le cnam

Apprentissage, réseaux de neurones et modèles graphiques

Chapitre 8 : Introduction aux réseaux bayesiens

Michel Crucianu

http://cedric.cnam.fr/~crucianm/ml.html

le cnam

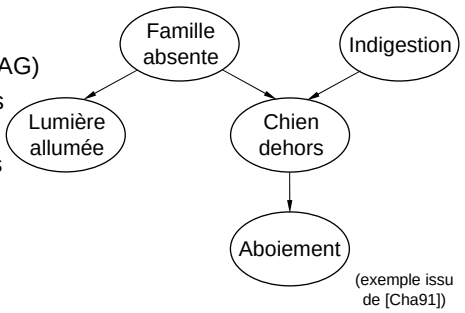
Contenu du chapitre

- Qu'est-ce qu'un réseau bayesien ?
- Inférence dans les réseaux bayesiens
  - Méthodes exactes
  - Méthodes approximatives
- Apprentissage des réseaux bayesiens

le cnam

Qu'est-ce qu'un réseau bayesien ?

- Réseau bayesien =
  1. Graphe orienté acyclique (DAG)
     - Nœuds : variables aléatoires (continues, booléennes…)
     - Arcs orientés : dépendances directes probabilistes
  +
  2. Probabilités conditionnelles

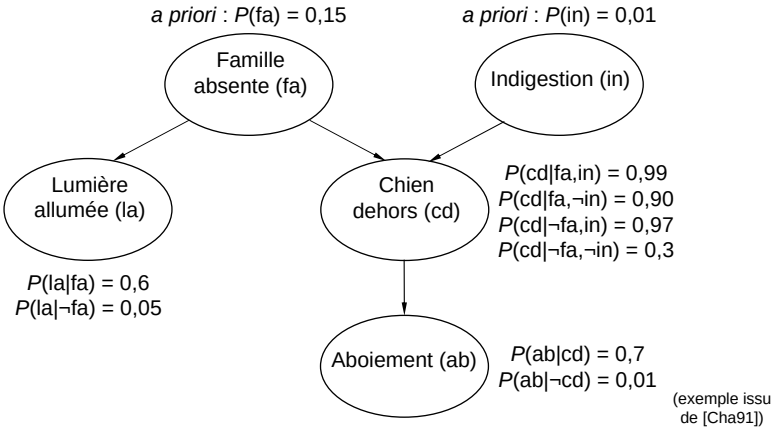


(exemple issu de [Cha91])

- Différentes interprétations et appellations correspondantes : réseaux de croyances, réseaux de causalité, modèles génératifs...
- Membres de la famille plus vaste des modèles graphiques, qui inclut aussi les graphes non orientés (champs de Markov) et mixtes

le cnam

Exemple simple



(exemple issu de [Cha91])

(notation : fa signifie *FA* = vrai, ¬fa signifie *FA* = faux)

le cnam

Quelques possibilités d'utilisation

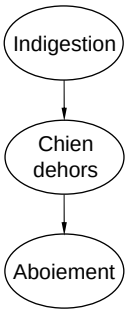
- Déterminer des probabilités *a posteriori*
  - Exemple (prédiction) : sachant que fa & ¬in → *P*(ab|fa,¬in) = ?
  - Exemple (diagnostic) : sachant que ab & ¬la → *P*(fa|ab,¬la) = ?
- Déterminer l'explication la plus vraisemblable
  - Exemple : sachant que ab, qui de fa et in est la plus probable ? Et si en plus ¬la ?
- Prise de décision
  - Exemple : pour réduire l'ambiguïté concernant fa, il vaut mieux se renseigner sur la ou sur ab ?

25 septembre 2013 RCP209 5

le cnam

Connexion linéaire

- L'ignorance de l'état du nœud intermédiaire *CD* (cd ou ¬cd) lie les deux autres nœuds :
  $P(AB|IN) \neq P(AB)$
- La connaissance de l'état du nœud intermédiaire *CD* (cd ou ¬cd) sépare les deux autres nœuds (le chemin est bloqué par l'évidence) :
  $P(AB|IN,CD) = P(AB|CD)$



(on ignore le reste du réseau)

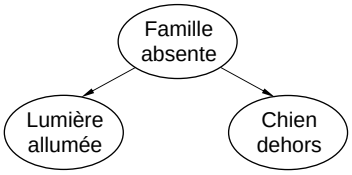
(si on sait que Chien dehors, alors connaître Indigestion n'apporte rien à l'évaluation de la probabilité de Aboiement)

25 septembre 2013 RCP209 6

le cnam

Connexion divergente

- L'ignorance de l'état du nœud intermédiaire *FA* (fa ou ¬fa) lie les deux autres nœuds :
  $P(CD|LA) \neq P(CD)$
- La connaissance de l'état du nœud intermédiaire *FA* (fa ou ¬fa) sépare les deux autres nœuds (le chemin est bloqué par l'évidence) :
  $P(CD|LA,FA) = P(CD|FA)$

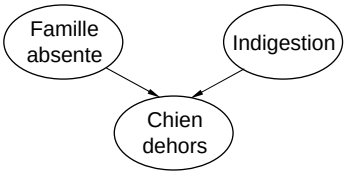


(si on sait que Famille absente, alors connaître l'état de la lumière n'apporte rien à l'évaluation de la probabilité de Chien dehors)

le cnam

Connexion convergente

- L'ignorance de l'état du nœud intermédiaire *CD* (cd ou ¬cd) sépare les deux autres nœuds :
  $P(FA|IN) = P(FA)$
- La connaissance de l'état du nœud intermédiaire *CD* (cd ou ¬cd) lie les deux autres nœuds :
  $P(FA|CD,IN) \neq P(FA|CD)$



(si on sait que Chien dehors, alors apprendre qu'il n'a pas d'indigestion rend plus probable Famille absente)

le cnam

D-dépendance, D-séparation

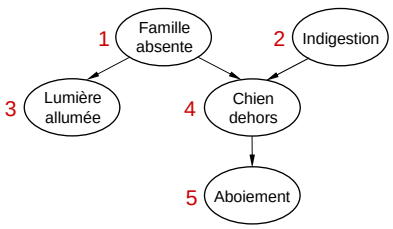
- Un chemin (non orienté) de la variable (nœud) *A* à *C* est une d-dépendance par rapport à un ensemble de variables $\mathcal{E}$ si, pour tout nœud *B* situé sur ce chemin,
  - le chemin est linéaire ou divergent en *B* et *B* n'est pas un membre de $\mathcal{E}$

  ou
  - le chemin est convergent en *B* et soit *B*, soit un de ses descendants est dans $\mathcal{E}$
- Deux variables *A* et *C* sont d-séparées par un ensemble de variables $\mathcal{E}$ s'il n'y a pas de d-dépendance par rapport à $\mathcal{E}$ entre elles

le cnam

Décomposition de la distribution jointe

- Pour toute distribution jointe et toute numérotation des variables,
  $$P(V_1,\ldots,V_n)=\prod_{i=1}^{n} P(V_i \mid V_1,\ldots,V_{i-1})$$
- En utilisant la notion de d-séparation et un ordonnancement des nœuds on peut montrer que, pour un réseau bayesien,
  $$P(V_1,\ldots,V_n)=\prod_{i=1}^{n} P(V_i \mid \text{parents}(V_i))$$
- Dans notre exemple :
  $P(FA,IN,LA,CD,AB)$ =
  $= P(FA) \cdot P(IN) \cdot$
  $\cdot P(LA|FA) \cdot$
  $\cdot P(CD|FA,IN) \cdot$
  $\cdot P(AB|CD)$



le cnam

Explication de l'exemple

- Sans tenir compte de la structure du réseau, avec la numérotation indiquée on peut écrire

  $P(FA,IN,LA,CD,AB) = P(FA) \cdot P(IN|FA) \cdot P(LA|FA,IN) \cdot$
  $\cdot P(CD|FA,IN,LA) \cdot P(AB|FA,IN,LA,CD)$

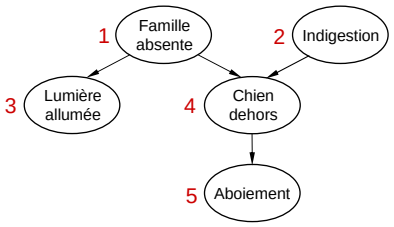
- Or,

$P(IN|FA) = P(IN)$ car le nœud convergent CD ne fait pas partie de la condition

$P(LA|FA,IN) = P(LA|FA)$

$P(CD|FA,IN,LA) = P(CD|FA,IN)$

$P(AB|CD,FA,IN,LA) = P(AB|CD)$



25 septembre 2013 RCP209 11

le cnam

Inférence dans un RB

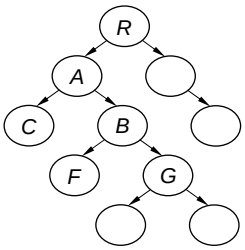
- Inférence : observations (évidence) → distribution sur les variables non directement observées
- Dans un cas général, le problème est NP-complet !
- Méthodes
  - Exactes
    - Complexité linéaire pour les arbres
    - Complexité supérieure en présence de boucles (non orientées)
  - Approximatives
    - Échantillonnage
    - Méthodes variationnelles
    - Propagation

25 septembre 2013 RCP209 12

le cnam

Méthode exacte pour les arbres

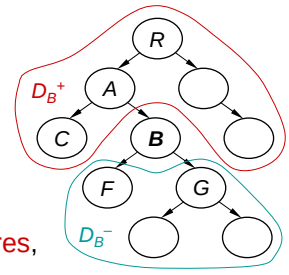
- Méthode basée sur « l'échange de messages », introduite dans [Pea86]
- Caractéristiques importantes d'un arbre
  - Chemin unique entre 2 nœuds, quels qu'ils soient
  - Aucune connexion convergente
- Notations employées dans la suite
  - Variables (nœuds) : A, B, C, …
  - Valeurs des variables : A_1, A_2, …
  - Évidence (données connues) : D
  - Probabilités dynamiques (conditionnées par l'évidence) : $Be(B_i) = P(B_i|D)$ (*belief* ou croyance : éviter la confusion avec $P(B_i)$)



le cnam

Séparation ancêtres – successeurs

- On note par D_B^- l'évidence présente dans les descendants de B et par D_B^+ l'évidence présente dans le reste de l'arbre
- $P(B_i|D_B^+,D_B^-) = P(D_B^+,D_B^-|B_i)\ P(B_i) / P(D_B^+,D_B^-) =$
  $= P(D_B^-|B_i)\ P(B_i|D_B^+)\ P(D_B^+) / P(D_B^+,D_B^-) =$
  $= \alpha\ P(D_B^-|B_i)\ P(B_i|D_B^+)$
  (où $\alpha = P(D_B^+) / P(D_B^+,D_B^-)$)

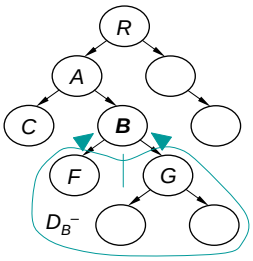


- Donc $Be(B_i) = \alpha\ P(D_B^-|B_i)\ P(B_i|D_B^+)$
  $= \alpha \cdot \lambda(B_i) \cdot \pi(B_i)$
  avec
  $\pi(B_i) = P(B_i|D_B^+)$:
  contribution causale des ancêtres,
  $\lambda(B_i) = P(D_B^-|B_i)$:
  contribution de diagnostic des descendants

le cnam

Calcul composante « successeurs »

- L'évidence présente dans D_B^- est à son tour composée de D_B^{1-}, …, D_B^{m-}, pour les m descendants directs de B
- Comme B sépare ces sous-arbres,
  $\lambda(B_i) = P(D_B^-|B_i) = \prod_k P(D_B^{k-}|B_i)$
- Si F est le k-ième descendant direct de B, alors
  $P(D_B^{k-}|B_i) = \sum_j P(D_F^-|B_i,F_j)\, P(F_j|B_i)$
  or, $P(D_F^-|B_i,F_j) = P(D_F^-|F_j) = \lambda(F_j)$
  et donc $P(D_B^{k-}|B_i) = \sum_j P(F_j|B_i)\, \lambda(F_j)$
- On notera $P(D_B^{k-}|B_i)$ par $\lambda_k(B_i)$

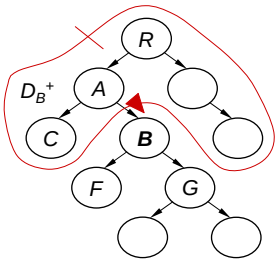


25 septembre 2013 RCP209 15

le cnam

Calcul composante « ancêtres »

- $\pi(B_i) = P(B_i|D_B^+) = \sum_j P(B_i|A_j,D_B^+)\, P(A_j|D_B^+)$
- Si A est l'unique parent direct de B, alors
  $P(B_i|A_j,D_B^+) = P(B_i|A_j)$
- Aussi, pour A on peut écrire
  $P(A_j|D_B^+) = \alpha\, P(A_j|D_A^+) \prod_k \lambda_k(A_j)$
  où $\lambda_k(A_j) = P(D_A^{k-}|A_j)$, pour le
  k-ième parmi les m frères de B
- On obtient donc
  $\pi(B_i) = \beta \sum_j P(B_i|A_j)\, \pi(A_j) \prod_k \lambda_k(A_j)$
- On notera $\pi(A_i) \prod_{l \neq B} \lambda_l(A_i)$ par $\pi_B(A_i)$



25 septembre 2013 RCP209 16

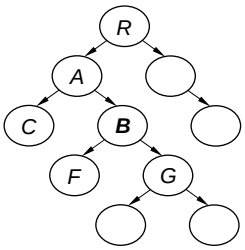
le cnam

Mécanisme de propagation

- Calculs réalisés pour chaque nœud (générique : *B*) :
  1. $\lambda(B_i) = \prod_{\text{descendants de } B} \lambda_F((\text{descendant } k \text{ de } B)_i)$
  2. Propagation vers l'unique ascendant *A*, pour tout *i*, de $\lambda_B(A_i) = \sum_j P(B_j|A_i)\ \lambda(B_j)$

  … (attente vague descendante)

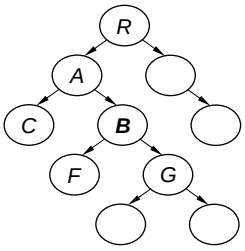
  3. $\pi(B_i) = \beta \sum_j P(B_i|A_j)\ \pi_B(A_j)$ (*A* étant l'unique parent de *B*)
  4. Propagation vers le *k*-ième descendant (ici *F*), pour tout *i*, de $\pi_F(B_i) = \pi(B_i) \prod_{l \neq k} \lambda_l(B_i)$
  5. Calcul croyance : $Be(B_i) = \alpha\ \lambda(B_i)\ \pi(B_i)$



le cnam

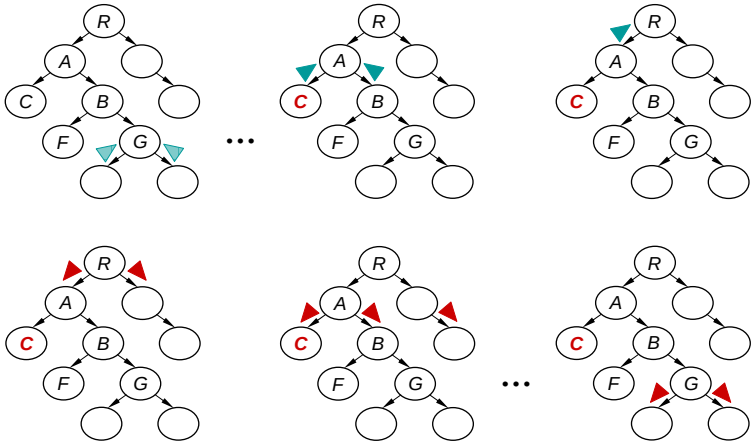
Mécanisme de propagation (2)

- Initialisations :
  - Nœud d'évidence (donnée), valeur *i* observée : $\pi(\text{nœud}_i) = \lambda(\text{nœud}_i) = 1$, $\pi(\text{nœud}_j) = \lambda(\text{nœud}_j) = 0$ pour $j \neq i$
  - Racine : $\pi(\text{racine}) = P(\text{racine})$ (*a priori*)
  - Feuille : $\lambda(\text{feuille}_i) = 1$
- Propagation globale :
  - Étape 1 : nœuds d'évidence et feuilles → ascendants,
  - Étape 2 : racine → descendants



le cnam

Illustration d'un parcours complet



le cnam

Extension aux réseaux sans boucles

- A sépare les sous-graphes $G_{TA}^{+} \cup G_{RA}^{+}$, G_{AC}^{-} et G_{AB}^{-}
- Comme dans le cas précédent, on peut montrer que

  $Be(A_i) = P(A_i|D_{TA}^{+}, D_{RA}^{+}, D_{AC}^{-}, D_{AB}^{-}) =$

  $= \alpha\ P(A_i|D_{TA}^{+}, D_{RA}^{+})\ P(D_{AC}^{-}|A_i)\ P(D_{AB}^{-}|A_i)$

  ou encore

  $Be(A_i) = \alpha\ P(D_{AC}^{-}|A_i)\ P(D_{AB}^{-}|A_i) \cdot$

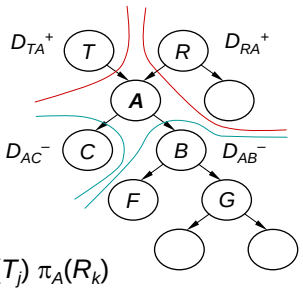
  $\cdot \sum_{jk} P(A_i|T_j,R_k)\ P(T_j|D_{TA}^{+})\ P(R_k|G_{RA}^{+})$
- Si on note

  $\lambda_C(A_i) = P(D_{AC}^{-}|A_i)$, $\lambda_B(A_i) = P(D_{AB}^{-}|A_i)$

  $\pi_A(T_j) = P(T_j|D_{TA}^{+})$, $\pi_A(R_k) = P(R_k|D_{RA}^{+})$

  on obtient

  $Be(A_i) = \alpha\ \lambda_C(A_i)\ \lambda_B(A_i) \sum_{jk} P(A_i|T_j,R_k)\ \pi_A(T_j)\ \pi_A(R_k)$



le cnam

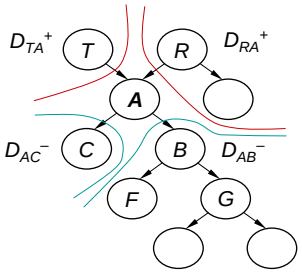
Extension aux réseaux sans boucles

- On démontre les relations suivantes [Pea86] :

  $\lambda_A(T_i) = \alpha \sum_j [\pi_A(R_j) \sum_k \lambda_C(A_k) \lambda_B(A_k) P(A_k|T_i,R_j)]$

  $\pi_C(A_i) = \alpha \lambda_B(A_i) [\sum_{jk} P(A_i|T_j,R_k) \pi_A(T_j) \pi_A(R_k)]$

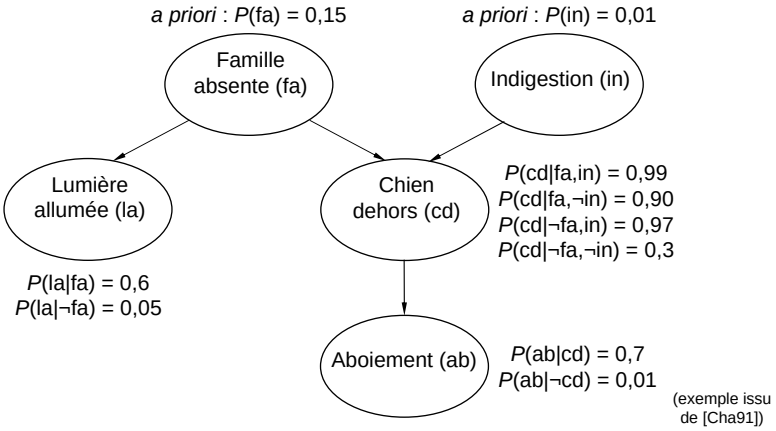
  $\Rightarrow$ propagation similaire à celle employée pour les arbres, mais il faut partir des nœuds où les calculs sont possibles !



25 septembre 2013 RCP209 21

le cnam

Retour à l'exemple simple



(exemple issu de [Cha91])

(notation : fa signifie *FA* = vrai, ¬fa signifie *FA* = faux)

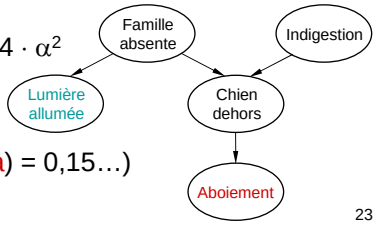
25 septembre 2013 RCP209 22

le cnam

Exemple : diagnostic

- Sachant que ab & ¬la, déterminer *Be*(fa) = *P*(fa|ab,¬la)
- Évidence : λ(ab) =1, λ(¬ab) = 0, λ(la) = 0, λ(¬la) = 1
- Propagation :
  $\lambda_{AB}(cd) = P(ab|cd)\cdot\lambda(ab) + P(\neg ab|cd)\cdot\lambda(\neg ab) = 0{,}7$, $\lambda_{AB}(\neg cd) = 0{,}01$
  $\pi_{CD}(in) = P(in|D_{IN\,CD}^{+}) = P(in) = 0{,}01$, $\pi_{CD}(\neg in) = P(\neg in) = 0{,}99$
  $\lambda_{CD}(fa) = \alpha \cdot (0{,}01 \cdot 0{,}6931 + 0{,}99 \cdot 0{,}631) = \alpha \cdot 0{,}69301$
  $\lambda_{CD}(\neg fa) = \alpha \cdot (0{,}01 \cdot 0{,}6793 + 0{,}99 \cdot 0{,}217) = \alpha \cdot 0{,}28276$
  $\lambda_{LA}(fa) = 0{,}4$, $\lambda_{LA}(\neg fa) = 0{,}95$

⇒ $Be(fa) = \alpha\cdot\lambda_{CD}(fa)\cdot\lambda_{LA}(fa)\cdot P(fa) \cong 0{,}04 \cdot \alpha^2$
et $Be(\neg fa) \cong 0{,}228 \cdot \alpha^2$
ce qui donne, en normalisant,
Be(fa) ≅ 0,15 (à comparer avec *P*(fa) = 0,15…)

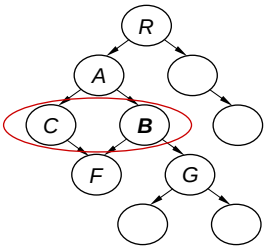


le cnam

Présence de boucles

- Principe de la méthode de [LS88] : pour chaque boucle, on remplace les variables intermédiaires de la boucle par une variable composée qui les regroupe
- Dans l'exemple ci-dessous : une variable *BC* regrouperait *B* et *C*
- Problème : *b* variables dans la boucle, avec *e* états chacune ⇒ e^b états pour la variable résultante !

⇒ méthode appliquée uniquement aux boucles de taille réduite



le cnam

Apprentissage d'un réseau bayesien

- Cas 1 : structure prédéfinie, observabilité complète
- Cas 2 : structure prédéfinie, observabilité partielle
- Cas 3 : ensemble de nœuds prédéfini, structure inconnue et observabilité complète
- Cas 4 : ensemble de nœuds prédéfini, structure inconnue et observabilité partielle

le cnam

Apprentissage : cas 1

- Structure prédéfinie, observabilité complète
  - On sait quelles variables interviennent et quelles variables dépendent directement de quelles autres
  - On cherche à déterminer les probabilités conditionnelles pour chaque dépendance directe
  - Toutes les variables (nœuds) sont directement observables

1. Estimation directe à partir des observations (maximum de vraisemblance, MV ou ML)
2. Utilisation de *a priori* si les observations sont en nombre très réduit (maximum *a posteriori*, MAP)

le cnam

Apprentissage : cas 2

- Structure prédéfinie, observabilité partielle
  - On sait quelles variables interviennent et quelles variables dépendent directement de quelles autres
  - On cherche à déterminer les probabilités conditionnelles pour chaque dépendance directe
  - Certaines variables (nœuds) ne sont pas observables
- Algorithme EM (*expectation-maximization*) :
  1. Étape E : en considérant que les paramètres (probabilités conditionnelles) correspondent à l'estimation courante, appliquer l'algorithme d'inférence pour déterminer l'espérance de chaque variable non observée
  2. Étape M : en considérant que les espérances calculées ont été observées, estimer les paramètres (probabilités conditionnelles) comme dans le cas 1 (approche MV ou MAP)

le cnam

Apprentissage : cas 3

- Ensemble de nœuds prédéfini, structure inconnue et observabilité complète (voir [Hec95])
  - On sait quelles variables interviennent, mais on ne sait pas quelles variables dépendent directement de quelles autres
  - On cherche à déterminer les dépendances directes et les probabilités conditionnelles pour chaque dépendance directe
  - Toutes les variables (nœuds) sont directement observables
- Exige
  1. Une méthode de génération de structures candidates
  2. Un critère de comparaison de graphes candidats (en général $P(D|G)$ ou $P(G|D)$)
- Souvent la structure est partiellement connue

le cnam

Apprentissage : cas 4

- Ensemble de nœuds prédéfini, structure inconnue et observabilité partielle (voir [Hec95])
  - On connaît une partie des variables qui interviennent mais d'autres peuvent exister et on ne sait pas quelles variables dépendent directement de quelles autres
  - On cherche à déterminer les dépendances directes et les probabilités conditionnelles pour chaque dépendance directe
  - Certaines variables (nœuds) ne sont pas observables
- EM structurel : à l'intérieur de EM, faire aussi une recherche locale dans l'espace des structures
- La création de variables (nœuds) intermédiaires peut aider à factoriser les dépendances

le cnam

Références

[Cha91] E. Charniak. Bayesian Networks without Tears, *AI Magazine*, pp. 50-63, 1991.

[Hec95] D. Heckerman. A tutorial on learning with Bayesian networks, *Microsoft Research tech. report*, MSR-TR-95-06.

[Jen01] F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer, 2001.

[JGJ98] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, L.K. Saul. An introduction to variational methods for graphical models. Dans M. Jordan (ed.) *Learning in Graphical Models*. MIT Press, 1998.

[LS88] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their applications to expert systems. *Proc. Royal Statistical Society*, Series B., 50, 154-227, 1988.

le cnam

Références

[NWL04] P. Naïm, P.-H. Wuillemin, Ph. Leray, O. Pourret, A. Becker. *Réseaux bayesiens*. Eyrolles, 2004.

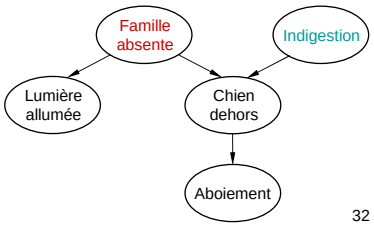
[Pea86] J. Pearl. Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, Vol. 29, No. 3, 241-288, September 1986.

[Wei00] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation* 12: 1-41, 2000.

le cnam

Exemple : prédiction

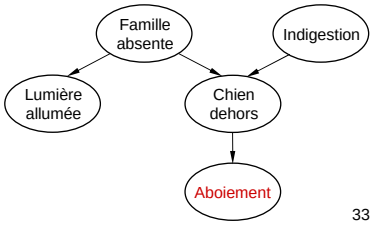
- Sachant que fa & ¬in, déterminer *Be*(ab) = *P*(ab|fa,¬in)



le cnam

Exemple : quelle explication

- Sachant que ab, qui de fa et in est la plus probable ?
  Et si en plus ¬la ?



le cnam

Exemple : quelle information chercher

- Pour réduire l'ambiguïté concernant fa, il vaut mieux se renseigner sur la ou sur ab ?