
COMPTE-RENDU DU DEVOIR 2

Vincent Matthys

1 Indépendance conditionnelle et factorisations

1.1

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z &\Leftrightarrow p(x, y \mid z) = p(x \mid z)p(y \mid z) \quad \forall x, y, z \text{ t.q. } p(z) > 0 \\ &\Leftrightarrow p(x, y, z) = p(x \mid z)p(y \mid z)p(z) \quad \forall x, y, z \text{ t.q. } p(z) > 0 \\ &\Leftrightarrow \frac{p(x, y, z)}{p(y, z)} = p(x \mid z) \frac{p(y \mid z)p(z)}{p(y, z)} \quad \forall x, y, z \text{ t.q. } p(y, z) > 0 \\ &\Leftrightarrow p(x \mid y, z) = p(x \mid z) \quad \forall x, y, z \text{ t.q. } p(y, z) > 0 \end{aligned} \tag{1}$$

1.2

Etant donné le modèle graphique orienté G :

$$p \in \mathcal{L}(G) \Leftrightarrow \forall x, y, z, t \quad p(x, y, z, t) = p(x)p(y)p(z \mid x, y)p(t \mid z) \tag{2}$$

1.3

2 Distributions factorisant sur un graphe

2.1

2.2

3 Entropie et information mutuelle

3.1 Entropie

3.1.a)

Avec les conventions définies :

$$p_X(x) = \mathbb{P}(X = x) < 1 \Rightarrow p_X(x) \log(p_X(x)) < 0 \Rightarrow - \sum_{x \in \mathcal{X}} p_X(x) \log(p_X(x)) = H(X) \geq 0$$

$$H(X) = 0 \Rightarrow \forall x \in \mathcal{X}, p_X(x) \log(p_X(x)) = 0 \Rightarrow \forall x \in \mathcal{X}, p_X(x) = 1 \Rightarrow X \text{ is constant with probability 1} \quad (3)$$

3.1.b)

Par définition de la Kullback-Leibler divergence il vient :

$$\begin{aligned} D(p_X \parallel q) &= \sum_{x \in \mathcal{X}} p_X(x) \log \frac{p_X(x)}{q(x)} \\ &= - \sum_{x \in \mathcal{X}} p_X(x) \log q(x) - H(X) \\ &= - \sum_{x \in \mathcal{X}} p_X(x) \log \frac{1}{|\mathcal{X}|} - H(X) \quad \text{puisque} \quad \forall x \in \mathcal{X}, q(x) = \frac{1}{k} = \frac{1}{k} \quad (4) \\ &= \log k \sum_{x \in \mathcal{X}} p_X(x) - H(X) \\ &= \log k - H(X) \quad \text{car} \quad \sum_{x \in \mathcal{X}} p_X(x) = 1 \end{aligned}$$

3.1.c)

Avec les équations (3) et (4), on a directement

$$\log k - H(X) = D(p_X \parallel q) \leq \log k \quad (5)$$

3.2 Information mutuelle

3.2.a)

Par définition de l'information mutuelle :

$$\begin{aligned} I(X_1, X_2) &= \sum_{(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) \log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1) p_2(x_2)} \quad (6) \\ I(X_1, X_2) &= D(p_{1,2} \parallel p_1 p_2) \quad \text{par définition de la Kullback-Leibler divergence} \end{aligned}$$

Or la Kullback-Leibler divergence est positive pour toute paire $(p_{1,2}, p_1 p_2)$ de distributions, donc $I(X_1, X_2) \geq 0$.

3.2.b)

Toujours avec la définition de l'information mutuelle :

$$\begin{aligned}
I(X_1, X_2) &= \sum_{(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) \log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1) p_2(x_2)} \\
I(X_1, X_2) &= \sum_{(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) \log p_{1,2}(x_1, x_2) \\
&\quad - \sum_{(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) \log (p_1(x_1) p_2(x_2)) \\
I(X_1, X_2) &= -H(X_1, X_2) \\
&\quad - \sum_{x_1 \in \mathcal{X}_1} \left(\sum_{x_2 \in \mathcal{X}_2} p_{1,2}(x_1, x_2) \right) \log (p_1(x_1)) \\
&\quad - \sum_{x_2 \in \mathcal{X}_2} \left(\sum_{x_1 \in \mathcal{X}_1} p_{1,2}(x_1, x_2) \right) \log (p_2(x_2)) \\
I(X_1, X_2) &= -H(X_1, X_2) \\
&\quad - \sum_{x_1 \in \mathcal{X}_1} p_1(x_1) \log (p_1(x_1)) \\
&\quad - \sum_{x_2 \in \mathcal{X}_2} p_2(x_2) \log (p_2(x_2)) \\
I(X_1, X_2) &= H(X_1) + H(X_2) - H(X_1, X_2)
\end{aligned} \tag{7}$$

Et ainsi l'information mutuelle peut s'écrire uniquement à partir des entropies de X_1 , X_2 et de (X_1, X_2) .

3.2.c)

D'après l'équation (7), on peut réécrire :

$$\begin{aligned}
H(X_1, X_2) - (H(X_1) + H(X_2)) &= -I(X_1, X_2) \\
H(X_1, X_2) - (H(X_1) + H(X_2)) &\leq 0 \quad \text{puisque, } I(X_1, X_2) \geq 0
\end{aligned} \tag{8}$$

D'après l'équation (8), pour p_1 et p_2 données, l'entropie maximale correspond à $I(X_1, X_2) = 0$, c'est-à-dire $D(p_{1,2} \parallel p_1 p_2) = 0$ d'après l'équation (4). Or la Kullback-Leibler divergence s'anulle si et seulement si les distributions sont identiques. On a alors $p_{1,2} = p_1 p_2$, ce qui correspond à $X_1 \perp\!\!\!\perp X_2$.

4 Gaussian Mixtures

4.1 K-means

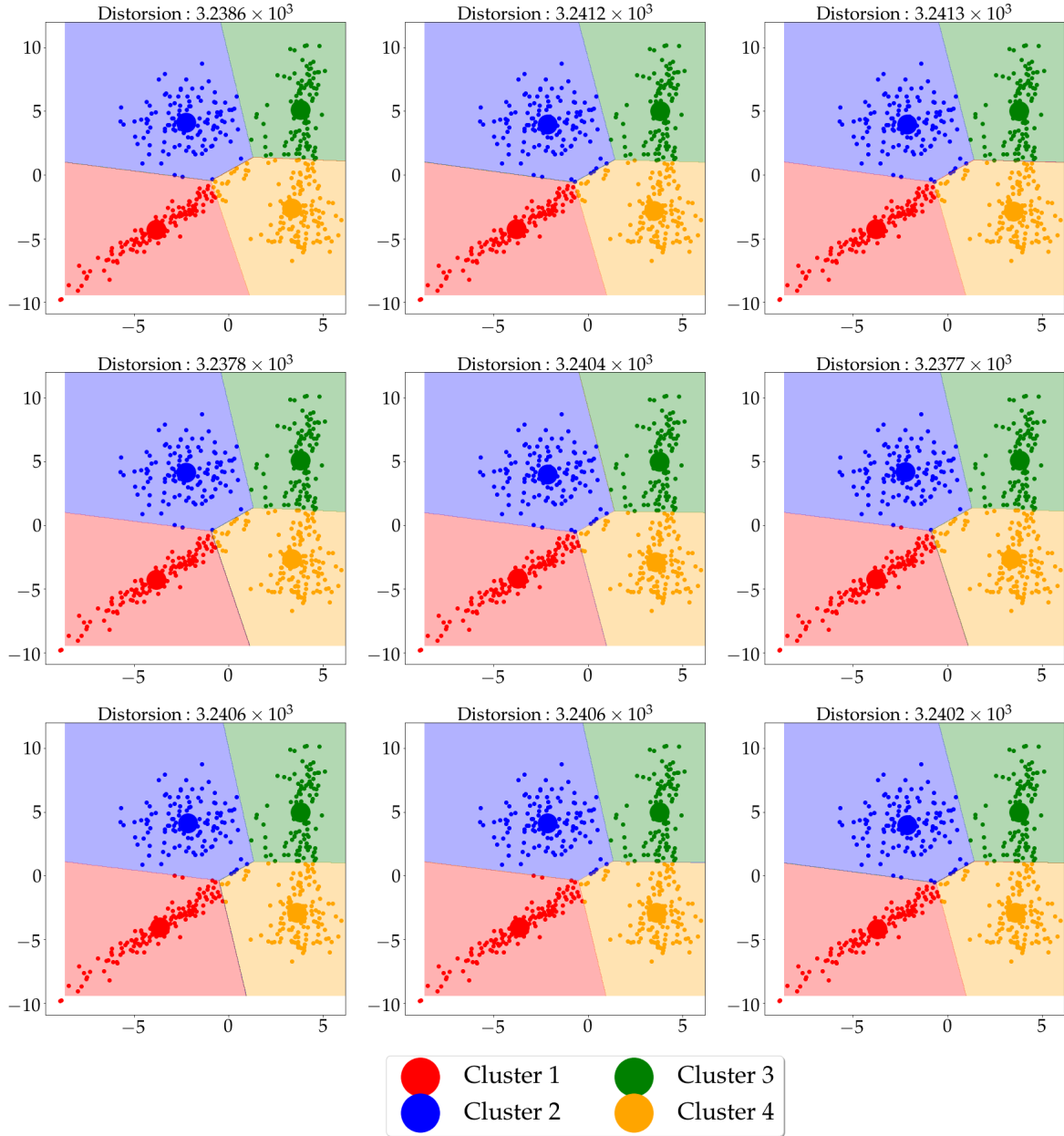


FIGURE 1 – Visualisation du clustering obtenu par k-means des données d’entraînement sur 9 initialisations aléatoires des centroïdes

En figure 1 sont présentés les résultats du clustering des données d’entraînement *EMGaussian.data* avec 9 initialisations aléatoires distinctes des centroïdes. La condition d’arrêt est remplie quand la variation de distorsion est inférieure à 1. On a utilisé l’angle des centroïdes finaux afin d’artificiellement repérer par les mêmes couleurs et mêmes chiffres les clusters situés dans des zones similaires du plan. On constate que la distorsion de ces 9 initialisations différentes ne diffère que par le 3^e chiffre significatif, avec une moyenne de 3.239×10^3 et avec un écart-type 1.335. Les minima locaux de cette distorsion ont donc des valeurs très semblables sur ces itérations, et les positions des centroïdes finaux sont regroupées dans la table 1. On

Centroïde	$\mu_x \pm \sigma_x$	$\mu_y \pm \sigma_y$
1	-3.747 ± 0.068	-4.193 ± 0.081
2	-2.177 ± 0.044	4.060 ± 0.078
3	3.794 ± 0.007	5.038 ± 0.048
4	3.483 ± 0.107	-2.795 ± 0.105

TABLE 1 – Positions des centroïdes sur les 9 initialisations aléatoires

peut constater que les variations autour des positions moyennes sont de l'ordre du dixième dans le pire des cas, correspondant à une variation en distance de l'ordre de 0.3 %. Néanmoins on constate que les points à la frontière des clusters ont tendance à basculer, tantôt vers l'un ou l'autre des clusters, notamment à la frontière des clusters jaune bleu et rouge, où les points proches de (0,0) sont facilement attribué par l'humain au cluster 1, rouge, qui présente une forme très allongée, différente des autres formes de clusters. Or la clusterisation par k-means ne permet pas de prendre en compte cette spécificité de forme. En définitif, les clusters trouvés par k-means sont, pour ce jeu de données d'entraînement, très stables pour la majorité des initialisations aléatoires et permette de séparer en première approximation, indépendamment de la forme du cluster, les points suivant un consensus proche de celui de l'homme, à l'exception de quelques points centraux.

4.2 Mixture de gaussiennes isotropiques

Dans le cas d'une mixture de gaussiennes isotropiques, on peut réécrire l'étape de maximisation suivant :

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \arg \max_{\boldsymbol{\theta}} (\mathbb{E}_{z|x, \boldsymbol{\theta}^{(t)}}) \\ \boldsymbol{\theta}^{(t+1)} &= \arg \max_{\boldsymbol{\theta}} \left(\sum_{n=1}^N \sum_{k=1}^4 q_{nk}^{(t)} \left(\log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \right) \right) \quad (9)\end{aligned}$$

où $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ et $q_{nk}^{(t)} = \mathbb{P}(z_n = k | x_n, \boldsymbol{\theta}^{(t)})$ avec z_n variable latente associée à x_n . On peut alors séparer la maximisation suivant $\boldsymbol{\pi}$ et suivant $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, de sorte que :

$$\begin{aligned}\boldsymbol{\pi}^{(t+1)} &= \arg \max_{\boldsymbol{\pi}} \left(\sum_{n=1}^N \sum_{k=1}^4 q_{nk}^{(t)} \log \pi_k \right) \quad \text{avec} \quad \forall n, \sum_{k=1}^4 q_{nk} = 1 \\ (\boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Sigma}^{(t+1)}) &= \arg \max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \left(\sum_{n=1}^N \sum_{k=1}^4 q_{nk} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \right) \right) \quad (10)\end{aligned}$$

Les deux problèmes de minimisation en (10) sont différentiables et concaves (le deuxième étant analogue au calcul du MLE pour une distribution normale) et on a :

$$\begin{aligned}\frac{\partial \left(\sum_{n=1}^N \sum_{k=1}^4 q_{nk}^{(t)} \log \pi_k^{(t+1)} \right)}{\partial \pi_k^{(t+1)}} &= \sum_{n=1}^N q_{nk}^{(t)} - \lambda \pi_k^{(t+1)} = 0 \quad \text{avec} \quad \lambda \in \mathbb{R} \\ \pi_k^{(t+1)} &\propto \sum_{n=1}^N q_{nk}^{(t)} \quad (11) \\ \pi_k^{(t+1)} &= \frac{\sum_{n=1}^N q_{nk}^{(t)}}{\sum_{k=1}^4 \sum_{n=1}^N q_{nk}^{(t)}} = \frac{1}{N} \sum_{n=1}^N q_{nk}^{(t)}\end{aligned}$$

$$\frac{\partial \left(\sum_{n=1}^N \sum_{k=1}^4 q_{nk}^{(t)} \left(-\frac{1}{2} (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \right) \right)}{\partial \mu_k} = \sum_{n=1}^N q_{nk}^{(t)} \left(\Sigma_k^{-1} (x_i - \mu_k^{(t+1)}) \right) = 0 \quad (12)$$

$$\mu_k^{(t+1)} = \frac{\sum_{n=1}^N q_{nk}^{(t)} x_i}{\sum_{n=1}^N q_{nk}^{(t)}}$$

Avec l'hypothèse $\Sigma_k = \sigma_k^2 \mathbf{I}$ d'isotropie, il vient :

$$\frac{\partial \left(\sum_{n=1}^N \sum_{k=1}^4 q_{nk}^{(t)} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \right) \right)}{\partial \sigma_k^2} = \sum_{n=1}^N q_{nk} \left(\frac{1}{\sigma_k^2} - \frac{1}{\sigma_k^4} (x_i - \mu_k)^\top (x_i - \mu_k) \right)$$

$$\sigma_k^2 = \frac{\sum_{n=1}^N q_{nk} (x_i - \mu_k)^\top (x_i - \mu_k)}{\sum_{n=1}^N q_{nk}} \quad (13)$$

4.3 Mixture de gaussiennes g n rales

4.4 subsection name