WWD_3

zhou ziliang

2019/11/12

```r
## load the packeage
  library(cclust)
  library(tidyverse)

## read file
  animals <- read.csv("Animals.csv")
## rename the animals
  animals <-rename(animals,Weight = 'x',Height = 'y',Species = 'z')
  animals$Species <- factor(animals$Species, levels = c("a","b","c","d"),
                            labels = c("Ostrich","Deer","Bear","Gaint tortise
"))
```

Question 1

```r
d <- cbind(animals["Weight"],animals["Height"]) %>%
    dist()
  hd <- hclust(d)
  group.4 <- cutree(hd,4)
  table(animals$Species,group.4)
```

```
##                 group.4
##                    1   2   3   4
##   Ostrich        541 383   7   0
##   Deer           210 434  72   0
##   Bear            32 478 263  38
##   Gaint tortise    0  81 262 225
```

the method cannot separate the animals very well. For example, the most of Deer and Bear are separated in group 2

the function of calculate the max curvature.

as we know, in a ~~parame~~ twice differentiable plane curve

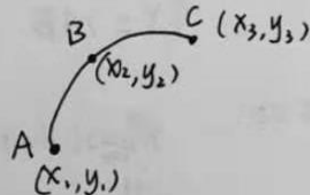
the curvature is:

$$K = \frac{x'y'' - y'x''}{(x'^2 + y'^2)^{\frac{3}{2}}}$$

assume, there are three point in a curve

A is starting point, C is ending point

assume, the parametric equation of it is



$$\begin{cases} x = a_1 + a_2 t + a_3 t^2 \\ y = \cancel{a} b_1 + b_2 t + b_3 t^2 \end{cases}$$

Let $t_a = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

$t_b = \sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2}$

so: $t \in (t_a, t_b)$

as we know

$(x, y)|_{t=-t_a} = (x_1, y_1)$

$(x, y)|_{t=0} = (x_2, y_2)$

$(x, y)|_{t=t_b} = (x_3, y_3)$

so:

$$\begin{cases} x_1 = a_1 - a_2 t_a + a_3 t_a^2 \\ x_2 = a_1 \\ x_3 = a_1 + a_2 t_b + a_3 t_b^2 \end{cases} \qquad \begin{cases} y_1 = b_1 - b_2 t_a + b_3 t_a^2 \\ y_2 = b_1 \\ y_3 = b_1 + b_2 t_a + b_3 t_a^2 \end{cases}$$

Let: $X = (x_1, x_2, x_3)'$, $A = (a_1, a_2, a_3)'$

$B = (b_1, b_2, b_3)^T$, $M = \begin{Bmatrix} 1, & ta, & ta^2 \\ 1, & 0, & 0 \\ 1, & ta, & ta^2 \end{Bmatrix}$

So: we can translate above equatin into matrix form.

So: $\begin{cases} X = MA \\ Y = MB \end{cases} \Rightarrow \begin{cases} A = M^{-1}X \\ B = M^{-1}Y \end{cases}$

when $t=0$,

$$x'|_{t=0} = (a_1 + a_2 ta + a_3 ta^2)'|_{ta=0} = a_2$$

$$x''|_{t=0} = (a_1 + a_2 ta + a_3 ta^2)''|_{ta=0} = 2a_3$$

$$y'|_{t=0} = (b_1 + b_2 ta + b_3 ta^2)'|_{ta=0} = b_2$$

$$y''|_{t=0} = (b_1 + b_2 t + b_3 t^2)''|_{t=0} = 2b_3.$$

So.

the curveture ρ

$$k = \frac{x''y' - x'y''}{\left((x')^2 + (y')^2\right)^{\frac{3}{2}}} = \frac{2a_3 \cdot b_2 - 2a_2 \cdot b_3}{(a_2^2 + b_2^2)^{\frac{3}{2}}}$$

```r
curv <- function(x,y){
    ta <- sqrt((x[2]-x[1])^2+(y[2]-y[1])^2)
    tb <- sqrt((x[3]-x[2])^2+(y[3]-y[2])^2)
    M_matrix <- matrix(c(1,ta,ta^2,1,0,0,1,tb,tb^2),3,3,byrow=TRUE)
    M_inverse <- solve(M_matrix)
    a <- M_inverse %*% x
    b <- M_inverse %*% y
    curvature <- 2*(a[3]*b[2]-a[2]*b[3])/sqrt(a[2]^2+b[2]^2)
    return(curvature)
  }
  max.curv <- function(x,y) {
      I <- c(2:(length(x)-1))
      cur <- I
    for (i in I){
```

```r
    s.x <- c(x[i-1],x[i],x[i+1])
    s.y <- c(y[i-1],y[i],y[i+1])
    cur[i] <- curv(s.x,s.y)
  }
    cur <- abs(cur)
    max <- max(cur)
    n <- which.max(cur)+1
    result <- list(max,n,cur)
    names(result) <-c("max value","number","curature")
    return(result)
}
```

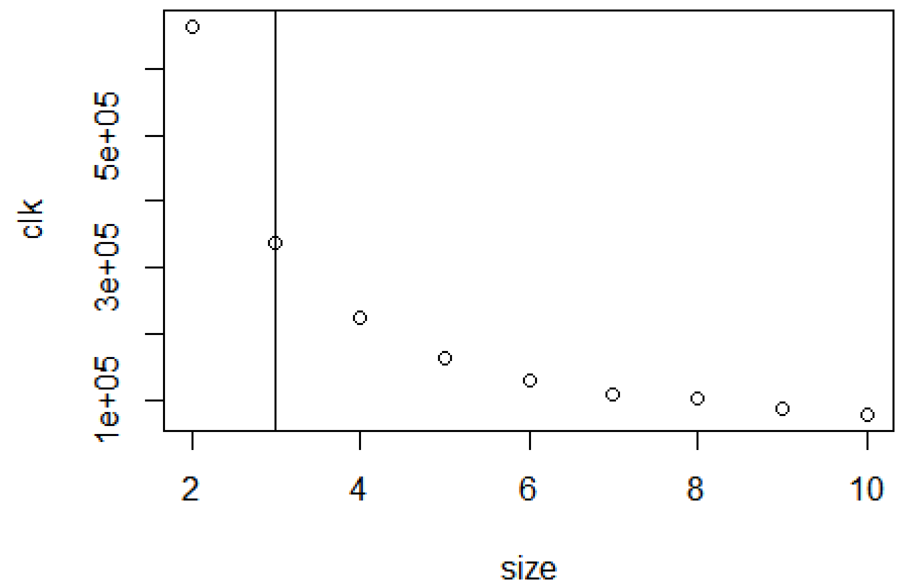
Question 2

```r
animals_df <- cbind(animals["Weight"],animals["Height"])

tot.withinss <- function(number,df){
  c1 <- kmeans(df,number)
  return(c1$tot.withinss)
}
size = c(2:10)
clk <- sapply(size,tot.withinss,df = animals_df)

# find the max curvature
max.curv(size,clk)

# plot the clk against the size, and plot the vertical line
plot(size,clk)
abline(v = 3)
```



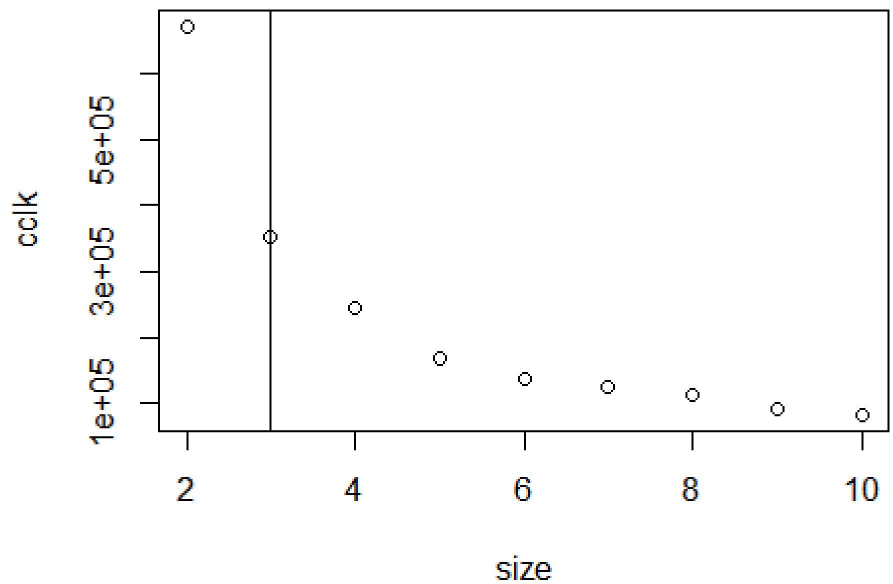

```
kc <- kmeans(animals_df,3)
table(animals$Species,kc$cluster)

##
##                  1   2   3
##   Ostrich        5 652 274
##   Deer          36 302 378
##   Bear         292  54 465
##   Gaint tortise 488   0  80
```

Question 3

```
ctot.withinss <- function(number,df,dist="manhattan"){
  df <- as.matrix(df)
  c2 <- cclust(df,number,dist = dist)
  return(sum(c2$withinss))
}
cclk <- sapply(size,ctot.withinss,df = animals_df)
max.curv(size,cclk)

plot(size,cclk)
abline(v=3)
```




```
  ckc <- animals_df %>% as.matrix() %>%
    cclust(3,dist = "manhattan")
  table(animals$Species,ckc$cluster)

##
##                      1   2   3
##   Ostrich            7 283 641
##   Deer              61 368 287
##   Bear             339 422  50
##   Gaint tortise    516  52   0
```

There are not obesely difference of this two cluster obtained.