Mélanges de lois et Algorithme EM

Master MIGS

Yoann Offret

1 Mélanges de lois

Soient X et Y deux v.a.r. et $C \sim \mathcal{B}(p)$ indépendante de (X,Y). On pose

$$Z = CX + (1-C)Y = X\mathbb{1}_{\{C=1\}} + Y\mathbb{1}_{\{C=0\}} = \left\{ \begin{array}{lll} X & \text{si} & C=1 \\ Y & \text{si} & C=0 \end{array} \right. .$$

On dit que la loi de Z est un mélange des lois de X et Y avec les coefficients de mélange p et $1-p$. On suppose maintenant que X et Y ont des densités g et h.

1. Déterminer la densité f de Z.
2. Déterminer la loi conditionnelle de C sachant $Z = z$.
3. Justifier que $\mathbb{E}[\varphi(C,Z)|Z=z] = \mathbb{E}[\varphi(C,z)|Z=z]$.
4. On suppose $X \sim \mathcal{E}(\lambda)$ et $Y \sim \mathcal{E}(\mu)$.
   (a) Simuler un N-échantillon de Z.
   (b) Vérifier expérimentalement que votre N-échantillon est bien distribué comme un mélange de ces deux lois.
5. Généraliser ce modèle de mélange à un mélange de K variables aléatoires réelles $X_1, \cdots, X_K$ avec les proportions $p_1, \cdots, p_K$.
6. Les données suivantes ont été simulées avec $\lambda = 1/2$, $\mu = 1/4$ et $p = 1/3$:

$$4.02, 0.33, 0.86, 0.67, 0.86, 0.80, 4.43, 1.09, 3.75, 0.85, 3.09, 5.17, 10.9, 2.85, 6.76$$

   Donner pour chacune de ces valeurs la probabilité qu'elle soit issue de X ou de Y.

2 Estimation des paramètres : algorithme EM

On reprend l'exemple précédent avec $X \sim \mathcal{E}(\lambda)$, $Y \sim \mathcal{E}(\mu)$ et $C \sim \mathcal{B}(p) \perp\!\!\!\perp (X,Y)$. On suppose ici que l'on ne connaît aucun paramètres du modèle, à savoir $\Theta = (\lambda, \mu, p)$ et que l'on observe un N-échantillon du mélange noté $\mathbf{z} = (z_1, \cdots, z_N)$. On veut retrouver les paramètres du modèle et en plus prédire/classifier les lois (celles de X ou Y) dont sont issues les observations.

2.1 Vraisemblance globale

Soient $(C_i, (X_i, Y_i))_{1 \leq i \leq N}$ i.i.d. de même loi que $(C, (X,Y))$ et $Z_i = C_i X_i + (1-C_i)Y_i$ pour $1 \leq i \leq N$. On pose $\mathbf{C} = (C_1, \cdots, C_N)$, $\mathbf{Z} = (Z_1, \cdots, Z_n)$. Quelque soit $(\mathbf{c}, \mathbf{z}) = (c_i, z_i)_{1 \leq i \leq N}$ on note $L_\Theta(\mathbf{c}, \mathbf{z})$ la vraisemblance de l'échantillon $(C_i, Z_i)_{1 \leq i \leq N}$. Justifier que

$$\log L_\Theta(c,z) = \log(p\lambda) \sum_{i=1}^N \mathbb{1}_{c_i=1} + \log((1-p)\mu) \sum_{i=1}^N \mathbb{1}_{c_i=0} - \lambda \sum_{i=1}^N z_i \mathbb{1}_{c_i=1} - \mu \sum_{i=1}^N z_i \mathbb{1}_{c_i=0}.$$

2.2 Phase Expectation

On note $\Theta' = (\lambda', \mu', p')$ un autre jeu de paramètres et on considère

$$\mathcal{E}_{\Theta'}(\Theta) = \mathbb{E}_{\Theta'}[\log L_\Theta(C, Z)|Z = z]. \tag{1}$$

Ici $\mathbb{E}_{\Theta'}$ représente l'espérance d'une fonction de $(\mathbf{C}, \mathbf{Z})$ en supposant que les paramètres sont donnés par Θ'.

1. Soit $1 \leq i \leq N$. Justifier que la loi de (C_i, Z_i) sachant $\mathbf{Z}$ est la même que celle de (C_i, Z_i) sachant Z_i.
2. Montrer que

$$\begin{aligned}\mathcal{E}_{\Theta'}(\Theta) = \log(p\lambda) \sum_{i=1}^{N} \frac{p'\lambda' e^{-\lambda' z_i}}{p'\lambda' e^{-\lambda' z_i} + (1-p')\mu' e^{-\mu' z_i}} \\ + \log((1-p)\mu) \sum_{i=1}^{N} \frac{(1-p')\mu' e^{-\mu' z_i}}{p'\lambda' e^{-\lambda' z_i} + (1-p')\mu' e^{-\mu' z_i}} \\ - \lambda \sum_{i=1}^{N} z_i \frac{p'\lambda' e^{-\lambda' z_i}}{p'\lambda' e^{-\lambda' z_i} + (1-p')\mu' e^{-\mu' z_i}} \\ - \mu \sum_{i=1}^{N} z_i \frac{(1-p')\mu' e^{-\mu' z_i}}{p'\lambda' e^{-\lambda' z_i} + (1-p')\mu' e^{-\mu' z_i}}.\end{aligned} \tag{2}$$

2.3 Phase Maximisation

On cherche maintenant à maximiser la quantité (1) en fonction de Θ. Notons que l'on a les contraintes $0 < p < 1$ et $\lambda, \mu > 0$.

Montrer que

$$\begin{cases} p &= \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \mathbb{P}_{\Theta'}(C_i = 1|Z_i = z_i). \\ \dfrac{1}{\lambda} &= \dfrac{\sum_{i=1}^{N} z_i P_{\Theta'}(C_i = 1|Z_i = z_i)}{\sum_{i=1}^{N} \mathbb{P}_{\Theta'}(C_i = 1|Z_i = z_i)} \\ \dfrac{1}{\mu} &= \dfrac{\sum_{i=1}^{N} z_i P_{\Theta'}(C_i = 0|Z_i = z_i)}{\sum_{i=1}^{N} \mathbb{P}_{\Theta'}(C_i = 0|Z_i = z_i)} \end{cases}.$$

2.4 Algorithme EM

On définit pour $0 < p_0 < 1$ et $\lambda_0, \mu_0 > 0$ la suite récurrente

$$\begin{cases} p_{n+1} &= \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \mathbb{P}_{\Theta_n}(C_i = 1|Z_i = z_i). \\ \dfrac{1}{\lambda_{n+1}} &= \dfrac{\sum_{i=1}^{N} z_i \mathbb{P}_{\Theta_n}(C_i = 1|Z_i = z_i)}{\sum_{i=1}^{N} \mathbb{P}_{\Theta_n}(C_i = 1|Z_i = z_i)} \\ \dfrac{1}{\mu_{n+1}} &= \dfrac{\sum_{i=1}^{N} z_i \mathbb{P}_{\Theta_n}(C_i = 0|Z_i = z_i)}{\sum_{i=1}^{N} \mathbb{P}_{\Theta_n}(C_i = 0|Z_i = z_i)} \end{cases}.$$

Le prince de l'algorithme est tout simplement d'itérer cette suite. Sous de bonnes conditions, cette suite converge vers un point fixe $(p_\infty, \lambda_\infty, \mu_\infty)$ et lorsque N est suffisamment grand cette limite est proche de la bonne valeur des paramètres.

> Implémenter cet algorithme sur un jeu de données simulé et vérifié qu'il vous permet d'estimer correctement les paramètres du modèle et de classifier les données.

2.5 Mélange de gaussiennes

La Figure 1 représente un échantillon de taille 5000 d'un mélange de deux vecteurs gaussiens du plan. On notera p le paramètre de mélange , $m_1 = (x_1, y_1)$ et $m_2 = (x_2, y_2)$ les moyennes des deux gaussiennes du mélange et enfin

$$K_1 = \begin{pmatrix} \sigma_1^2 & c_1 \\ c_1 & \tau_1^2 \end{pmatrix} \quad \text{et} \quad K_2 = \begin{pmatrix} \sigma_2^2 & c_2 \\ c_2 & \tau_2^2 \end{pmatrix},$$

les matrices de covariances de ces vecteurs gaussiens.

> Implémenter l'algorithme EM sur le jeu de données *Brute.txt* et retrouver les paramètres du modèle. On procédera alors à une classification des données et comparera avec le fichier *Classification.txt* donnant la bonne classification de ces données.

On pourra importer les données avec les commandes :

```
NDbrute = []
with open("Brute.txt", "r") as fichier:
    for l in fichier:
        T = [str(d) for d in l.split(",")]
        NDbrute.append([float(T[0]), float(T[1])])

Label=[]
with open("Classification.txt","r") as fichier:
    for l in fichier:
    S=[str(d) for d in l.split(",")]
    Label.append(float(S[0]))
```

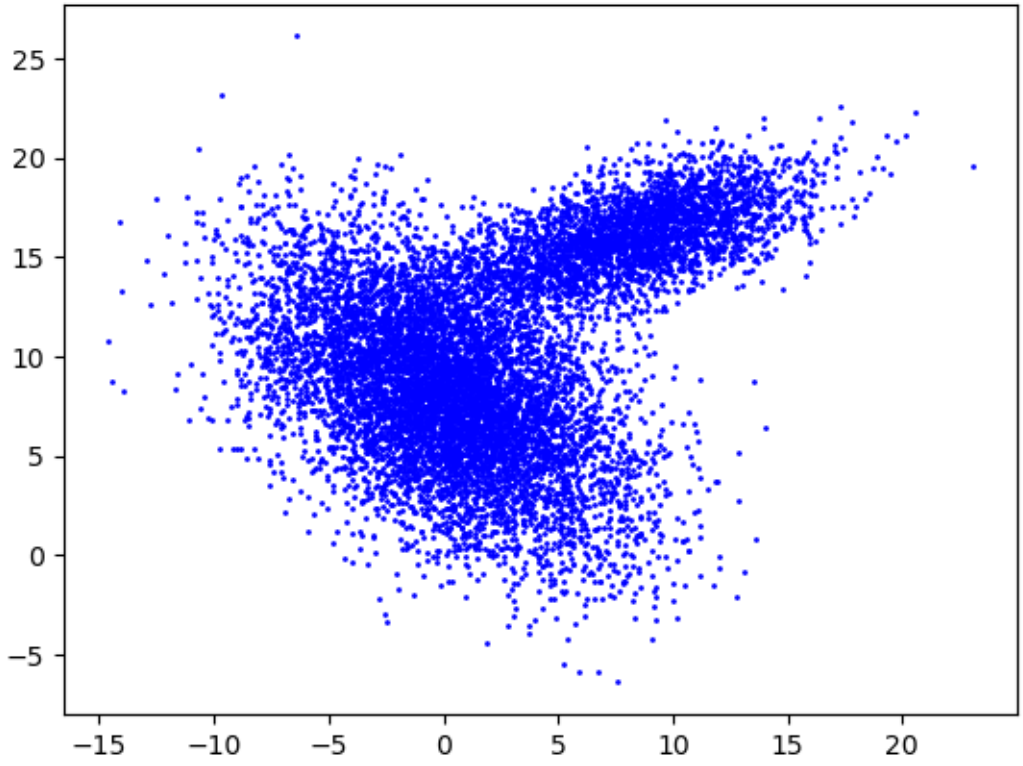

FIGURE 1 – Mélange de deux gaussiennes de $\mathbb{R}^2$