, l'univers ne contient qu'environ 1+e80 électrons...

Les deux chiens et les n puces

Cet exemple permettra d'illustrer la notion de chaîne de Markov, de distribution stationnaire, de périodicité, de stationnarité et d'ergodicité, et l'inférence par propagation.

Nous définissons un nombre infini de variables aléatoires:

- $\mathcal{X}_t$: désigne le nombre de puces sur le dos du premier chien au temps $t = 0, 1, ..., \infty$.

On a:

- $\mathcal{X}_t \in \{0, 1, 2, \ldots, n\}, \forall t$.
- $P(\mathcal{X}_0 = n) = 1$, et donc $P(\mathcal{X}_0 = k) = 0, \forall k \in \{0, \ldots, n-1\}$.
- On peut se convaincre que $\mathcal{X}_{t+1} \perp (\mathcal{X}_{t-1}, \ldots, \mathcal{X}_0) | \mathcal{X}_t$

$\Rightarrow P(\mathcal{X}_{t+1}, \mathcal{X}_t, \ldots, \mathcal{X}_0) = P(\mathcal{X}_{t+1}|\mathcal{X}_t) P(\mathcal{X}_t|\mathcal{X}_{t-1}) \ldots P(\mathcal{X}_0)$

Le cas particulier des deux puces

Désignons

par π_t le vecteur ligne $[P(\mathcal{X}_t = 0), \ldots, P(\mathcal{X}_t = n)]$, et

par Π_t la matrice dont l'élément i, j vaut $P(\mathcal{X}_{t+1} = j|\mathcal{X}_t = i)$.

Lorsque $n = 2$ on a

- $\pi_0 = [0,\ 0,\ 1]$
- $\Pi_t = \Pi = \begin{bmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{bmatrix}, \forall t \in \{0, 1, \ldots, \infty\}$ *Invariance!*

Calculs pour $t = 1, 2, \ldots$

- On a $P(\mathcal{X}_t = j) = \sum_{i=0}^{n} P(\mathcal{X}_{t-1} = i) P(\mathcal{X}_t = j|\mathcal{X}_{t-1} = i)$.
- En notation matricielle on a donc $\pi_t = \pi_{t-1}\Pi_{t-1}$.
- Par conséquent $\pi_t = \pi_0\Pi_0\Pi_1 \cdots \Pi_{t-1} = \pi_0\Pi^t$

Calculer π_t pour $t = 1, 2, 3 \ldots$: $[0, 1, 0], [0.5, 0, 0.5], [0, 1, 0] \ldots$
Quid si en $t = 0$, on a 1 puce sur le dos de chaque chien ?
Quid si en $t = 0$ on ne connaît pas la répartition des n puces ?

Notion de distribution stationnaire:

- Solution de $q\Pi = q$ avec $q_1 + q_2 + q_3 = 1$
- Ici $[q_1, q_2, q_3] \begin{bmatrix} -1 & 1 & 0 & 1 \\ 0.5 & -1 & 0.5 & 1 \\ 0 & 1 & -1 & 1 \end{bmatrix} = [0, 0, 0, 1]$
- Ou encore $[q_1, q_2, q_3] \begin{bmatrix} -1 & 1 & 1 \\ 0.5 & -1 & 1 \\ 0 & 1 & 1 \end{bmatrix} = [0, 0, 1]$
- On trouve $q = [0.25, 0.5, 0.25]$.

La prédiction du futur en général

En général on définit une chaîne de Markov à p états par

- Une distribution initiale: π_0 (vecteur de dim. p)
- Une suite de matrices de transition $p \times p$: $\Pi_0, \Pi_1, \Pi_2, \ldots$

NB. Si la suite est constante on dit que la chaîne est invariante dans le temps, et on note la matrice de transition par Π.

On peut alors calculer les matrices de transition à k pas de temps

- $\Pi_{t,t+k} = \Pi_t \Pi_{t+1} \cdots \Pi_{t+k-1}$.
- L'élément i, j de $\Pi_{t,t+k}$ vaut $P(\mathcal{X}_{t+k} = j | \mathcal{X}_t = i)$.
- De π_t on calcule $\pi_{t+k} = \pi_t \Pi_{t,t+k}$ en kp^2 opérations $+\times$.

NB. Le même calcul par marginalisation nécessite p^{k+1} additions.
$P(\mathcal{X}_{t+k}=i) = \sum_{i_1=1}^{p} \cdots \sum_{i_k=1}^{p} P(\mathcal{X}_{t+k}=i, \mathcal{X}_{t+k-1}=i_1, \ldots, \mathcal{X}_t=i_k)$

Algorithme de propagation gauche$\longrightarrow$droite

Calcul de $P(\mathcal{X}_{t+1})$ à partir de $P(\mathcal{X}_t)$ en multipliant le vecteur π_t à gauche de la matrice Π_t. Chaque propagation prend p^2 opérations.

Si on connaît la valeur de $\mathcal{X}_0$ on remplace π_0 par le vecteur I_{X_0} pour calculer $P(\mathcal{X}_t|X_0), \forall t > 0$.

Si on connaît $\mathcal{X}_k$ alors à partir de $\mathcal{X}_k$ on propage le vecteur I_{X_k} vers la droite pour calculer $P(\mathcal{X}_t|X_k), \forall t > k$.

Question: Comment calculer $P(\mathcal{X}_t|X_k), \forall t < k$?

Question: Comment calculer $P(\mathcal{X}_t|X_0, X_k), \forall t \in \{1, \ldots, k-1\}$?

De la prédiction du futur au retour dans le passé

Supposons que nous connaissions le nombre de puces sur le dos du premier chien en $t = 0$ et en $t = T$. (I.e. on se donne X_0 et X_T). Que vaut alors, $P(\mathcal{X}_t|X_0, X_T)$?

- $P(\mathcal{X}_t|X_0, X_T) = \frac{P(\mathcal{X}_t, X_0, X_T)}{\sum_{X_t} P(X_t, X_0, X_T)} = \frac{P(\mathcal{X}_t, X_T|X_0)P(X_0)}{\sum_{X_t} P(X_t, X_T|X_0)P(X_0)}$.
- Or $P(\mathcal{X}_t, X_T|X_0) = P(\mathcal{X}_t|X_0)P(X_T|\mathcal{X}_t)$
- $P(\mathcal{X}_t|X_0) = X_0$-ème ligne de la matrice $\Pi_{0,t}$
- $P(X_T|\mathcal{X}_t) = X_T$-ème colonne de la matrice $\Pi_{t,T}$
- $P(\mathcal{X}_t, X_T|X_0) = (l_{X_0}\Pi_{0,t}) \times (\Pi_{t,T}c_{X_T})$,

  où l_i (resp. c_j) désigne le vecteur ligne (resp. colonne) dont la composante i (resp. j) vaut 1 (les autres sont nulles). Le symbole $\times$ représente le produit terme à terme d'un vecteur ligne par un vecteur colonne.

Discussion

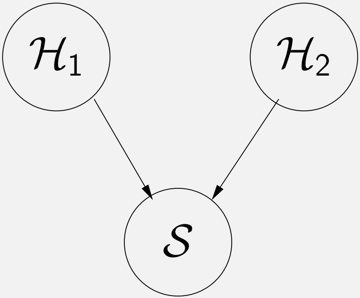
Si X_T n'est pas observé, on ne propage rien à partir de la droite. Cela revient à propager à partir de la droite le vecteur colonne $\mathbf{1}$, puisque $\Pi_t \mathbf{1} = \mathbf{1}$ (les lignes des matrices Π_t somment à 1).

Si on n'observe pas X_0 il faut propager vers la droite un vecteur π_0.

Si en plus d'observer $\mathcal{X}_0$ et $\mathcal{X}_T$, on observe aussi $\mathcal{X}_k$ $(0 < k < T)$ on propage l_{X_0} vers la droite jusque k puis l_{X_k}; on propage c_{X_T} vers la gauche jusque k puis c_{X_k}.

Cet algorithme peut se généraliser directement au cas où on fait des observations à un nombre quelconque de pas de temps et qu'on souhaite calculer les distributions de probabilités conditionnelles de chaque $\mathcal{X}_t$ étant données les observations.

Pile ou face



Chaque noeud correspond à une v.a. A chaque v.a. correspond une loi conditionnelle $P(\mathcal{X}_i|pa(\mathcal{X}_i))$.

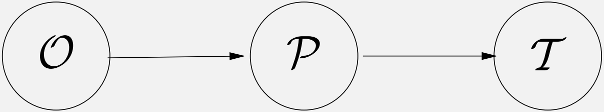
Le graphe indique que la distribution conjointe peut se factoriser sous la forme $\prod_{i=1}^{N} P(\mathcal{X}_i|pa(\mathcal{X}_i))$.

Ce graphe indique que $P(\mathcal{H}_1, \mathcal{H}_2, \mathcal{S}) = P(\mathcal{H}_1)P(\mathcal{H}_2)P(\mathcal{S}|\mathcal{H}_1, \mathcal{H}_2)$. *Il s'agit d'un exemple de réseau bayesien. Le graphe est dirigé et acyclique.*

L'ours, Pierre et Tom

Pour ce problème, la loi de probabilité conjointe se factorise de la manière suivante: $P(\mathcal{O}, \mathcal{P}, \mathcal{T}) = P(\mathcal{O})P(\mathcal{P}|\mathcal{O})P(\mathcal{T}|\mathcal{P})$.

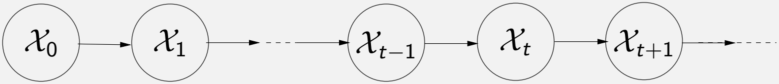
Graphiquement cette factorisation est donc représentée par



Ce graphe indique essentiellement que $\mathcal{O} \perp \mathcal{T}|\mathcal{P}$.

La chaîne de Markov

Le cas général de la chaîne de Markov correspond au graphe suivant:

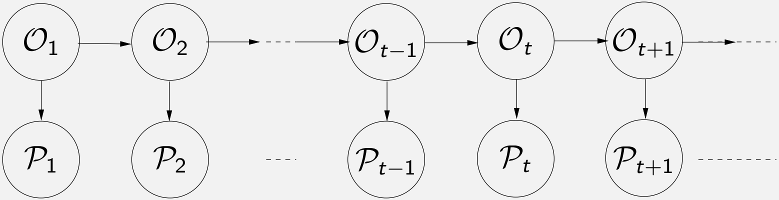


ce qui correspond à la factorisation

$$P(\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_{t-1}, \mathcal{X}_t, \mathcal{X}_{t+1}, \ldots) =$$
$$P(\mathcal{X}_0)P(\mathcal{X}_1|\mathcal{X}_0)\cdots P(\mathcal{X}_{t-1}|\mathcal{X}_{t-2})P(\mathcal{X}_t|\mathcal{X}_{t-1})P(\mathcal{X}_{t+1}|\mathcal{X}_t)\cdots$$

La balade des ours

Il s'agit d'une chaîne de Markov cachée



Le graphe exprime la factorisation: $P(\mathcal{O}_1, \mathcal{P}_1, \ldots, \mathcal{O}_t, \mathcal{P}_t) =$
$P(\mathcal{O}_1)P(\mathcal{O}_2|\mathcal{O}_1)\cdots P(\mathcal{O}_t|\mathcal{O}_{t-1})P(\mathcal{P}_1|\mathcal{O}_1)\ldots P(\mathcal{P}_t|\mathcal{O}_t)$

Synthèse

La notion d'indépendance conditionnelle (d'ensembles) de v.a. est une notion fondamentale dans le domaine des processus aléatoires

- pour la construction de modèles à partir d'hypothèses physiques
- pour la mise au point d'algorithmes efficaces d'inférence.

Les réseaux bayesiens permettent

- de représenter graphiquement une factorisation d'une distribution conjointe d'un ensemble de variables aléatoires
- de déduire des relations d'indépendances conditionnelles qui sont une conséquence de cette factorisation à partir de la structure du graphe (cf propriété de d-séparation)
- de construire des algorithmes d'inférence probabiliste qui propagent de l'information le long des arcs du réseau bayesien.