Rapport MVA : Graphical models - HWK 2

Plassier Vincent

1 Conditional independance and factorizations

1.1

Pour tout $p \in \mathcal{L}(G)$, nous avons $p(x,y,z,t) = p(x)p(y)p(z|x,y)p(t|z)$. Donnons un exemple pour lequel nous n'avons pas $X \perp\!\!\!\perp Y|Z$. Supposons que $X \sim \mathcal{B}(0.5)$ et $X \sim \mathcal{B}(0.5)$. Posons $Z = X + Y$ ainsi que $T = Z$. Nous avons immédiatement que

$$p(x,y,z,t) = p(x)p(y))p(z|x,y)p(t|z)$$

Cependant,

$$\mathbb{P}(X = 0, Y = 0|T = 1) = 0$$

alors que

$$\mathbb{P}(X = 0|T = 1) = \mathbb{P}(Y = 0|T = 1) = 1/2$$

D'où

$$\mathbb{P}(X = 0|T = 1) \times P(Y = 0|T = 1) \neq \mathbb{P}(X = 0, Y = 0|T = 1)$$

Par définition de l'espérance conditionnelle, on en déduit que X et Y ne sont pas indépendantes sachant T.

1.2

1.2.1

Tout d'abord, on remarque que

$$p(x) = p(z = 0)p(x|z = 0) + p(z = 1)p(x|z = 1) \qquad (\star)$$

Ecrivons les deux hypothèses. Nous avons :

$$\begin{aligned} &1) \quad p(x,y) \overset{X \perp\!\!\!\perp Y}{=} p(x)p(y) \\ &2) \quad p(x,y) = \sum_z p(x,y,z) = \sum_z p(z)p(x,y|z) \\ &\qquad \overset{X \perp\!\!\!\perp Y|Z}{=} \sum_z p(z)p(x|z)p(y|z) \end{aligned}$$

Or $p(z)p(y|z) = p(y)p(z|y)$. En combinant les deux points précédents, on obtient

$$p(x)p(y) = p(y)\sum_z p(x|z)p(z|y)$$

On en déduit que

$$0 = p(y)\left[p(x|z = 0)p(z = 0|y) + p(x|z = 1)p(z = 1|y) - p(x)\right]$$

En remplaçant $p(x)$ par son expression dans $(\star)$, on obtient :

$$\begin{aligned} 0 &= p(y)\left[p(x|z = 0)[1 - p(z = 1|y) - (1 - p(z = 1))] + p(x|z = 1)[p(z = 1|y) - p(z = 1)]\right] \\ &= p(y)\left(p(x|z = 0)[p(z = 1) - p(z = 1|y)] + p(x|z = 1)[p(z = 1|y) - p(z = 1)]\right) \\ &= p(y)[p(z = 1) - p(z = 1|y)][p(x|z = 0) - p(x|z = 1)] \end{aligned}$$

On obtient que

1. soit $x \longmapsto p(x|z=0) - p(x|z=1)$ est constante.
2. soit $y \longmapsto p(y)\,[p(z=1) - p(z=1|y)]$ est constante.

Cas 1 : supposons que $x \longmapsto p(z=1) - p(z=1|x)$ est constante. D'où $\alpha := p(z=1|x)$ est constante. Nous avons $\forall z \in \{0,1\}$:

$$\begin{aligned} p(z|x) &= zp(z=1|x) + (1-z)p(z=0|x) \\ &= zp(z=1|x) + (1-z)(1-p(z=1|x)) \end{aligned}$$

D'où $p(z|x) = z\alpha + (1-z)(1-\alpha)$. En sommant sur x' la quantité $p(x')p(z|x')$, on obtient :

$$\begin{aligned} p(z) = \sum_{x'} p(x')p(z|x') = (\sum_{x'} p(x'))\,(z\alpha + (1-\alpha)(1-z)) = \\ = z\alpha + (1-\alpha)(1-z) = p(z|x) \end{aligned}$$

On en déduit que $p(z) = p(z|x)$. D'où $X \perp\!\!\!\perp Z$.

Cas 2 : supposons que $y \longmapsto p(y)\,[p(z=1) - p(z=1|y)]$ est constante. On en déduit qu'il existe $c \in \mathbb{R}$ tel que $p(z=1,y) = p(z=1)p(y) + c$ En sommant sur y on obtient que $c = 0$. Comme $p(z=0,y) = p(y) - p(z=1,y)$, on a immédiatement que $p(z,y) = p(z)p(y)$. D'où $Y \perp\!\!\!\perp Z$.
Dans les deux cas on a montré que $X \perp\!\!\!\perp Z$ ou $Y \perp\!\!\!\perp Z$.

1.2.2

Prenons $X \sim \mathcal{B}(0.5)$ et $X \sim \mathcal{B}(0.5)$ tels que $X \perp\!\!\!\perp Y$. Définissons :

$$Z = X + 2^{Y+1}$$

Nous avons $Z(\Omega) = \{2,3,4,5\}$. De plus, $\forall z \in Z(\Omega)$, nous avons un unique couple (x,y) tel que

$$z = x + 2^{y+1}$$

On en déduit que

$$\begin{aligned} \mathbb{P}(X = x', Y = y'|Z = z) &= \mathbb{1}_{x'=x}\mathbb{1}_{y'=y} \\ &= \mathbb{P}(X = x'|Z = z)\mathbb{P}(Y = y'|Z = z) \end{aligned}$$

Donc $X \perp\!\!\!\perp Y|Z$ alors que ni X, ni Y n'est indépendant de Z.

2 Distributions factorizing in a graph

2.1

Soit $p \in L(G)$, par définition $\exists f_i \geq 0, \quad \sum_i f_i = 0$ et

$$\begin{aligned} \forall x, \quad p(x) &= \prod_{k=1}^{n} f_k(x_k, x_{pi_k}) \\ &= \left[\prod_{k \notin \{i,j\}} f_k(x_k, x_{\pi_k})\right] f_i(x_i, x_{\pi_i}) f_j(x_j, x_{\pi_j}) \end{aligned}$$

Posons $\forall k \in \{1,\dots,n\}, \pi'_k$ l'ensemble des parents de k dans (V, E'). Prenons $\forall k \notin \{i,j\}, \quad g_k = f_k$. D'après le cours, nous avons

$$\begin{aligned} f_i(x_i, x_{\pi_i}) f_j(x_j, x_{\pi_j}) &= p(x_i|x_{\pi_i})p(x_j|x_{\pi_j}) \\ &= p(x_i|x_{\pi_i})p(x_j|x_{\pi_i}, x_i) \\ &= p(x_i, x_j|x_{\pi_i}) \end{aligned}$$

De même,

$$\begin{aligned} p(x_i|x'_{\pi_i})p(x_j|x'_{\pi_j}) &= p(x_i|x_{\pi_i}, x_j)p(x_j|x_{\pi_i}) \\ &= p(x_i, x_j|x_{\pi_i}) \end{aligned}$$

On en déduit que

$$f_i(x_i, x_{\pi_i}) f_j(x_j, x_{\pi_j}) = p(x_i|x'_{\pi_i}) p(x_j|x'_{\pi_j})$$

En posant $g_i(x_i, x_{\pi'_i}) = p(x_i|x'_{\pi_i})$, ainsi que $g_j(x_j, x_{\pi'_j}) = p(x_j|x'_{\pi_j})$, le calcul précédent montre que

$$\forall x, \quad p(x) = \prod_{k=1}^{n} f_k(x_k, x_{pi_k})$$

et on vérifie immédiatement que

$$\forall k \in \{1, \dots, n\}, g_k \geq 0 \text{ et } \sum_k g_k = 1$$

On en déduit que $p \in L(G')$, d'où $L(G) \subset L(G')$, par symétrie on a $L(G') \subset L(G)$. On en déduit que $L(G) = L(G')$.

2.2

Commençons par démontrer que G ne peut pas contenir de v-structure. Raisonnons par l'absurde en supposant qu'il existe u, v, w tels que $u \longrightarrow w \longleftarrow v$. Par définition du "directed tree",$\exists r$ la racine de l'arbre ainsi que deux suites notées : $(a_i)_{1 \leq i \leq q}$ et $(b_j)_{1 \leq j \leq p}$ telles que :

$$\begin{cases} a_0 = r \longrightarrow a_1 \longrightarrow \dots \longrightarrow a_q = u \\ b_0 = r \longrightarrow b_1 \longrightarrow \dots \longrightarrow b_p = v \end{cases}$$

Ainsi, en posant $a_{q+1} = b_{p+1} = w$, on obtient deux chemins distincts allant de r vers w, cela contredit la définition du "directed tree".

Désormais, montrons que $\mathcal{L}(G) = \mathcal{L}(G')$. Prenons $p(x) \in \mathcal{L}(G)$. Par définition, nous avons des fonctions f_i telles que :

$$p(x) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i})$$

Posons $\Psi_i(x_i, x_{\pi_i}) = f_i(x_i, x_{\pi_i})$. Comme G' est non-orienté, nous avons que (x_i, x_{π_i}) est une clique. On en déduit que $\mathcal{L}(G) \subset \mathcal{L}(G')$. Attaquons-nous à la réciproque. Prenons $p(x) \in \mathcal{L}(G')$, nous avons

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \Psi_C(x_C)$$

Comme G ne contient aucune V-structure, on en déduit que $\forall C \in \mathcal{C}$, $\exists i$, $C = (x_i, x_{\pi_i})$.
D'où $\exists I_x \subset \{1, \dots, n\}$ tel que $\mathcal{C} = (x_i, x_{\pi_i})_{i \in I_x}$. Notons $\Psi_C(x_C) = \Psi_i(x_i, x_{\pi_i})$.
Quitte à poser

$$\varphi_i(x_i, x_{\pi_i}) = \begin{cases} \prod_{j \in I_x \cap \{i\}} \Psi_j(x_j, x_{\pi_j}) \text{ si } I_x \cap \{i\} \neq \emptyset \\ 1 \text{ sinon} \end{cases}$$

on peut supposer que $I_x = \{1, \dots, n\}$. De plus, comme G est fini, dirigé et acyclique il existe $k \in \{1, \dots, n\}$ qui ne soit le parent de personne. Quitte à inverser les rôles, on peut supposer que $k = n$. En posant $Z_n = \sum_{x_{V \setminus \{n\}}} \prod_{i=1}^{n-1} \Psi_i(x_i, x_{\pi_i}) \left(\sum_{x_n} \Psi_n(x_n, x_{\pi_n}) \right)$, nous avons :

$$\begin{aligned} p(x_1, \dots, x_{n-1}) &= \sum_{x_n} p(x_1, \dots, x_n) \\ &= \frac{1}{Z_{n-1}} \prod_{i=1}^{n-1} \Psi_i(x_i, x_{\pi_i}) \left(\frac{Z_{n-1}}{Z_n} \sum_{x_n} \Psi_n(x_n, x_{\pi_n}) \right) \quad (\star) \end{aligned}$$

Posons $f_n(x_n, x_{\pi_n}) = \frac{Z_{n-1}}{Z_n} \sum_{x_n} \Psi_n(x_n, x_{\pi_n}) \geq 0$. Supposons par récurrence sur le nombre de sommets du graphe G que $\mathcal{L}(G) = \mathcal{L}(G')$. Si $|G| = 1$, le résultat est immédiat. Désormais, supposons le résultat vrai au rang n-1. Comme $\frac{1}{Z_{n-1}} \Pi_{i=1}^{n-1} \Psi_i(x_i, x_{\pi_i}) \in \mathcal{L}(G')$. On a par hypothèse de récurrence qu'il existe $f_1, \ldots, f_{n-1}$ telles que :

$$\prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) = \frac{1}{Z_{n-1}} \prod_{i=1}^{n-1} \Psi_i(x_i, x_{\pi_i})$$

De plus, en sommant $(\star)$, on obtient :

$$\begin{aligned} 1 &= \sum_{x_1, \ldots, x_{n-1}} p(x_1, \ldots, x_{n-1}) \\ &= \sum_{x_1, \ldots, x_{n-1}} \left[\frac{1}{Z_{n-1}} \prod_{i=1}^{n-1} \Psi_i(x_i, x_{\pi_i}) \right] \left(\frac{Z_{n-1}}{Z_n} \sum_{x_n} \Psi_n(x_n, x_{\pi_n}) \right) \\ &= \sum_{x_n} f_n(x_n, x_{\pi_n}) \end{aligned}$$

On en déduit que $p(x) \in \mathcal{L}(G)$. D'où $\mathcal{L}(G) = \mathcal{L}(G')$.

3 Implementation - Gaussian mixtures

3.1 (a)

En initialisant l'algorithme K-means avec des valeurs choisies aléatoirement parmies nos données, on trouve une distorsion très variable. Celle-ci se situe genéralement aux alentours de 6000 mais peut parfois monter jusqu'à 11000. On en conclut que les résultats sont très changeants.

3.2 (b)

D'après la formule de Bayes, nous avons $\forall j \in [\![1, K]\!]$:

$$p_\theta(z = j|x) = \frac{p_\theta(x|z = j) p_\theta(z = j)}{p_\theta(x)} = \frac{\pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}{\sum_{j'=1}^{n} \pi_{j'} \mathcal{N}(x_i|\mu_{j'}, \Sigma_{j'})}$$

Désormais, écrivons le "complete likelihood" :

$$\begin{aligned} l_{c,t} &= \ln(p_\theta(x, z)) = \ln \left[\prod_{i=1}^{n} p_\theta(x_i, z_i) \right] \\ &= \ln \left[\prod_{i=1}^{n} p_\theta(x_i|z_i) p_\theta(z_i) \right] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{K} [\mathbb{I}(z_i = j) \ln \pi_j + \mathbb{I}(z_i = j) \ln p_\theta(x_i|z = j)] \\ &= \sum_{j=1}^{K} \left(\sum_{i=1}^{n} \mathbb{I}(z_i = j) \right) \ln \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{K} \mathbb{I}(z_i = j) \ln \left(\mathcal{N}(x_i|\mu_j, \Sigma_j) \right) \end{aligned}$$

1. Pour terminer le E-step, passons à l'espérance conditionnelle sachant X dans l'expression précédente. Nous remarquons que seules les fonctions indicatrices $\mathbb{I}(z_i = j)$ dépendent de X. Par linéarité de l'espérance, en gardant les notations du cours on obtient

$$\mathbb{E}_{Z|X,\theta_t} [\log(p_\theta(x, z))] = \sum_{j=1}^{K} \left(\sum_{i=1}^{n} \tau_i^j(\theta_t) \right) \ln \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{K} \tau_i^j(\theta_t) \ln \left(\mathcal{N}(x_i|\mu_j, \Sigma_j) \right)$$

2. Désormais, attaquons-nous à l'étape M-step. Cela consiste à maximiser en θ_t l'expression précédente. De plus, comme nous considérons un modèle gaussien dont la matrice est proportionnelle à l'identité $\theta_t = (\pi_{j,t}, \mu_{j,t}, v_{j,t}Id)_{j=1}^K$. Nous cherchons à maximiser

$$\sum_{j=1}^{K}\left(\sum_{i=1}^{n}\tau_i^j(\theta_t)\right)\ln\pi_j + \sum_{j=1}^{K}\sum_{i=1}^{n}\tau_i^j(\theta_t)\left[\frac{-d}{2}\ln(v_{j,t}) - \frac{1}{2v_{j,t}}\|x_i-\mu_{j,t}\|^2\right]$$

Comme la quantité à maximiser est une somme de quantités indépendantes en les variables à optimiser, on peut majorer chacune des parties de la somme indépendamment. Comme dans le cours, en écrivant le lagrangien on trouve :

$$\pi_{j,t+1} = \frac{1}{n}\sum_{i=1}^{n}\tau_i^j(\theta_t)$$

De plus, on constate que $\forall j$,

$$\mu_j \longmapsto -\frac{1}{2}\sum_{i=1}^{n}\tau_i^j(\theta_t)\|x_i-\mu_{j,t}\|^2$$

est concave et tend vers $-\infty$ en l'infini (sauf pour le cas trivial où tous les $\tau_i^j(\theta_t)$ sont nuls. Dans la suite nous utiliserons la convention $0/0 = 0$). Pour trouver le maximum de cette fonction, il suffit donc de trouver un point d'annulation du gradient. Comme dans le cours, nous trouvons :

$$\mu_{j,t+1} = \frac{\sum_i \tau_i^j(\theta_t)x_i}{\sum_i \tau_i^j(\theta_t)}$$

Définissons les fonctions

$$f_j : v \longmapsto \sum_i \tau_i^j(\theta_t)\left[\frac{-d\ln(v)}{2} - \frac{1}{2}(x_i-\mu_{j,t})^T(x_i-\mu_{j,t})\right]$$

Pour tout $j \in \{1,\dots,k\}$, f_j est convexe et coercive. Il suffit donc de trouver un point d'annulation de la dérivée pour trouver un maximum. Comme :

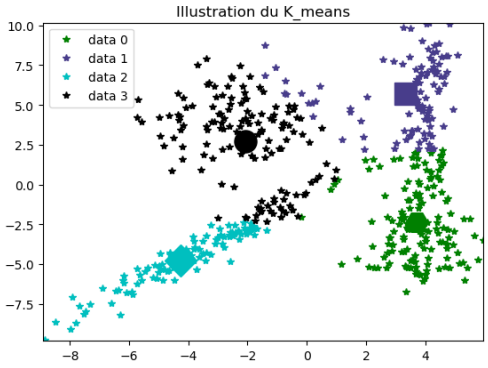
$$f_j'(v) = \frac{-d}{v}(\sum_i \tau_i^j(\theta_t)) + \frac{1}{v^2}(\sum_i \tau_i^j(\theta_t)\|x_i-\mu_{j,t}\|^2)$$

On obtient que f_j est minimale en

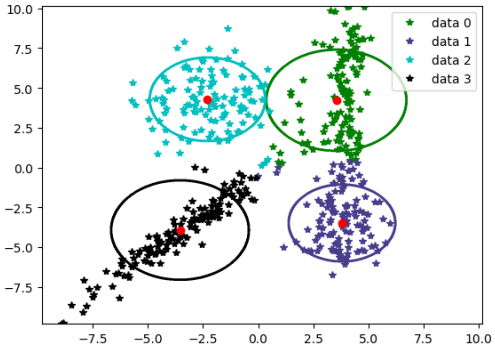
$$v_{j,t+1} = \frac{\sum_i \tau_i^j(\theta_t)\|x_i-\mu_{j,t}\|^2)}{d\sum_i \tau_i^j(\theta_t)}$$

4 Annexes - Figures

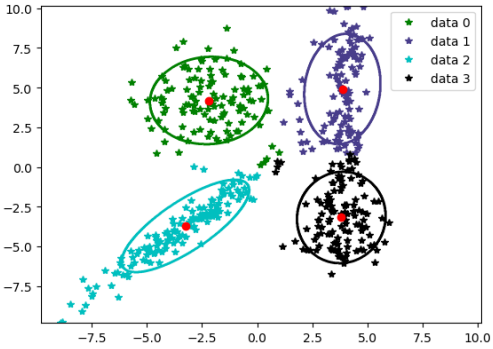
Log-likelihood	Entraînement	Test
isotropic model	-2645.389	-2681.55
general model	-2362.84	-2438.21



(a) Illustration du K-means



(b) Gaussian mixture, isotropic model



(c) Gaussian mixture, general model

FIGURE 1: Représentation des données d'entraînement A

Pour les deux modèles, nous trouvons que l'entropie des données d'entraînements est plus faible que celle données tests. En effet, le log-likelihood décrit la plausibilité du modèle en fonction des valeurs observées. Il est donc logique que le log-likelihood soit plus faible pour les données sur lesquelles le modèle s'est entraîné. De plus, nous constatons que l'entropie du modèle isotropique est plus importante que celle du modèle général. Cela est cohérent puisque le modèle général a moins de contraintes que celui isotropique, il peut donc mieux s'adapter aux observations.