SY19 Automne 2013

TP 2 : Classification et mélange

Exercice 1.

Le jeu de données considéré est constitué de 150 iris décrits par quatre variables : longueur des sépales, largeur des sépales, longueur des pétales et largeur des sépales. Charger le jeu de données, sélectionner les variables quantitatives et normaliser en utilisant le code R suivant :

```
library(MASS)
data(iris)
donnees <- NULL
donnees$num <- iris[,c(1:4)]
donnees$cls <- iris[,5]
```

1. Tenter une partition en $K \in \{2, 3, 4\}$ classes avec la fonction `kmeans` ; visualiser et commenter.
2. Étudier la stabilité du résultat de la partition : effectuer $n = 100$ classifications en trois classes du jeu de données, et comparer les valeurs d'inertie intra-classes obtenues. Commenter.
3. Choix du nombre de classes optimal : calculer la valeur moyenne d'inertie intra-classe obtenue sur $n = 100$ classifications, pour $K \in \{2, 3, 4, 5\}$ classes. Représenter la variation de l'inertie moyenne en fonction de K. Proposer un nombre de classes en se basant sur cette courbe.
4. Comparer les résultats de la partition obtenue par les centres mobiles avec la partition réelle des iris en trois groupes.

Exercice 2.

L'objectif de cet exercice est d'implémenter, d'étudier et de comparer les algorithmes EM et CEM dans le cas de données monodimensionnelles ($p = 1$).

Données synthétiques

On s'intéressera dans un premier temps à un jeu de données synthétique x issu d'un mélange Gaussien de deux classes ω_1 et ω_2 présentes en proportions identiques $\pi_1 = \pi_2 = \frac{1}{2}$, de paramètres $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 6$, $\sigma_2 = 5$. Les données peuvent être générées comme suit :

```
x <- c(rnorm(1000),rnorm(1000,mean=6,sd=5))
```

En reprenant les exemples, vus en cours, de l'algorithme EM :

1. rappeler l'expression des équations de mise à jour des paramètres μ_k et σ_k^2, pour $k \in \{1, 2\}$;
2. implémenter les algorithmes EM et CEM en complétant la fonction gmixtmono dont l'ossature est fournie dans le fichier mixtmono.r (disponible sur le site de SY19) ;
3. comparer les valeurs des paramètres du modèle de mélange (μ_k et σ_k^2, pour $k \in \{1, 2\}$) calculées avec l'algorithme EM, celles obtenues avec l'algorithme CEM, et les vraies valeurs des paramètres utilisées pour générer les données synthétiques ;
4. comparer les partitions obtenues avec les *k-means* et les algorithmes EM et CEM (en utilisant la règle du MAP, ou *maximum a posteriori*), à la partition réelle des données ; discuter.

Un indicateur numérique de l'adéquation entre deux partitions est l'indice de Rand. Pour calculer cet indice, on pourra utiliser la fonction randindex, disponible sur le site de SY19.

Application à des données réelles

On souhaite appliquer l'algorithme développé au jeu de données Crabs (bibliothèque MASS). Charger le jeu de données et effectuer les prétraitements suivants :

```
library(MASS)
data(crabs)
crabsquant <- crabs[,4:8]
crabsquant2 <- crabsquant/crabsquant[,4]
crabsquant2 <- crabsquant2[,-4]
```

Dans cette partie, on utilisera les algorithmes EM et CEM pour déterminer les paramètres d'un modèle de mélange à $K = 2$ composantes. Effectuer une classification des données avec chacun des algorithmes, tout d'abord à partir de la 1re variable, puis à partir de la 2e variable. Comment interpréter les résultats obtenus ?

Exercice 3.

L'objectif de cet exercice est d'implémenter et d'étudier les algorithmes EM et CEM dans le cas plus général d'un mélange gaussien multidimensionnel, sous l'hypothèse d'indépendance des variables conditionnellement aux classes (matrices de variance-covariance diagonales, mais spécifiques à chaque classe), pour un nombre K de classes quelconque. On dispose pour cela de la fonction mvdnorm, et de la fonction gmixtmulti à compléter (l'ossature de cette fonction est fournie dans le fichier mixtmult.r).

Données synthétiques

On s'intéressera dans un premier temps à un jeu de données synthétique X de $\mathbb{R}^2$ généré suivant un mélange Gaussien de trois classes ω_1, ω_2 et ω_3. Ces classes sont présentes en proportions $\pi_1 = 0.35$, $\pi_2 = 0.25$ et $\pi_3 = 0.4$, et sont caractérisées par les paramètres suivants :

$$\mu_1 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \mu_2 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}, \mu_3 = \begin{pmatrix} 1 \\ -3 \end{pmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Les données peuvent être générées comme suit :

```
library(MASS)
prop <- rmultinom(1,3000,c(.35,.25,.4))
X <- rbind(mvrnorm(prop[1],mu=c(3,1),Sigma=diag(c(1,1))),
  mvrnorm(prop[2],mu=c(-1,2),Sigma=diag(c(2,1))),
  mvrnorm(prop[3],mu=c(1,-3),Sigma=diag(c(1,2))))
```

1. Donner l'expression des équations de mise à jour des paramètres π_k, μ_k, et Σ_k, dans le cas des modèles de mélange multidimensionnels, sous l'hypothèse d'indépendance des variables conditionnellement aux classes (matrices de covariance diagonales). Détailler les calculs et justifier.
2. Implémenter les algorithmes EM et CEM en complétant la fonction mixtmult (disponible sur le site de SY19).
3. Appliquer l'algorithme des k-means et les algorithmes EM et CEM au jeu de données synthétique généré comme décrit ci-dessus. Comparer les résultats obtenus.

Application à des données réelles

On souhaite appliquer l'algorithme développé au jeu de données Crabs (bibliothèque MASS). Charger le jeu de données et effectuer les prétraitements comme dans l'exercice précédent.

Effectuer une classification du jeu de données en $K = 4$ classes, avec l'algorithme des K-means, et au moyen d'un modèle de mélange dont les paramètres seront estimés avec les algorithmes EM et CEM. Interpréter les résultats obtenus en vous appuyant sur votre connaissance des données. Comparer les différents algorithmes de classification.