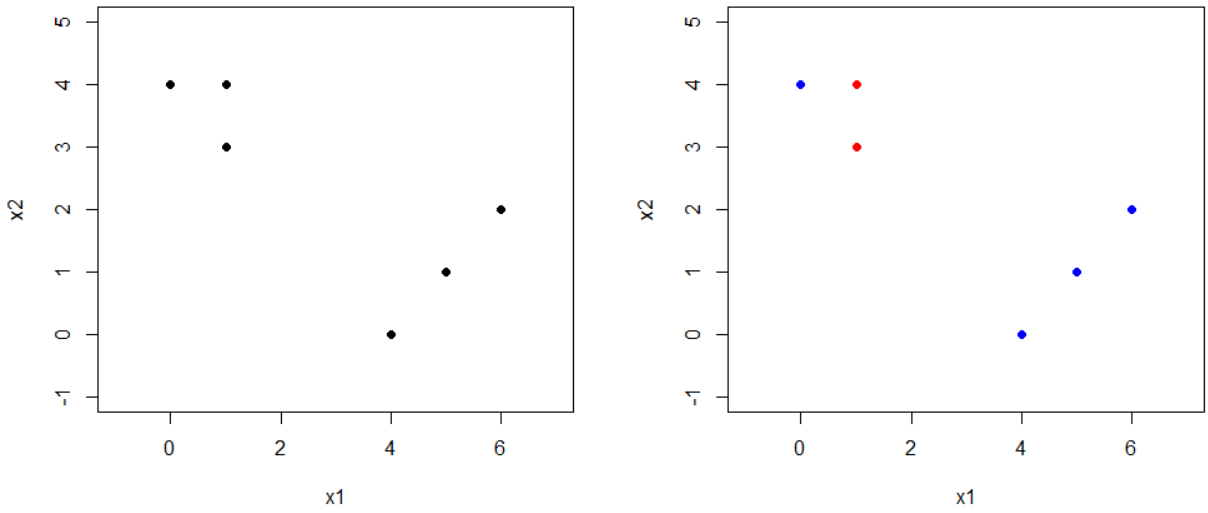
ESI 4606: Engineering Analytics I
Homework 7 Solution

Problem 1 (1.5 points)
Solution: (a) Recall root node, there is "9Y5N", thus,
$\text{Entropy}_{\text{R}} = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) \approx 0.94.$
If splitting variable is "Temp", it yields "Temp=Hot: 2Y2N", "Temp=Mild: 4Y2N" and "Temp=Cool: 3Y1N". The corresponding entropies are given by
$\text{Entropy}_{\text{Hot}} = -\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4}) = 1.$
$\text{Entropy}_{\text{Mild}} = -\frac{4}{6}\log_2(\frac{4}{6}) - \frac{2}{6}\log_2(\frac{2}{6}) \approx 0.92$
$\text{Entropy}_{\text{Cool}} = -\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2(\frac{1}{4}) \approx 0.81$
Gain(R,"Temp")$=0.94 - (\frac{4}{14}\times 1 + \frac{6}{14}\times 0.92 + \frac{4}{14}\times 0.81) \approx 0.03$
(b) If splitting variable is "Hum", it yields "Hum=High: 3Y4N" and "Hum=Normal: 6Y1N". The corresponding entropies are given by
$\text{Entropy}_{\text{High}} = -\frac{3}{7}\log_2(\frac{3}{7}) - \frac{4}{7}\log_2(\frac{4}{7}) \approx 0.99.$
$\text{Entropy}_{\text{Normal}} = -\frac{6}{7}\log_2(\frac{6}{7}) - \frac{1}{7}\log_2(\frac{1}{7}) \approx 0.59$
Gain(R,"Hum")$=0.94 - (\frac{7}{14}\times 0.99 + \frac{7}{14}\times 0.59) \approx 0.15$
(c) If splitting variable is "Wind", it yields "Wind=Weak: 6Y2N" and "Wind=Strong: 3Y3N". The corresponding entropies are given by
$\text{Entropy}_{\text{Weak}} = -\frac{6}{8}\log_2(\frac{6}{8}) - \frac{2}{8}\log_2(\frac{2}{8}) \approx 0.81.$
$\text{Entropy}_{\text{Strong}} = -\frac{3}{6}\log_2(\frac{3}{6}) - \frac{3}{6}\log_2(\frac{3}{6}) = 1$
Gain(R,"Wind")$=0.94 - (\frac{8}{14}\times 0.81 + \frac{6}{14}\times 1) \approx 0.05$

Problem 2 (1.5 points)
Solution: (a) Figure 1a shows plot of observations.

(b) Figure 1b shows plot of observations with initialization. Observations labeled with Cluster 1 are colored in red and observations labeled with Cluster 2 are colored in blue.

(c-e): Results of iterations 1 and 2 are shown in Tables 1 and **??**. Since reassigned clusters remain unchanged at iteration 2, the algorithm stops.



(a) Plot for problem 1(a)

(b) Plot for problem 1(b)

Figure 1: Plots for observations and clustering initialization

Table 1: K-means results at the iteration 1

Obs.	X_1	X_2	Initialized Cluster No.	Centroid	Distance to (1,3.5)	Distance to (3.75,1.75)	Reassigned Cluster No.
1	1	4	1	(1,3.5)	0.50	3.55	1
2	1	3	1	(1,3.5)	0.50	3.02	1
3	0	4	2	(3.75,1.75)	1.12	4.37	1
4	5	1	2	(3.75,1.75)	4.72	1.46	2
5	6	2	2	(3.75,1.75)	5.22	2.26	2
6	4	0	2	(3.75,1.75)	4.61	1.77	2

(f) Figure 2 shows plot of observations for the final clustering results. Observations labeled with Cluster 1 are colored in red and observations labeled with Cluster 2 are colored in blue.

Problem 3 (2 points)

Solution: (a) Using hierarchical clustering with complete linkage, calculation results are shown in Table 2 and the corresponding dendrogram is shown in Figure 3.

(b) Using hierarchical clustering with single linkage, calculation results are shown in Table 3 and the corresponding dendrogram is shown in Figure 4.

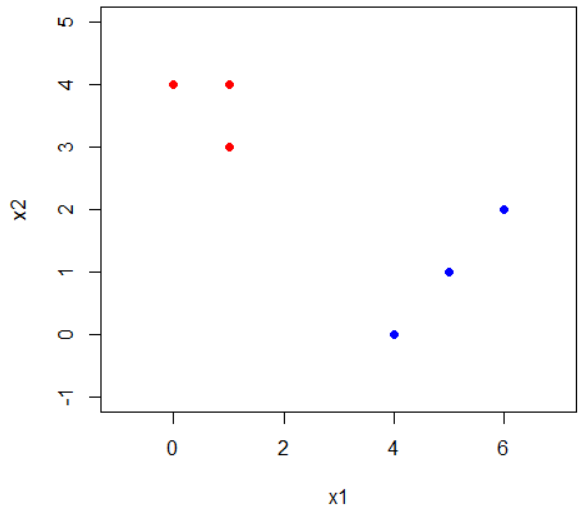


Figure 2: Plot for problem 1 (f)

Table 2: Stepwise results of hierarchical clustering using complete linkage

Obs.	1	2	3	4
1	0	**0.3**	0.4	0.7
2		0	0.5	0.8
3			0	0.45
4				0

$\rightarrow$

Obs.	{1,2}	3	4
{1,2}	0	0.5	0.8
3		0	**0.45**
4			0

$\rightarrow$

Obs.	{1,2}	{3,4}
{1,2}	0	**0.8**
{3,4}		0

$\rightarrow$

Obs.	{1,2,3,4}
{1,2,3,4}	0

Table 3: Stepwise results of hierarchical clustering using single linkage

Obs.	1	2	3	4
1	0	**0.3**	0.4	0.7
2		0	0.5	0.8
3			0	0.45
4				0

$\rightarrow$

Obs.	{1,2}	3	4
{1,2}	0	**0.4**	0.7
3		0	0.45
4			0

$\rightarrow$

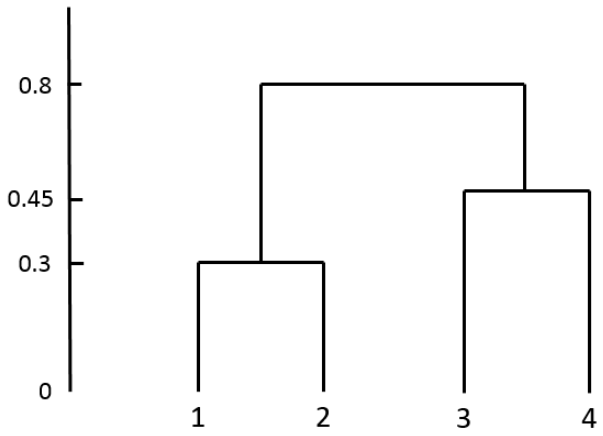


Figure 3: Dendrogram of hierarchical clustering using complete linkage

Obs.	{1,2,3}	4
{1,2,3}	0	**0.45**
4		0

$\rightarrow$

Obs.	{1,2,3,4}
{1,2,3,4}	0

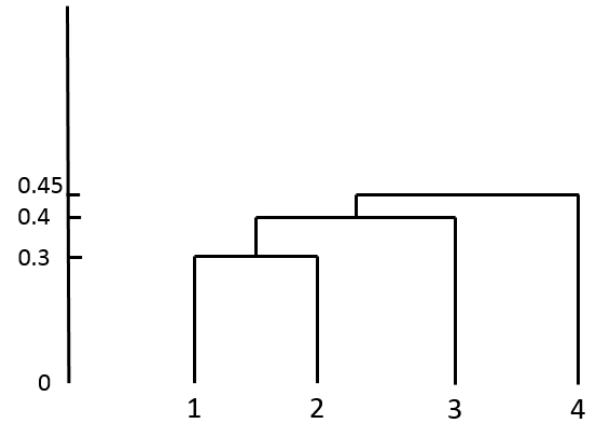


Figure 4: Dendrogram of hierarchical clustering using single linkage

(c) If we cut the dendrogram obtained in Figure 3 such that two clusters result, the clusters are: cluster 1 with observations 1,2 and cluster 2 with observations 3,4.

(d) If we cut the dendrogram obtained in Figure 4 such that two clusters result, the clusters are: cluster 1 with observations 1,2,3 and cluster 2 with observations 4.