

rd-8. クラスタリング

データサイエンス演習

(R システムを使用)

<https://www.kkaneko.jp/de/rd/index.html>

金子邦彦



rd-8. クラスタリング

(Rシステムでデータサイエンス演習)

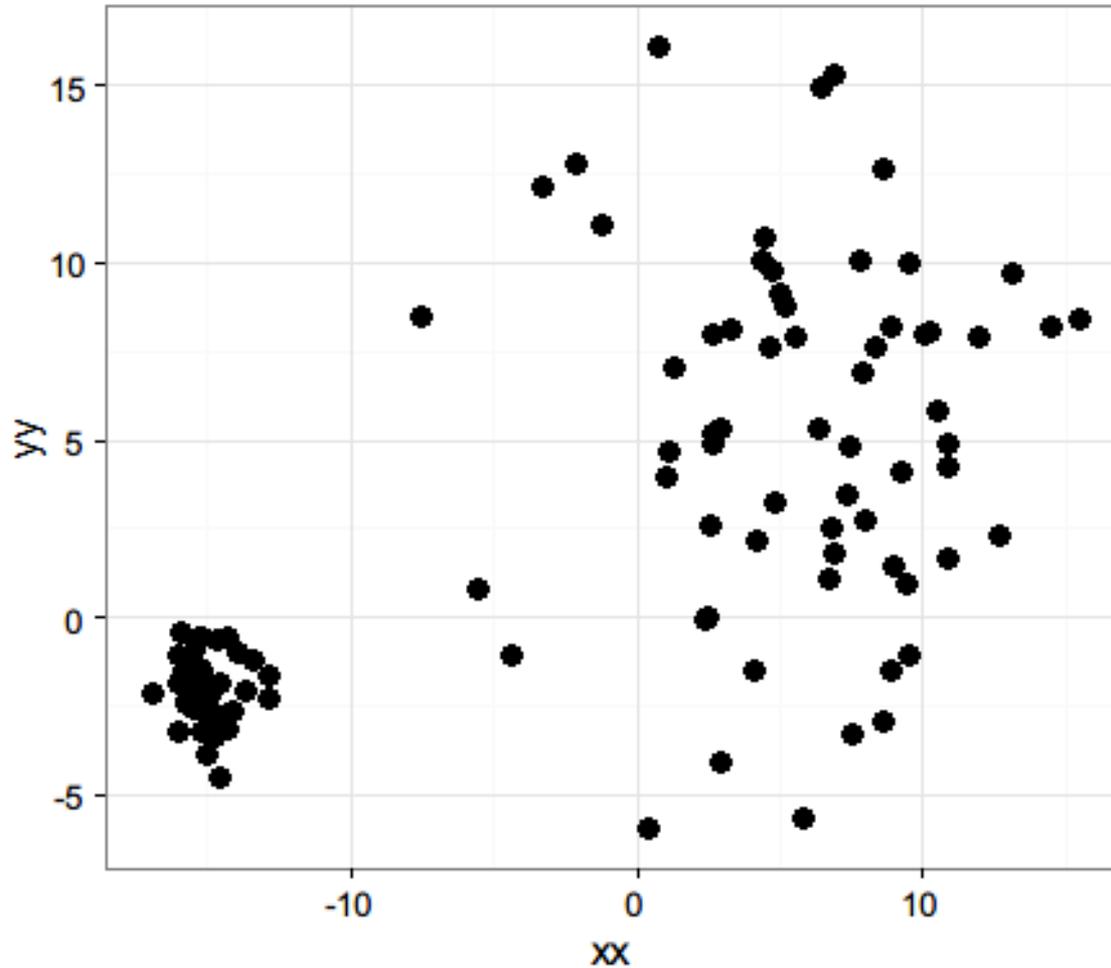
<https://www.kkaneko.jp/cc/rd/index.html>

金子邦彦



クラスタリング

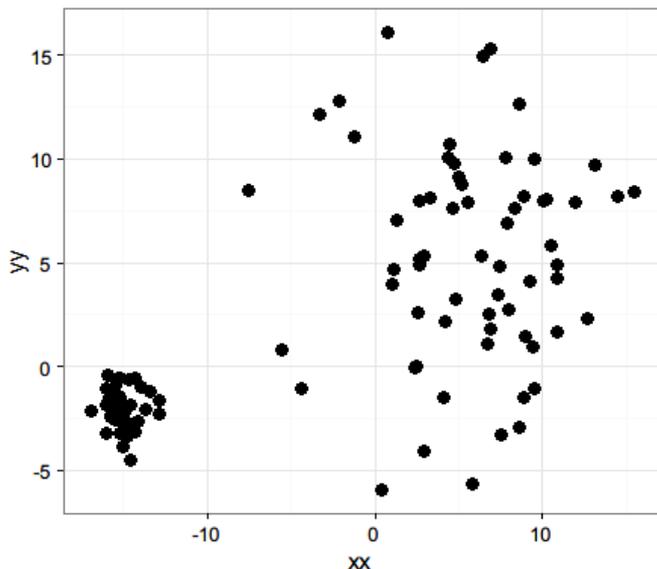
- 集団が2つ以上 (外れ値とは違う)



クラスタデータの合成の例



集団が2つ



```
x <- rnorm(100000, mean=5, sd=5)
y <- rnorm(100000, mean=5, sd=5)
x2 <- rnorm(80000, mean=-15, sd=1)
y2 <- rnorm(80000, mean=-2, sd=1)
x3 <- c( x, x2 )
y3 <- c( y, y2 )
n <- floor( runif(100, 1, 180000+1) )
d12 <- data.frame( xx=x3[n], yy=y3[n] )
library(ggplot2)
ggplot(d12, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + theme_bw()
```

※ **c** は、ベクトルデータの連結

mclust パッケージを使うには



- 準備

インターネット接続が必要

- インストール

`install.packages("mclust")`

```
> install.packages("mclust")
Installing package into 'C:/Users/user/Documents/R/win-library/3.2'
(as 'lib' is unspecified)

There is a binary version available but the source version is later
  binary source needs_compilation
mclust    5.1    5.2                TRUE

Do you want to install from sources the package which needs compilation?
y/n: y
installing the source package 'mclust'

trying URL 'https://mran.microsoft.com/snapshot/2016-04-01/src/contrib/mclust_5.2.tar.gz'
Content type 'application/octet-stream' length 2749850 bytes (2.6 MB)
downloaded 2.6 MB
```

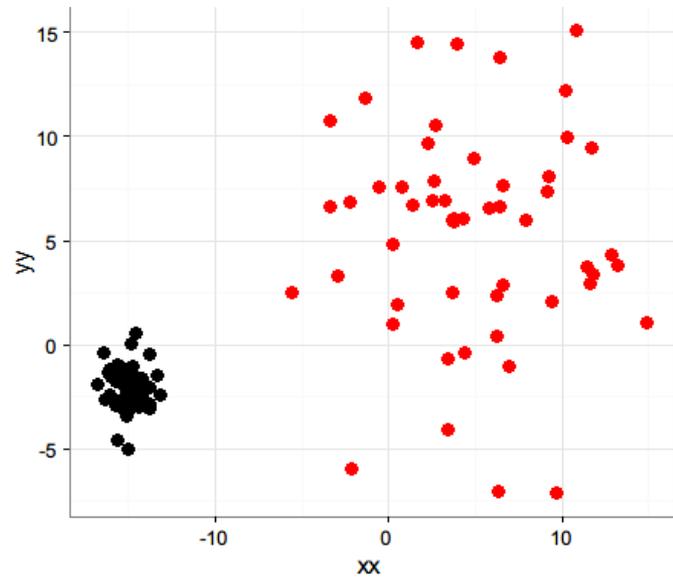
クラスタリングの例



```
> c <- Mclust(d12, modelNames="VVV")
> print(c$classification)
 [1] 1 2 1 2 1 1 1 1 2 2 2 1 2 1 2 1 2 1 1 2 2 2 1 2 2 2 1 2
[29] 1 1 1 1 2 2 1 2 1 2 2 1 1 1 1 2 1 1 2 2 2 2 1 1 1 1 1 1
[57] 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 2 1 2 1 1
[85] 2 1 1 2 1 1 2 1 1 1 2 2 1 1 1 1
>
```

```
library(mclust)
c <- Mclust(d12, modelNames="VVV")
print(c$classification)
```

クラスタリングの結果を、色付きの散布図でプロット



```
library(mclust)
c <- Mclust(d12, modelNames="VVV")
print(c$classification)
library(ggplot2)
ggplot(d12, aes(x=xx)) +
  geom_point( aes(y=yy), size=3, col=c$classification ) + theme_bw()
```