hw2

hyz

2015年11月18日

library(ggplot2) ##read data setwd("E:\hw2") data<-read.table(file="GeneMatrix.txt") data<-t(data) data<-as.data.frame(data)

## problem1-1

d<-dist(data,method="euclidean") model<-hclust(d,method = "average") mod1<-as.dendrogram(model) plot(mod1,labels=FALSE)

result<-cutree(model,k=2)

heatmap(as.matrix(data),Rowv =mod1, Colv = F,scale= "none")

## load the result data

resultdata <- read.delim(file = "clinical\_data.txt") state <- resultdata[1,8]

data <- cbind(state,data) names(data)[which(names(data[,])=="PH-4")]=c("PH4") for(i in 1:3) { resultdata$sampleID=sub("-",".",resultdata$sampleID) } for( i in 1:length(data$state)) { data[i,1]=resultdata[which(resultdata$sampleID==row.names(data[i,])),8] }

## count the error%

count = 0 result <- as.data.frame(result) for(i in 1:length(data[,1])) { if((as.character(data[i,1])=="Positive" && result[i,1]=="2") || (as.character(data[i,1])=="Negative" && result[i,1]=="1")) count=count+1 }

show("Error Rate:") show(count/length(data[,1]))

## problem1-2:pca

dataset<-read.table("GeneMatrix.txt") m\_matrix <- data.matrix(dataset) ouj<-eigen((m\_matrix)%*%t(m\_matrix)) pca<-t(m\_matrix)%*%ouj$vector[,1:20]

d\_pca<-dist(pca) model\_pca<-hclust(d\_pca,method = "average") mod\_pca<-as.dendrogram(model\_pca) plot(model\_pca,labels=FALSE)

result\_pca=cutree(model\_pca,k=2) err\_pca=0 result\_pca <- as.data.frame(result\_pca) for(i in 1:length(data[,1])) { if((as.character(data[i,1])=="Positive" && result\_pca[i,1]=="2") || (as.character(data[i,1])=="Negative" && result\_pca[i,1]=="1")) err\_pca=err\_pca+1 }

show("Error Rate for PCA:") show(err\_pca/length(data[,1]))

## result:

# 用所有特征聚类的错误率是6.32%

# 用PCA提取前20个主成分后进行聚类的错误率是6.70%

## problem2

d=matrix(c(1,0,2,0,2,0,2, 1,1,1,0,1,0,2, 1,2,1,1,1,1,1, 0,1,0,2,0,1,1, 0,2,1,2,0,1,0),nrow=7,ncol=5) eig<-eigen(t(d)%\*%d) #第一主成分 eig$vectors[,1]