高级统计方法 第10次作业:

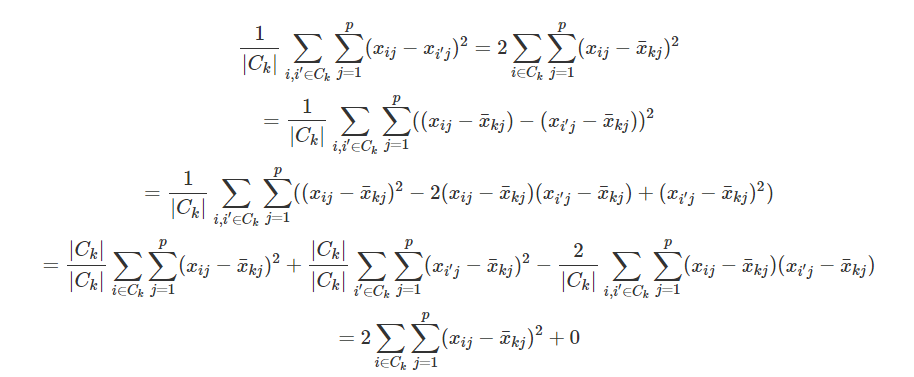
序号： 08 姓名： 潘晨楷 学号： 20212241116 班级：软2107

**作业评阅：**

**概念**

1.问题（略）

（a）问题（略）



（b）问题（略）

由式(10.12)可知，最小化每个聚类的欧氏距离平方和与最小化每个聚类的聚类内方差是一样的

2.问题（略）

（a）问题（略）

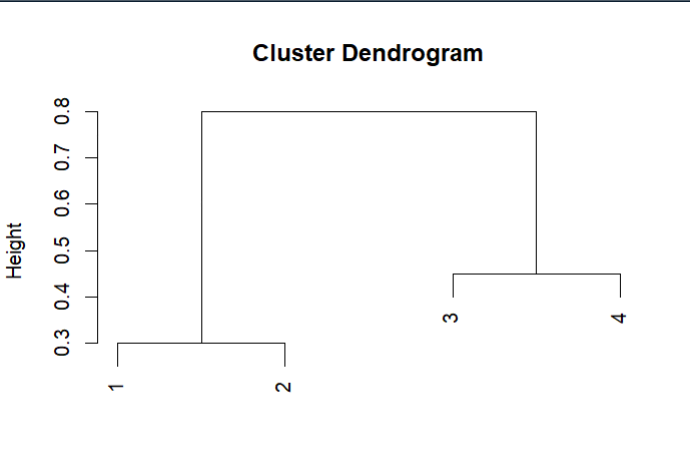
d = as.dist(matrix(c(0, 0.3, 0.4, 0.7,

0.3, 0, 0.5, 0.8,

0.4, 0.5, 0.0, 0.45,

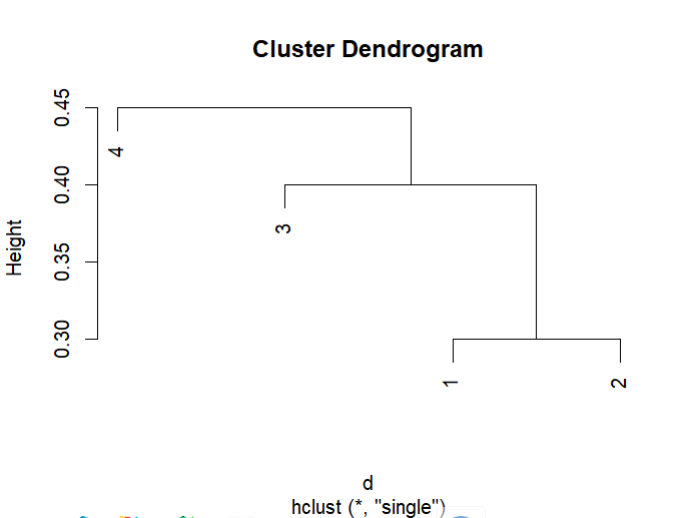
0.7, 0.8, 0.45, 0.0), nrow=4))

plot(hclust(d, method="complete"))



（b）问题（略）

plot(hclust(d, method="single"))



（c）问题（略）

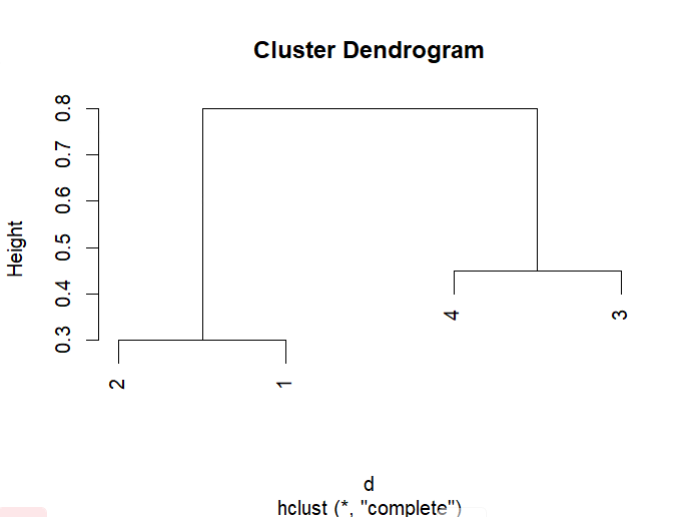
(1, 2) (3, 4)

（d）问题（略）

(1, 2, 3) (4)

（e）问题（略）

plot(hclust(d, method="complete"), labels=c(2,1,4,3))



3.问题（略）

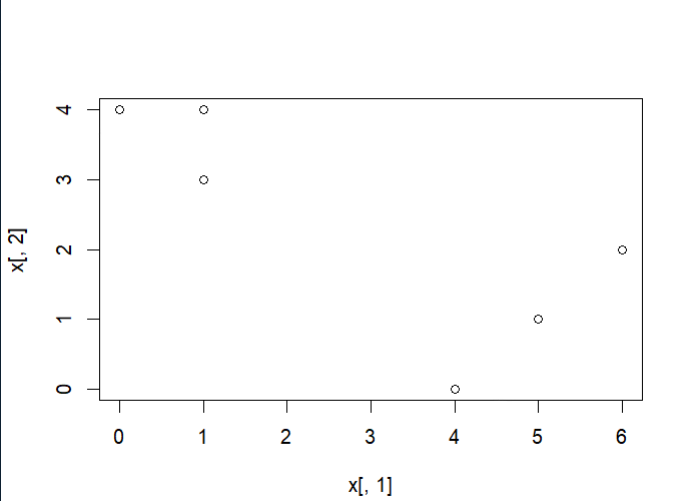
set.seed(1)

x = cbind(c(1, 1, 0, 5, 6, 4), c(4, 3, 4, 1, 2, 0))

x

（a）问题（略）

plot(x[,1], x[,2])



（b）问题（略）

labels = sample(2, nrow(x), replace=T)

Labels



（c）问题（略）

centroid1 = c(mean(x[labels==1, 1]), mean(x[labels==1, 2]))

centroid2 = c(mean(x[labels==2, 1]), mean(x[labels==2, 2]))

centroid1



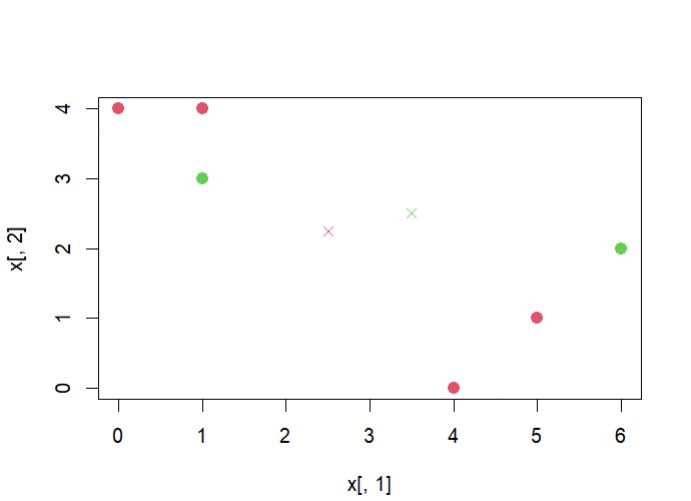
centroid2



plot(x[,1], x[,2], col=(labels+1), pch=20, cex=2)

points(centroid1[1], centroid1[2], col=2, pch=4)

points(centroid2[1], centroid2[2], col=3, pch=4)



（d）问题（略）

euclid = function(a, b) {

return(sqrt((a[1] - b[1])^2 + (a[2]-b[2])^2))

}

assign\_labels = function(x, centroid1, centroid2) {

labels = rep(NA, nrow(x))

for (i in 1:nrow(x)) {

if (euclid(x[i,], centroid1) < euclid(x[i,], centroid2)) {

labels[i] = 1

} else {

labels[i] = 2

}

}

return(labels)

}

labels = assign\_labels(x, centroid1, centroid2)

Labels



（e）问题（略）

last\_labels = rep(-1, 6)

while (!all(last\_labels == labels)) {

last\_labels = labels

centroid1 = c(mean(x[labels==1, 1]), mean(x[labels==1, 2]))

centroid2 = c(mean(x[labels==2, 1]), mean(x[labels==2, 2]))

print(centroid1)

print(centroid2)

labels = assign\_labels(x, centroid1, centroid2)

}



Labels

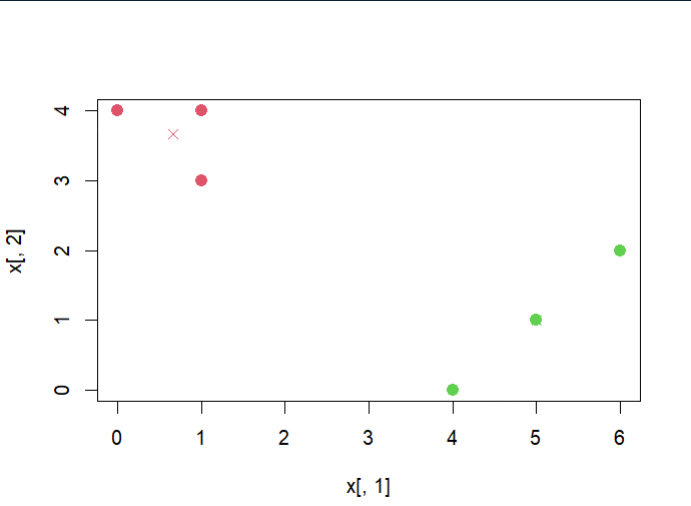


（f）问题（略）

plot(x[,1], x[,2], col=(labels+1), pch=20, cex=2)

points(centroid1[1], centroid1[2], col=2, pch=4)

points(centroid2[1], centroid2[2], col=3, pch=4)



4.问题（略）

（a）问题（略）

没有足够的信息。最大簇间不相似度可以等于或不等于最小簇间不相似度。如果差异相等，它们就会在同一高度融合。如果它们不相等，它们的单键树状图将在较低的高度融合。

（b）问题（略）

两种汇合的高度相同，因为连锁不影响叶片与叶片的融合。

、、、、、、

**应用**

7.问题（略）

library(ISLR)

set.seed(1)

dsc = scale(USArrests)

a = dist(dsc)^2

b = as.dist(1 - cor(t(dsc)))

summary(b/a)



10.问题（略）

（a）问题（略）

set.seed(2)

x = matrix(rnorm(20\*3\*50, mean=0, sd=0.001), ncol=50)

x[1:20, 2] = 1

x[21:40, 1] = 2

x[21:40, 2] = 2

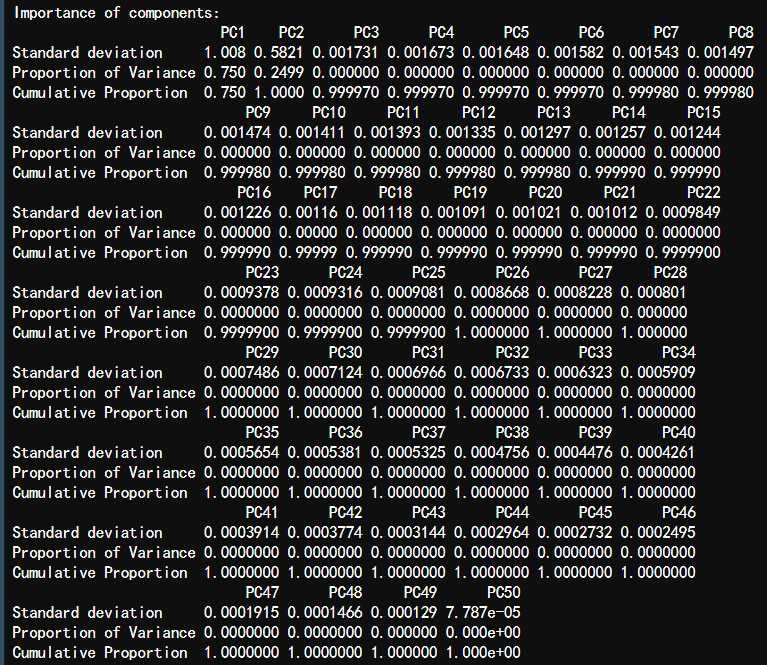
x[41:60, 1] = 1

这里的概念是在两个维度中分离这三个类。

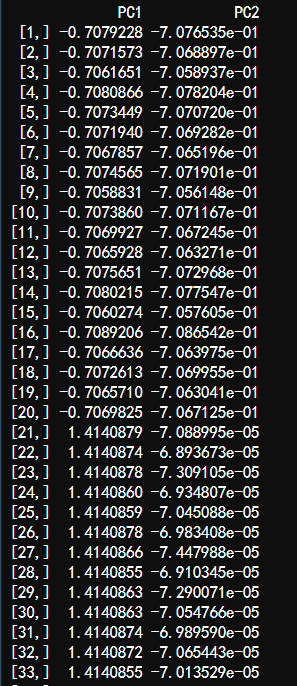
（b）问题（略）

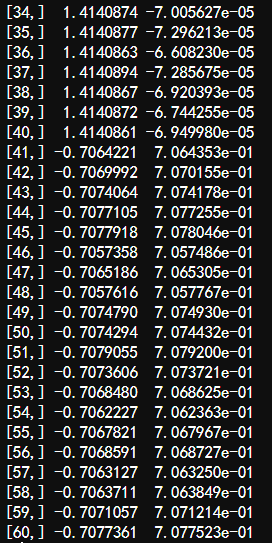
pca.out = prcomp(x)

summary(pca.out)

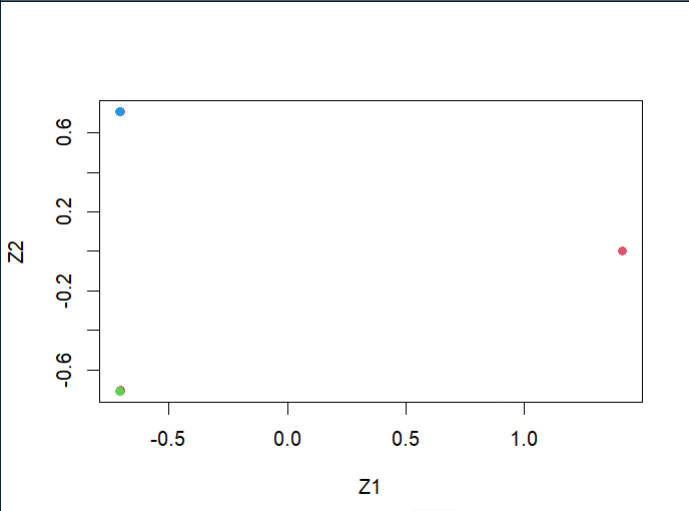


pca.out$x[,1:2]





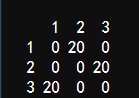
plot(pca.out$x[,1:2], col=2:4, xlab="Z1", ylab="Z2", pch=19)



（c）问题（略）

km.out = kmeans(x, 3, nstart=20)

table(km.out$cluster, c(rep(1,20), rep(2,20), rep(3,20)))



匹配结果很好

（a）问题（略）

km.out = kmeans(x, 2, nstart=20)

km.out$cluster



之前一个类的所有还是集中在一个类里。

（b）问题（略）

km.out = kmeans(x, 4, nstart=20)

km.out$cluster

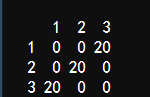


之前的一个集群全部拆分为两个集群。

（c）问题（略）

km.out = kmeans(pca.out$x[,1:2], 3, nstart=20)

table(km.out$cluster, c(rep(1,20), rep(2,20), rep(3,20)))



匹配得非常好

（d）问题（略）

km.out = kmeans(scale(x), 3, nstart=20)

km.out$cluster



结果比（b）差：观测值的缩放会影响它们之间的距离。