Semaine 4 - l'algorithme EM

Exercice 1 – EM et mixture de gaussiennes

Les prix fonciers d'un quartier suivent une mixture de 2 gaussiennes, de paramètres respectifs (μ_1, σ_1^2) et (μ_2, σ_2^2). Le tableau ci-dessous recense les prix en 100K€ de quelques transactions immobilières :

8	1	4	3	3	5	7	5	4	5

On appellera π_1 et π_2 les coefficients des 2 gaussiennes dans la mixture.

Q 1.1 Triez les éléments de l'échantillon par ordre croissant et servez-vous des 5 plus petites valeurs pour estimer par maximum de vraisemblance (μ_1, σ_1^2) et des 5 plus grandes pour estimer, toujours par maximum de vraisemblance (μ_2, σ_2^2). Dans ces conditions, quelles valeurs faudrait-il logiquement affecter aux poids π_1 et π_2 ? On rappelle que la fonction de densité de la loi normale est :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

Q 1.2 En partant du $\Theta^0 = \{\mu_1, \sigma_1^2, \pi_1, \mu_2, \sigma_2^2, \pi_2\}$ obtenu à la question précédente, estimez la valeur de Q_i^1 selon l'algorithme EM.

Q 1.3 Estimez la valeur du paramètre Θ^1.

Exercice 2 – EM et loi exponentielle

Une entreprise s'intéresse à la durée de vie d'un composant informatique. Pour cela, elle a fait fonctionné 10 composants et a noté les temps (en mois) au bout desquels lesdits composants ont cessé de fonctionner. Les résultats sont répertoriés dans le tableau ci-dessous. Pour ces tests, l'entreprise a imposé un timeout de 20 mois et, lorsque les composants continuaient à fonctionner après ce délai, elle a noté dans le tableau un « ? ».

1	2	2	3	3	7	10	?	?	?

La loi classiquement utilisée en statistiques pour modéliser la durée de vie de composants est la loi exponentielle (de paramètre λ) dont la fonction de densité est $f(x|\lambda) = \lambda e^{-\lambda x}$, pour tout $x \geq 0$.

Q 2.1 En ne prenant en compte que les données numériques du tableau (c'est-à-dire sans tenir compte des « ? »), estimez la valeur du paramètre λ par maximum de vraisemblance.

Q 2.2 La troncature de la loi exponentielle sur l'intervalle $[20, +\infty[$ est la loi $g(x|\lambda) = \begin{cases} 0 & \text{si } x < 20 \\ \mu e^{-\lambda x} & \text{si } x \geq 20 \end{cases}$
Donnez une expression du paramètre μ en fonction de λ (suggestion : l'intégrale d'une fonction de densité sur tout son domaine de définition est égale à 1). On rappelle que la dérivée de $e^{\alpha x}$ par rapport à x est égale à $\alpha e^{\alpha x}$.

Q 2.3 On va maintenant exécuter l'algorithme EM afin de déterminer le paramètre λ de la loi exponentielle $f(x|\lambda)$ en tenant compte des « ? ». Donnez une expression de $Q_i^1(x_i^h) = p(x_i^h|x_i^o, \lambda)$, $i = 8, 9, 10$, en fonction d'une valeur λ_0 (qui sera, par la suite égale au λ estimé à la question 2.1). On comprendra le conditionnement par x_i^o comme « étant donné que l'on n'a pas observé l'arrêt du composant ».

Q 2.4 Que vaut $p(x_i^o|\lambda)$ pour $i = 8, 9, 10$, c'est-à-dire la probabilité de ne pas observer l'arrêt du composant. Sachant que $p(x_i^o, x_i^h|\lambda) = p(x_i^o|\lambda) \times p(x_i^h|x_i^o, \lambda)$, donnez l'expression de $p(x_i^o, x_i^h|\lambda)$.

Q 2.5 Donnez une expression de $Q_i^1(x_i^h) \log\left(\frac{p(x_i^o, x_i^h|\lambda)}{Q_i^1(x_i^h)}\right)$, $i = 8, 9, 10$, en fonction de λ et de l'expression obtenue

dans la question précédente.

Q 2.6 Donnez une expression de $\int Q_i^1(x_i^h) \log \left(\frac{p(x_i^o, x_i^h|\lambda)}{Q_i^1(x_i^h)} \right) dx_i^h$, pour $i = 8, 9, 10$.

On rappelle que, pour $\alpha > 0$, $\int_{20}^{+\infty} xe^{-\alpha x} dx = \frac{1}{\alpha} \left(20 + \frac{1}{\alpha} \right) e^{-20\alpha}$ (on le démontre aisément par intégration par parties).

Q 2.7 En déduire une expression pour $\log L^{t+1}(\mathbf{x}^o, \lambda)$

Q 2.8 On peut maintenant appliquer EM. En supposant que $\Theta^0 = \lambda_{ML}$, le lambda estimé par maximum de vraisemblance à la question 2.1, estimez Θ^1.

Q 2.9 En utilisant l'expression obtenue à la question précédente, pour quelle valeur de λ aura-t-on convergence ?