Week 8 Assignment

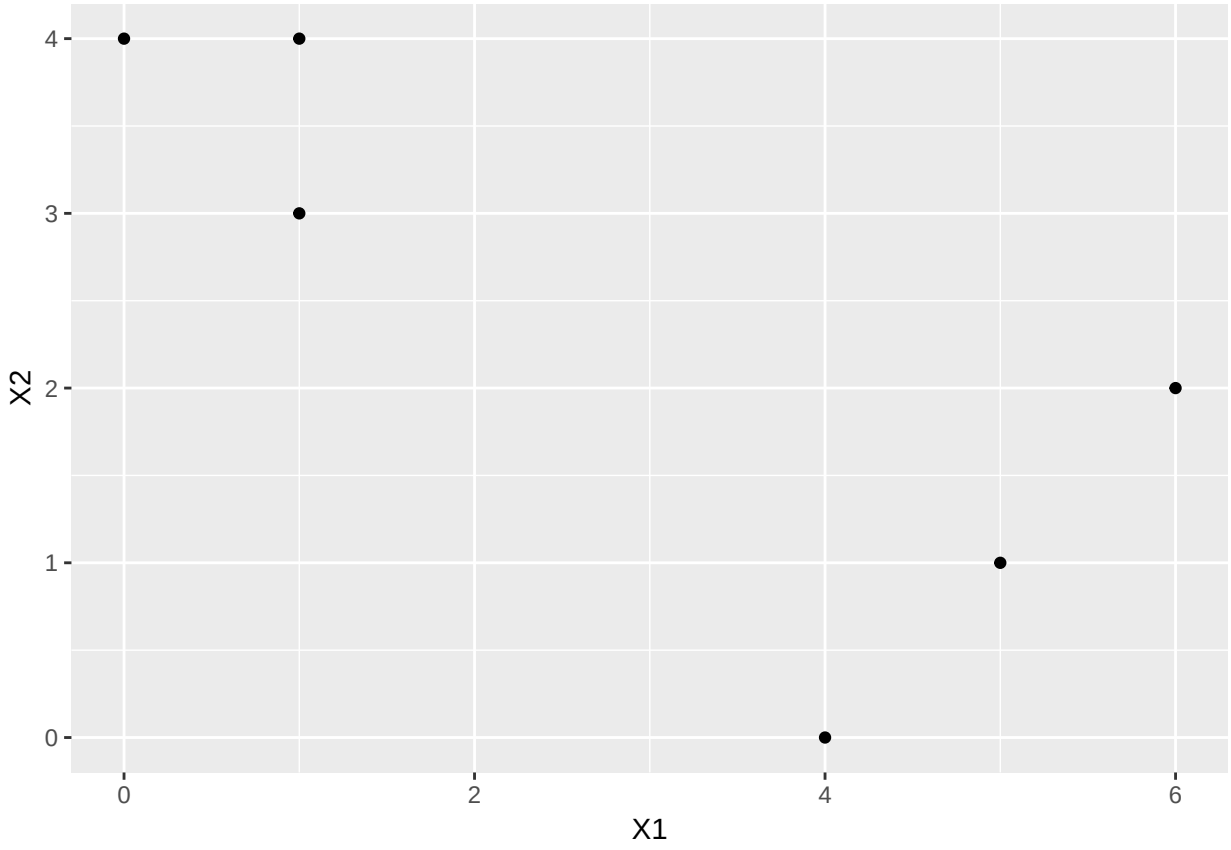
David Russo

3/9/2017

10.7 #3

- a)

```
obs <- data.frame(
  X1 = c(1, 1, 0, 5, 6, 4),
  X2 = c(4, 3, 4, 1, 2, 0)
)

obs %>%
  ggplot(aes(x = X1, y = X2)) +
  geom_point()
```



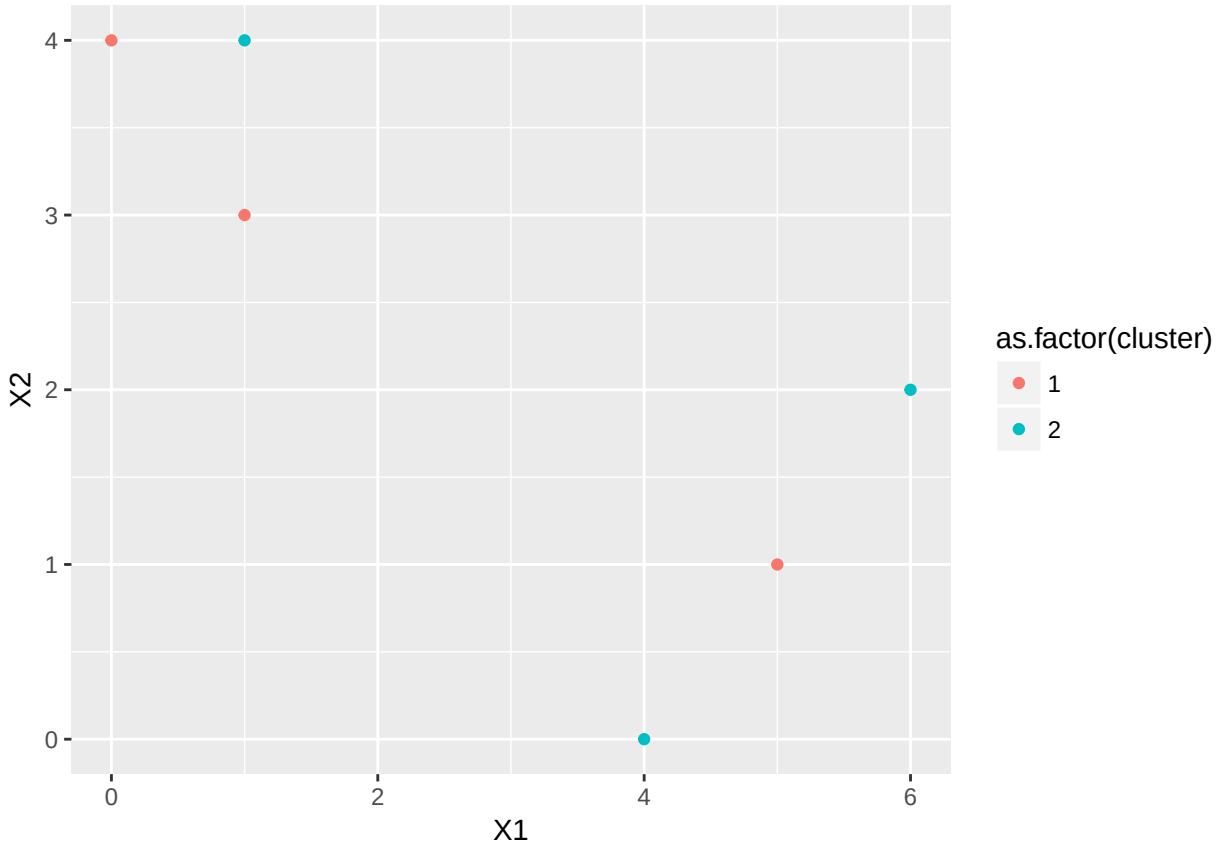
- b)

```
# set random number seed for replicable results
set.seed(2017)

# assign the labels
obs$cluster <- sample(rep(c(1, 2), each = 3), 6, replace = FALSE)
```

```r
# print the labels
obs
```

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

```r
# plot the labels
obs %>%
  dplyr::select(X1, X2, cluster) %>%
  ggplot(aes(x = X1, y = X2, color = as.factor(cluster))) +
  geom_point()
```



- c)

```r
centroids <-
  obs %>%
  dplyr::group_by(cluster) %>%
  dplyr::summarise(centroid_X1 = round(mean(X1), 2), centroid_X2 = round(mean(X2), 2)) %>%
  as.data.frame()

centroids
```

```
##   cluster centroid_X1 centroid_X2
## 1       1        2.00        2.67
```

```
## 2       2        3.67        2.00
```

- d)

```
# create distance from cluster1 centroid
obs$dist_from_cluster1_centroid <- round(
  sqrt((obs$X1 - centroids$centroid_X1[centroids$cluster == 1])^2 +
       (obs$X2 - centroids$centroid_X2[centroids$cluster == 1])^2), 4)

# create distance from cluster2 centroid
obs$dist_from_cluster2_centroid <- round(
  sqrt((obs$X1 - centroids$centroid_X1[centroids$cluster == 2])^2 +
       (obs$X2 - centroids$centroid_X2[centroids$cluster == 2])^2), 4)

obs$new_cluster <-
  ifelse(obs$dist_from_cluster1_centroid <= obs$dist_from_cluster2_centroid, 1, 2)

obs
```

```
##   X1 X2 cluster dist_from_cluster1_centroid dist_from_cluster2_centroid
## 1  1  4       2                     1.6640                     3.3360
## 2  1  3       1                     1.0530                     2.8511
## 3  0  4       1                     2.4019                     4.1796
## 4  5  1       1                     3.4335                     1.6640
## 5  6  2       2                     4.0557                     2.3300
## 6  4  0       2                     3.3360                     2.0270
##   new_cluster
## 1           1
## 2           1
## 3           1
## 4           2
## 5           2
## 6           2
```

- e)

```
cluster_difference <- FALSE

while(cluster_difference == FALSE){

# reset obs$cluster to obs$new_cluster
obs$cluster <- obs$new_cluster

# create distance from cluster1 centroid
obs$dist_from_cluster1_centroid <- round(
  sqrt((obs$X1 - centroids$centroid_X1[centroids$cluster == 1])^2 +
       (obs$X2 - centroids$centroid_X2[centroids$cluster == 1])^2), 4)

# create distance from cluster2 centroid
obs$dist_from_cluster2_centroid <- round(
  sqrt((obs$X1 - centroids$centroid_X1[centroids$cluster == 2])^2 +
       (obs$X2 - centroids$centroid_X2[centroids$cluster == 2])^2), 4)

obs$new_cluster <-
  ifelse(obs$dist_from_cluster1_centroid <= obs$dist_from_cluster2_centroid, 1, 2)
```

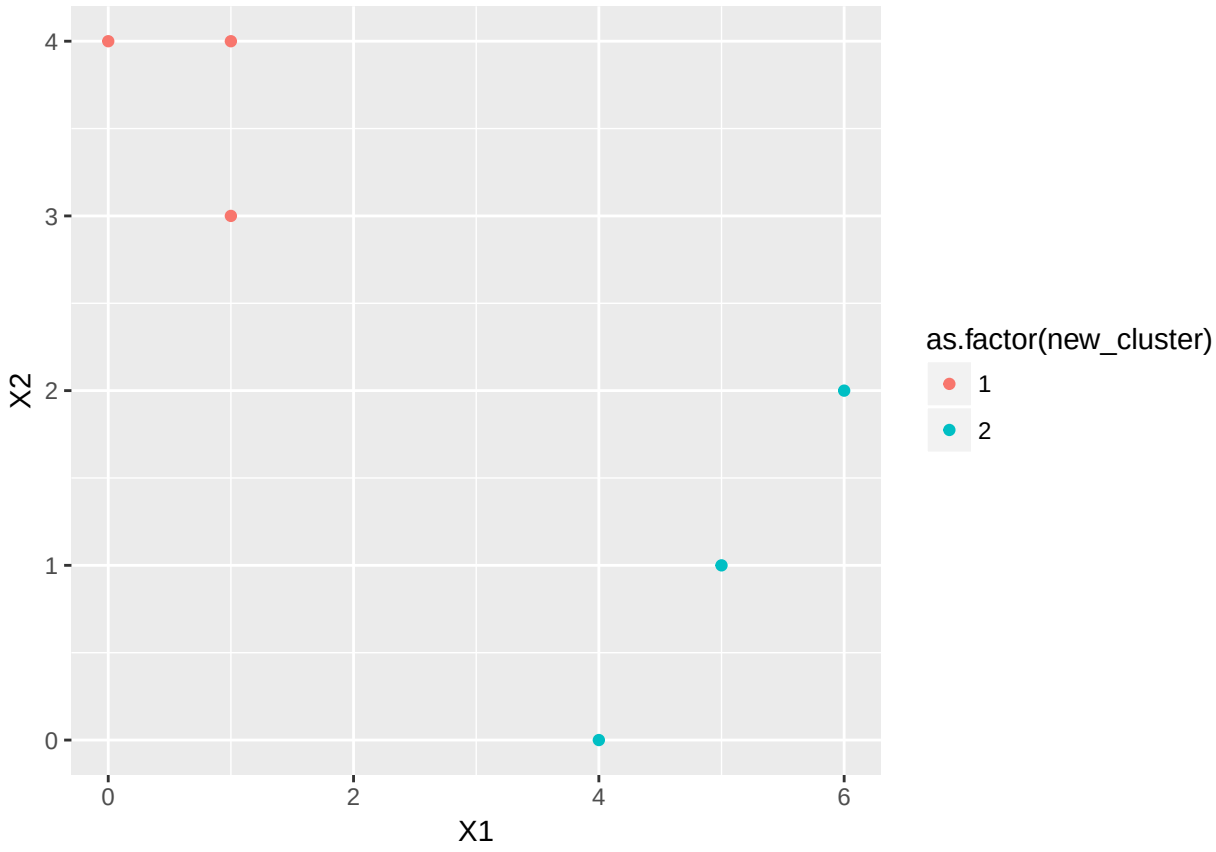
```
cluster_difference <- all(obs$cluster == obs$new_cluster)

print(obs)

}
```

```
##   X1 X2 cluster dist_from_cluster1_centroid dist_from_cluster2_centroid
## 1  1  4       1                      1.6640                      3.3360
## 2  1  3       1                      1.0530                      2.8511
## 3  0  4       1                      2.4019                      4.1796
## 4  5  1       2                      3.4335                      1.6640
## 5  6  2       2                      4.0557                      2.3300
## 6  4  0       2                      3.3360                      2.0270
##   new_cluster
## 1           1
## 2           1
## 3           1
## 4           2
## 5           2
## 6           2
```

The clustering only took one iteration until the algorithm converged.

- f)

```
obs %>%
  dplyr::select(X1, X2, new_cluster) %>%
  ggplot(aes(x = X1, y = X2, color = as.factor(new_cluster))) +
  geom_point()
```


4
3
2
1
0
X2
0
2
4
6
X1
as.factor(new_cluster)
1
2