Homework 4

Michael Carrion

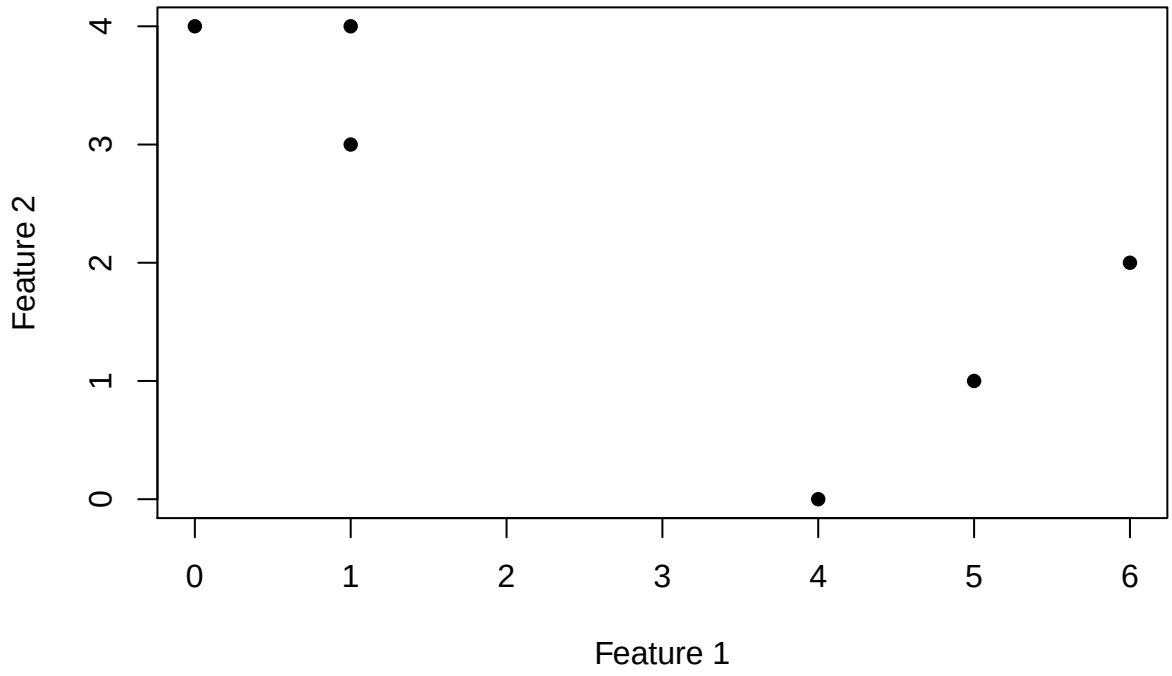
March 2, 2020

```
load("~/problem-set-4/Data and Codebook/legprof-components.v1.0.RData")
myDf <- x
```

Performing k-Means by Hand

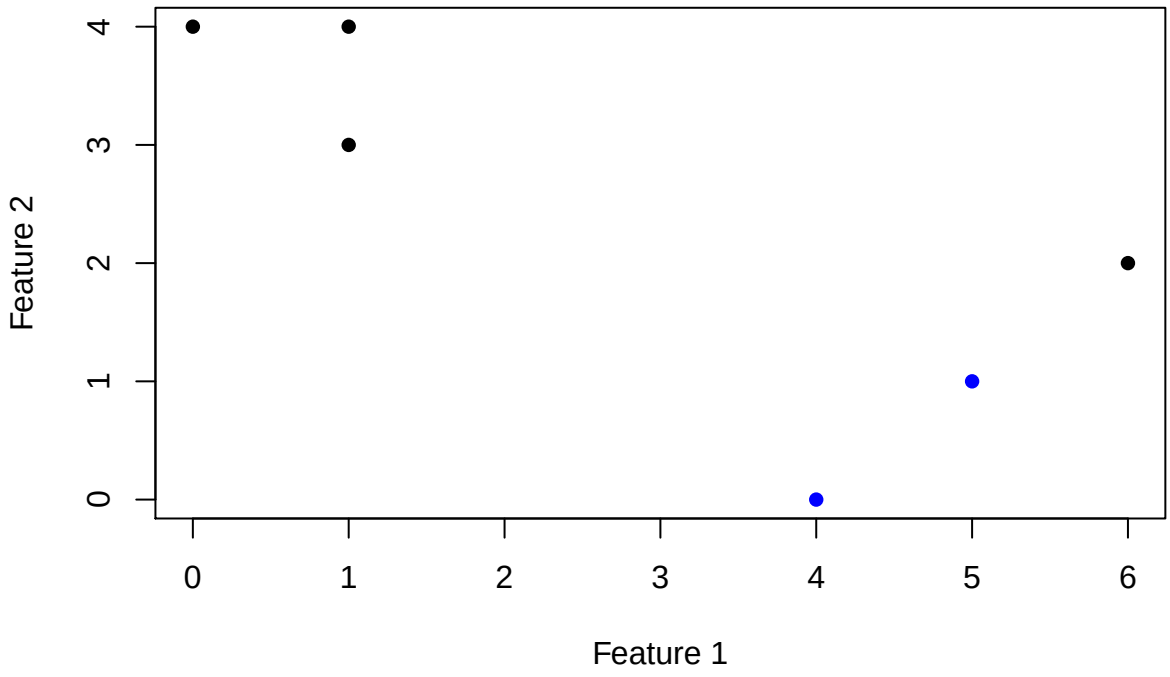
1.

```
x <- cbind(c(1, 1, 0, 5, 6, 4), c(4, 3, 4, 1, 2, 0))
k <- 2
plot(x,xlab="Feature 1",ylab="Feature 2",pch=16)
```



2.

```
set.seed(123)
cluster <- sample(seq(0,1), size=6, replace=TRUE)
xMod <- cbind(x,cluster)
plot(x[,1],x[,2],col=rgb(0,0,cluster),pch=16,xlab="Feature 1",ylab="Feature 2")
```



3.

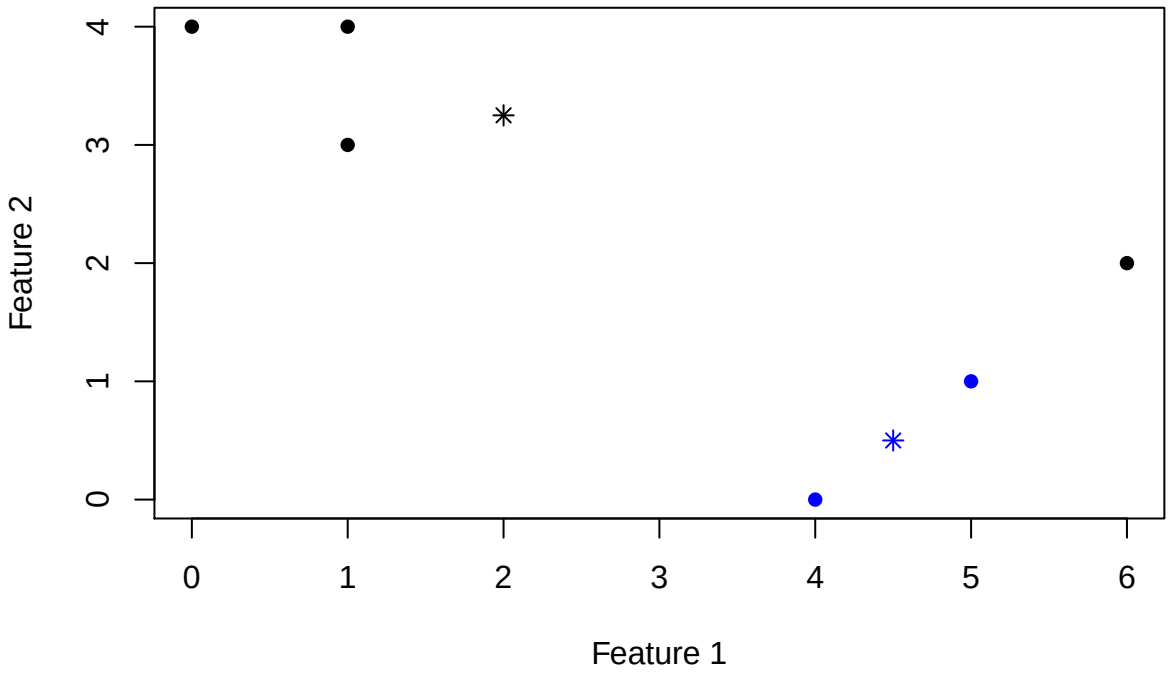
```
x0 <- mean(xMod[,1][xMod[,3]==0])
y0 <- mean(xMod[,2][xMod[,3]==0])

x1 <- mean(xMod[,1][xMod[,3]==1])
y1 <- mean(xMod[,2][xMod[,3]==1])

c(x0,y0) #Centroid for cluster 0
```

```
## [1] 2.00 3.25
```

```
c(x1,y1) #Centroid for cluster 1
```

```
## [1] 4.5 0.5
```

```
plot(x[,1],x[,2],col=rgb(0,0,cluster),pch=16,xlab="Feature 1",ylab="Feature 2")
points(x0,y0,col="black",pch=8)
points(x1,y1,col="blue",pch=8)
```



Note that the black and blue $*$ denote the centroid for the two clusters.

4.

```
dist <- function(x1, x2) sqrt(sum((x1 - x2) ^ 2))
out <- NULL
c0 <- c(x0,y0)
c1 <- c(x1,y1)
for(i in 1:nrow(x)){
    out[i] <- if (dist(x[i,],c0) <= dist(x[i,],c1)) 0 else 1
}
xMod2 <- cbind(x,out)
xMod2
```

```
##          out
## [1,] 1 4   0
## [2,] 1 3   0
## [3,] 0 4   0
## [4,] 5 1   1
## [5,] 6 2   1
## [6,] 4 0   1
```

Thus, we see the fifth observation was relabled from cluster 0 to cluster 1.

5.

```
x0 <- mean(xMod2[,1][xMod2[,3]==0])
y0 <- mean(xMod2[,2][xMod2[,3]==0])

x1 <- mean(xMod2[,1][xMod2[,3]==1])
y1 <- mean(xMod2[,2][xMod2[,3]==1])

out <- NULL
c0 <- c(x0,y0)
c1 <- c(x1,y1)
```

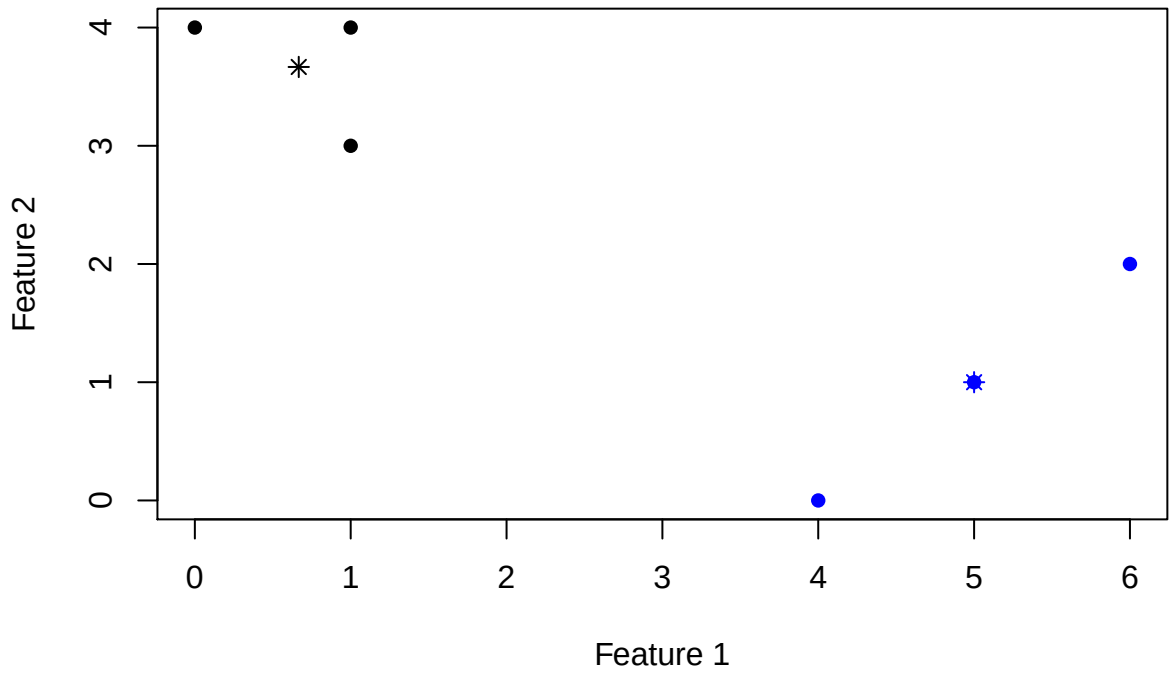
```
for(i in 1:nrow(x)){
    out[i] <- if (dist(x[i,],c0) <= dist(x[i,],c1)) 0 else 1
}
xMod3 <- cbind(x,out)
xMod3
```

```
##          out
## [1,] 1 4   0
## [2,] 1 3   0
## [3,] 0 4   0
## [4,] 5 1   1
## [5,] 6 2   1
## [6,] 4 0   1
```

Thus, we see that the the labels to the clusters stop changing after our first re-labeling (in Part 4).

6.

```
plot(x[,1],x[,2],col=rgb(0,0,out),pch=16,xlab="Feature 1",ylab="Feature 2")
points(x0,y0,col="black",pch=8)
points(x1,y1,col="blue",pch=8)
```



Note: As above, $*$ denotes the cluster centroids.