SPLEX TME 3

Mélange de Gaussiens et l'algorithme EM

Implementer l'algorithme EM et tester le sur les données "obeses_temoins" pour clusteriser les patients en deux classes. Comparez vous resultats avec ceux obtenus avec un package R, par exemple, le package R "mclust".

Mélange de Gaussiens

Si le événements sont issus d'une composition de N phénomènes, la densité p(X) prendra la forme d'une composition de lois normales. On peut l'approximer par une somme pondérée de densité normales, "mélange de Gaussiens"

$$p(x) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(x; \mu_m, \Sigma_m), \tag{1}$$

ou

$$\mathcal{N}(x, \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} \det(\Sigma)^{\frac{1}{2}}} \exp\Big(-1/2(x-\mu)^T \Sigma^{-1} (x-\mu) \Big). \tag{2}$$

. Dans le mĺange de Gaussiens il faut estimer trois paramètres

$$(\alpha, \mu, \Sigma). \tag{3}$$

Si toutes les μ et Σ etait fixe, on pourrait calculer les α_m directement. Mait ils sont libres et il faut donc les esimter par un processus iterative. Un tel processus est composé de deux étapes.

Algorithme EM

Soit une ensembe "training set" de N observation $X_{n}{}_{n=1}^{N}$, $z^n \in \{0,1\}^m$ est l'indicateur de classe. La log-vraisemblence est

$$L(\alpha, \mu, \Sigma) = \sum_n \sum_m z_m^n \Big(\ln \alpha_m - \frac{1}{2} \ln |\Sigma_m| - \frac{1}{2} \mathrm{tr}(\Sigma_m^1 (x^n - \mu_m)(x^n - \mu_m)^T) \Big) \tag{4}$$

On fait une premier estimation des paramètres (α, μ, Σ) et puis on altern "Expectation" et "Maximisation" pendant plusieurs iterations t, tant que l'algorithme n'a pas convergé.

- “Expectation step”: Faire une estimation des valeurs manquantes (“hidden”)

$$z_m^n = p(z_m^n = 1 | x^n, \alpha^{\text{old}}, \mu^{\text{old}}, \Sigma^{\text{old}}) = \frac{\alpha_m^t \mathcal{N}(x_n; \mu_m^t, \Sigma_m^t)}{\sum_{j=1}^{M} \alpha_j^t \mathcal{N}(x_n; \mu_j^t, \Sigma_j^t)}. \tag{5}$$

- “Maximisation step”:

$$\alpha_m = \frac{1}{N} \sum_n z_m^n, \tag{6}$$

$$\mu_m = \frac{\sum_n z_m^n x^n}{\sum_n z_m^n}, \tag{7}$$

$$\Sigma_m = \frac{\sum_n z_m^n (x^n - \mu_m)(x^n - \mu_m)^T}{\sum_n z_m^n}. \tag{8}$$