COM S 474/574 Spring 2020
Introduction to Machine Learning

Homework 6 Solutions – Problem 1

1 Problems

Problem 1. [25 points] Ch 10 # 2 "Suppose that we have four..."

Show your calculations. Recall that "single linkage" means we define the distance of two clusters C_1 and C_2 as the minimum distance of any pair of elements

$$dist(C_1, C_2) := \min_{a \in C_1, b \in C_2} dist(a, b)$$

and that "complete linkage" means we define the distance of two clusters C_1 and C_2 as the maximum distance of any pair of elements

$$dist(C_1, C_2) := \max_{a \in C_1, b \in C_2} dist(a, b).$$

(the following is written more thoroughly than was needed for credit, to make the process clear)

We begin with four clusters, each with a single sample, {1}, {2}, {3}, and {4}. We do not know the actual feature values of these samples (eg, we cannot make a scatter plot of them). We only know their distances between each other. But hierarchical clustering with single and complete linkage only need the distance matrix.

	{1}	{2}	{3}	{4}
{1}		0.3	0.4	0.7
{2}	0.3		0.5	0.8
{3}	0.4	0.5		0.45
{4}	0.7	0.8	0.45	

a. Starting from threshold of 0 and increasing, the first merger is at a distance threshold of 0.3 at which clusters {1} and {2} merge. Now we have clusters {1, 2}, {3}, and {4}. We should rewrite the distance matrix for these three clusters. The distance between {3} and {4} is still 0.45. But what about the others? Since we are using complete linkage,

$$\begin{aligned} dist(\{1,2\}, \{3\}) &= \max\{dist(1,3),\ dist(2,3)\} \\ &= \max\{0.4,\ 0.5\} \\ &= 0.5 \end{aligned}$$

Likewise,

$$\begin{aligned} dist(\{1,2\},\{4\}) &= \max\{dist(1,4),\ dist(2,4)\} \\ &= \max\{0.7,\ 0.8\} \\ &= 0.8 \end{aligned}$$

We can now write down the new distance matrix

$$\begin{array}{c} \\ \{1,2\} \\ \{3\} \\ \{4\} \end{array} \begin{array}{c} \begin{array}{ccc} \{1,2\} & \{3\} & \{4\} \end{array} \\ \begin{bmatrix} & 0.5 & 0.8 \\ 0.5 & & 0.45 \\ 0.8 & 0.45 & \end{bmatrix} \end{array}$$

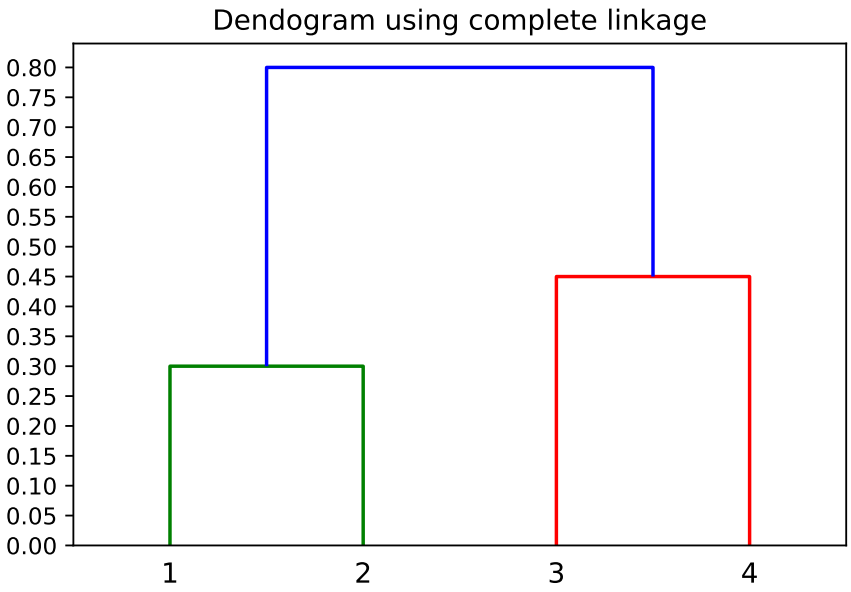
and we repeat that process. The next merger occurs at a threshold of 0.45, where clusters $\{3\}$ and $\{4\}$ merge. We need to re-calculate cluster distances,

$$\begin{aligned} dist(\{1,2\},\{3,4\}) &= \max\{dist(1,3),\ dist(2,3),\ dist(1,4),\ dist(2,4)\} \\ &= \max\{0.4,\ 0.5,\ 0.7,\ 0.8\} \\ &= 0.8 \end{aligned}$$

The new distance matrix is

$$\begin{array}{c} \\ \{1,2\} \\ \{3,4\} \end{array} \begin{array}{c} \begin{array}{cc} \{1,2\} & \{3,4\} \end{array} \\ \begin{bmatrix} & 0.8 \\ 0.8 & \end{bmatrix} \end{array}$$

and the next (final) merger is at a threshold of 0.8. The dendogram is shown below.



b. With single linkage, the procedure is almost identical except for how cluster distances are computed.
Starting from threshold of 0 and increasing, the first merger is at a distance threshold of 0.3 at which clusters $\{1\}$ and $\{2\}$ merge. Now we have clusters $\{1,2\}$, $\{3\}$, and $\{4\}$. We should rewrite the distance matrix for these three clusters. The distance between $\{3\}$ and $\{4\}$ is still 0.45. But what about the others?

$$\begin{aligned} dist(\{1,2\},\{3\}) &= \min\{dist(1,3),\ dist(2,3)\} \\ &= \min\{0.4,\ 0.5\} \\ &= 0.4 \end{aligned}$$

Likewise,

$$\begin{aligned} dist(\{1,2\},\{4\}) &= \min\{dist(1,4),\ dist(2,4)\} \\ &= \min\{0.7,\ 0.8\} \\ &= 0.7 \end{aligned}$$

We can now write down the new distance matrix

$$\begin{array}{c} \\ \{1,2\} \\ \{3\} \\ \{4\} \end{array} \begin{array}{c} \begin{array}{ccc} \{1,2\} & \{3\} & \{4\} \end{array} \\ \begin{bmatrix} & 0.4 & 0.7 \\ 0.4 & & 0.45 \\ 0.7 & 0.45 & \end{bmatrix} \end{array}$$

and we repeat that process. The next merger occurs at a threshold of 0.4, where clusters

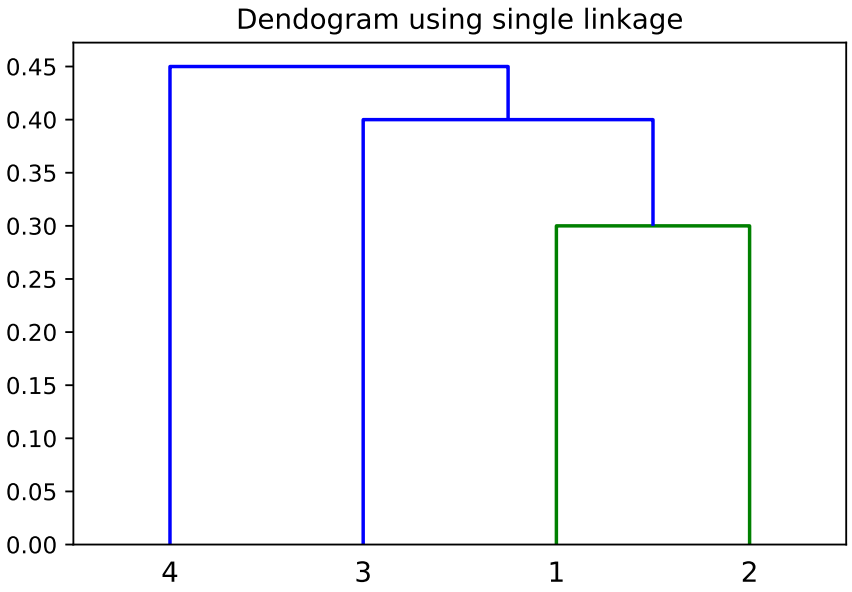
$\{1,2\}$ and $\{3\}$ merge. We need to re-calculate cluster distances,

$$\begin{aligned} dist(\{1,2,3\},\{4\}) &= \min\{dist(1,4),\ dist(2,4),\ dist(3,4)\} \\ &= \max\{0.7,\ 0.8,\ 0.45\} \\ &= 0.45 \end{aligned}$$

The new distance matrix is

$$\begin{array}{c} \\ \{1,2,3\} \\ \{4\} \end{array} \begin{array}{c} \begin{array}{cc} \{1,2,3\} & \{4\} \end{array} \\ \begin{bmatrix} & 0.45 \\ 0.45 & \end{bmatrix} \end{array}$$

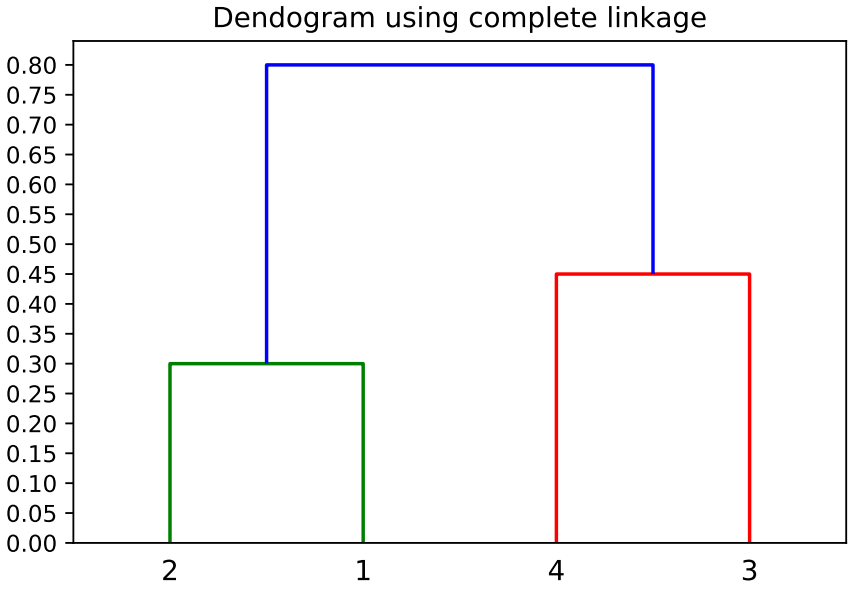
and the next (final) merger is at a threshold of 0.45. The dendogram is shown below.



c. If we cut the dendogram in (a) as described, the two clusters would be $\{1,2\}$ and $\{3,4\}$.

d. If we cut the dendogram in (b) as described, the two clusters would be $\{1,2,3\}$ and $\{4\}$.

e. The ordering of samples along the horizontal axis of the dendogram has no special meaning – they are ordered to make the diagrams look visually pleasing (namely, few intersecting lines).
Below is one example where we changed the label orderings (though the colored lines stay the same). You can order the nodes any way you want and it would not change the meaning, though if you did 1, then 3, then 2, then 4, some of the lines would cross which without color coding or some halfcircles at the intersection could be confusing.


Dendogram using complete linkage
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2
1
4
3