hw6_prob1

Haoyu_Zhao

May 29, 2017

Problem 1

First put the result and the code of this problem

```
#two ways to do the MDS, the first uses the center of each region and channel,
#the second uses the whole data
library(MASS)
library(distances)
#loading the data from the data set
readdata <- read.csv('./Wholesale customers data.csv')
#a new data frame without the first 2 attribute, Channel and Region
data_new <- readdata[,3:8]
dim(data_new)
```

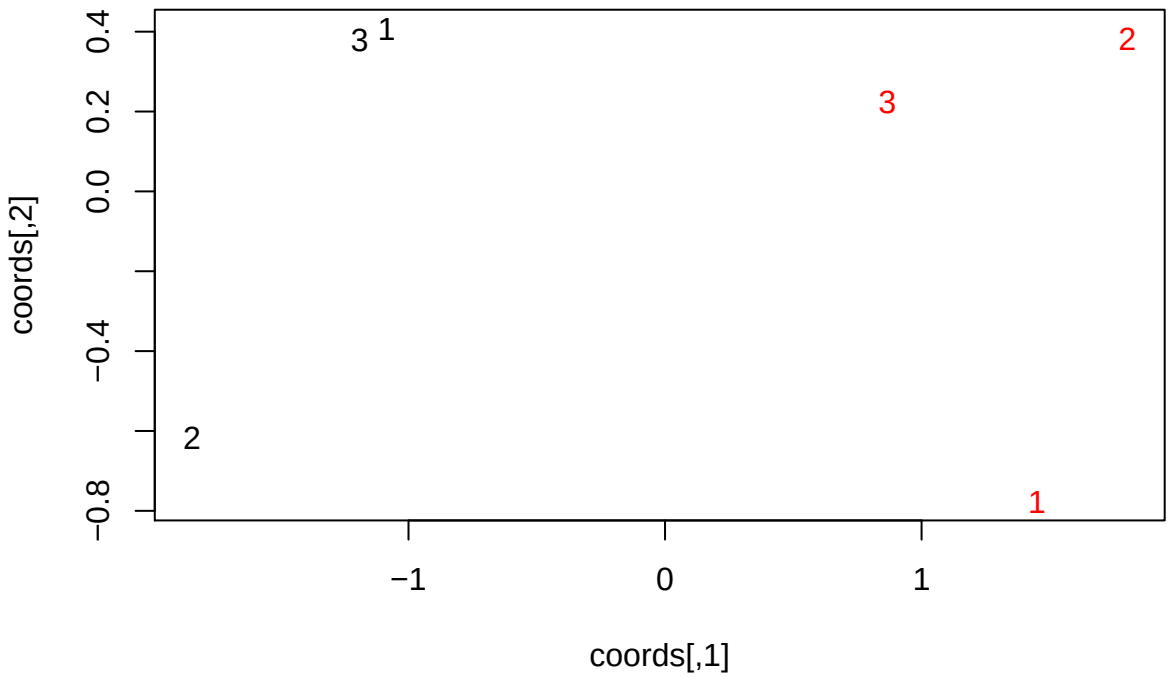
```
## [1] 440   6
```

```
data_new <- as.matrix(data_new)
dim(data_new)
```

```
## [1] 440   6
```

```
#label the records, easy to do the MDS
labs <- rep(0,440)
for (i in 1:440) {
  labs[i] <- (readdata$Channel[i]-1) * 3 + readdata$Region[i]
}

labs <- matrix(labs,440,1)

data_new <- cbind(labs, data_new)

#declare the variables
centres <- matrix(0,6,6)
S <- as.matrix(var(data_new[,-1]))

mahal <- matrix(0,6,6)

#compute the centers of the each of the classes
for(i in 1:6) {
  centres[i,] <- apply(data_new[labs==i,-1],2,mean)
}

#compute the mahalanobis distance for the centers
for(i in 1:6) {
  mahal[i,] = mahalanobis(centres, centres[i,], S)
}

#plot the centers by MDS
```

```
coords <- cmdscale(mahal)
plot(coords, type = 'n')
text(coords, labels = c(1,2,3,1,2,3),
     col = c(1,1,1,2,2,2))
```




```
#the second way to do the MDS, using the 440 points
library(MASS)
library(distances)
#loading the data from the data set
readdata <- read.csv('./Wholesale customers data.csv')

#a new data frame without the first 2 attribute
data_new <- readdata[,3:8]
dim(data_new)
```

```
## [1] 440   6
```

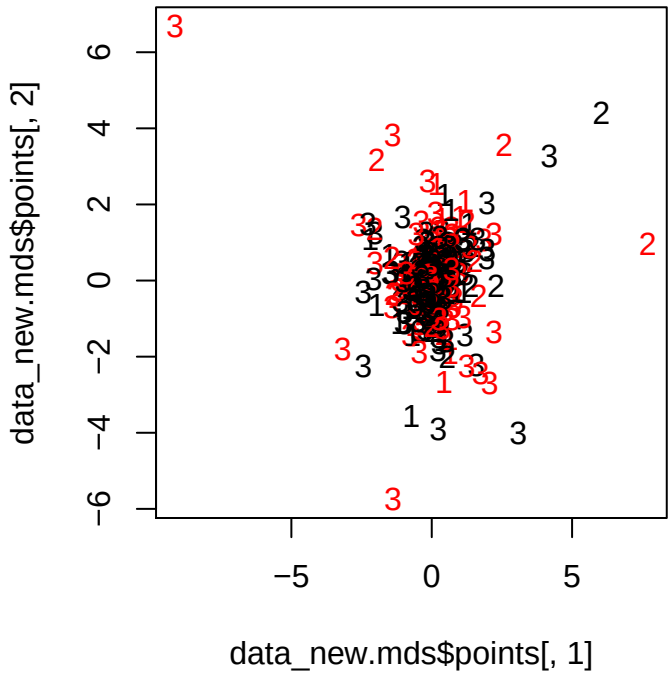
```
data_new <- as.matrix(data_new)
dim(data_new)
```

```
## [1] 440   6
```

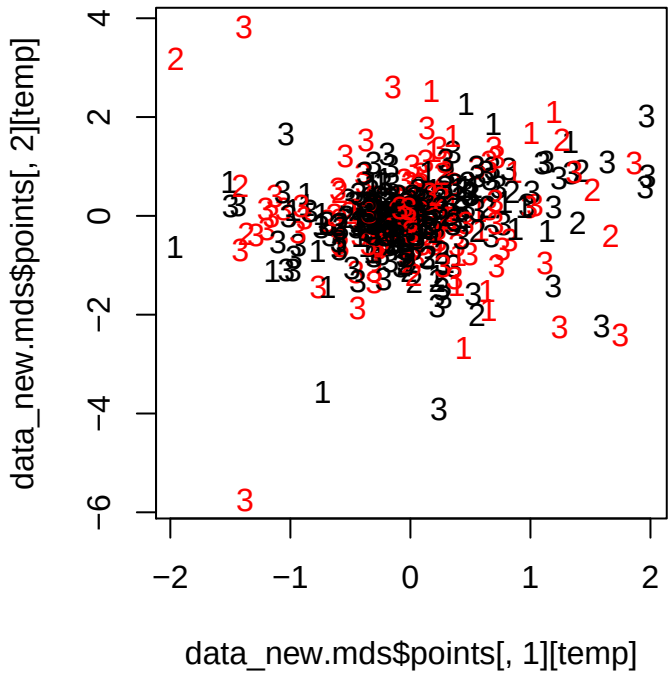
```
#calculating the distance
#using the mahalanobis distance
#the distance_matrix method is in the package 'distances'
data_new.mahal <- distances(data_new, normalize='mahalanobis')
data_new.dist <- as.matrix(distance_matrix(data_new.mahal))

#do the MDS,with k=399
data_new.mds <- cmdscale(data_new.dist, k=339, eig=T)
```

```
## Warning in cmdscale(data_new.dist, k = 339, eig = T): only 219 of the first
## 339 eigenvalues are > 0
```

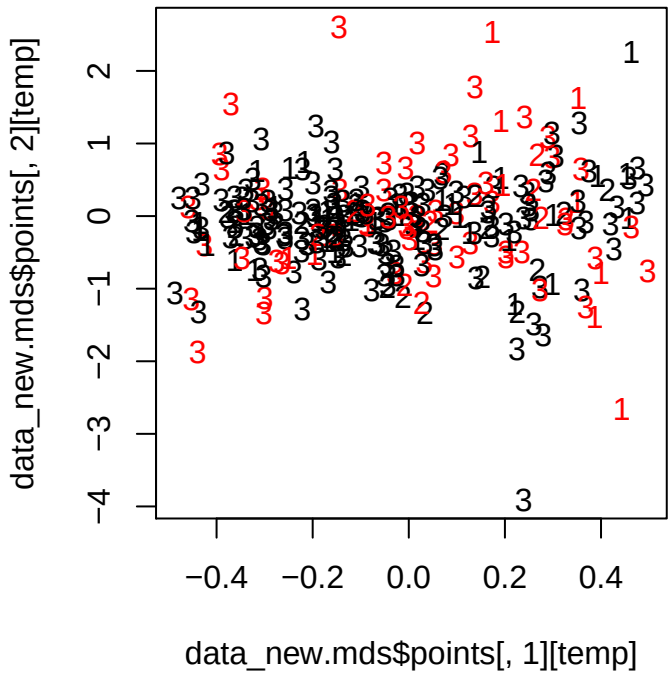
```
#plot the result
par(pty='s')
plot(data_new.mds$points[,1], data_new.mds$points[,2],type='n')
text(data_new.mds$points[,1], data_new.mds$points[,2],
     labels=readdata$Region,col=readdata$Channel)
```




```
#zoom part of the graph, in which the points are much more than other parts
temp <- c()
for (i in 1:440) {
  if (data_new.mds$points[,1][i] < 2 & data_new.mds$points[,1][i] > -2) {
    temp <- append(temp,i)
  }
}

#plot the result
par(pty='s')
plot(data_new.mds$points[,1][temp], data_new.mds$points[,2][temp],type='n')
text(data_new.mds$points[,1][temp], data_new.mds$points[,2][temp],
     labels=(readdata$Region)[temp],col=readdata$Channel[temp])
```




```
#zoom part of the graph again, in which the points are much more than other parts
temp <- c()
for (i in 1:440) {
  if (data_new.mds$points[,1][i] < 0.5 & data_new.mds$points[,1][i] > -0.5) {
    temp <- append(temp,i)
  }
}

#plot the result
par(pty='s')
plot(data_new.mds$points[,1][temp], data_new.mds$points[,2][temp],type='n')
text(data_new.mds$points[,1][temp], data_new.mds$points[,2][temp],
     labels=(readdata$Region)[temp],col=readdata$Channel[temp])
```



a. In the first method, $k = 2$, because the points are too few.

In the second method, the k is selected as 399, $399 = 440 - 1$. There are 440 records in total, and after calcualting the mahalanobis distance, the matrix is 440×440. So we select $k = 399$ and do the MDS, and the MDS will stop when the eigenvalue is less than 0. In this way, it will give the most precise result of all the details.

b. The distance used in this problem is mahalanobis distance.

For the first method, we compute the centers of each class and then compute the mahalanobis distance for each center by the covariance matrix computed by the raw data.

For the second method, we use the mahalanobis distance for all the data. We first use the records to approximate the covariance matrix of the data, and then normalize the data by the approximated covariance matrix. After normalized, we can compute the points by euclidean distance, and the final result is the mahalanobis distance of the original data.

c. In the first method, we can see that the red dots are on the right and the black points are on the left of the plot. The label 3 is nearest and the label 1 and 2 is far away.

In the second method, there are 3 plots in total. The first is the origianl graph, the second is a zoomed graph of the first, and the third is a zoomed graph or part of the second graph. From the whole graph, we can see that there are some outliers in the data. From the zoomed graph we can see that the red points(Channel 1) is more scattered than the black points(Channel 2). We can also find that the red points tend to lies in the right, but the black points tend to lies in the left. In total, the red and black points just interwined with each other and is hard to seperate them. From the courtier of the points, no matter black or red, the center has larger density and the margin has lower density, and the courtier has ecllipse shape, so it is reasonable to assume it as normal.

But it is really hard to distinguish the region of the points in the plot. From the plot, we can see that all of them are really scattered. We can find many points with label '3', but that is due to the fact the number of the region '3' is much more than the orthes. So in this method that we print all the points seperately on the graph, it is really hard to find the features to distinguish the region feature. If we first take the mean of each (Channel,Region) pair and plot the 6 possible pairs on the graph, it may help find the relations of the Region features.