Homework 8

Yi Chen

3/17/2020

## Homework 8

### problem one

There are eight ingredients for each type of pizza, which ingredients are binary scored (contain or not contain).

Without any other prior information, it’s safe to assume that each ingredients measure the characteristic of pizza equally. Consequently, the similarity coefficient can be defined as the aggrement (proportation of the same ingredient requirement) of their ingredients:

Thus, the agreement level is in the range of 0 to 1. When all the ingredients requirement status are the same for the pair of pizza and pizza , the agreement level is 1. Similarly, if all the ingredient requirement status are different for the pair of pizza and , the aggreement level is 0.

Bascially, it measure the proportation of ingredient among the 8 ingredients that the pair of pizza share the same requirement (both contain or both uncontain).

It is a unique format of construct model of feature-matching:

Here, I take , as the number of indentical elements, and .

Aggreement is also a unique Gower’s coefficient for binary variables and weight for all ingredient variables as the same as 1.

pizza\_1 <- c(1,1,0,0,0,0,0,0)  
pizza\_2 <- c(0,1,1,0,0,0,0,0)  
pizza\_3 <- c(1,1,1,0,0,0,0,0)  
pizzas <- list(pizza\_1,pizza\_2,pizza\_3)  
  
for (i in 1:ncol(combn(3,2))){  
 print(paste("similarity coefficient between ",as.character(combn(3,2)[1,i]), " and ",as.character(combn(3,2)[2,i])))  
   
 print(mean((pizzas[[combn(3,2)[1,i]]]==pizzas[[combn(3,2)[2,i]]])))  
}

## [1] "similarity coefficient between 1 and 2"  
## [1] 0.75  
## [1] "similarity coefficient between 1 and 3"  
## [1] 0.875  
## [1] "similarity coefficient between 2 and 3"  
## [1] 0.875

### problem two

In this example, the data is the multivariate data which contains four main variables for measuring the similiarity. All four variables are continuous variable, which is bigger than 0.

There are two possible choice for these ratio variables: Gower’s coefficient and Mahalanobis distance.

The Gowers’ coefficient can be calculated by:

, where .

The Mahalanobis distance can be calculated by:

To test which approch performance better. I test both methods with the given IRIS\_MLT.txt dat. The one makes the within group distance small and between group distance big will be the one perform better.

iris <- read.table("IRIS\_MLT.txt")  
## for the Gower’s coefficient  
range\_v <- apply(iris[1:4],2,max) - apply(iris[1:4],2,min)  
gower <- function(x,y){  
 s <- 0  
 for (k in 1:4){  
 s <- s + (1 - abs(x[k]-y[k])/range\_v[k])  
 }  
 s <- s / 4  
 return(s)  
}  
  
within\_group\_distance <- 0  
between\_group\_distance <- 0  
for (i in 1:(nrow(iris)-1)){  
 for (j in i:nrow(iris)){  
 x <- as.vector(iris[i,1:4])  
 y <- as.vector(iris[j,1:4])  
 distance <- (1 - as.numeric(gower(x,y)))^(1/2)  
 if (iris[i,5] == iris[j,5]){  
 within\_group\_distance = within\_group\_distance + distance   
 }else{  
 between\_group\_distance = between\_group\_distance + distance  
 }  
}  
}  
  
within\_group\_distance / between\_group\_distance

## [1] 0.2735307

## for the Mahalanobis distance  
covariance\_matrix <- cov(iris[,1:4])  
within\_group\_distance <- 0  
between\_group\_distance <- 0  
for (i in 1:(nrow(iris)-1)){  
 for (j in i:nrow(iris)){  
 x <- iris[i,1:4]  
 y <- iris[j,1:4]  
 distance <- sqrt(as.matrix((x - y)) %\*% as.matrix(solve(covariance\_matrix)) %\*% as.matrix(t(x-y)))  
 if (iris[i,5] == iris[j,5]){  
 within\_group\_distance = within\_group\_distance + distance   
 }else{  
 between\_group\_distance = between\_group\_distance + distance  
 }  
}  
}  
  
within\_group\_distance / between\_group\_distance

## 1  
## 1 0.365216

The ratio of sum of within group distance to the sum of between group distance for Mahalanobis distance is bigger than Gower’s coefficient. This means Gower’s coefficient has a better performance, according to our requirement, which gives a relative smaller within group distance compared with the between group distance. Consequently, I pick the Gower’s coefficient.

x\_list <- list(iris[1,1:4],iris[51,1:4],iris[101,1:4])  
  
all\_comb <- combn(3,2)  
for (i in 1:ncol(all\_comb)){  
 print(paste("similarity coefficient between ",as.character(all\_comb[1,i]), " and ",as.character(all\_comb[2,i])))  
 print((1 - gower(x\_list[[all\_comb[1,i]]],x\_list[[all\_comb[2,i]]]))^(1/2))  
}

## [1] "similarity coefficient between 1 and 2"  
## V1  
## 1 0.6542362  
## [1] "similarity coefficient between 1 and 3"  
## V1  
## 1 0.7339382  
## [1] "similarity coefficient between 2 and 3"  
## V1  
## 51 0.4782216

### problem three

In this multivariate data set, all the vaiables are comtinuous ratio variables. Again, Gower’s coefficient and Mahalanobis distance are appropriate for handeling this data set. Here, I recommend the use of Mahalanobis distance because it is usually related to the other multivariate analysis techniques like mutltivariate linear regression which is based on the normal distribution. Besides, the calculation of Mahalanobis distance require relatively less computation.

##### Optional question

The data set in the document which contian additional categorial (nomial) data set should not be incorperated in to the distance calculation directly since they are using the different scale.