クラスタリングで分類

```
Data(iris)
x <- iris[,-5]
(cl <- kmeans(x, 3))
plot(x, col = cl$cluster)
points(cl$centers, col = 1:3, pch = 8, cex=2)

# クラスタリング結果をそのまま分類結果と考えたときの正解率をみてみよう
# permutations という関数を持つライブラリ
library(gtools)
# クラスタリングは(自律的な)分類と考えることができる。
# しかし、クラスタには名前がない。整数が用いられるが、真の分類とは全く別の名前である。
# そこで、クラスタと真のカテゴリとの全ての組合せを試し、正解率が一番高くなる組み合わせ
# (クラスタの並べ方)を発見することを感が合える。

( tbl <- table(iris[,5],cl$cluster) )
(res <- apply(permutations(3,3), 1, function(x) sum(diag(tbl[,x]))))
res.sorted <- sort.int( res, decreasing=T, index.return=T )
# accuracy
( acc <- res.sorted$x[1]/sum(tbl) )
# そのaccuracyを与える順列(その結果のconfusion matrix)
tbl[, permutations(3,3)[res.sorted$ix[1],]]
```

R でEM

```
# EM algorithm for clustering
library(mclust)
# data sample in mlbench will be used
library(mlbench)
smly <- mlbench.smiley()
colnames(smly$x) <-c("x1","x2")

(gm4 <- Mclust(smly$x,G=4))
mclust2Dplot(smly$x, parameters=gm4$parameters,
     z=gm4$z, what="classification")
title("Clustering: 4 clusters")

dev.new()
(gm4to10 <- Mclust(smly$x,G=4:10))
mclust2Dplot(smly$x, parameters=gm4to10$parameters,
     z=gm4to10$z, what="classification")
title("Clustering: best in 4 to 10 clusters")
```

K-means と EM の比較

```
# EM algorithm for clustering
library(mclust)
# data sample in mlbench will be used
library(mlbench)
smly <- mlbench.smiley()
colnames(smly$x) <-c("x1","x2")

cl <- kmeans( smly$x, 4)
plot(smly$x, col = cl$cluster)
points(cl$centers, col = 1:4, pch = 8, cex=2)
title("Clustering: 4 clusters")

dev.new()
cl <- kmeans( smly$x, 10 )
plot(smly$x, col = cl$cluster)
points(cl$centers, col = 1:10, pch = 8, cex=2)
title("Clustering: 10 clusters")
```

本日の課題2

- Animals のデータを k-means でクラスタリング分析してください。その結果を、hclust でのクラスタリング結果と比較してください。
  - Animals のデータを直接使うなら、例えば、`x <- Animals[,2:3]` として、iris と同様にクラスタリングします。クラスタ数は2や3で試してください。
- 芳しくないなら、恐らく、スケールが違うからでしょうから、`x <- myAnimals` としてみてください。
- 散布図から分かることですが、点の分布が小さい方にたくさん集まっています。それが原因からもしれません。では、対数をとってみましょう。`log(Animals[,2:3])` などとすると対数がとれます。

本日の課題3(余裕があれば)

- library(mlbench)に含まれる mlbench.spirals を対象として、EMアルゴリズムと k-means アルゴリズムを用いたクラスタリングを行って下さい。
  - 前のスライドと同じ手続きで進めて下さい。
  - クラスタ数はいくつぐらいが妥当でしょうか？
  - それは、実際と符合しますか？
  - 符合しないとしたら、それはどうしてでしょうか？