Département MIDO
Module DM/ML
2015 / 2016

Travaux Pratiques n° 2 : Clustering

Objectifs : s'initier à la classification non-supervisée : K-moyennes vs EM.

Nous allons considérer dans ce TP deux algorithmes de classification non-supervisée : l'algorithme de k-moyennes et l'algorithme EM. Dans un premier temps nous allons travailler sur un jeu de données artificiel puis appliquerons ces méthodes aux données d'un futur TP portant sur le Text Mining.

1 Avant goût

Le but ici est de vous donner un exemple simple d'utilisation de l'algorithme k-moyennes :

```
x = c(−2,−2,0,2,−2,3)
y= c(2, −1,−1,2,3,0)
don = matrix(data=c(x,y), nr=6, nc=2)
ctre = c(−1,2,−1,3)
ctre1 =matrix(data=ctre, nr=2, nc=2)
cl1 = kmeans(don,ctre1,algorithm="Lloyd")
plot(don, col = cl1$cluster)
points(cl1$centers, col = 1 : 2, pch = 8, cex=2)
```

2 Et si on jouait !

2.1 Données simples

Le but dans cette section est de générer un ensemble de données artificiel permettant de mieux comprendre le fonctionnement des algorithmes.

1. Tirer aléatoirement deux sous-ensembles de données en 2D de 20 points chacun selon une loi normale de moyenne $(1, 1)$ et $(5, 5)$ respectivement et d'écart-type 1.
   fonction `R` : `rnorm`
2. Visualiser les deux ensembles de points séparément.
   fonctions `R` : `range, plot`
3. Construire une matrice de données à partir des points tirées aléatoirement ci-dessus.
   fonctions `R` : `matrix, c`
4. Effectuer une classification k-moyennes (voir plus haut pour un exemple simple)
5. Effectuer une classification EM.
   Paquet `R` : `mclust`. Méthode : `Mclust`
   Commencer par lire la documentation de la méthode `Mclust`.
6. Comparer les résultats des deux méthodes.

2.2 Données moins simples

1. Reprenez les étapes de la section précédente, mais en augmentant le nombre de sous ensembles de données (plusieurs sous-ensembles de 20 points par exemple) avec des moyennes et écart-types différents. Comparer les résultats des algorithmes k-moyennes et EM.
2. Reprenez le même traitement mais en considérant d'autres lois de probabilités. Pour connaître les différentes lois sous `R` : `help(Distributions)`