Homework 5

Sean Eva

April 2022

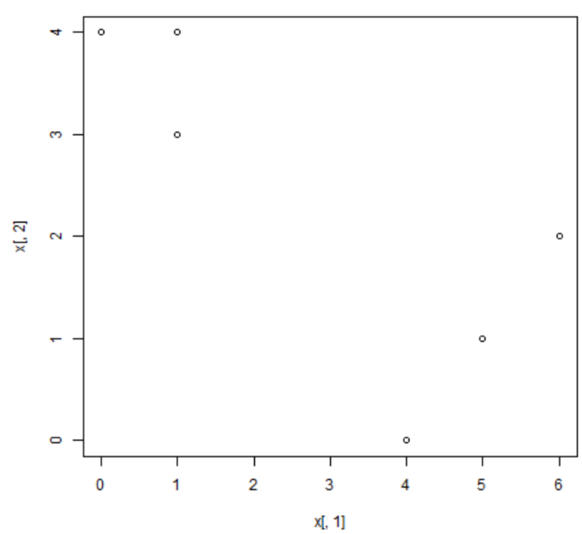
1 Theoretical Problems

12.6: Exercise 1. (a) 12.18 states: $\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$

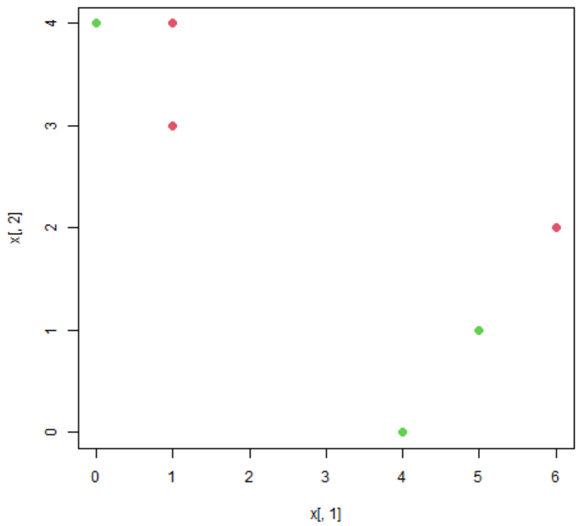
Proof. We may write $\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = \sum_i\sum_j x_{ij}^2 - 2\sum_i\sum_j x_{ij}\bar{x}_k j + \sum_i\sum_j x_{ij}^2 = 2\sum_i\sum_j x_{ij}^2 - 2|C_k|\sum_j \bar{x}_{kj}^2$ which implies that $2\sum_i\sum_j x_{ij}^2 - 4\sum_i\sum_j x_{ij}\bar{x}_{kj} + 2\sum_i\sum_j \bar{x}_{kj}^2 = 2\sum_i\sum_j x_{ij}^2 - 2|C_k|\sum_j \bar{x}_{kj}^2$. Which therefore proves 12.18 as desired. □

(b) This identity shows us that when we assign each observation to the cluster whose centroid is closest, we actually decrease the right member of the identity. So, we also will decrease the left member of the identity which is our objective. Another way of seeing this is that the identity shows that minimizes the sum of the squared Euclidean distance for each cluster is the same as minimizing the within-cluster variance for each cluster.

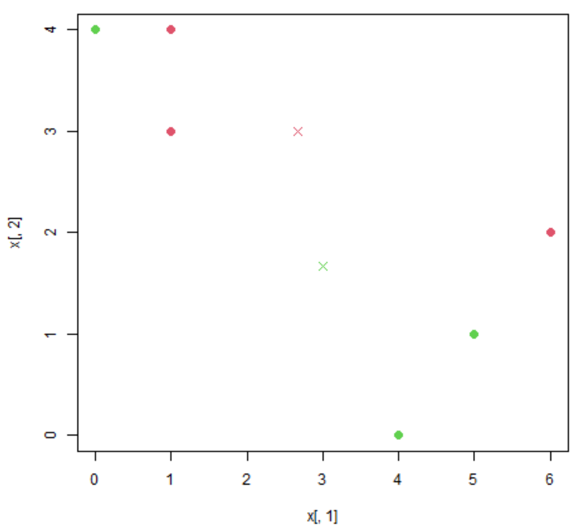
12.6: Exercise 3. (a) Graph with points:



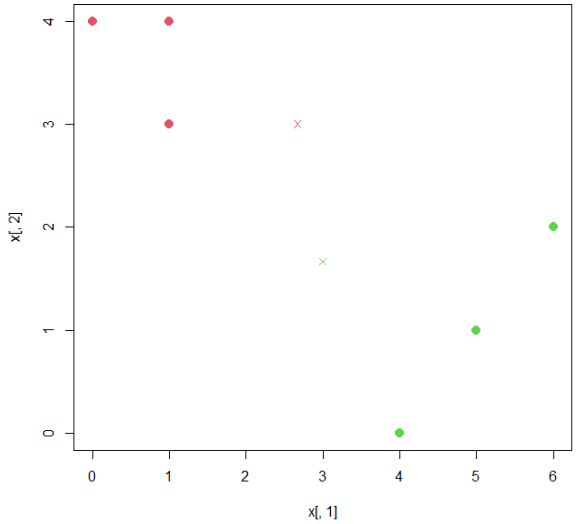
(b) Cluster 1: Points 1, 2, 5. Cluster 2: 3, 4, 6.



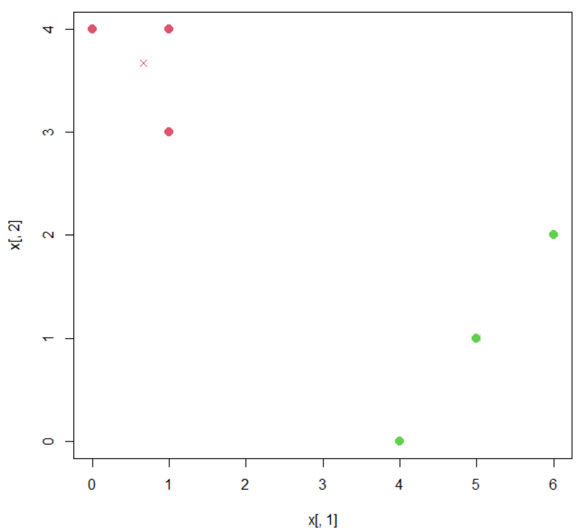
(c) Calculation:
$\bar{x}_{11} = \frac{1}{3}(1+1+6) = \frac{8}{3}, \bar{x}_{12} = \frac{1}{3}(2+4+3) = 3.$
$\bar{x0}_{21} = \frac{1}{3}(0+4+5) = 3, \bar{x}_{22} = \frac{1}{3}(4+0+1) = \frac{5}{3}$
On graph marked with x's.



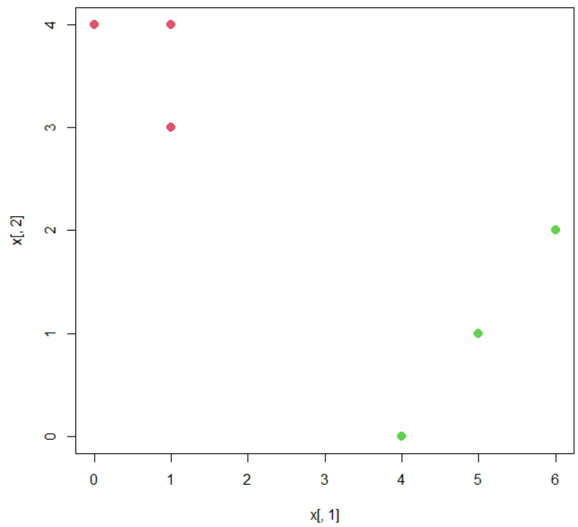
(d) Cluster 1: Points 1, 2, 3. Cluster 2: 4, 5, 6. New graph with reassigned locations.



(e) New Calculations:
$\bar{x}_{11} = \frac{1}{3}(1+1+0) = \frac{2}{3}, \bar{x}_{12} = \frac{1}{3}(4+3+4) = \frac{11}{3}$
$\bar{x}_{21} = \frac{1}{3}(5+6+4) = 5, \bar{x}_{22} = \frac{1}{3}(1+2+0) = 1$
If we reassign and perform again we will arrive to the same centroids and groupings again so the process will end here. New graph with new centroid locations:



(f) Graph with final labels.



n-complete graph: 1. $\begin{bmatrix} n-1 & -1 & -1 & \cdots & -1 \\ -1 & n-1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \ddots & -1 \\ -1 & -1 & -1 & \cdots & n-1 \end{bmatrix}$

2. The eigenvalues are $\lambda = 0, n$ where n has a geometric multiplicity of $n-1$. The eigenvector corresponding to $\lambda = 0$ is $\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ and the eigenvectors corresponding to $\lambda = n$ are $\begin{bmatrix} -1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, ..., \begin{bmatrix} -1 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$.

2 Programming

1. All in coding file.

2. Sorry I ran out of time on this question.