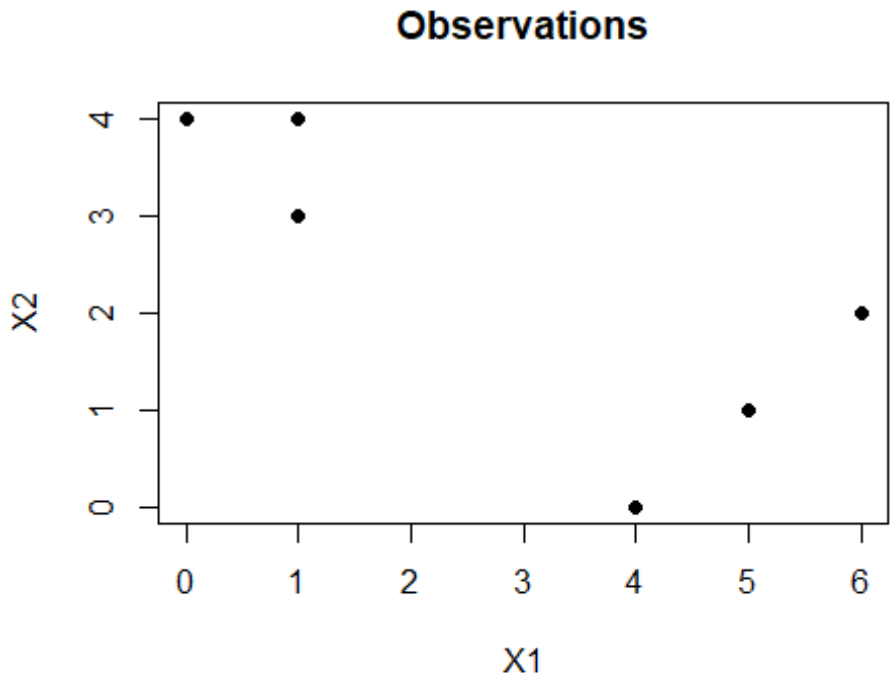
Q42

John Bute

2024-11-29

1. Plot the observations.

```
data <- data.frame(
  X1 = c(1, 1, 0, 5, 6, 4),
  X2 = c(4, 3, 4, 1, 2, 0)
)
rownames(data) <- c(1, 2, 3, 4, 5, 6)
plot(data$X1, data$X2, pch = 19, col = "black", xlab = "X1", ylab = "X2",
main = "Observations")
```



2. Randomly assign a cluster label to each observation. You can use the sample() command in R to do this. Report the cluster labels for each observation.

```
set.seed(5)
data$Cluster <- sample(c(1, 2), size = nrow(data), replace = TRUE)
print(data)
```

```
##   X1 X2 Cluster
## 1  1  4       2
## 2  1  3       1
```

```
## 3  0  4       1
## 4  5  1       1
## 5  6  2       1
## 6  4  0       1
```

There are two clusters, one with only 1 observation and the other with 5. 3. Compute the centroid for each cluster.

```
centroids <- aggregate(cbind(X1, X2) ~ Cluster, data = data, FUN = mean)
print(centroids)
```

```
##   Cluster  X1 X2
## 1       1 3.2  2
## 2       2 1.0  4
```

4. Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

```
reassign_clusters <- function(data, centroids) {
  data$Cluster <- apply(data[, c("X1", "X2")], 1, function(point) {
    dists <- apply(centroids[, -1], 1, function(center) sqrt(sum((point -
center)^2)))
    return(which.min(dists))
  })
  return(data)
}

data <- reassign_clusters(data, centroids)
print(data)
```

```
##   X1 X2 Cluster
## 1  1  4       2
## 2  1  3       2
## 3  0  4       2
## 4  5  1       1
## 5  6  2       1
## 6  4  0       1
```

Our observations have changed, as now three of them report to the centroid of cluster 1, while the others are closer to the centroid of cluster 2.

5. Repeat 3. and 4. until the answers obtained stop changing.

```
repeat {
  old_clusters <- data$Cluster
  centroids <- aggregate(cbind(X1, X2) ~ Cluster, data = data, FUN = mean)
  data <- reassign_clusters(data, centroids)
  if (all(old_clusters == data$Cluster)) break
}
print(data)
```

```
##   X1 X2 Cluster
## 1  1  4       2
```

```
## 2  1  3       2
## 3  0  4       2
## 4  5  1       1
## 5  6  2       1
## 6  4  0       1
```

6. In your plot from 1., color the observations according to the cluster labels obtained.

```
plot(data$X1, data$X2, pch = 19, col = data$Cluster, xlab = "X1", ylab =
"X2", main = "Final Clustering")
text(data$X1, data$X2, labels = rownames(data), pos = 3, cex = 0.8)
legend("topright", legend = c("Cluster 1", "Cluster 2"), col = 1:2, pch = 19)
```