cours 2
Apprentissage de paramètres de RB

Aix*Marseille université Initiative d'excellence — Master SID — Raisonnement dans l'incertain

©CG(2026)

Généralités sur l'apprentissage

Apprentissage de réseaux bayésiens

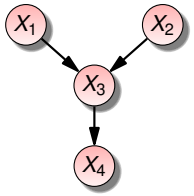
- Objectif : estimer
  - la structure $\mathcal{G}$ du réseau bayésien
  - les paramètres $P(X|\mathbf{pa}(X))$ du réseau bayésien
- En se fondant sur :
  - une ou plusieurs base(s) de données
    - complètes ou avec données manquantes
  - des connaissances *a priori*
    - contraintes sur la structure 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 **D**
  - en calculant $\text{Argmax}_{\Theta_i} \mathcal{L}(\Theta_i : \mathbf{D})$

Apprentissage de paramètres (3/4)

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

- X : variable aléatoire, domaine : $\Omega_X = \{x_1, \ldots, x_n\}$
- $\Theta = \{\theta_1, \ldots, \theta_n\}$ $\quad P(X = x_i|\Theta) = \theta_i$
- 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 **D**
  $N_{ij} = \sum_k N_{ijk}$
- $\Theta_i = \{\theta_{ijk} : 1 \le j \le q_i, 1 \le k \le r_i\}$: 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 !

Exemple d'apprentissage de $P(X|Y)$

- $\Omega_X = \{x_1, x_2\}, \quad \Omega_Y = \{y_1, y_2\}$
- 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

Apprentissage de paramètres avec *a priori* (1/2)

- *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

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

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

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

A priori classique

Distribution de Dirichlet

- Y : variable de domaine le simplexe k-dimensionnel :

  $\{k\text{-uplets } (y_1, \ldots, y_k) \text{ t.q. } y_i \geq 0 \text{ pour tout } i, \text{ et } \sum_{i=1}^{k} y_i = 1\}$

  $\Longrightarrow Y =$ 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 $\theta_{ijk}^* = \dfrac{N_{ijk} + \alpha_{ijk} - 1}{N_{ij} + \alpha_{ij} - r_i}$

Bibliographie

- 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