

Statistique II

M2 Radiophysique médicale, INSTN, 2023

Clément GAUCHY (clement.gauchy@cea.fr) Blog: clgch.github.io

CEA SACLAY

Sommaire

1. Algorithme *Expectation-Maximization*
2. Application à la Tomographie à Émission de Positons (TEP)
3. Introduction aux méthodes Bayésiennes
4. Modèle Bayésien pour la segmentation d'image TEP



Introduction

- Méthode EM (**Expectation-Maximization**) (proposée par Dempster en 1977) dans le cas où la mise en oeuvre du maximum de vraisemblance peut être complexe

- Exemple : données issues d'un mélange de gaussiennes

- Soit un n-échantillon $(x_i)_{1 \leq i \leq n}$ où chaque x_i est issu d'une loi $\mathcal{N}(\mu_j, \sigma_j^2)$ parmi m avec la probabilité α_j telle que $\sum_{j=1}^m \alpha_j = 1$ et pour tout j , $0 \leq \alpha_j \leq 1$. En notant $\Theta = (\alpha_j, \mu_j, \sigma_j^2)_{1 \leq j \leq m}$ les $3m$ paramètres inconnus, la densité de probabilité s'écrit :

$$p(x|\Theta) = \sum_{j=1}^m \frac{\alpha_j}{\sqrt{2\pi\sigma_j^2}} \exp - \frac{(x - \mu_j)^2}{2\sigma_j^2}$$

- la log-vraisemblance à partir d'un n-échantillon $\mathcal{D}_n = (x_i)_{1 \leq i \leq n}$

$$\ell(\Theta|\mathcal{D}_n) = \ln \prod_{i=1}^n p(x_i|\Theta) = \sum_{i=1}^n \ln \left[\sum_{j=1}^m \frac{\alpha_j}{\sqrt{2\pi\sigma_j^2}} \exp - \frac{(x_i - \mu_j)^2}{2\sigma_j^2} \right]$$

- Information manquante : on ne connaît pas la loi (j parmi m) qui a généré la réalisation x_i
- L'estimation des paramètres Θ par maximum de vraisemblance est un problème complexe notamment dans le cas de la grande dimension (nombre m de lois, dimension des $x \in \mathbb{R}^d$).

Algorithme EM pour un mélange de lois (1)

- Variable aléatoire X distribuée selon un mélange (combinaison convexe) de lois :

$$p_{\Theta}(X) = \sum_{j=1}^m \alpha_j p_j(X; \theta_j) \text{ où } \Theta = (\alpha_j, \theta_j)_{1 \leq j \leq m}, \sum_{j=1}^m \alpha_j = 1, \forall j \in \{1, \dots, m\}, 0 \leq \alpha_j \leq 1.$$

- Log vraisemblance d'un n-échantillon $\mathcal{D}_n = (x_i)_{1 \leq i \leq n}$ (tirages i.i.d.)

$$\begin{aligned} \ell(\Theta | \mathcal{D}_n) &= \ln \prod_{i=1}^n p_{\Theta}(x_i) \\ &= \sum_{i=1}^n \ln \sum_{j=1}^m \alpha_j p_j(x_i; \theta_j) \text{ difficile à optimiser} \end{aligned}$$

Algorithme EM pour un mélange de lois (1)

- Variable aléatoire X distribuée selon un mélange (combinaison convexe) de lois :

$$p_{\Theta}(X) = \sum_{j=1}^m \alpha_j p_j(X; \theta_j) \text{ où } \Theta = (\alpha_j, \theta_j)_{1 \leq j \leq m}, \sum_{j=1}^m \alpha_j = 1, \forall j \in \{1, \dots, m\}, 0 \leq \alpha_j \leq 1.$$

- Log vraisemblance d'un n -échantillon $\mathcal{D}_n = (x_i)_{1 \leq i \leq n}$ (tirages i.i.d.)

$$\begin{aligned} \ell(\Theta | \mathcal{D}_n) &= \ln \prod_{i=1}^n p_{\Theta}(x_i) \\ &= \sum_{i=1}^n \ln \sum_{j=1}^m \alpha_j p_j(x_i; \theta_j) \text{ difficile à optimiser} \end{aligned}$$

- Introduisons Z variable aléatoire discrète (*variable latente*) à valeurs dans $\{1, 2, \dots, m\}$, définissant le rang de la loi qui a généré x :

$$\underbrace{p_{\Theta}(X, Z = j)}_{\text{loi jointe}} = \alpha_j p_j(X; \theta_j)$$

- Log-vraisemblance du n -échantillon $(x_i, z_i)_{1 \leq i \leq n}$ (tirages i.i.d.)

$$\ell(\Theta | (x_i, z_i)_{1 \leq i \leq n}) = \ln \prod_{i=1}^n p_{\Theta}(X = x_i, Z = z_i) = \sum_i \sum_j \ln \alpha_j p_{z_j}(x_i; \theta_{z_j})$$

- La log-vraisemblance ne peut pas être évaluée (on ne connaît pas z_i !).

Algorithme EM pour un mélange de lois (2)

- Dans l'EM, la log-vraisemblance $\ell(\Theta|(x_i, z_i)_{1 \leq i \leq n})$ est remplacée par son espérance conditionnelle aux observations x_i .
- Espérance de la log-vraisemblance est calculée par rapport à la variable aléatoire Z de probabilité $q_{\Theta}(Z|(x_i)_{1 \leq i \leq n})$ conditionnelle à l'observation x et à la valeur courante des paramètres Θ

$$\ell(\Theta|(x_i, z_i)_{1 \leq i \leq n}) \rightarrow \ell_c(\Theta|(x_i)_{1 \leq i \leq n}; \Theta_k) := \mathbb{E}_{Z \sim q_{\Theta_k}(Z|(x_i)_{1 \leq i \leq n})}[\ell(\Theta|Z, (x_i)_{1 \leq i \leq n}; \Theta_k)]$$

- Algorithme itératif : Θ_0 (initialisation) puis calcul des $\Theta_k = (\alpha_j^{(k)}, \theta_j^{(k)})_{1 \leq j \leq m}$ pour $k = 1, 2, \dots$ par
- Première étape : **Espérance**

$$\ell_c(\Theta|(x_i)_{1 \leq i \leq n}; \Theta_k) = \mathbb{E}_{Z \sim q_{\Theta_k}(Z|(x_i)_{1 \leq i \leq n})}[\ell(\Theta|Z, (x_i)_{1 \leq i \leq n}; \Theta_k)]$$

- la loi $q_{\Theta_k}(Z|(x_i)_{1 \leq i \leq n})$ est conditionnelle par rapport aux observations $(x_i)_{1 \leq i \leq n}$ et en supposant la loi de mélange définie par les paramètres Θ_k

Algorithme EM pour un mélange de lois (2)

- Dans l'EM, la log-vraisemblance $\ell(\Theta|(x_i, z_i)_{1 \leq i \leq n})$ est remplacée par son espérance conditionnelle aux observations x_i .
- Espérance de la log-vraisemblance est calculée par rapport à la variable aléatoire Z de probabilité $q_{\Theta}(Z|(x_i)_{1 \leq i \leq n})$ conditionnelle à l'observation x et à la valeur courante des paramètres Θ

$$\ell(\Theta|(x_i, z_i)_{1 \leq i \leq n}) \rightarrow \ell_c(\Theta|(x_i)_{1 \leq i \leq n}; \Theta_k) := \mathbb{E}_{Z \sim q_{\Theta_k}(Z|(x_i)_{1 \leq i \leq n})}[\ell(\Theta|Z, (x_i)_{1 \leq i \leq n}; \Theta_k)]$$

- Algorithme itératif : Θ_0 (initialisation) puis calcul des $\Theta_k = (\alpha_j^{(k)}, \theta_j^{(k)})_{1 \leq j \leq m}$ pour $k = 1, 2, \dots$ par
- Première étape : **Espérance**

$$\ell_c(\Theta|(x_i)_{1 \leq i \leq n}; \Theta_k) = \mathbb{E}_{Z \sim q_{\Theta_k}(Z|(x_i)_{1 \leq i \leq n})}[\ell(\Theta|Z, (x_i)_{1 \leq i \leq n}; \Theta_k)]$$

- la loi $q_{\Theta_k}(Z|(x_i)_{1 \leq i \leq n})$ est conditionnelle par rapport aux observations $(x_i)_{1 \leq i \leq n}$ et en supposant la loi de mélange définie par les paramètres Θ_k
- Seconde étape : **Maximisation**

$$\Theta_{k+1} = \arg \max_{\Theta} \ell_c(\Theta|(x_i)_{1 \leq i \leq n}, \Theta_k)$$

- la valeur des paramètres à l'itération $k + 1$ est la valeur de Θ qui maximise $\ell_c(\Theta|(x_i)_{1 \leq i \leq n}, \Theta_k)$
- **Avantage:** Propriété théorique $\ell(\Theta_{k+1}|(x_i)_{1 \leq i \leq n}) \geq \ell(\Theta_k|(x_i)_{1 \leq i \leq n}) \rightarrow$ on maximise la log-vraisemblance conditionnellement aux observations $(x_i)_{1 \leq i \leq n}$.
- **Inconvénients:** Minima locaux \rightarrow solution numérique dépend de la condition initiale Θ_0 des paramètres

Algorithme EM pour le mélange de Gaussiennes (1)

- La probabilité conditionnelle $q_{\Theta}(Z|X)$ s'obtient par la règle classique de Bayes

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)} \Rightarrow q_{\Theta}(Z|X) = \frac{p_{\Theta}(Z, X)}{p_{\Theta}(X)}$$

- Cas du mélange de m Gaussiennes de paramètres $\Theta = (\alpha_j, \mu_j, \sigma_j^2)_{1 \leq j \leq m}$

- A l'étape k , les paramètres Θ_k sont connus :

$$\begin{aligned} q_{\Theta_k}(Z = j|X = x) &= \frac{\alpha_j^{(k)} p_j(X = x | \mu_j^{(k)}, (\sigma^2)_j^{(k)})}{\sum_{s=1}^m \alpha_s^{(k)} p_s(X = x | \mu_s^{(k)}, (\sigma^2)_s^{(k)})} \\ \ell_c(\Theta|(x_i)_{1 \leq i \leq n}; \Theta_k) &= \sum_{i=1}^n \sum_{j=1}^m q_{\Theta_k}(Z = j|X = x_i) \ln(\alpha_j p_j(x_i | \theta_j)) \\ &= \sum_{i=1}^n \sum_{j=1}^m q_{\Theta_k}(Z = j|X = x_i) \ln \frac{\alpha_j}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right] \\ &= \sum_{i=1}^n \sum_{j=1}^m q_{\Theta_k}(Z = j|X = x_i) \left[\ln(\alpha_j) - \frac{\ln \sigma_j^2}{2} - \frac{(x_i - \mu_j)^2}{2\sigma_j^2} - \frac{n}{2} \ln(2\pi) \right] \end{aligned}$$

- Rappel de EM : à l'itération $k + 1$, Θ_{k+1} est obtenu en maximisant $\ell_c(\Theta|(x_i)_{1 \leq i \leq n}, \Theta_k)$:

$$\text{Valeurs qui annulent le gradient} \rightarrow \frac{\partial}{\partial \Theta} \ell_c(\Theta|(x_i)_{1 \leq i \leq n}; \Theta_k) = 0 \text{ pour } \Theta = \Theta_{k+1}$$

Algorithme EM pour le mélange de Gaussiennes (2)

Solution analytique pour $1 \leq j \leq m$

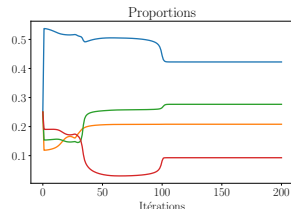
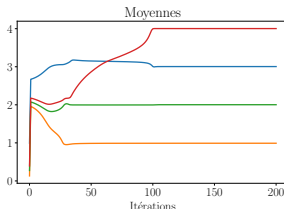
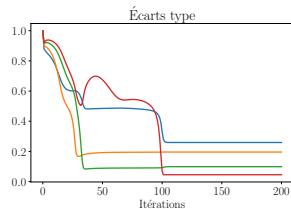
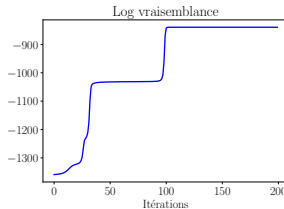
$$\begin{aligned}\Theta_k &= (\alpha_j^{(k)}, \mu_j^{(k)}, (\sigma^2)_j^{(k)})_{1 \leq j \leq m} \\ \alpha_j^{(k+1)} &= \frac{1}{n} \sum_{i=1}^n q_{\Theta_k}(Z = j | X = x_i) \\ \mu_j^{(k+1)} &= \frac{\sum_{i=1}^n x_i q_{\Theta_k}(Z = j | X = x_i)}{\sum_{i=1}^n q_{\Theta_k}(Z = j | X = x_i)} \\ (\sigma^2)_j^{(k+1)} &= \frac{\sum_{i=1}^n (x_i - \mu_j^{(k+1)})^2 q_{\Theta_k}(Z = j | X = x_i)}{\sum_{i=1}^n q_{\Theta_k}(Z = j | X = x_i)}\end{aligned}$$

Exemple numérique

Exemple d'un mélange de 4 Gaussiennes

$$\begin{aligned} X &\sim 0.2 \times \mathcal{N}(1, 0.2^2) \\ &+ 0.3 \times \mathcal{N}(2, 0.1^2) \\ &+ 0.4 \times \mathcal{N}(3, 0.25^2) \\ &+ 0.1 \times \mathcal{N}(4, 0.05^2) \end{aligned}$$

Application de l'algorithme EM sur un échantillon de taille 1000, initialisation des moyennes à 0, écarts type à 1 et des poids uniformes $\alpha_j = 0.25$



Sommaire

1. Algorithme *Expectation-Maximization*
2. Application à la Tomographie à Émission de Positons (TEP)
3. Introduction aux méthodes Bayésiennes
4. Modèle Bayésien pour la segmentation d'image TEP



Algorithme EM pour la Tomographie par Emission de Positons

- Images réalisées à partir de la détection de photons issus de l'annihilation de positons émis par un produit radioactif injecté au patient.
- Sources de photons $1 \leq j \leq m$, loi de Poisson de paramètre λ_j : $X_j \sim \mathcal{P}(\lambda_j)$. X_j est le nombre total de photons émis par la source j .
- Capteurs des photons, $1 \leq i \leq m$ et p_{ij} = probabilité que le capteur i détecte le photon j avec $\sum_{j=1}^m p_{ij} = 1$ $\sum_{i=1}^n p_{ij} = 1$ si chaque photon est détecté par un et un seul capteur.
- N_{ij} : Nombre de photons émis par la source j puis détectés par le capteur i :

$$(N_{ij}|X_j) \sim \mathcal{B}(X_j, p_{ij})$$

↪ Loi conditionnelle $N_{ij}|X_j$ binomiale

- Nombre de photons émis par toutes les sources puis détectés par le capteur i

$$Y_i = \sum_{j=1}^m N_{ij}$$

Algorithme EM pour la Tomographie par Emission de Positons

- Images réalisées à partir de la détection de photons issus de l'annihilation de positons émis par un produit radioactif injecté au patient.
- Sources de photons $1 \leq j \leq m$, loi de Poisson de paramètre λ_j : $X_j \sim \mathcal{P}(\lambda_j)$. X_j est le nombre total de photons émis par la source j .
- Capteurs des photons, $1 \leq i \leq m$ et p_{ij} = probabilité que le capteur i détecte le photon j avec $\sum_{j=1}^m p_{ij} = 1$ $\sum_{i=1}^n p_{ij} = 1$ si chaque photon est détecté par un et un seul capteur.
- N_{ij} : Nombre de photons émis par la source j puis détectés par le capteur i :

$$(N_{ij}|X_j) \sim \mathcal{B}(X_j, p_{ij})$$

↪ Loi conditionnelle $N_{ij}|X_j$ binomiale

- Nombre de photons émis par toutes les sources puis détectés par le capteur i

$$Y_i = \sum_{j=1}^m N_{ij}$$

- Objectif : estimer pour chaque source j le nombre moyen d'émission de photon $\lambda_j = \mathbb{E}(X_j)$ à partir des seules mesures Y_i sans connaissance des N_{ij} . Question: quel est le lien avec le problème du modèle de mélange de Gaussiennes ? (Car il y en a un !)
- On suppose les proportions p_{ij} connues → algorithme **MLEM** : Maximum Likelihood Expectation Maximization. Matrice P de taille $(n \times m)$, d'éléments p_{ij} :

Résolution d'un problème inverse $Y = PX$ où X est inconnu

- En TEP, plus d'inconnues que d'équations ($m > n$) et de plus les mesures sont bruitées

Modélisation du problème

- Nombre de photons \rightarrow loi de Poisson indépendantes $X_j \sim \mathcal{P}(\lambda_j)$, $1 \leq j \leq m$
- Capteurs $1 \leq i \leq n$: $N_{ij} \sim \mathcal{P}(p_{ij}\lambda_j)$ (pas trivial mais cela peut se démontrer). \triangle Ce n'est pas pareil que $N_{ij} = p_{ij}X_j$!!! (Pourquoi ?)

$$\mathbb{P}(N_{ij} = n) = e^{-p_{ij}\lambda_j} \frac{(\lambda_j p_{ij})^n}{n!}$$

- Données manquantes (n_{ij}). Seules données disponibles $y_i = \sum_j n_{ij}$. La loi de Poisson est stable par sommation: $\sum_j N_{ij} \sim \mathcal{P}(\sum_j p_{ij}\lambda_j)$. On retombe sur un problème de mélanges de lois, cette fois ci de lois de Poisson.
- En posant $\mathbf{y} := (y_i)_{1 \leq i \leq n}$, la log vraisemblance s'écrit:

$$\ell(\Lambda|\mathbf{y}) = \ln \prod_{i=1}^n e^{-\sum_j p_{ij}\lambda_j} \frac{(\sum_j \lambda_j)^{y_i}}{y_i!}$$

Le MLE $\hat{\Lambda}_n = (\hat{\lambda}_{j,n})_{1 \leq j \leq m}$ doit vérifier:

$$\sum_{i=1}^n p_{ij} = \sum_{i=1}^n \frac{p_{ij} y_i}{\sum_{s=1}^m p_{is} \hat{\lambda}_{s,n}}$$

Algorithme MLEM (1)

- Log-vraisemblance des $\Lambda := (\lambda_j)$ conditionnellement aux réalisations $N := (n_{ij})$:

$$\ell(\Lambda|N) = \ln \prod_{i=1}^n \prod_{j=1}^m e^{-p_{ij}\lambda_j} \frac{(\lambda_j p_{ij})^{N_{ij}}}{N_{ij}!} = \sum_{i=1}^n \sum_{j=1}^m (-p_{ij}\lambda_j + N_{ij} \ln(p_{ij}\lambda_j) - \ln(N_{ij}!))$$

- On applique le principe EM en calculant l'espérance de la vraisemblance conditionnellement à un jeu de paramètres $\Lambda^{(k)}$ fixé et aux seules mesures disponibles \mathbf{y} :

$$\begin{aligned} \ell_c(\Lambda|\mathbf{y}, \Lambda^{(k)}) &= \mathbb{E}_{N \sim p(N|\mathbf{y}, \Lambda^{(k)})} \ell(\Lambda|N, \mathbf{y}, \Lambda^{(k)}) \\ &= \sum_{i=1}^n \sum_{j=1}^m \mathbb{E}_{N_{ij}} [-p_{ij}\lambda_j + N_{ij} \ln(p_{ij}\lambda_j) - \ln(N_{ij}!)] \end{aligned}$$

Algorithme MLEM (2)

- La valeur optimale de λ_j est obtenue en annulant le gradient de la vraisemblance conditionnellement (en simplifiant la notation de l'espérance)

$$\begin{aligned}\frac{\partial}{\partial \lambda_j} \ell_c(\Lambda | \mathbf{y}, \Lambda^{(k)}) &= \frac{\partial}{\partial \lambda_j} \sum_{i=1}^n \sum_{s=1}^m \mathbb{E}_{N_{ij}} [-p_{is} \lambda_s + N_{is} \ln(p_{is} \lambda_s) - \ln(N_{is}!)] | \mathbf{y}, \Lambda^{(k)} \\ &= \frac{\partial}{\partial \lambda_j} \sum_{i=1}^n \mathbb{E}_{N_{ij}} [-p_{ij} \lambda_j + N_{ij} \ln(p_{ij} \lambda_j) - \ln(N_{ij}!)] | \mathbf{y}, \Lambda^{(k)} \\ &= \sum_{i=1}^n -p_{ij} + \mathbb{E}(N_{ij} | y_i, \Lambda^{(k)}) \frac{\partial \ln(p_{ij} \lambda_j)}{\partial \lambda_j} \\ &= \sum_{i=1}^n -p_{ij} + \mathbb{E}(N_{ij} | y_i, \Lambda^{(k)}) \frac{1}{\lambda_j} \\ &= \sum_{i=1}^n -p_{ij} + \frac{1}{\lambda_j} \sum_{i=1}^n \mathbb{E}(N_{ij} | y_i, \Lambda^{(k)})\end{aligned}$$

- Le gradient s'annule pour la valeur :

$$\lambda_j^{(k+1)} = \frac{\sum_{i=1}^n \mathbb{E}(N_{ij} | y_i, \Lambda^{(k)})}{\sum_{i=1}^n p_{ij}}$$

- Il *suffit* donc de calculer les n moyennes conditionnelles $\mathbb{E}(N_{ij} | y_i, \Lambda^{(k)})$.

Algorithme MLEM (3)

Lemme (Calcul de la loi $N_{ij}|y_i, \Lambda$)

Soient 2 lois de Poisson X_1, X_2 indépendantes de paramètres respectifs λ_1, λ_2 . La loi conditionnelle $X_1|X_1 + X_2$ est la loi binomiale $\mathcal{B}(X_1 + X_2, \lambda_1/(\lambda_1 + \lambda_2))$.

- A partir de Y_i somme de m lois de Poisson N_{ij} de paramètres $p_{ij}\lambda_j$, on en déduit :

$$N_{ij}|y_i, \Lambda \sim \mathcal{B}\left(y_i, \frac{p_{ij}\lambda_j}{\sum_{s=1}^m p_{is}\lambda_s}\right)$$

- La moyenne d'une loi binomiale $\mathcal{B}(n, p)$ étant égale à np , on obtient :

$$\mathbb{E}(N_{ij}|y_i, \Lambda^{(k)}) = \frac{y_i p_{ij} \lambda_j^{(k)}}{\sum_{s=1}^m p_{is} \lambda_s^{(k)}}$$

- D'où l'algorithme itératif MLEM pour le calcul des paramètres λ_j :

$$\lambda_j^{(k+1)} = \frac{1}{\sum_{i=1}^n p_{ij}} \sum_{i=1}^n \frac{y_i p_{ij} \lambda_j^{(k)}}{\sum_{s=1}^m p_{is} \lambda_s^{(k)}} = \frac{\lambda_j^{(k)}}{\sum_{i=1}^n p_{ij}} \sum_{i=1}^n \frac{y_i p_{ij}}{\sum_{s=1}^m p_{is} \lambda_s^{(k)}}$$

Analyse de l'algorithme MLEM

■ Rappel de l'algorithme MLEM

$$\lambda_j^{(k+1)} = \frac{\lambda_j^{(k)}}{\sum_{i=1}^n p_{ij}} \sum_{i=1}^n \frac{y_i p_{ij}}{\sum_{s=1}^m p_{is} \lambda_s^{(k)}}$$

■ Vérification des points fixes $\lambda_j^{(k+1)} = \lambda_j^{(k)}$ (hyp. $\lambda_j \neq 0$) :

$$\lambda_j = \frac{\lambda_j}{\sum_{i=1}^n p_{ij}} \sum_{i=1}^n \frac{y_i p_{ij}}{\sum_{s=1}^m p_{ij} \lambda_s} \Rightarrow \sum_{i=1}^n p_{ij} = \sum_{i=1}^n p_{ij} \frac{y_i}{\sum_{s=1}^m p_{is} \lambda_s}$$

■ Mais on vérifie que toute valeur λ' telle que $P\lambda' = P\lambda$ est également solution de MLEM

- ☹ Non unicité : plusieurs solutions si le nombre lignes m de la matrice P (nombre de capteurs) est inférieur au nombre de ses colonnes (nombre de sources)
- ☹ Convergence assez lente.

Sommaire

1. Algorithme *Expectation-Maximization*
2. Application à la Tomographie à Émission de Positons (TEP)
3. Introduction aux méthodes Bayésiennes
4. Modèle Bayésien pour la segmentation d'image TEP



Méthodes Bayésiennes

- Thomas Bayes (XVIII^{ème} siècle), mathématicien britannique, pasteur connu par son théorème

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \times \mathbb{P}(A)}{\mathbb{P}(B)}, \quad \text{noter la symétrie}$$

- $\mathbb{P}(A)$ probabilité *a priori* de A
- $\mathbb{P}(B|A)$ la vraisemblance, probabilité conditionnelle de B sachant A
- $\mathbb{P}(B)$ probabilité *a priori* de B , loi marginale de B
- $\mathbb{P}(A|B)$ probabilité *a posteriori* car postérieure à la connaissance de B , probabilité conditionnelle de A sachant B

"Oubli de la fréquence de base"

- Dans une population, 1 personne sur 1000 est malade.
- Si la personne est malade, le test de dépistage est positif dans 99 cas sur 100 et si elle ne l'est pas il est positif dans 2 cas sur 1000.
- Quelle est la probabilité que la personne soit réellement malade lorsque le test est positif ?

"Oubli de la fréquence de base"

- Dans une population, 1 personne sur 1000 est malade.
- Si la personne est malade, le test de dépistage est positif dans 99 cas sur 100 et si elle ne l'est pas il est positif dans 2 cas sur 1000.
- Quelle est la probabilité que la personne soit réellement malade lorsque le test est positif ?
- Notons P l'évènement "le test positif" et M "la personne est malade"

$$\begin{aligned}\mathbb{P}(M|P) &= \frac{\mathbb{P}(P|M) \times \mathbb{P}(M)}{\mathbb{P}(P)} \\ &= \frac{\mathbb{P}(P|M) \times \mathbb{P}(M)}{\mathbb{P}(P|M)\mathbb{P}(M) + \mathbb{P}(P|\bar{M})\mathbb{P}(\bar{M})} \\ &= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.002 \times 0.999} \simeq \frac{1}{3}\end{aligned}$$

- Conclusion: les données du problème laissent supposer un test plus fiable. **Biais cognitif appelé "l'oubli de la fréquence de base" ou "négligence de la taille de l'échantillon"**

Interprétation des probabilités

■ Probabilités objectives

- pour décrire le hasard physique, la **variabilité naturelle** de certains phénomènes : jeux de roulettes, atomes radioactifs
- probabilité vu comme le taux du nombre d'occurrences d'un évènement particulier parmi l'ensemble des évènements
- modélisation des **incertitudes aléatoires** : simulation des signaux sismiques

■ Probabilités subjectives

- pour décrire un degré de croyance sur une loi physique ou la valeur particulière d'un paramètre d'un modèle physique : existence du boson de Higgs
- probabilité comme quantification du degré de certitude d'une proposition
- modélisation des **incertitudes épistémiques** : en neutronique, représentation de la valeur de la section efficace d'une interaction entre neutron/noyau

■ Méthodes Bayésiennes sont fondées sur l'interprétation subjective

- dans l'exemple sur le test de dépistage, la personne est soit malade ou non
- la probabilité $\mathbb{P}(M|P)$ quantifie le degré de croyance sur la proposition "la personne est malade lorsque le test est positif"

En statistique fréquentiste...

- Le paramètre θ est une grandeur à estimer à partir des données observées X
- Seul les données X sont aléatoires
- Les principaux objectifs sont:
 - Proposer un **estimateur ponctuel** $\hat{\theta}$ de θ ;
 - Construire un **intervalle de confiance** pour quantifier l'incertitude sur θ
 - Faire des test d'hypothèse sur θ (Rendez vous le 30 Novembre !)

Notations et définitions

- Soit X une variable aléatoire dont la distribution de probabilités dépend de paramètre(s) θ
- Notre **a priori** sur ce paramètre est représentée par une variable aléatoire θ de densité $\pi(\theta)$ $\triangle \theta$ désigne à la fois le paramètre à estimer et la variable aléatoire le représentant !
- La **loi a priori** $\pi(\theta)$ traduit le degré de méconnaissance sur la valeur réelle du coefficient θ et non pas sa variabilité
- On se donne une loi conditionnelle de X sachant θ de densité de probabilité $p(X|\theta)$
- À partir de la règle de Bayes, on obtient la **loi a posteriori** sur θ noté $\pi(\theta|X)$

$$\pi(\theta|X) = \frac{p(X|\theta) \times \pi(\theta)}{p(X)}$$

- Le terme $p(X)$ (loi marginale (ou évidence) de X) est un facteur de normalisation qui ne dépend pas du paramètre θ

$$p(x) = \int p(X = x|\theta)\pi(\theta)d\theta$$

Et en statistique bayésienne

- Le paramètre θ est **empreint d'incertitudes**
- Le paramètre θ est une variable aléatoire
- Une distribution de probabilité sur θ , la loi *a priori*, est assignée à θ **avant d'observer les données X** .
- - La loi a priori décrit notre état des connaissances sur $\theta \implies$ **Vision subjective de la probabilité**
 - Définition **conditionnelle** à un état des connaissances ! \implies calcul de la loi *a posteriori*

Règle de Bayes en statistique

Soit un n -échantillon $(X_i = x_i)_{1 \leq i \leq n}$ i.i.d tel que $X_1 \sim p(\cdot|\theta)$.

La vraisemblance s'écrit de la même façon dans le cadre Bayésien:

$$\mathcal{L}(x_1, \dots, x_n|\theta) = \prod_{i=1}^n p(X = x_i|\theta)$$

La loi *a posteriori* s'écrit:

$$\pi(\theta|X_1 = x_1, \dots, X_n = x_n) = \frac{\pi(\theta)\mathcal{L}(\theta|x_1, \dots, x_n)}{p(X)} = \frac{\pi(\theta) \prod_{i=1}^n p(X = x_i|\theta)}{p(X)}$$

Exercice: Que se passe t'il quand on observe une nouvelle donnée X_{n+1} et qu'on utilise comme loi a priori $\pi(\theta|X_1, \dots, X_n)$?

Loi a posteriori non normalisée

- En général, compte tenu de la difficulté du calcul de la densité marginale $p(x)$, la loi *a posteriori* est traitée à partir de la forme non normalisée :

$$\underbrace{\pi(\theta|X)}_{\text{densité a posteriori}} \propto \underbrace{p(X|\theta)}_{\text{vraisemblance}} \times \underbrace{\pi(\theta)}_{\text{densité a priori}}$$

- On étudiera l'**algorithme de Métropolis-Hasting** (méthode **MCMC : Markov Chain Monte Carlo**) permettant de simuler (échantillonner) la loi *a posteriori* à partir de sa forme non normalisée
- A partir de la loi *a posteriori*, estimation Bayésienne du paramètre θ par
 - la moyenne *a posteriori* $\mathbb{E}[\theta|X] = \int \theta \times \pi(\theta|X) d\theta$
 - le maximum *a posteriori* (MAP) $\arg \max_{\theta} \pi(\theta|X)$ (valeur de θ qui maximise la densité *a posteriori*)

- Estimation par domaine (intervalle si θ est scalaire) de niveau de confiance α

$$\mathbb{P}(\theta \in I|X) = \int_I \pi(\theta|X) d\theta = \alpha$$

- Lorsqu'on retient un intervalle centré sur la médiane (cas où θ est scalaire)

$$I_{\text{centré}} = [Z_{(1-\alpha)/2}, Z_{(1+\alpha)/2}]$$

où $Z_{(1-\alpha)/2}, Z_{(1+\alpha)/2}$ sont les quantiles de la loi *a posteriori* $\pi(\theta|X)$. On parle d'**intervalle de crédibilité**

- ⚠ Ne pas confondre avec les intervalles de confiance !

Estimateurs Bayésien

Qu'est ce qu'un estimateur dans le paradigme Bayésien ? \Rightarrow il est associé à une fonction de coût.

Estimateurs Bayésien

Qu'est ce qu'un estimateur dans le paradigme Bayésien ? \implies il est associé à une fonction de coût.

Un estimateur de θ est une fonction des données $\hat{\theta}(X)$. Pour quantifier la qualité de l'estimation on associe une fonction de coût $L(\theta, \hat{\theta}(X))$.

Estimateurs Bayésien

Qu'est ce qu'un estimateur dans le paradigme Bayésien ? \implies il est associé à une fonction de coût.

Un estimateur de θ est une fonction des données $\hat{\theta}(X)$. Pour quantifier la qualité de l'estimation on associe une fonction de coût $L(\theta, \hat{\theta}(X))$.

Le *risque a posteriori* est la moyenne *a posteriori* de la fonction de coût

$$\rho(\pi, \hat{\theta}|X) = \mathbb{E}_{\theta \sim \pi(\cdot|X)}[L(\theta, \hat{\theta}(X))|X]$$

Estimateurs Bayésien

Qu'est ce qu'un estimateur dans le paradigme Bayésien ? \implies il est associé à une fonction de coût.

Un estimateur de θ est une fonction des données $\hat{\theta}(X)$. Pour quantifier la qualité de l'estimation on associe une fonction de coût $L(\theta, \hat{\theta}(X))$.

Le *risque a posteriori* est la moyenne *a posteriori* de la fonction de coût

$$\rho(\pi, \hat{\theta}|X) = \mathbb{E}_{\theta \sim \pi(\cdot|X)}[L(\theta, \hat{\theta}(X))|X]$$

Version "facile": L'estimateur de Bayes associé à la loi a priori π et à la fonction de coût L est l'estimateur minimisant le risque a posteriori:

$$\forall x \in \mathcal{X}, \hat{\theta}_{\text{Bayes}}(X) = \underset{\hat{\theta}(x)}{\operatorname{argmin}} \rho(\pi, \hat{\theta}(x))$$

Lien avec la théorie de la décision.

Estimateurs Bayésien: exemples

L'estimateur de Bayes associé à la loi a priori π et à la fonction de coût quadratique $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ est la moyenne a posteriori:

$$\hat{\theta}_{\text{Moy}} = \mathbb{E}_{\theta}[\theta|X]$$

L'estimateur de Bayes associé à la loi a priori π et à la fonction de coût L^1 $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ est la médiane a posteriori

Pour une fonction de coût quelconque, l'estimateur de Bayes se détermine par simulation Monte-Carlo (cf. le cours du 30 Octobre)

Cas particulier du maximum a posteriori (MAP)

Soit un n -échantillon $\mathcal{D} = (X_i)_{1 \leq i \leq n}$ i.i.d. de loi $p(\cdot|\theta)$ et avec $\theta \sim \pi$.

Le MAP est défini par:

$$\hat{\theta}_{\text{MAP}} = \operatorname{argmax} p(\theta|\mathcal{D}) = \operatorname{argmax} p(\mathcal{D}|\theta)\pi(\theta)$$

Le MAP ne dépend pas de la constante de normalisation !

$$\hat{\theta}_{\text{MAP}} = \underbrace{\sum_{i=1}^n \log p(X_i|\theta)}_{\text{log vraisemblance}} + \log \pi(\theta)$$

Le terme $\log \pi(\theta)$ peut s'interpréter comme une régularisation de la log vraisemblance

Paramètre d'une loi de Bernoulli

- Estimation du paramètre θ d'une loi de Bernoulli (jeu de pile/face)

$$X \in \{0, 1\}, \quad \mathbb{P}(X = x) = \theta^x (1 - \theta)^{(1-x)}$$

- On dispose d'un n-échantillon $\mathcal{D} = (x_1, x_2, \dots, x_n)$ issu de n tirages indépendants de X

$$p(\mathcal{D}|\theta) = \prod_{i=1}^n \mathbb{P}(X_i = x_i|\theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \underbrace{\theta^s (1 - \theta)^{n-s}}_{\propto \text{loi binomiale}}, \quad \text{où } s = \sum_{i=1}^n x_i$$

- Modélisation *a priori* sur θ par une loi beta dont la densité s'exprime en fonction de 2 paramètres (a, b) positifs

$$\pi(\theta; a, b) = \frac{1}{B(a, b)} \theta^{a-1} (1 - \theta)^{b-1} \mathbf{1}_{[0,1]}(\theta) \quad \text{où } B = \int_0^1 u^{a-1} (1 - u)^{b-1} du (\text{fonction beta})$$

- D'où la loi *a posteriori*

$$\pi(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta) \times \pi(\theta) = \theta^{a+s-1} (1 - \theta)^{b+n-s-1} \times \mathbf{1}_{[0,1]}(\theta)$$

- La loi *a posteriori* est donc une loi beta de paramètres $a + s$ et $b + n - s$

- a et b peuvent s'interpréter comme un nombre "virtuel" de pile ou face. → La loi *a posteriori* appartient à la même famille que la loi *a priori*. On dit que les deux lois sont *conjugués*.

Paramètre d'une loi de Bernoulli (suite)

- La loi *a posteriori* de θ s'exprime donc conditionnellement à $S(X_1, \dots, X_n) = \sum_{i=1}^n x_i$

$$\underbrace{\pi(\theta|\mathcal{D}) = \pi(\theta|S(X_1, \dots, X_n) = s)}_{S(X_1, \dots, X_n) \text{ statistique exhaustive}} = \frac{1}{B(a+s-1, b+n-s-1)} \theta^{a+s-1} (1-\theta)^{b+n-s-1} \mathbf{1}_{[0,1]}(\theta)$$

- La moyenne d'une loi bêta(a, b) est $a/(a+b)$ et son mode est $(a-1)/(a+b-2)$ pour $a > 0, b > 0$
- Estimations Bayésiennes de θ par:

l'espérance *a posteriori* $\hat{\theta}_{\text{EP}} = \frac{a+s}{a+b+n}$. On retrouve l'estimateur de la moyenne empirique pour le cas limite $a = b = 0$!

le mode *a posteriori* : $\hat{\theta}_{\text{MAP}} = \frac{a+s-1}{a+b+n-2}$

Paramètre d'une loi de Poisson

- Loi de Poisson de paramètre $\lambda > 0$ notée $X \sim \mathcal{P}(\lambda)$:

$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \in \mathbb{N}$$

- Calcul de la vraisemblance à partir d'un n-échantillon $\mathbf{x} = (x_1, x_2, \dots, x_n)$ (i.i.d.)

$$f(\mathbf{x}|\lambda) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \times \lambda^{\sum_{i=1}^n x_i} \times \frac{1}{\prod_{i=1}^n x_i!}$$

- La statistique $S = \sum X_i$ est exhaustive: factorisation de $f(\mathbf{x}|\lambda) = g(\lambda, s)h(\mathbf{x})$
- Estimation par maximum de vraisemblance :

$$\hat{\lambda}_{\text{MV}}(\mathbf{x}) = \hat{\lambda}_{\text{MV}}(s) = \arg \max_{\lambda} f(s|\lambda) = \arg \max_{\lambda} e^{-n\lambda} \lambda^s = \frac{s}{n}$$

Paramètre d'une loi de Poisson (suite)

- Estimation Bayésienne en postulant une loi *a priori* de type Gamma $\Gamma(\alpha > 0, \beta > 0)$ de densité

$$f(x; \alpha, \beta) = x^{\alpha-1} e^{-\beta x} \times \frac{\beta^\alpha}{\Gamma(\alpha)}, \quad x > 0$$

- Calcul de la loi *a posteriori* avec $s = \sum_{i=1}^n x_i$:

$$\begin{aligned}\pi(\lambda|s) &\propto f(s|\lambda) \times \pi(\lambda) \\ &\propto e^{-n\lambda} \lambda^s \times \lambda^{\alpha-1} e^{-\beta\lambda} \\ &\propto \lambda^{s+\alpha-1} e^{-\lambda(n+\beta)} \\ &\propto \text{densité de la loi } \Gamma(\alpha + s, \beta + n)\end{aligned}$$

- La moyenne d'une loi $\Gamma(\alpha, \beta)$ est α/β et son mode est $(\alpha - 1)/\beta$ pour $\alpha > 1, \beta > 0$.

- Estimations Bayésiennes de λ par

- l'espérance *a posteriori* $\hat{\lambda}_{EP}(s) = \frac{\alpha+s}{\beta+n}$

- ou le mode *a posteriori* : $\hat{\lambda}_{MP}(s) = \frac{\alpha+s-1}{\beta+n}$

- On remarque que la loi *a posteriori* $\Gamma(\alpha + s, \beta + n)$ est de la même famille que la loi *a priori* $\Gamma(\alpha, \beta)$. → **Loi a priori $\Gamma(\alpha, \beta)$ est dite conjugué à la distribution de Poisson.**

Sommaire

1. Algorithme *Expectation-Maximization*
2. Application à la Tomographie à Émission de Positons (TEP)
3. Introduction aux méthodes Bayésiennes
4. Modèle Bayésien pour la segmentation d'image TEP



Segmentation d'image TEP

Modèle Bayésien de segmentation d'image TEP pour la localisation de tumeur proposé dans la thèse de Z. Irace, Chapitre 3

On considère une image TEP (x_1, \dots, x_n) tel que x_i est le nombre de photons reçue par le i ème voxel.

On va considérer que l'image TEP est partitionné en K tissus biologique distinct, justifiant qu'ils aient chacun leur propre distribution statistique.

Modélisation statistique intra-classe

On a vu dans la section 2 que:

$$X_i \sim \mathcal{P}(\lambda)$$

L'hypothèse Poissonienne est discutable, on modélise λ par une variable aléatoire tel que $\lambda \sim \Gamma(\alpha, \beta)$.

La distribution marginale $p(X_i) = \int_0^{+\infty} p(X_i|\lambda)p(\lambda)d\lambda$ est une **binomiale négative**

$$P(X_i = x_i | \alpha, \beta) = \binom{x_i + \alpha - 1}{x_i} \left(\frac{1}{1 + \beta} \right)^\alpha \left(\frac{\beta}{1 + \beta} \right)^{x_i} = \frac{\Gamma(x_i + \alpha)}{x_i! \Gamma(\alpha)} \left(\frac{1}{1 + \beta} \right)^\alpha \left(\frac{\beta}{1 + \beta} \right)^{x_i}$$

On note $X_i \sim \mathcal{BN}(\mu, \kappa)$ avec la moyenne $\mu = \alpha$ et l'inverse dispersion $\kappa = \alpha\beta$.

$$\mathbb{E}(X_i) = \mu \quad \text{Var}(X_i) = \mu + \mu^2 / \kappa$$

Modélisation inter classes

L'image TEP est partitionnée en classes (C_1, \dots, C_k) , on peut considérer que le nombre de photons suit une loi binomiale négative sur chaque zone C_j .

$$\forall X \in C_j, X \sim \mathcal{BN}(\mu_j, \kappa_j)$$

La distribution de probabilité de l'image est donc **un modèle de mélange**

$$X \sim \sum_{j=1}^K \omega_j \mathcal{BN}(\mu_j, \kappa_j)$$

avec $(\omega)_{1 \leq j \leq K}$ une combinaison convexe

Objectif: Estimer $(\mu_j, \kappa_j, \omega_j)_{1 \leq j \leq J}$

Modèle Bayésien

On note $\mathbf{x} = (x_1, \dots, x_n)$ l'image TEP, et $\boldsymbol{\theta} = (\mu_j, \kappa_j)_{1 \leq j \leq J}$.

On définit la variable latente Z à valeurs dans $\{1, \dots, K\}$ tel que $z_i = j \iff x_i \in C_j$. On note $\mathbf{z} = (z_1, \dots, z_n)$

On définit des lois a priori sur $\boldsymbol{\theta}$ et Z pour pouvoir appliquer la règle de Bayes:

$$p(\boldsymbol{\theta}, \mathbf{z} | \mathbf{x}) \propto p(\mathbf{x} | \boldsymbol{\theta}, \mathbf{z}) \pi(\boldsymbol{\theta}) \pi(\mathbf{z})$$

⚠ L'échantillonnage de la loi a posteriori est complexe vu que la constante de normalisation est inconnue ! On verra des techniques d'échantillonnage lors du cours sur les méthodes Monte-Carlo.

Lois a priori sur (μ_j, κ_j)

Les lois a priori sur les paramètres $(\mu_j, \kappa_j)_{1 \leq j \leq K}$ sont des lois Gamma:

$$\mu_j \sim \Gamma(1 + a_\mu, -1/b_\mu)$$

$$\kappa_j \sim \Gamma(1 + a_\kappa, -1/b_\kappa)$$

Loi a priori sur \mathbf{z} : le champ de Potts

La loi a priori sur $\mathbf{z} = (z_1, \dots, z_n)$ est le **champ de Potts**, lui même une extension du modèle d'Ising:

$$\pi(\mathbf{z}) = \frac{1}{C(\gamma)} \exp \left[\sum_{i=1}^n \sum_{i' \in \mathcal{V}(i)} \gamma \mathbf{1}_{z_i = z_{i'}} \right]$$

- $\mathcal{V}(\cdot)$ désigne les points voisins du voxel i
- γ est le paramètre de granularité: plus γ est grand, plus les régions correspondant à chaque classe seront connexes
- $C(\gamma)$ est la constante de normalisation (appelé *fonction de partition* par les physiciens)

Références

- Manuscrit de thèse de Zacharie Irace. *Modélisation statistique et segmentation d'images TEP : application à l'hétérogénéité et au suivi de tumeurs*. INP Toulouse, 2014. (En ligne)
- X. de Scheemaekere, *Les fondements philosophiques du concept de probabilité*, Université Libre de Bruxelles, 2012. (En ligne)