cours 3
Apprentissage d’un réseau bayésien

Aix Marseille université Initiative d'excellence

Master SID — Raisonnement dans l’incertain

Généralités sur l'apprentissage

Apprentissage de réseaux bayésiens

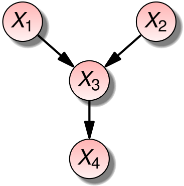
- Objectif : estimer
  - la structure $\mathcal{G}$ du réseau bayésien
  - les paramètres $P(X|\mathbf{pa}(X))$ du réseau bayésien
- En se fondant sur :
  - une ou plusieurs base(s) de données
    - complètes ou avec données manquantes
  - des connaissances *a priori*
    - contraintes sur la structure du RB
    - *A priori* sur les paramètres $P(X|\mathbf{pa}(X))$
    - connaissances expertes, *etc.*

1. Apprendre la structure du RB
2. Apprendre les paramètres sachant cette structure

Apprentissage de paramètres (1/4)

- Base de données **D** complète : pas de valeur manquante
  D $=$ N lignes : **D** $= \langle d^{(1)}, \ldots, d^{(N)} \rangle$
  ligne $d^{(i)}$: instanciation/observation de toutes les variables
- Structure du RB $\mathcal{G}$ connue

X_1	X_2	$\cdots$	X_n
1	toto	$\cdots$	0
$\cdots$	$\cdots$	$\cdots$	$\cdots$
2	titi	$\cdots$	0



- Θ : ensemble des paramètres du RB
  valeurs des tables $P(X_i | \mathbf{Pa}(X_i))$

Objectif : Estimer Θ qui « colle » le mieux aux données **D**

- « colle » le mieux $\Longrightarrow$ le plus vraisemblable

Apprentissage de paramètres (2/4)

Vraisemblance : $\mathcal{L}(\Theta : \mathbf{D}) = P(\mathbf{D}|\Theta)$

- Structure $\mathcal{G} \Longrightarrow$ indépendances

Estimation par maximum de vraisemblance

- Θ_i : les paramètres de $P(X_i|\mathbf{Pa}(X_i))$
- Estimer indépendamment chaque Θ_i :
  - en ne tenant compte que des colonnes X_i et $\mathbf{Pa}(X_i)$ de $\mathbf{D}$
  - en calculant $\mathrm{Argmax}_{\Theta_i} \mathcal{L}(\Theta_i : \mathbf{D})$

Apprentissage de paramètres (3/4)

Estimation d'une distribution marginale $P(X)$:

- X : variable aléatoire, domaine : $\Omega_X = \{x_1, \dots, x_n\}$
- $\Theta = \{\theta_1, \dots, \theta_n\}$ $\qquad P(X = x_i | \Theta) = \theta_i$
- N_i : nombre d'occurrences de x_i dans **D**

Théorème

Si $\Theta^* = \text{Argmax}_\Theta\, \mathcal{L}(\Theta : \mathbf{D})$, alors $\theta_i^* = \dfrac{N_i}{N}$

- Démonstration : Optimum obtenu pour $\frac{\partial \log \mathcal{L}(\Theta:\mathbf{D})}{\partial \theta_i} = 0$

⚠ contrainte : $\sum_{i=1}^n \theta_i = 1$

Exemple d'apprentissage de $P(X)$

- $X =$ « dé à 6 faces » 

- Base de données **D** :

X	1	3	2	4	2	5	1	1	2	3

- $N = 10$

- N_i :

N_1	N_2	N_3	N_4	N_5	N_6
3	3	2	1	1	0

- Estimation de $P(X = x_i)$:

x_1	x_2	x_3	x_4	x_5	x_6
0,3	0,3	0,2	0,1	0,1	0

Apprentissage de paramètres (4/4)

Estimation de $P(X_i|\mathbf{Pa}(X_i))$ par max de vraisemblance (MLE)

- r_i : taille du domaine de X_i
  domaine de $X_i = \{x_{i1}, \ldots, x_{ir_i}\}$
- q_i : taille du domaine de $\mathbf{Pa}(X_i)$
  domaine de $\mathbf{Pa}(X_i) = \{w_{i1}, \ldots, w_{iq_i}\}$
- N_{ijk} : nombre d'occurrences de $(X_i = x_k, \mathbf{Pa}(X_i) = w_{ij})$ dans $\mathbf{D}$
  $N_{ij} = \sum_k N_{ijk}$
- $\Theta_i = \{\theta_{ijk} : 1 \leq j \leq q_i, 1 \leq k \leq r_i\}$: paramètres de $P(X_i|\mathbf{Pa}(X_i))$
  $\theta_{ijk} = P(X_i = x_k|\mathbf{Pa}(X_i) = w_{ij}, \Theta_i)$
- Si $\Theta_i^* = \text{Argmax}_{\Theta_i} \mathcal{L}(\Theta_i : \mathbf{D})$, alors $\theta_{ijk}^* = \dfrac{N_{ijk}}{N_{ij}}$

N_{ij} peut être égal à 0 !

- $\Omega_X = \{x_1, x_2\}, \quad \Omega_Y = \{y_1, y_2\}$
- Base de données **D** :

X	x_1	x_2	x_2	x_2	x_1	x_1	x_1	x_1	x_2	x_2
Y	y_1	y_1	y_1	y_1	y_2	y_2	y_2	y_2	y_2	y_2

- N_{ijk} $\quad$ N_{ij}

	x_1	x_2	
y_1	1	3	4
y_2	4	2	6

- Estimation de $P(X|Y)$:

	x_1	x_2
y_1	$1/4$	$3/4$
y_2	$4/6$	$2/6$

- *A priori* : distribution $\pi(\Theta)$ sur les paramètres

  $\Longrightarrow$ Estimation par maximum *a posteriori* (MAP)

$$\Theta^* = \text{Argmax}_\Theta\, P(\Theta|\mathbf{D})$$

- Formule de Bayes : $P(\Theta|\mathbf{D}) = \dfrac{P(\mathbf{D}|\Theta) \times \pi(\Theta)}{P(\mathbf{D})}$

- $P(\mathbf{D}) = \sum_\theta P(\mathbf{D}|\Theta = \theta) \times \pi(\theta)$ = constante pour le Argmax

  $\Longrightarrow \Theta^* = \text{Argmax}_\Theta\, P(\mathbf{D}|\Theta) \times \pi(\Theta) = \text{Argmax}_\Theta \displaystyle\prod_{i=1}^{n} \mathcal{L}(\Theta_i : \mathbf{D})\pi(\Theta)$

- Hypothèse : indépendance des paramètres : $\pi(\Theta) = \displaystyle\prod_{i=1}^{n} \pi(\Theta_i)$

  $\Longrightarrow \Theta^* = \text{Argmax}_\Theta \displaystyle\prod_{i=1}^{n} \mathcal{L}(\Theta_i : \mathbf{D})\pi(\Theta_i)$

$$\Longrightarrow \Theta_i^* = \text{Argmax}_{\Theta_i}\, P(\mathbf{D}|\Theta_i) \times \pi(\Theta_i)$$

A priori classique

Distribution de Dirichlet

- Y : variable de domaine le simplexe k-dimensionnel :
  $\{k$-uplets $(y_1, \ldots, y_k)$ t.q. $y_i \geq 0$ pour tout i, et $\sum_{i=1}^k y_i = 1\}$
  $\Longrightarrow Y =$ ensemble de distributions de probabilité
- Soit $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_k\}$ t.q. $\alpha_i > 0$ pour tout i
- $Dir(Y, \boldsymbol{\alpha})$ = distribution de probabilité définie sur Ω_Y par :

$$Dir(Y = y, \boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{k} y_i^{\alpha_i - 1}$$

avec $\mathrm{B}(\cdot) =$ constante de normalisation $=$ fonction Beta

- Justification : Geiger & Heckerman (1997)

Apprentissage de paramètres avec *a priori* (2/2)

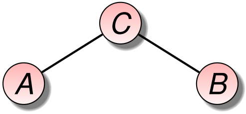
Estimation par Max a posteriori (MAP)

- *A priori* de Dirichlet d'hyperparamètres α_{ijk}
- Soit $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$
- Si $\Theta_i^* = \text{Argmax}_{\Theta_i} P(\mathbf{D}|\Theta_i) \times \pi(\Theta_i)$ alors $\theta_{ijk}^* = \dfrac{N_{ijk} + \alpha_{ijk} - 1}{N_{ij} + \alpha_{ij} - r_i}$

2 Indépendances et graphe

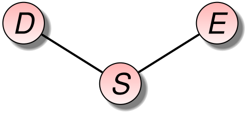
$\text{non}(A \perp\!\!\!\perp B)$

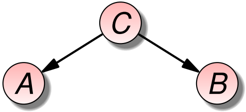
$A \perp\!\!\!\perp B \mid C$

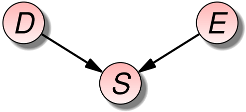


$D \perp\!\!\!\perp E$

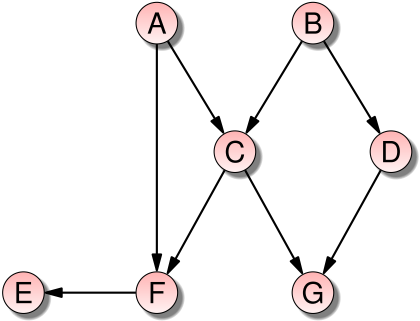
$\text{non}(D \perp\!\!\!\perp E \mid S)$







Indépendances et modèle graphique



- $A \perp\!\!\!\perp G$?
- $E \perp\!\!\!\perp G$?
- $A \perp\!\!\!\perp D$?
- $A \perp\!\!\!\perp D | G$?

$\Longrightarrow$ Raisonnement sur la partie graphique du modèle

$\Longrightarrow$ Vérifications d'indépendances conditionnelles sans connaître les valeurs des probabilités !

d-séparation

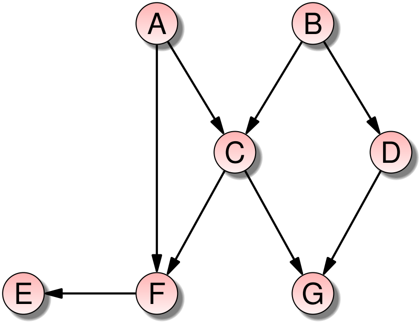
Chaîne $\langle X_1, \ldots, X_n \rangle$

- ensemble de nœuds $\{X_1, \ldots, X_n\}$
- pour tout $i \in \{1, \ldots, n-1\}$, le graphe contient l'arc $X_i \to X_{i+1}$ ou $X_{i+1} \to X_i$

Chaîne $\langle X_1, \ldots, X_n \rangle$ *bloquée par un ensemble* $\mathbf{Z}$

- Bloquée si et seulement si $\exists\ i \in \{2, \ldots, n-1\}$ tel que l'une des 2 propriétés ci-dessous est vérifiée :
  1. (X_{i-1}, X_i, X_{i+1}) est une V-structure : $X_{i-1} \to X_i \leftarrow X_{i+1}$, et ni X_i ni ses descendants ne sont dans $\mathbf{Z}$
  2. (X_{i-1}, X_i, X_{i+1}) n'est pas une V-structure et $X_i \in \mathbf{Z}$

d-séparation – bis



- Chaîne $\langle D, B, C, A\rangle$ bloquée par $\emptyset$?
- Chaîne $\langle D, B, C, A\rangle$ bloquée par $\{E\}$?
- Chaîne $\langle D, B, C, F, A\rangle$ bloquée par $\{E\}$?
- Chaîne $\langle D, G, C, A\rangle$ bloquée par $\{E\}$?
- Chaîne $\langle D, G, C, F, A\rangle$ bloquée par $\{E\}$?

d-séparation

- A, B, C trois variables aléatoires ou groupes de variables disjoints
- A est d-séparé de B par C si toute chaîne entre A et B est bloquée par C.

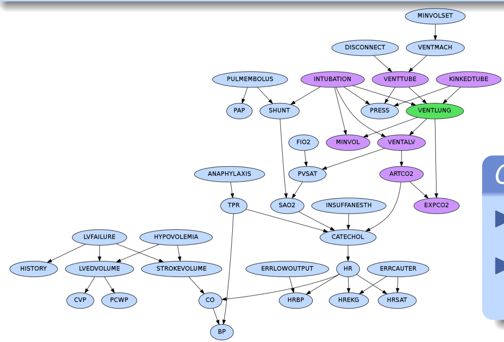
Théorème

- A, B, C trois variables aléatoires ou groupes de variables disjoints
- A est d-séparé de B par $C \Longrightarrow A \perp\!\!\!\perp B | C$

Conséquence de la *d*-séparation

Couverture de Markov d'un nœud X_i

- $\mathcal{X} = \{X_1, \dots, X_n\}$: nœuds du réseau bayésien
- $MB(X_i)$ = couverture de Markov de X_i
  $$= \text{ensemble de nœuds t.q. } X_i \perp\!\!\!\perp (\mathcal{X} \backslash (MB(X_i) \cup \{X_i\})) | MB(X_i)$$
- $MB(X_i) = \{\text{parents de } X_i\} \cup \{\text{enfants de } X_i\} \cup \{\text{parents des enfants de } X_i\} \backslash \{X_i\}$



- X_i : nœud en vert
- $MB(X_i)$: nœuds en violet

Conséquences

- $P(X_i | \mathcal{X} \backslash \{X_i\}) = P(X_i | MB(X_i))$
- Classification : observer $MB(X_i)$ suffit pour « classifier » X_i

Séparation, indépendance et maps

- Notation : $\langle A \perp_{\mathcal{G}} B|C\rangle$: A d-séparé de B par C dans graphe $\mathcal{G}$
  $A \perp\!\!\!\perp_P B|C$: selon la distribution de probabilité P,
  A indépendant de B conditionnellement à C

Définitions

- Soit $\mathcal{G}$ un graphe et P une distribution de probabilité
- $\mathcal{G}$ I-map (independence map) de P ssi $\langle A \perp_{\mathcal{G}} B|C\rangle \Longrightarrow A \perp\!\!\!\perp_P B|C$
  Interprétation : absence d'arc dans $\mathcal{G} \Longrightarrow$ indépendance dans P
- $\mathcal{G}$ D-map (dependence map) de P ssi $A \perp\!\!\!\perp_P B|C \Longrightarrow \langle A \perp_{\mathcal{G}} B|C\rangle$
  Interprétation : présence d'arc dans $\mathcal{G} \Longrightarrow$ dépendance
- $\mathcal{G}$ P-map (perfect map) de P ssi $A \perp\!\!\!\perp_P B|C \Longleftrightarrow \langle A \perp_{\mathcal{G}} B|C\rangle$
  Interprétation : présence d'arc dans $\mathcal{G}$ ssi dépendance
  absence d'arc dans $\mathcal{G}$ ssi indépendance

Réseaux bayésiens et maps

Définition d'un réseau bayésien

- $\mathcal{G}$ un graphe, muni de la d-séparation
- $\mathcal{G}$ est une I-map d'une distribution P

$\Longrightarrow$ $\mathcal{G}$ est une structure de réseau bayésien pour P

$\Longrightarrow$ P est factorisable selon le graphe $\mathcal{G}$

Propriété de Markov globale (PMG)

$\mathcal{G}$ vérifie la PMG pour P ssi $\mathcal{G}$ est une I-map pour P.

$\Longrightarrow$ les réseaux bayésiens vérifient la PMG

Propriété de Markov locale (PML)

$\mathcal{G}$ vérifie la PML pour P ssi pour toute variable X :

$$X \perp\!\!\!\perp_P \text{non descendants}(X) \mid \text{parents}(X).$$

$\Longrightarrow$ les réseaux bayésiens vérifient la PML

PMG $\Longleftrightarrow$ PML

3. Apprentissage de structure

Apprentissage de structure – une 1ère idée

- Objectif : déterminer la structure $\mathcal{G}$ à partir de données **D**
- Rappel : $\mathcal{G}$ P-map (perfect map) de P ssi
$$A \perp\!\!\!\perp_P B | C \Longleftrightarrow \langle A \perp_{\mathcal{G}} B | C \rangle$$
- Algorithme « naïf » :
  - créer toutes les structures $\mathcal{G}$ possibles
  - $\forall\ \mathcal{G}$, calculer tous les triplets (A, B, C) t.q. $\langle A \perp_{\mathcal{G}} B | C \rangle$
  - tester si $A \perp\!\!\!\perp_P B | C$ (par exemple, test du χ^2 en utilisant **D**)
  - si vrai pour tout triplet (A, B, C), structure $\mathcal{G}$ trouvée

Problèmes de l'algorithme naïf (1/3)

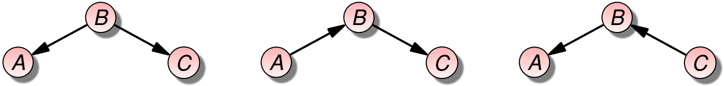
Théorème – Robinson(1977)

Le nombre de structures $\mathcal{G}$ à n nœuds est super-exponentiel :

$$\#(n) = \begin{cases} 1 & \text{si } n \leq 1 \\ \displaystyle\sum_{i=1}^{n} (-1)^{i+1} 2^{i(n-1)} C_n^i \times \#(n-1) & \text{si } n > 1 \end{cases}$$

$\Longrightarrow$ on ne peut pas tester toutes les structures

3 réseaux bayésiens équivalents (mêmes indep.) :



$\Longrightarrow$ ne les compter que pour 1 seul réseau !

$\Longrightarrow$ Appliquer l'algorithme dans l'espace des classes d'équivalence de Markov

Classe d'équivalence de Markov

Définition : équivalence de Markov

- $\mathcal{G}_1, \mathcal{G}_2$ deux structures de RB contenant les mêmes nœuds/variables aléatoires
- $\mathcal{G}_1, \mathcal{G}_2 \in$ même classe d'équivalence de Markov ssi, pour tous ensembles de variables disjoints A, B, C :

$$\langle A \perp_{\mathcal{G}_1} B | C \rangle \Longleftrightarrow \langle A \perp_{\mathcal{G}_2} B | C \rangle$$

$$A \leftarrow B \rightarrow C \qquad A \rightarrow B \rightarrow C \qquad A \leftarrow B \leftarrow C$$

Définitions

- $\mathcal{G}$ structure de RB. Squelette de $\mathcal{G}$ obtenu en remplaçant les arcs $X \rightarrow Y$ par des arêtes $X - Y$.
- $\{$nœuds $X, Y, Z\} =$ v-structure $\Longleftrightarrow$ dans $\mathcal{G}$, $\exists\, X \rightarrow Y \leftarrow Z$ et $\nexists X \rightarrow Z$ et $\nexists Z \rightarrow X$

Théorème – Verma et Pearl (1991)

$\mathcal{G}_1, \mathcal{G}_2 \in$ même classe d'équivalence de Markov ssi même squelette et mêmes v-structures.

Problèmes de l'algorithme naïf (2/3)

Appliquer l'algo dans l'espace des classes d'équivalence

Propriété expérimentale – Gillispie et Perlman (2002)

Ratio $\dfrac{\text{taille de l'espace des DAG}}{\text{taille de l'espace des classes d'équivalence}} \approx 3,7$

$\Longrightarrow$ Pas d'avantage à utiliser les classes d'équivalence

Théorème – Chickering, Heckerman, Meek (2004)

L'apprentissage de structure de RB est NP-hard.

- 2 alternatives :
  - Apprentissage « exact » pour des « petits » RB
  - Apprentissage « approché » $\Longrightarrow$ heuristiques

- 2ème problème : tester si $A \perp\!\!\!\perp_P B | C$ impossible si **D** petite ou $\Omega_{A \cup B \cup C}$ de grande taille

$$A \perp_{\mathcal{G}} B | C \Longleftrightarrow (X \perp_{\mathcal{G}} Y | C \; \forall X \in A, \forall Y \in B)$$

- Perfect map $\Longrightarrow A \perp\!\!\!\perp_P B | C \Longleftrightarrow (X \perp\!\!\!\perp_P Y | C \;\; \forall X \in A, \forall Y \in B)$

Tests d'indépendance

- Hypothèse (**DAG-faithfulness**) : P représentable par une perfect map $\mathcal{G}$
- Ne tester que l'indépendance conditionnelle de couples de variables

Théorème d'Hammersley-clifford

P distribution strictement positive $\Longrightarrow$ P représentable par une perfect map.

⚠ Relations déterministes entre variables $\Longrightarrow$ P non strictement positive

Idée générale de l'apprentissage sous contraintes

1. Apprendre le squelette via des tests d'indépendance

2a. Orienter les v-structures

2b. Propager ces orientations afin qu'elles ne créent pas de nouvelle v-structure

3. Orienter le reste des arêtes sans créer de nouvelle v-structure

- Algorithmes classiques :
  - Inductive causation (IC) – Verma et Pearl (1990)
  - PC – Spirtes, Glymour et Scheines (2000)

Bibliographie

- Chickering D., Heckerman D. et Meek C. (2004) « Large-Sample Learning of Bayesian Networks is NP-Hard », Journal of Machine Learning Research, 5 :1287–1330
- Geiger D. et Heckerman D. (1997) « A Characterization of the Dirichlet Distribution through Global and Local Parameter Independence », The Annals of Statistics, 25(3) :1344–1369
- Gillispie S. et Perlman M. (2002) « The size distribution for Markov equivalence classes of acyclic digraph models », Artificial Intelligence, 141 :137–155
- Robinson, R. (1977) « Counting unlabeled acyclic digraphs », Combinatorial Mathematics V, 622 : :28–43
- Spirtes E., Glymour C. et Scheines R. (2000) Causation, Prediction and Search, 2nd edition, Springer-Verlag
- Verma T. et Pearl J. (1990) « Equivalence and synthesis of causal models », Proceedings of UAI, 220–227