Hans Trautlein
Due: April 15, 2016

Problem Set 8
Andrew Bray

Chapter 10

1. *This problem involves the K-means clustering algorithm.*

   (a) *Prove the equation below.*

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

$$= \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}((x_{ij}-\bar{x}_{kj})-(x_{i'j}-\bar{x}_{kj}))^2$$

$$= \frac{|C_k|}{|C_k|}\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2+\frac{|C_k|}{|C_k|}\sum_{i'\in C_k}\sum_{j=1}^{p}(x_{i'j}-\bar{x}_{kj})^2-\frac{2}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})(x_{i'j}-\bar{x}_{kj})$$

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

   (b) *On the basis of this identity, argue that the K-means clustering algorithm (Algorithm 10.1) decreases the objective (10.11) at each iteration.*

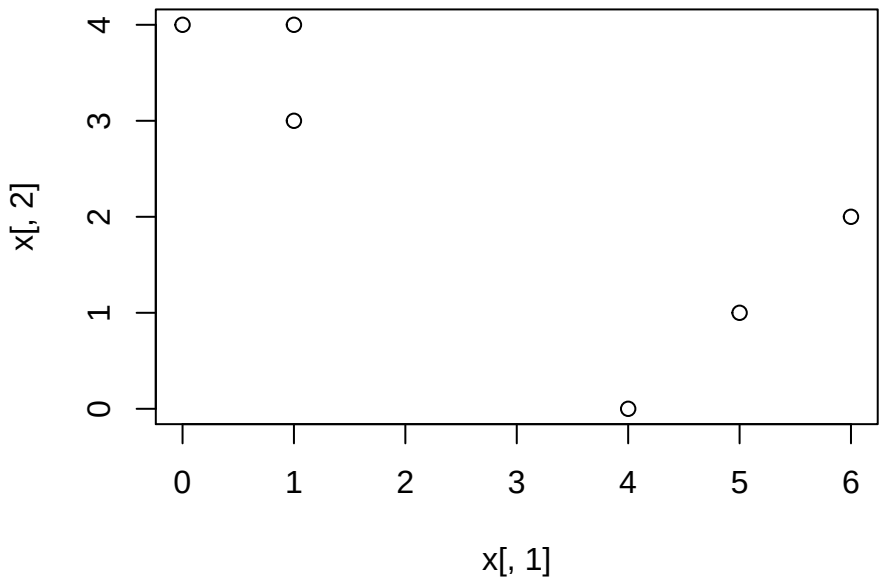
   Minimizing the in-cluster variance across clusters, as seen above, is the same thing as minimizing the Euclidean distance for each cluster.

3. *In this problem, you will perform K-means clustering manually, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ features. The observations are as follows.*

Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

   (a) *Plot the observations.*

```
plot(x[,1], x[,2])
```



(b) *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.*

```
labs = sample(2, nrow(x), replace=TRUE)
labs
```

```
## [1] 1 1 2 2 2 2
```

(c) *Compute the centroid for each cluster.*

```
cent1 = c(mean(x[labs == 1, 1]), mean(x[labs == 1, 2]))
cent2 = c(mean(x[labs == 2, 1]), mean(x[labs == 2, 2]))
cent1
```

```
## [1] 1.0 3.5
```

```
cent2
```

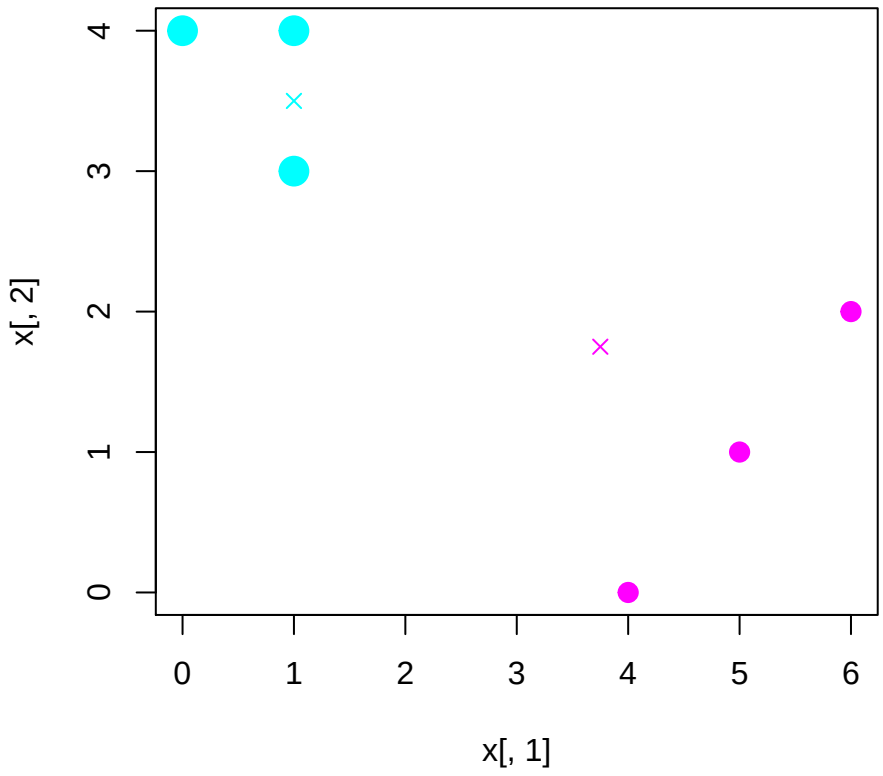
```
## [1] 3.75 1.75
```

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

```
# in code have function that finds labs, echo set to F
labs <- assign_labs(x, cent1, cent2)
labs
```

```
## [1] 1 1 1 2 2 2
```

(e) *In your plot from (a), color the observations according to the cluster labels obtained.*



5. *In words, describe the results that you would expect if you performed K-means clustering of the eight shoppers in Figure 10.14, on the basis of their sock and computer purchases, with $K = 2$. Give three answers, one for each of the variable scalings displayed. Explain.*

   (a) More computers and socks (1, 2, 7, 8) versus least socks and computer (3, 4, 6, 8).

   (b) No computer (1, 2, 3, 4) versus a computer (5, 6, 7, 8). Distance is smaller on socks dimension whereas it is larger on the computer dimension.

   (c) (Similar to above) No computer (1, 2, 3, 4) against a computer (5, 6, 7, 8).