MADI

Pierre-Henri WUILLEMIN

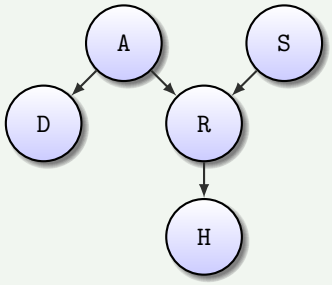
DESIR
LIP6
pierre-henri.wuillemin@lip6.fr

Bayesian networks : definition

Définition (Bayesian Network (BN))

A Bayesian network is a joint distribution over a set of random (discrete) variables. A Bayesian network is represented by a directed acyclic graph (DAG) and by a conditional probability table (CPT) for each node $P(X_i|parents_i)$

Altitude, Rain, Happy, Demography et Sea



$$P(A,R,H,D,S) = P(A) \cdot P(S) \cdot P(R|A,S) \cdot p(H|R) \cdot P(D|A)$$

Fondations

Rapides rappels : indépendances (conditionnelles)

Soit X, Y, Z trois variables aléatoires (ou groupes de variables)

➥ si $P(Y) > 0$, $P(X|Y) = \dfrac{P(X,Y)}{P(Y)}$

$$\Longrightarrow P(X,Y) = P(X|Y)P(Y)$$

$$\Longrightarrow P(X,Y,Z) = P(X|Y,Z)P(Y,Z) = P(X|Y,Z)P(Y|Z)P(Z)$$

Indépendance marginale

$X \perp\!\!\!\perp Y$ si et seulement si $P(X,Y) = P(X)P(Y)$
si et seulement si $P(X \mid Y) = P(X)$

Indépendance conditionnelle

$X \perp\!\!\!\perp Y | Z$ si et seulement si $P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$
si et seulement si $P(X \mid Y,Z) = P(X \mid Z)$

Modèle probabiliste décomposable

Modèle probabiliste complexe

Dans un modèle décomposable, une probabilité sur Ω sera donc représentée par une loi jointe des variables de $\mathfrak{X}$.

$$\forall \omega \in \Omega, p(\omega) = p\left(X_1 = X_1(\omega), X_2 = X_2(\omega), \cdots, X_n = X_n(\omega)\right)$$

 Explosion combinatoire : Si toutes les variables sont binaires, un système factorisé en n variables nécessitent $\approx 2^n$ valeurs !

La factorisation peut-elle permettre d'améliorer la compacité ? Grâce à l'indépendance conditionnelle ! !

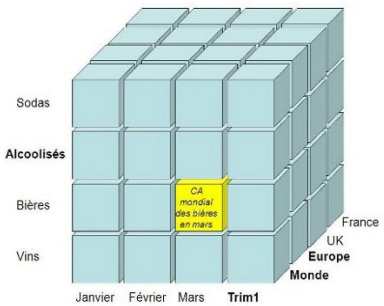
2^3 $\quad p(X,Y,Z) = p(X) \cdot p(Y \mid X) \cdot p(Z \mid X,Y)$ $\quad 2+2^2+2^3$

Avec $X \perp\!\!\!\perp Y$ et $Z \perp\!\!\!\perp X, Y$:

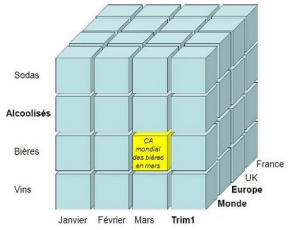
2^3 $\quad p(X,Y,Z) = p(X) \cdot p(Y) \cdot p(Z)$ $\quad 2+2+2$

Modèles complexes factorisable

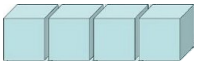
		Janvier	Février	Mars	Trim1
France	Bières	70	70	80	220
	Vins	100	110	90	300
	Total	170	180	170	520
UK	Bières	250	220	240	710
	Vins	50	40	60	150
	Total	300	260	300	860
Total Europe		*470*	*440*	*470*	*1380*



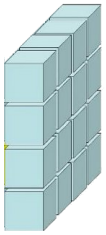
Comment voir dans ce modèle que $Mois \perp\!\!\!\perp \{Pays, Boisson\} | ?$



$=$



$*$



Modèle d'indépendances

Soit notre loi jointe $p(X_1, \cdots, X_n)$ (hyper-cube). Il existe des indépendances conditionnelles testables (χ^2) dans cette loi. Comment les représenter aisément ?
Il serait intéressant de fournir un outil basé sur les variables aléatoires du modèle, qui permettrait de manipuler les indépendances conditionnelles de manière plus naturelle qu'indirectement dans l'hyper-cube de la loi jointe :
Quel objet manipule-t-on lorsqu'on parle de l'**ensemble** des indépendances condionnelles de $p(\,)$?.

Définition (Séparabilité)

Soit $\mathfrak{I} \subset \mathcal{P}(\mathfrak{X}) \times \mathcal{P}(\mathfrak{X}) \times \mathcal{P}(\mathfrak{X})$. $\forall U, V, W \subset \mathfrak{X}$, on dit que U et V sont séparés par W ($\ll U \perp\!\!\!\perp V \mid W \gg_{\mathfrak{I}}$) si et seulement si $(U, V, W) \in \mathfrak{I}$.

Définition

Modèle d'indépendance On nomme $\mathfrak{I}$ un modèle d'indépendance.

Relation entre $\mathfrak{I}$ et p : modèle d'indépendance probabiliste

L'ensemble $\mathfrak{I}_p = \{(U, V, W) \in \subset \mathcal{P}(\mathfrak{X}) \times \mathcal{P}(\mathfrak{X}) \times \mathcal{P}(\mathfrak{X}), U \perp\!\!\!\perp V \mid W\}$ est un modèle d'indépendance.

$$U \perp\!\!\!\perp V \mid W \Longleftrightarrow \ll U \perp\!\!\!\perp V \mid W \gg_{\mathfrak{I}_p}$$

...D'autres modèles d'indépendance ?

Structuration d'un modèle d'indépendance : Semi-graphoïde et graphoïde

Définition (semi-graphoïde)

Un modèle d'indépendance $\mathfrak{I}$ est un semi-graphoïde s'il satisfait $\forall A, B, S, P \subset \mathfrak{X}$:

1. *Indépendance triviale* $\quad \ll A \perp\!\!\!\perp \emptyset \mid S \gg_{\mathfrak{I}}$
2. *Symétrie* $\quad \ll A \perp\!\!\!\perp B \mid S \gg_{\mathfrak{I}} \Rightarrow \ll B \perp\!\!\!\perp A \mid S \gg_{\mathfrak{I}}$
3. *Décomposition* $\quad \ll A \perp\!\!\!\perp (B \cup P) \mid S \gg_{\mathfrak{I}} \Rightarrow \ll A \perp\!\!\!\perp B \mid S \gg_{\mathfrak{I}}$
4. *Union faible* $\quad \ll A \perp\!\!\!\perp (B \cup P) \mid S \gg_{\mathfrak{I}} \Rightarrow \ll A \perp\!\!\!\perp B \mid (S \cup P) \gg_{\mathfrak{I}}$
5. *Contraction* $\quad \left\{ \begin{array}{c} \ll A \perp\!\!\!\perp B \mid (S \cup P) \gg_{\mathfrak{I}} \\ \ll A \perp\!\!\!\perp P \mid S \gg_{\mathfrak{I}} \end{array} \right\} \Rightarrow \ll A \perp\!\!\!\perp (B \cup P) \mid S \gg_{\mathfrak{I}}$

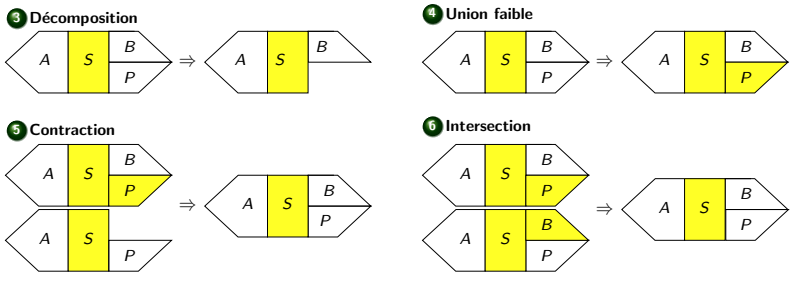
Définition (graphoïde)

Un modèle d'indépendance $\mathfrak{I}$ est un graphoïde s'il satisfait $\forall A, B, S, P \subset \mathfrak{X}$:
$\mathfrak{I}$ est un semi-graphoïde

6. *Intersection* $\quad \left\{ \begin{array}{c} \ll A \perp\!\!\!\perp B \mid (S \cup P) \gg_{\mathfrak{I}} \\ \ll A \perp\!\!\!\perp P \mid (S \cup B) \gg_{\mathfrak{I}} \end{array} \right\} \Rightarrow \ll A \perp\!\!\!\perp (B \cup P) \mid S \gg_{\mathfrak{I}}$

Semi-graphoïde et graphoïde - représentation des axiomes

3. Décomposition $\quad \ll A \perp\!\!\!\perp (B \cup P) \mid S \gg_{\mathfrak{I}} \Rightarrow \ll A \perp\!\!\!\perp B \mid S \gg_{\mathfrak{I}}$
4. Union faible $\quad \ll A \perp\!\!\!\perp (B \cup P) \mid S \gg_{\mathfrak{I}} \Rightarrow \ll A \perp\!\!\!\perp B \mid (S \cup P) \gg_{\mathfrak{I}}$
5. Contraction $\quad \left\{ \begin{array}{c} \ll A \perp\!\!\!\perp B \mid (S \cup P) \gg_{\mathfrak{I}} \\ \ll A \perp\!\!\!\perp P \mid S \gg_{\mathfrak{I}} \end{array} \right\} \Rightarrow \ll A \perp\!\!\!\perp (B \cup P) \mid S \gg_{\mathfrak{I}}$
6. Intersection $\quad \left\{ \begin{array}{c} \ll A \perp\!\!\!\perp B \mid (S \cup P) \gg_{\mathfrak{I}} \\ \ll A \perp\!\!\!\perp P \mid (S \cup B) \gg_{\mathfrak{I}} \end{array} \right\} \Rightarrow \ll A \perp\!\!\!\perp (B \cup P) \mid S \gg_{\mathfrak{I}}$

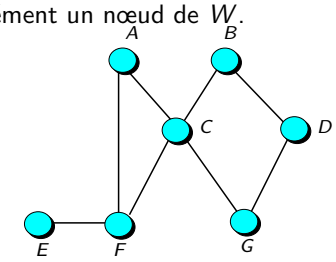


Semi-graphoïde et graphoïde - Utilisation

Théorème (loi de probabilité et graphoïde)

$\mathfrak{I}_p$ possède une structure de semi-graphoïde.
Si $p > 0$ alors $\mathfrak{I}_p$ possède une structure de graphoïde.

Soit un graphe $G = (\mathfrak{X}, \mathfrak{E})$,
$\forall U, V, W \subset X, \langle U \mid W \mid V \rangle_G$ indique que toute chaîne d'un nœud de U vers un nœud de V contient forcément un nœud de W.



Théorème (graphe non orienté et graphoïde)

$\{\langle U \mid W \mid V \rangle_G, U, V, W \subset X\}$ rest un modèle d'indépendance et possède une structure de graphoïde.

Modèle graphique

Définition (Modèle graphique)

Un modèle graphique est un modèle probabiliste factorisé qui se sert d'un graphe entre les variables aléatoires pour représenter des indépendances conditionnelles.

Est-ce que ça se passe bien ? Peut-on toujours avoir $(X \perp\!\!\!\perp Y \mid Z)_p \Leftrightarrow \langle X \mid Z \mid Y \rangle_G$?

Définition (I-map, D-map, P-map, graphe-isomorphisme)

soit $G = (\mathfrak{X}, \mathfrak{E})$ un graphe et une loi de probabilité p.
G est une $D_{\text{ependency}}$-map de p ssi $(X \perp\!\!\!\perp Y \mid Z)_p \Rightarrow \langle X \mid Z \mid Y \rangle_G$.
G est une $I_{\text{ndependency}}$-map de p ssi $(X \perp\!\!\!\perp Y \mid Z)_p \Leftarrow \langle X \mid Z \mid Y \rangle_G$.
G est une P_{erfect}-map de p ssi $(X \perp\!\!\!\perp Y \mid Z)_p \Leftrightarrow \langle X \mid Z \mid Y \rangle_G$.
Un loi de probabilité p est dite graphe-isomorphe si et seulement s'il existe un graphe G qui soit une P-map de p.

- Le graphe vide, sans arc est une **D-map** de toute distribution p.
- Le graphe complet est une **I-map** de toute distribution p.

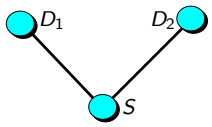
Modèle graphique non orienté : exemple 1

exemple 1

Soit un modèle probabiliste du système composé de 3 variables aléatoires : le tirage de deux dé D_1, D_2 et $S = D_1 + D_2$ qui est la somme des tirages des 2 dés.

indépendances et dépendances du modèle de l'exemple 1

- $D_1 \not\perp\!\!\!\perp S$ et $D_2 \not\perp\!\!\!\perp S$
- $D_1 \perp\!\!\!\perp D_2$ mais $D_1 \not\perp\!\!\!\perp D_2 \mid S$



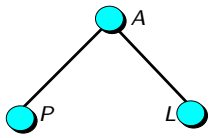
Modèle graphique non orienté : exemple 2

exemple 2

Dans un sondage, on s'aperçoit qu'il y a une forte corrélation entre l'aptitude à lire d'un individu et sa pointure...
On s'aperçoit rapidement que l'âge de l'individu est la variable qui explique cette corrélation bizarre.

indépendances et dépendances du modèle de l'exemple 2

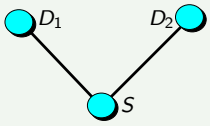
- $L \not\perp\!\!\!\perp A$ et $P \not\perp\!\!\!\perp A$
- $L \not\perp\!\!\!\perp P$ mais $L \perp\!\!\!\perp P \,|\, A$



Modèle graphique orienté

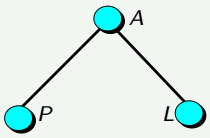
exemple 1

$D_1 \perp\!\!\!\perp D_2$ mais $D_1 \not\perp\!\!\!\perp D_2 \,|\, S$



exemple 2

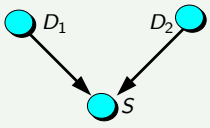
$L \not\perp\!\!\!\perp P$ mais $L \perp\!\!\!\perp P \,|\, A$



Lever l'ambiguïté en ajoutant de l'information qualitative sur les arcs : l'orientation.

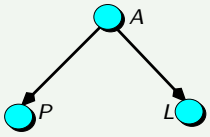
exemple 1 orienté

$D_1 \perp\!\!\!\perp D_2$ mais $D_1 \not\perp\!\!\!\perp D_2 \,|\, S$



exemple 2 orienté

$L \not\perp\!\!\!\perp P$ mais $L \perp\!\!\!\perp P \,|\, A$



Reste à donner un critère de séparation sur les graphes orientés : la d-séparation.

Modèle graphique orienté et d-séparation

Soit une chaîne $C = (x_i)_{i \in I}$ dans un graphe orienté $\overrightarrow{G}$. On dira que x_i est un puits de la chaîne C (ou C-puits) s'il est du type : $x_{i-1} \rightarrow x_i \leftarrow x_{i+1}$.

Définition (Chaîne active, bloquée)

Soit une chaîne $C = (x_i)_{i \in I}$ dans $\overrightarrow{G}$ et Z un sous-ensemble de nœuds de $\overrightarrow{G}$. C est une chaîne active par rapport à Z si :

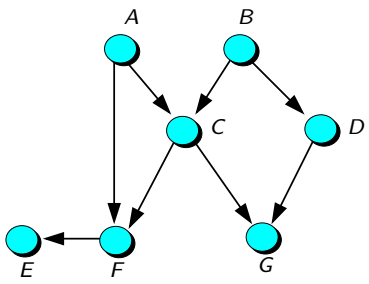
- *Tout C-puits a l'un de ses descendants ou lui-même dans Z.*
- *Aucun élément de C qui n'y est pas un C-puits n'appartient à Z.*

Une chaîne non active par rapport à Z est dite bloquée par Z.

Définition (d-séparation)

Soit $\overrightarrow{G} = (\mathfrak{X}, \mathfrak{E})$ un graphe orienté,
$\forall (X, Y, Z) \subset \mathfrak{X}$, X est d-séparé de Y par Z dans $\overrightarrow{G}$ ($\langle X \,|\, Z \,|\, Y \rangle_{\overrightarrow{G}}$) si et seulement si toute chaîne d'un élément de X vers un élément de Y est bloquée par Z.

Exemple de d-séparation



Réseau bayésien et propriétés de Markov

Définition (réseau bayésien)

Soit un graphe $\overrightarrow{G}$, muni de la d-séparation. Si $\overrightarrow{G}$ est I-map d'une loi p alors $\overrightarrow{G}$ est un réseau bayésien pour p.

Définition (Propriété de Markov globale)

$\overrightarrow{G}$ *vérifie la PMG pour* $p \Leftrightarrow \forall A, B, S \subset \mathfrak{X}$,

$$\langle A \,|\, S \,|\, B \rangle_{\overrightarrow{G}} \Rightarrow A \perp\!\!\!\perp B \,|\, S.$$

i.e. $\overrightarrow{G}$ est une I-map pour p.

Définition (Propriété de Markov locale)

$\overrightarrow{G}$ *vérifie la PML pour* $p \Leftrightarrow \forall x \in \mathfrak{X}$,

$$\{x\} \perp\!\!\!\perp \mathrm{nd}(x) \,|\, \Pi_x.$$

où $\mathrm{nd}(x)$ représente les nœuds non descendants de x et Π_x ses parents.

Réseau bayésien et propriétés de Markov (2)

Théorème

$$PMG \Longleftrightarrow PML$$

Un graphe est un réseau bayésien pour p si et seulement si chaque nœud est indépendant de ses non-descendants, conditionnellement à ses parents pour p.

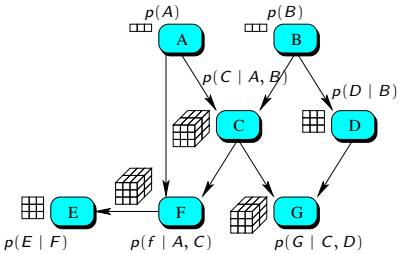
(quand la loi est positive)

Théorème (Factorisation récursive)

Soit $\overrightarrow{G} = (\mathfrak{X}, \mathfrak{E})$ *un réseau bayésien pour p, alors :*

$$p(\mathfrak{X}) = \prod_{X \in \mathfrak{X}} p\left(X \mid \Pi_X\right)$$

réseau bayésien : exemple et définition



$p(A, B, C, D, E, F, G,) = ?$
$p(A) \cdot p(B)$
$\cdot p(C \mid A, B) \cdot p(D \mid B) \cdot p(F \mid A, C)$
$\cdot p(E \mid F) \cdot p(G \mid C, D)$

Tout se passe comme si l'information était localisée dans les nœuds !

$P(A, B, C, D, E, F, G)$: $3^7 = 2187$ paramètres vs 105 paramètres dans le BN !

Définition (Réseau bayésien (BN))

Un réseau bayésien est une représentation compacte d'une distribution de probabilité sur un ensemble de variables aléatoires. Il s'appuie sur un graphe orienté sans circuit (DAG) pour représenter son modèle d'indépendance.
La décomposition de la loi jointe suivant le graphe s'écrit :

$$P(\mathfrak{X}) = \prod_i P(X_i \mid \Pi_i)$$

1er exemple de construction d'un RB (1/6)

Exemple de la dyspnée (Lauritzen & Spiegelhalter (88))

La dyspnée peut être engendrée par une tuberculose, un cancer des poumons, une bronchite, par plusieurs de ces maladies, ou bien par aucune.
Un séjour récent en Asie augmente les chances de tuberculose, tandis que fumer augmente les risques de cancer des poumons. Des rayons X permettent de détecter une tuberculose ou un cancer.
Un patient éprouve des difficultés à respirer. Dans quelle mesure peut-on dire qu'il est atteint de dyspnée ?

Variables aléatoires :

- D : dyspnée : oui/non
- C : cancer : oui/non
- A : Asie : oui/non
- R : rayons X : positif/négatif
- T : tuberculose : oui/non
- B : bronchite : oui/non
- F : fumer : oui/non

1er exemple de construction d'un RB (2/6)

Exemple de la dyspnée (Lauritzen & Spiegelhalter (88))

La dyspnée peut être engendrée par une tuberculose, un cancer des poumons, une bronchite, par plusieurs de ces maladies, ou bien par aucune. Un séjour récent en Asie augmente les chances de tuberculose, tandis que fumer augmente les risques de cancer des poumons. Des rayons X permettent de détecter une tuberculose ou un cancer. Un patient éprouve des difficultés à respirer. Dans quelle mesure peut-on dire qu'il est atteint de dyspnée ?

$P(D, R, T, C, B, A, F) = P(D|R, T, C, B, A, F) \times P(R, T, C, B, A, F)$

Or $P(D|R, T, C, B, A, F) = P(D|T, C, B)$

$\Longrightarrow P(D, R, T, C, B, A, F) = P(D|T, C, B) \times P(R, T, C, B, A, F)$

1er exemple de construction d'un RB (3/6)

Exemple de la dyspnée (Lauritzen & Spiegelhalter (88))

La dyspnée peut être engendrée par une tuberculose, un cancer des poumons, une bronchite, par plusieurs de ces maladies, ou bien par aucune. Un séjour récent en Asie augmente les chances de tuberculose, tandis que fumer augmente les risques de cancer des poumons. Des rayons X permettent de détecter une tuberculose ou un cancer. Un patient éprouve des difficultés à respirer. Dans quelle mesure peut-on dire qu'il est atteint de dyspnée ?

$P(D, R, T, C, B, A, F) = P(D|T, C, B) \times P(R, T, C, B, A, F)$

or $P(R, T, C, B, A, F) = P(R|T, C, B, A, F) \times P(T, C, B, A, F)$

et $P(R|T, C, B, A, F) = P(R|T, C)$

$\Longrightarrow P(D, R, T, C, B, A, F) = P(D|T, C, B) \times P(R|T, C) \times P(T, C, B, A, F)$

1er exemple de construction d'un RB (4/6)

Exemple de la dyspnée (Lauritzen & Spiegelhalter (88))

La dyspnée peut être engendrée par une tuberculose, un cancer des poumons, une bronchite, par plusieurs de ces maladies, ou bien par aucune. Un séjour récent en Asie augmente les chances de tuberculose, tandis que fumer augmente les risques de cancer des poumons. Des rayons X permettent de détecter une tuberculose ou un cancer. Un patient éprouve des difficultés à respirer. Dans quelle mesure peut-on dire qu'il est atteint de dyspnée ?

$P(D, R, T, C, B, A, F) = P(D|T, C, B) \times P(R|T, C) \times P(T, C, B, A, F)$

$P(T|C, B, A, F) = P(T|A)$

$P(D, R, T, C, B, A, F) = P(D|T, C, B) \times P(R|T, C) \times P(T|A) \times P(C, B, A, F)$

.....

$= P(D|T, C, B) \times P(R|T, C) \times P(T|A) \times P(C|F) \times P(B|F) \times P(A) \times P(F)$

1er exemple de construction d'un RB (5/6)

$P(D, R, T, C, B, A, F) =$
$P(D|T, C, B) \times P(R|T, C) \times P(T|A) \times P(C|F) \times P(B|F) \times P(A) \times P(F)$

Si toutes les variables ont 10 valeurs possibles :

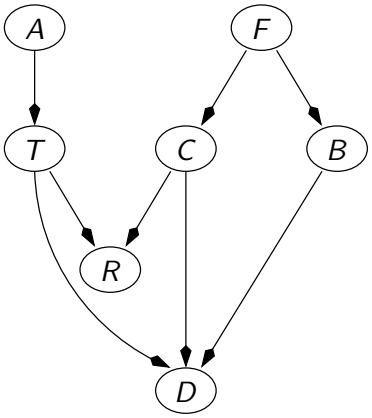
$P(D, R, T, C, B, A, F)$ nécessite une table de 10^7 éléments

formule décomposée nécessite :

$10000 + 1000 + 3 \times 100 + 2 \times 10 = 11320$ éléments

1er exemple de construction d'un RB (6/6)

$P(D, R, T, C, B, A, F) =$
$P(D|T, C, B) \times P(R|T, C) \times P(T|A) \times P(C|F) \times P(B|F) \times P(A) \times P(F)$



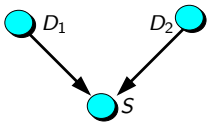
Autres modèles

Annexe : Autres modèles que les BNs ?

Modèle d'indépendance

- $D_1 \perp\!\!\!\perp D_2$

(donc $D_1 \not\perp\!\!\!\perp D_2 | S$)

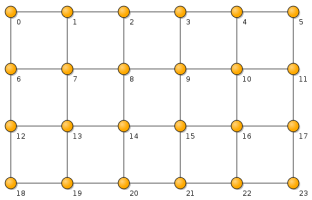


Modèle d'indépendance

- $A \perp\!\!\!\perp B | \{C, D\}$
- $C \perp\!\!\!\perp D | \{A, B\}$

?

Réseaux de Markov



- Séparation dans les réseaux de Markov : séparation dans les graphes non-orientés.
- Propriété de Markov locale dans les réseaux de Markov : $X \perp\!\!\!\perp \text{non-voisin}(X) \,|\, \text{voisin}(X)$

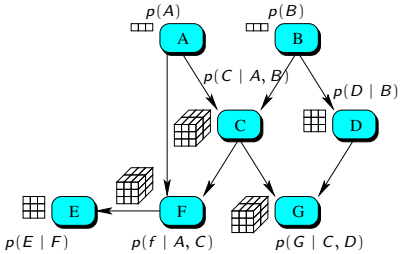
Définition (Réseau de Markov)

Un réseau de Markov est une représentation compacte d'une distribution de probabilité sur un ensemble de variables aléatoires. Il s'appuie sur un graphe non-orienté pour représenter son modèle d'indépendance.
La décomposition de la loi jointe suivant le graphe s'écrit :

$$P(\mathfrak{X}) = \frac{1}{Z} \cdot \prod_{C \in clique(G)} \Phi(C)$$

Applications

Utilisations des BNs : inférence probabiliste



diagnostic : $P(A \mid F)$

- *diagnostic de panne*
- *sûreté de fonctionnement*
- *filtrage de spams*

prédiction $P(E \mid B, A)$

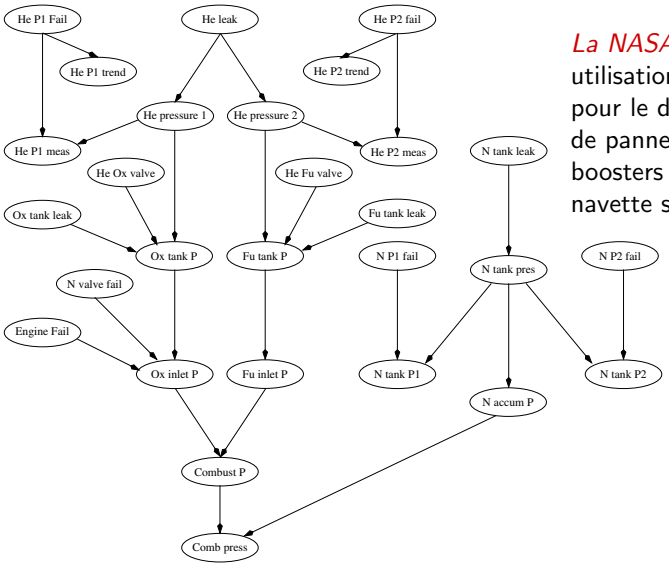
- *Simulation de process (industriels)*
- *prévisions boursières*
- *modélisation de joueurs*

Autres tâches

- *Cas le plus probable :* $\arg\max P(\mathfrak{X} \mid D)$
- *Analyse de sensibilité, information mutuelle, etc.*
- *Troubleshooting :* $\arg\max \frac{P(.)}{C(.)}$

Application 1 : Diagnostic de panne

Diagnostic de panne à la NASA

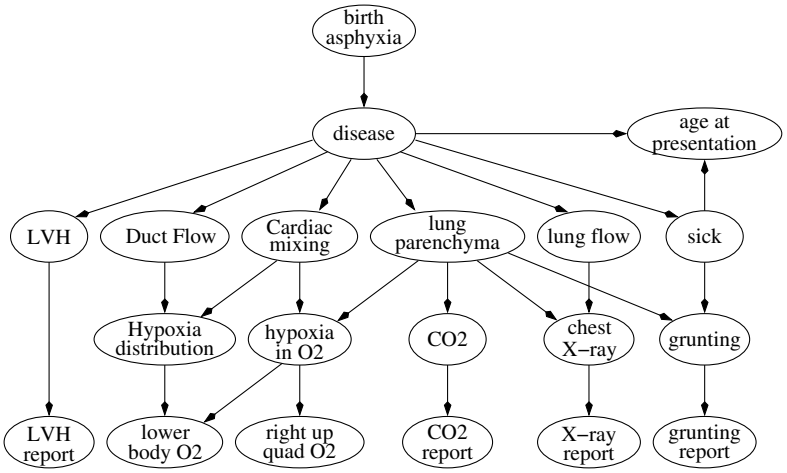


La NASA :
utilisation d'un RB pour le diagnostic de pannes des boosters de la navette spatiale

Application 2 : Diagnostic médical

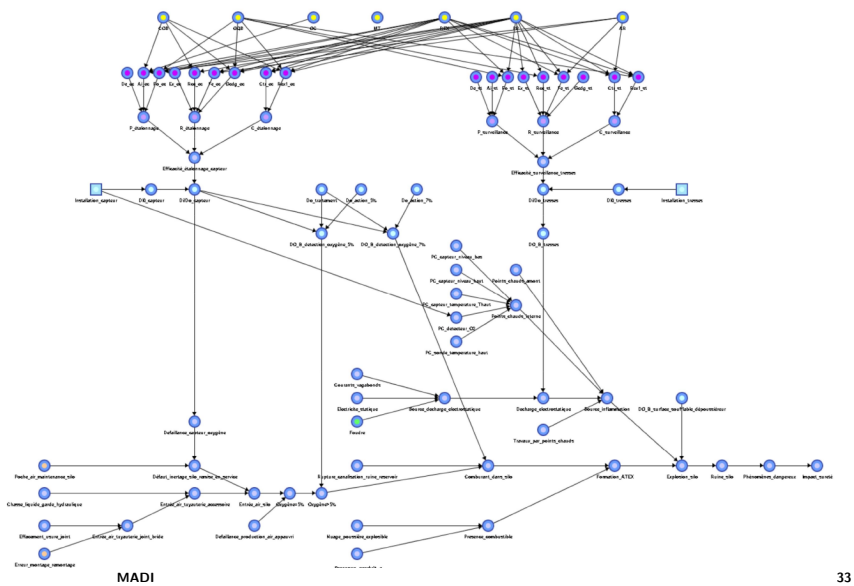
Le Great Ormond Street hospital for sick children

Aide au diagnostic des causes d'une cyanose ou d'une crise cardiaque chez le nourrisson juste après sa naissance.



Application 3 : analyse du risque

Modélisation des phénomèes de risque par réseau bayésien : **approche modulaire**.



Application 4 : classification bayésienne

Soient deux v.a. X (de dimension d) discrète et Y (de dimension 1) discrète (*pas forcément binaire*).
Sur une base d'apprentissage (supervisé) Π_a, on peut estimer les probabilités par des fréquences pour $P(X, Y)$.

Classification

Pour une instantiation x de X, on cherche à prédire sa classe (valeur de Y) : $\widehat{y}$.

1 Maximum de vraisemblance (ML)

$$\widehat{y} = \arg\max_{y_i} P(x \mid y_i)$$

2 Maximum a posteriori (MAP)

$$\widehat{y} = \arg\max_{y_i} P(y_i \mid x) = \arg\max_{y_i} P(y_i) \cdot P(x \mid y_i)$$

D'après la règle de Bayes, $P(Y \mid X) \propto P(X \mid Y) \cdot P(Y)$, on comprend que l'intérêt du MAP est de prendre en compte un *a priori* sur la fréquence de chaque classe.

⚠ Il peut être difficile d'obtenir ces distributions.
Particulièrement : $P(X \mid Y)$ peut demander beaucoup d'observation ! !

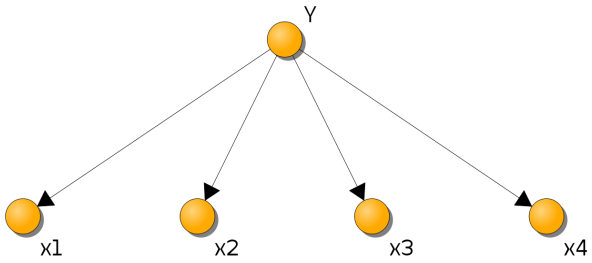
classification bayésienne (2) : Naif Bayes

Comment calculer $P(X \mid Y)$?

Classifieur bayésien naïf

$$\forall k \neq l, X^k \perp\!\!\!\perp X^l \mid Y \qquad \text{et} \qquad P(x, y) = P(y) \cdot \prod_{k=1}^{d} P(x^k \mid y)$$

Cette hypothèse est très forte. Elle a peu de chance de s'avouer exacte dans un cas réel. Néanmoins cette approximation donne des résultats souvent satisfaisants.

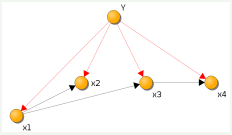


- Estimation des paramètres : trivial (si Π_a sans valeurs manquantes)
- ML : $\prod_{k=1}^{d} P(x^k \mid y)$ ...
- MAP : $P(y \mid x_1, \cdots, x_d)$: inférence dans le BN !

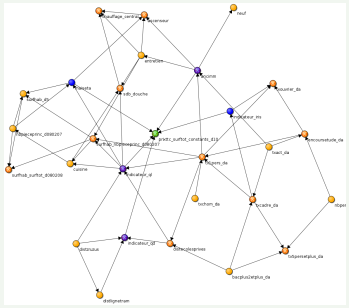
Classification bayésienne (3) : modèles plus complexes

Tree-Augmented Naive Models

Toute variable X_i peut avoir un parent autre que Y (mais un seul !).



Réseau bayésien complet



Dans un BN composé de Y et (X_i), calculer $P(Y \mid X_1, \cdots, X_n)$.
Note : on n'a pas besoin de tous les X_i : Markov Blanket $MB(.)$).

$$P(Y \mid X) = P(Y \mid MB(Y))$$

Application 5 : modèles séquentiels

données séquentielles

Un ensemble de données séquentielles est un ensemble de données dont la séquence est porteur de sens.

- $\{PressionPneuGauche, PressionPneuDroit, NiveauBatterie\}$ est un ensemble de variables dont une instantiation n'est pas un ensemble de données séquentielles.
- $\{Euro_{1999}, Euro_{2000}, Euro_{2001}, Euro_{2002}\}$ est un ensemble de variables dont une instantiation est un ensemble de données séquentielles.

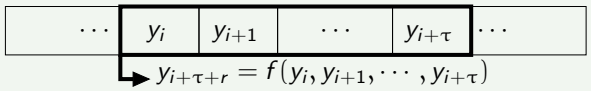
Un grand nombre d'applications peuvent se présenter sous la forme d'un modèle de données séquentielles :

- Données temporelles
  - Reconnaissance de la parole
  - Données sismiques
  - Données financières
  - ...
- Données générées par un processus mono-dimensionnel
  - Bio-séquences
  - ...

Prédiction de données séquentielles

Approches classiques

Principe : Prédire directement en fonction d'une fenêtre de valeurs.

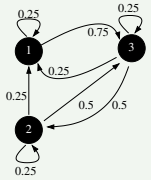


- Modèles linéaires : ARIMA (*auto-regressive integrated moving average*), ARMAX (*auto-regressive moving average exogenous variables*), etc.
- Modèles non linéaires : réseaux de neurones, arbres de décisions, etc.

inconvénients : Fenêtre limitée , peu adaptable (connaissance a priori ?) , mauvais comportement quand Y est multidimensionnel.

Modèle direct probabiliste : chaîne de Markov

- Une variable d'état discrète (X^n) (à l'instant n).
- Paramètres du modèle :
  - Condition initiale : $P(X^O)$
  - Modèle de transition : $P(X^n \mid X^{n-1})$

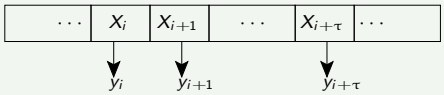


Dans le cas des CdM : fenêtre de taille 1

Représentation par espace d'états

Modèles à espace d'états (*state-space models*)

Principe : Les données temporelles (y_t) sont générées par un système dont l'état X_t évolue dans le temps et est non observé.



Une représentation par espace d'état implique une représentation d'une connaissance incertaine : l'état du système. Cette connaissance incertaine sera **probabilisée** ici.

On notera X_i ou X_t la (ou les) variable(s) aléatoire(s) représentant l'espace d'état.
On notera Y_i ou Y_t la (ou les) variable(s) aléatoire(s) représentant les observations.

Modèles connus à espace d'état

- HMM : Modèle de Markov caché
- KFM : filtre de Kalman
- dBN : réseaux bayésiens dynamiques

On peut voir les CdM, les HMMs et les KFMs comme des cas particulier de dBN. Voir plus loin.

Représentation par espace d'états (2)

Principes des modèles à espace d'états

- Les variables d'état forment une chaîne de Markov :
  - $P(X_1)$: les conditions initiales
  - $P(X_{t+1} \mid X_t)$ les probabilité de transition d'état.
- Les variables d'observations dépendent des variables d'état :
  - $P(Y_t \mid X_t)$ fonctions (probabilistes) d'observation.
- Le modèle doit vérifier certaines hypothèses :
  - Propriété de Markov sur X_t : $P(X_{t+1} \mid X_1, \cdots, X_t) = P(X_{t+1} \mid X_t)$
  - Propriété de Markov conditionnelle sur Y_t : $P(Y_t \mid X_t, Y_{t-1}) = P(Y_t \mid X_t)$
  - Invariance séquentielle (homogénéité) : $P(X_{t+1} \mid X_t)$ et $P(Y_t \mid X_t)$ ne dépendent pas de t.

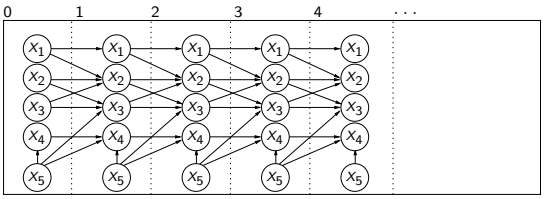
- HMM : X_t est un vecteur de variables aléatoires discrètes.
- KFM : X_t est un vecteur de variables aléatoires continues.
- dBN : Généralisation du modèle.

Les réseaux bayésiens dynamiques

dBN (dynamic BN)

Un réseau bayésien dynamique est un réseau bayésien dont les variables sont indicées par le temps t et par i : $X^{(t)} = X_1^{(t)}, \cdots, X_N^{(t)}$ et dont la distribution vérifie certaines propriétés :

- Markov ordre 1 :
  $P(X^{(t)} \mid X^{(0)}, \cdots, X^{(t-1)}) = P(X^{(t)} \mid X^{(t-1)})$,
- Homogénéité :
  $P(X^{(t)} \mid X^{(t-1)}) = \cdots = P(X^{(1)} \mid X^{(0)})$.



- L'adjectif *dynamique* n'est pas forcément bien choisi (puisqu'il y a homogénéité). *Temporel* ou *séquentiel* eût été de meilleur goût.
- Formellement, un réseau bayésien dynamique peut être considéré comme virtuellement infini.
- La définition ci-dessus est celle d'un dBN du premier ordre (t ne dépend que de $t-1$). On pourrait, bien évidemment, définir des dBNs d'ordre supérieur.
- On remarque que d'après la définition, les arcs d'un dBN vont de $X^{(t-1)}$ à X^t ou restent dans le même X^t (le même *timeslice*).

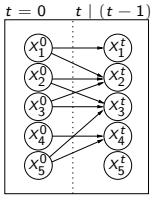
Les réseaux bayésiens dynamiques (2-TBN)

2-TBN

Un réseau bayésien dynamique est défini

- par les conditions initiales ($P(X^{(0)})$)
- par les relations entre des variables à l'instant $t-1$ et ces même variables à l'instant t (*timeslice*).

Cette représentation, appelée 2TBN (2 timeslice BN) permet de modéliser un BN virtuellemet infini qui en est le développement dans le temps, à partir d'un instant 0.

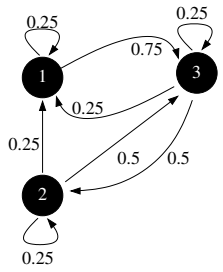


$P(x_1^{(t)}, \ldots, x_5^{(t)} \mid x_1^{(t-1)}, \ldots, x_5^{(t-1)})$

1024 contre 4+16+16+8+2=46 ! !

$= P(x_1^{(t)} \mid x^{(t-1)})$
$P(x_2^{(t)} \mid x_1^{(t-1)}, x_2^{(t-1)}, x_3^{(t-1)})$
$P(x_3^{(t)} \mid x_2^{(t-1)}, x_3^{(t-1)}, x_4^{(t-1)})$
$P(x_4^{(t)} \mid x_4^{(t-1)}, x_5^{(t-1)})$
$P(x_5^{(t)})$

Chaîne de Markov et réseaux bayésiens dynamiques

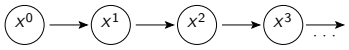


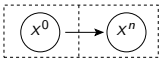
$$P(X^n \mid X^{n-1}) = \begin{pmatrix} 0.25 & 0 & 0.75 \\ 0.25 & 0.25 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{pmatrix}$$

Chaîne de Markov

- Une variable d'état discrète (X^n) (à l'instant n).
- Paramètres du modèle :
  - Condition initiale : $P(X^O)$
  - Modèle de transition : $P(X^n \mid X^{n-1})$

Réseau bayésien dynamique équivalent :

dBN : 

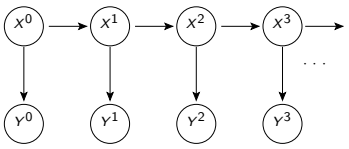
2TBN : 

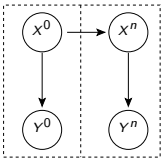
HMM et réseaux bayésiens dynamiques

HMM simple

- Une variable d'état discrète (X^n) (à l'instant n).
- Une variable d'observation discrète (Y^n)
- Paramètres du modèle :
  - Condition initiale : $P(X^O)$
  - Modèle de transition : $P(X^n \mid X^{n-1})$
  - Modèle d'observation : $P(Y^n \mid X^n)$

Ce qui donne, modélisé comme un réseau bayésien dynamique :

dBN : 

2TBN : 

Inférences dans les réseaux bayésiens dynamiques

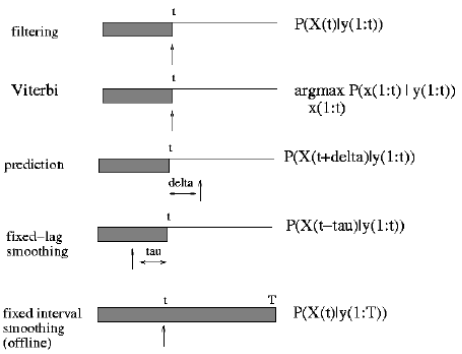
- A priori, très complexe : nombre de nœuds importants
- A priori, très complexe : "causes" communes dans un passé (lointain)

Complexité

NP-difficile

$P(X_t^i \mid y_{[1:r]})$?

- $r = t$: *Filtering*
- $r > t$: *Smoothing*
- $r < t$: *Prediction*
- MPE : *Viterbi*



rappels rapide : MDP

- Temps : t
- État : $S_t \in \mathcal{S}$, état à l'instant t ;
- Action : $a_t \in \mathcal{A}$, action à l'instant t ;
- Récompense : $r_t \in \mathbb{R}$, récompense à l'instant t ;
- Politique : π est une fonction de $\mathcal{S}$ dans $\mathcal{A}$:à chaque état, on associe une action à effectuer ($\pi(s_t) = a_t$).

Dynamique d'un processus markovien

Le système représenté évolue dans le temps, en fonctions de la dynamique propre du système et de la séquence de décisions de l'agent.
La dynamique est non-déterminisite et donc représentée par une probabilité de transition d'état à état, dépendant de l'action effectuée.

$$p(s_{t+1} \mid s_t, a_t) = p_{a_t}(s_{t+1} \mid s_t)$$

Trouver π^* : Value Iteration (VI) ou Policy Iteration (PI)

Rappels de l'algorithme Policy Iteration

Algorithme 6: Algorithme d'itération de la politique modifié - Critère γ-pondéré

initialiser $V_0 \in \mathcal{V}$ tel que $LV_0 \geq V_0$
$flag \leftarrow 0$
$n \leftarrow 0$
répéter
  pour $s \in S$ **faire**
    $\pi_{n+1}(s) \in \text{argmax}_{a \in A} \{r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V_n(s')\}$
    $(\pi_{n+1}(s) = \pi_n(s)$ si possible$)$
    $V_n^0(s) = \max_{a \in A} \{r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V_n(s')\}$
  $m \leftarrow 0$
  si $\| V_n^0 - V_n \| < \epsilon$ **alors** $flag \leftarrow 1$
  sinon
    répéter
      pour $s \in S$ **faire**
        $V_n^{m+1}(s) = r(s, \pi_{n+1}(s)) + \gamma \sum_{s' \in S} p(s' \mid s, \pi_{n+1}(s)) V_n^m(s')$
      $m \leftarrow m + 1$
    jusqu'à $\| V_n^{m+1} - V_n^m \| < \delta$
    $V_{n+1} \leftarrow V_n^m$
    $n \leftarrow n + 1$
jusqu'à $flag = 1$
retourner V_n, π_{n+1}

(origine : Frédérick Garcia, INRA, 2001)

Problèmes de calculs des probabilités de transition

Algorithme 6: Algorithme d'itération de la politique modifié - Critère γ-pondéré

initialiser $V_0 \in \mathcal{V}$ tel que $LV_0 \geq V_0$
$flag \leftarrow 0$
$n \leftarrow 0$
répéter
  pour $s \in S$ **faire**
    $\pi_{n+1}(s) \in \text{argmax}_{a \in A} \{r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V_n(s')\}$
    $(\pi_{n+1}(s) = \pi_n(s)$ si possible$)$
    $V_n^0(s) = \max_{a \in A} \{r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V_n(s')\}$
  $m \leftarrow 0$
  si $\| V_n^0 - V_n \| < \epsilon$ **alors** $flag \leftarrow 1$
  sinon
    répéter
      pour $s \in S$ **faire**
        $V_n^{m+1}(s) = r(s, \pi_{n+1}(s)) + \gamma \sum_{s' \in S} p(s' \mid s, \pi_{n+1}(s)) V_n^m(s')$
      $m \leftarrow m + 1$
    jusqu'à $\| V_n^{m+1} - V_n^m \| < \delta$
    $V_{n+1} \leftarrow V_n^m$
    $n \leftarrow n + 1$
jusqu'à $flag = 1$
retourner V_n, π_{n+1}

Si l'état est factorisé

- Comment calculer $P(s_{t+1} \mid s_t, a)$?
  Dans un espace factorisé, on peut noter $s = (x_1, \ldots, x_n)$.
  On utilise le $'$ pour indiquer le futur

- Comment calculer $P(x'_1, \ldots, x'_n \mid x_1, \ldots, x_n, a)$?
  En supposant chaque variable binaire, pour chaque action possible,
  Hyper-matrice à $2^{(2n)}$ paramètres ! ! Explosion combinatoire ! !

Curse of dimensionality (Bellman 1961)

Curse of dimensionality refers to the exponential growth of hypervolume as a function of dimensionality.

- Une solution pour limiter les dégâts :

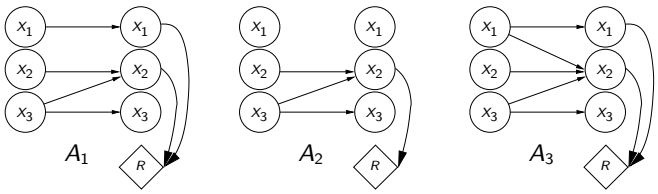
les réseaux bayésiens dynamiques

les FMDPs

Factored MDPs

Les FMDPs sont des MDPs dont l'espace d'état est factorisé. Cette factorisation permet de représenter les transitions comme des DBNs.

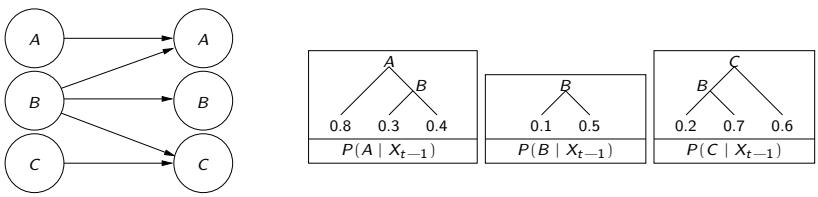
- La structure du DBN dépendant de l'action choisie.
- La récompense est une fonction de l'action et de l'ensemble des variables composants l'état. Mais se simplifie souvent (ne dépend que de certains des variables de l'état).



- Pour notre problème, l'inférence est facile : on connaît **complètement** $t-1$ et on veut t.

Améliorer la représentation ?

Peut on encore compacter la représentation ? OUI !
En utilisant les "symétries contextuelles", c'est-à-dire en utilisant une représentation des CPTs sous la forme d'arbre.



Il faut alors mettre à jour les algorithmes pour travailler sur les arbres (cf. SVI et SPI) :

SVI, SPI

Boutilier, C., Dearden, R., & Goldszmidt, M. **(2000)**. *Stochastic Dynamic Programming with Factored Representations*. Artificial Intelligence, 121.