```
#################################
##第十四回「クラスタリング」
#################################

############################
## 階層的クラスタリング
library(cluster)
library(RColorBrewer)

# MNIST手書き文字認識データの階層的クラスタリング
mnist <- read.csv("./mnist/train_data.csv",header=FALSE) #データの取得
dmnist <- dim(mnist) #dmnist[2]が変数の数, dmnist[2]番目の列にラベルが格納されて
いる
x <- matrix(1:(10*(dmnist[2]-1)),nrow=10)

#数字0から9, 各文字画像の平均値を計算
for(i in 1:10){
    x[i,] <- colMeans(mnist[which(mnist[,dmnist[2]]==i-1),1:(dmnist[2]-1)]) #数
字(i-1)の平均画像
}
label <- as.character(0:9)

hc <- hclust(dist(x),method="ward")  #階層的クラスタリングを実行
plot(hc,labels=label)

par(mfrow=c(3,4))
for(i in 1:10){
    img <- matrix(x[i,],nrow=28)  #平均画像をベクトルから行列へ変換
    img <- img[,rev(1:28)]  #そのままでは上下反転して表示されるので, 上下を反転
    image(img,col=topo.colors(256))
}
par(mfcol=c(1,1))


## ヒートマップ
# 行列データがあった場合, 行間の階層的クラスタリングと
# 列間の階層的クラスタリングを同時に行い, プロットできる.

# ジヒドロ葉酸還元酵素阻害剤データ (L1正則化の講義を参照)
data(dhfr,package="caret")
dim(dhfr) # 228次元, サンプル数325:高次元小標本データ
names(dhfr)

library(gplots)

#生データのヒートマップ
heatmap(t(as.matrix(dhfr[,2:228])),scale="col",col=redgreen(256))  # redgreenはl
ibrary(gplots)に入っている.

#見にくいので, データを小さくしてプロット
heatmap(t(as.matrix(dhfr[1:20,2:60])),scale="col",col=redgreen(256))

# サンプル間の距離関数とグループ間の距離関数を変更
dsf <- function(x) dist(x,method="maximum")
hcf <- function(d) hclust(d, method = "complete")
heatmap(t(as.matrix(dhfr[1:20,2:60])),scale="column",col=redgreen(256),distfun=d
sf,hclustfun=hcf)


##########################################
## K-meansと混合ガウスクラスタリング

## 人工データ
## ３つの要素からなる混合ガウス分布
mus <- matrix(c(1,-1,0,0,0,2),ncol=2)
xx <- matrix(1:(2*60),ncol=2)
yy <- c(1:60)
for(i in 1:3){
    x <- matrix(rnorm(2*20),ncol=2)
    S <- rWishart(1, 10, diag(2))/10  #分散共分散行列の生成
    S <- S[,,1]   #そのままでは３次元アレイなので, 行列になおす
    xx[((i-1)*20+1):(i*20),] <- (x%*%S)*0.45 + matrix(1, nrow=20, ncol=1)%*%mus[
```

```
i,]
    yy[((i-1)*20+1):(i*20)] <- i
}
plot(xx)
for(i in 1:3) points(xx[yy==i,],col=i,lwd=3)

## K-meansクラスタリング
Knum = 3
res <- kmeans(xx, Knum) #k-meansの実行
plot(xx)
for(i in 1:3) points(xx[res$cluster==i,],col=i,lwd=3)

table(yy,res$cluster) #そこそこ間違いなくクラスタリングできている

## 混合ガウス: EM-アルゴリズム
result  <- Mclust(xx) #EM-アルゴリズムの実行
summary(result)
par(mfrow=c(2,2))
plot(result)  #結果の要約をプロット
par(mfrow=c(1,1))

# 結果の2Dプロット
mclust2Dplot(xx,parameters=result$parameters,z=result$z,what="classification")

image(t(log(result$z)),col=hmcol) #各サンプルの各クラスタへの寄与率

table(yy,apply(result$z,1,which.max)) #ほぼ間違いなくクラスタリングできている.

# いくつかのモデルでGMMクラスタリング
model_names <- c('EII','VII','EEE','VVV') #分散共分散行列のモデル４つ
par(mfrow=c(2,2))
# クラスタ数固定=3
for(i in 1:4){
    result  <- Mclust(xx,G=3,modelNames=model_names[i]) #G=3でクラスタ数固定でGM
Mをあてはめ
    mclust2Dplot(xx,parameters=result$parameters,z=result$z,what="classification
")
    title(main=model_names[i])
}

# クラスタ数固定せず
for(i in 1:4){
    result  <- Mclust(xx,modelNames=model_names[i])
    mclust2Dplot(xx,parameters=result$parameters,z=result$z,what="classification
")
    title(main=model_names[i])
}
# VVVはちゃんとクラスタリングできている.

par(mfrow=c(1,1))


#######################################################
## TOPIX CORE 30銘柄のクラスタリング

#TPIX CORE 30に含まれる銘柄のインデックスと名前を取得
names <- read.csv("./topix/TOPIX_CORE_30_NAMES.csv",header=TRUE)

#各銘柄の時系列データの取得, 2013/6/28--2014/7/4の250日分
topix30 <- matrix(1:(250*30),ncol=30)
for(i in 1:30){
    index <- names[i,2]
    tmp <- read.csv(paste("./topix/",index,"-T.csv",sep=""),skip=2,header=FALSE)

    topix30[,i] <- as.matrix(tmp[,2]/tmp[,5]) #始値と終値の比を計算 (日足)
}
topix30 <- t(topix30) #銘柄×時間の並びにする
rownames(topix30) <- names[,3] #行に銘柄の名前を付与

# 時系列の距離を用いて階層的クラスタリング
hc <- hclust(dist(topix30),method="ward.D")
plot(hc)  #ファイナンス系が固まっている他, 似た業種が近くに分類されていることが
```

わかる.

```
## 混合ガウスモデルでクラスタリング
## 混合ガウスを当てはめるのに250次元は高すぎるので,
## 主成分分析で低次元に落としてからクラスタリング

## 主成分分析
cov <- var(topix30) #分散
res <- eigen(cov)  #固有値固有ベクトル分解
V <- res$vectors  #固有ベクトルの抽出
y =  topix30 %*% V #主成分スコアの計算

#主成分スコアのプロット
sy <- scale(y) #平均0, 分散1にスケーリング
plot(sy[,1:2],type="n",xlab="PC1",ylab="PC2")
text(sy[,1],sy[,2],names[,3],cex=1.3)

## 主成分スコアの上で混合ガウスクラスタリング
result  <- Mclust(scale(y[,1:3]),G=5)
plot(result)
heatmap(result$z,col=redgreen(256))
```