Chapter 14 – Cluster Analysis 聚类分析

Instructor: Zach Zhizhong ZHOU,
Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

Protein kMeans - 1

```
library(stats) #needed for kmeans
setwd("C:/BA/Clustering")
### *** European Protein Consumption, in grams/person-
day *** ###
## read in the data
food <- read.csv("protein.csv")
## first, clustering on just Red and White meat (p=2) and
k=3
## clusters
set.seed(1) ## to fix the random starting clusters
grpMeat <- kmeans(food[,c("RedMeat","WhiteMeat")],
centers=3,   nstart=10)
# nstart: if centers is a number, how many random sets
should be chosen?
```

Protein kMeans - 2

```
grpMeat

## list of cluster assignments

o=order(grpMeat$cluster)

data.frame(food$Country[o],grpMeat$cluster[o])

## plotting cluster assignments on Red and White meat scatter plot

plot(food$RedMeat, food$WhiteMeat, type="n", xlab="Red Meat", ylab="White Meat")

text(x=food$RedMeat, y=food$WhiteMeat, labels=food$Country, col=grpMeat$cluster+1)
```

Protein kMeans - 3

```
## same analysis, but now with clustering on all
## protein groups
## change the number of clusters to 7
set.seed(1)
grpProtein <- kmeans(food[,-1], centers=7, nstart=10)
o=order(grpProtein$cluster)
data.frame(food$Country[o],grpProtein$cluster[o])
## plotting cluster assignments on Red and White meat
scatter plot
plot(food$RedMeat, food$WhiteMeat, type="n",
xlim=c(3,19), xlab="Red Meat",     ylab="White Meat")
text(x=food$RedMeat, y=food$WhiteMeat,
labels=food$Country,     col=grpProtein$cluster+1)
```

Protein Hierachical - 1

```
library(cluster) #needed for hierachical clustering
### *** European Protein Consumption, in grams/person-
day *** ## read in the data
setwd("C:/BA/Clustering")
food <- read.csv("protein.csv")
## we use the program agnes in the package cluster
## argument diss=FALSE indicates that we use the
dissimilarity
## matrix that is being calculated from raw data.
## argument metric="euclidian" indicates that we use
Euclidian
## distance. no standardization is used as the default
## the default is "average" linkage
## Using data on all nine variables (features)
## Euclidean distance and average linkage
```

Protein Hierachical - 2

```
foodagg=agnes(food,diss=FALSE,metric="euclidian")
```

```
plot(foodagg) ## dendrogram
```