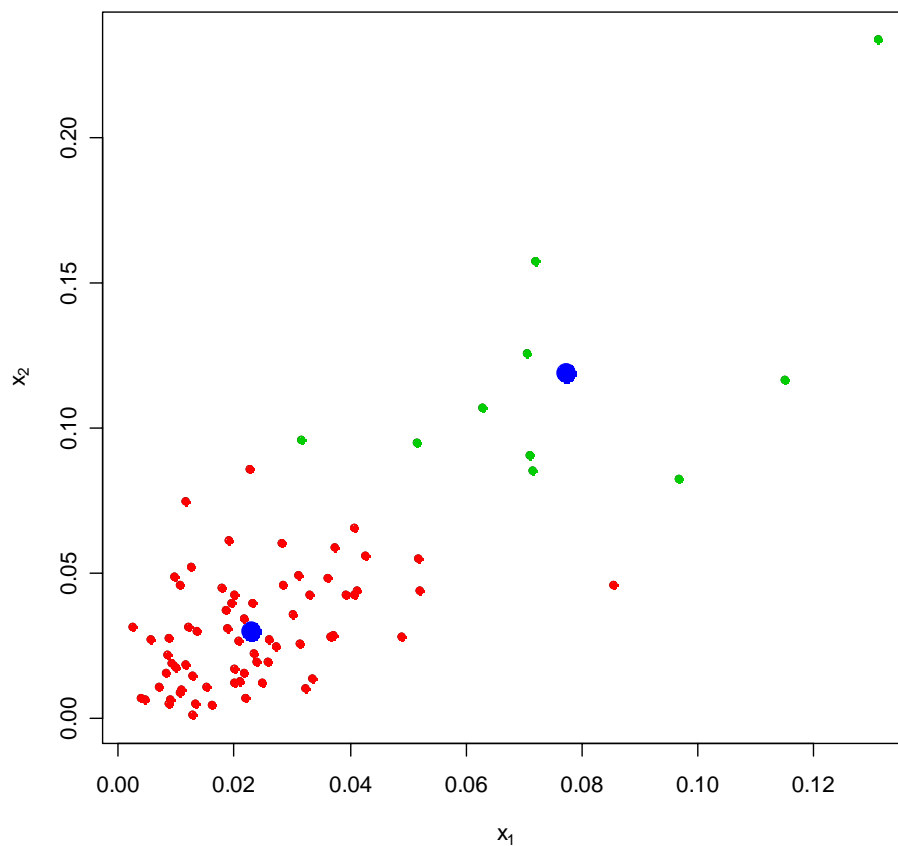


Homework 4 Solutions

1) Read Chapter 8 (Sections 8.1 and 8.2) and Chapter 2 (Section 2.4).

2)

```
data<-read.csv("sonar_test.csv",header=FALSE)
x<-data[,1:2]
plot(x,pch=19,xlab=expression(x[1]),ylab=expression(x[2]))
fit<-kmeans(x, 2)
points(fit$centers,pch=19,col="blue",cex=2)
library(class)
knnfit<-knn(fit$centers,x,as.factor(c(-1,1)))
points(x,col=1+1*as.numeric(knnfit),pch=19)
```



3)

```
y<-data[,61]  
sum(knnfit==y)/length(y)
```

You get 47% misclassification error.

4)

```
x<-data[,1:60]  
fit<-kmeans(x, 2)  
library(class)  
knnfit<-knn(fit$centers,x,as.factor(c(-1,1)))  
sum(knnfit==y)/length(y)
```

You get 44% misclassification error.

6)

```
x<-c(1,2,2.5,3,3.5,4,4.5,5,7,8,8.5,9,9.5,10)
```

```
center1<-1
```

```
center2<-2
```

```
for (k in 2:10){
```

```
  cluster1<-x[abs(x-center1[k-1])<=abs(x-center2[k-1])]
  cluster2<-x[abs(x-center1[k-1])>abs(x-center2[k-1])]
  center1[k]<-mean(cluster1)
```

```
  center2[k]<-mean(cluster2)
```

```
  center1[k]<-mean(cluster1)
```

```
  center2[k]<-mean(cluster2)
```

```
}
```

```
center1
```

```
center2
```

The final values for center1 and center2 match our answers from the previous problem.

7)

```
kmeans(x,2)
```

gives

Cluster means:

```
  [,1]
```

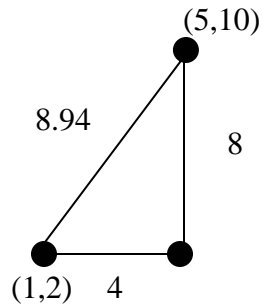
```
1 8.666667
```

```
2 3.187500
```

which matches our answer from the previous problem.

8)

a) The distance is the square root of $(1-5)^2+(2-10)^2$ which is the square root of 4^2+8^2 which is the square root of 80 which is 8.94.



b)

```
x1<-c(1,2)
x2<-c(5,10)
```

```
data<-matrix(c(x1,x2),nrow=2,byrow=T)
```

```
dist(data)
```

This gives the same answer of 8.94.

9)

a) The distance is the square root of $(1-5)^2+(2-10)^2+(3-4)^2+(6-12)^2$ which is 10.81665.

b)

```
x1<-c(1,2,3,6)
```

```
x2<-c(5,10,4,12)
```

```
data<-matrix(c(x1,x2),nrow=2,byrow=T)
```

```
dist(data)
```

This gives the same answer of 10.81665.

10) Read Chapter 10.

11) There are no outliers for midterm 1 by the $z = \pm 3$ rule. The smallest z is -2.28 and the largest z is 1.85.

```
data<-read.csv("spring2008exams.csv")  
  
exam1mean<-mean(data[,2],na.rm=TRUE)  
  
exam1sd<-sd(data[,2],na.rm=TRUE)  
  
z<-(data[,2]-exam1mean)/exam1sd  
  
sort(z)
```

12) There are no outliers for midterm 2 by the $z = \pm 3$ rule. The smallest z is -2.40 and the largest z is 1.30.

```
data<-read.csv("spring2008exams.csv")  
  
exam2mean<-mean(data[,3],na.rm=TRUE)  
  
exam2sd<-sd(data[,3],na.rm=TRUE)  
  
z<-(data[,3]-exam2mean)/exam2sd  
  
sort(z)
```

13) The user agent

Mozilla/5.0 (X11; U; Linux i686 (x86_64); en-US; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3

has a z score of 8.0 and the user agent

Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.1.4) Gecko/20070515 Firefox/2.0.0.4

has a z score of 3.7 so these two are outliers by the $z = \pm 3$ rule.

14) This rule says the grade score of 64 is an outlier for midterm 2.

```
data<-read.csv("spring2008exams.csv")
```

```
q1<-quantile(data[,3],.25,na.rm=TRUE)
```

```
q3<-quantile(data[,3],.75,na.rm=TRUE)
```

```
iqr<-q3-q1
```

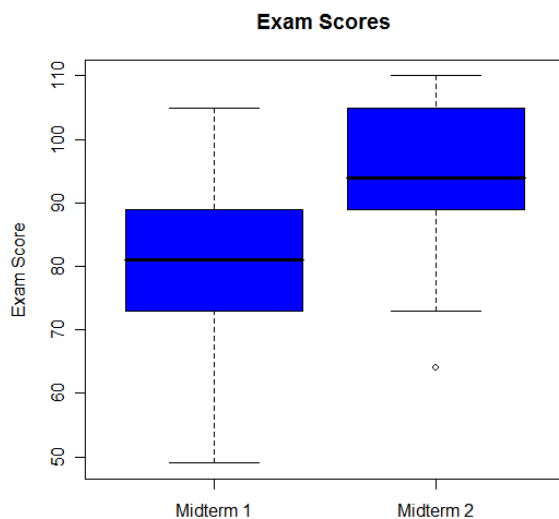
```
data[(data[,3]>q3+1.5*iqr),3]
```

```
data[(data[,3]<q1-1.5*iqr),3]
```

```
boxplot(data[,2],data[,3],col="blue",
```

```
main="Exam Scores",
```

```
names=c("Midterm 1","Midterm 2"),ylab="Exam Score")
```



15) Student #5 has the largest positive residual. The residual value was 18.17.

```
data<-read.csv("spring2008exams.csv")
model<-lm(data[,3]~data[,2])
plot(data[,2],data[,3],pch=19,xlab="Exam 1",
ylab="Exam2",xlim=c(60,110),ylim=c(60,110))
abline(model)
sort(model$residuals)
```

