AMAT583 (843X) Midterm practice

Problem 1. What is the edit distance between

(a) AA and BBB?

(b) $SVANE$ and $SAVNE$?

Explain your answers.

Solution:

(a) 3. An optimal edit sequence (out of several) is $AA \to AB \to BB \to BBB$.

(b) 2. An optimal edit sequence (out of several) is $SVANE \to SANE \to SAVNE$.

Problem 2. Compute the Wasserstein distance between $(5, 4, 2)$ and $(3, 3, 5)$. (You can also view these as functions $f, g\colon \{1, 2, 3\} \to [0, \infty)$ given by $f(1) = 4$, $f(2) = 4$, $f(3) = 2$ and $g(1) = 3, g(2) = 3, g(3) = 4$.) Write a transport plan realizing the this distance on matrix form.

Solution: An optimal transport plan is given by the matrix $\begin{bmatrix} 3 & 0 & 2 \\ 0 & 3 & 1 \\ 0 & 0 & 2 \end{bmatrix}$.[1] (This is not the only solution; for instance $\begin{bmatrix} 3 & 2 & 0 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix}$ is another.) The cost of the plan is determined by the nonzero off-diagonal elements: $2 \cdot 2 + 1 \cdot 1 = 5$.

Problem 3. Let $X = \{0, 2, 3, 4\} \subseteq \mathbb{R}$. Find the 2-means clustering of X. Explain your answer.

Solution: There are two realistic candidates: $\{\{0\}, \{2, 3, 4\}\}$ and $\{\{0, 2\}, \{3, 4\}\}$. The means of the former are 0 and 3, which gives a cost of

$$(0-0)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 = 0 + 1 + 0 + 1 = 2.$$

The means of the latter are 1 and 3.5, which gives a cost of

$$(0-1)^2 + (2-1)^2 + (3-3.5)^2 + (4-3.5)^2 > 1 + 1 = 2.$$

Thus, $\{\{0\}, \{2, 3, 4\}\}$ has the lowest cost, so it is the 2-means clustering of X.

Problem 4. Let $X = \{0, 1, 6, 7, 9.5\}$ be equipped with the standard metric.

(a) Find the single linkage dendrogram of X.

(b) Find the average linkage dendrogram of X.

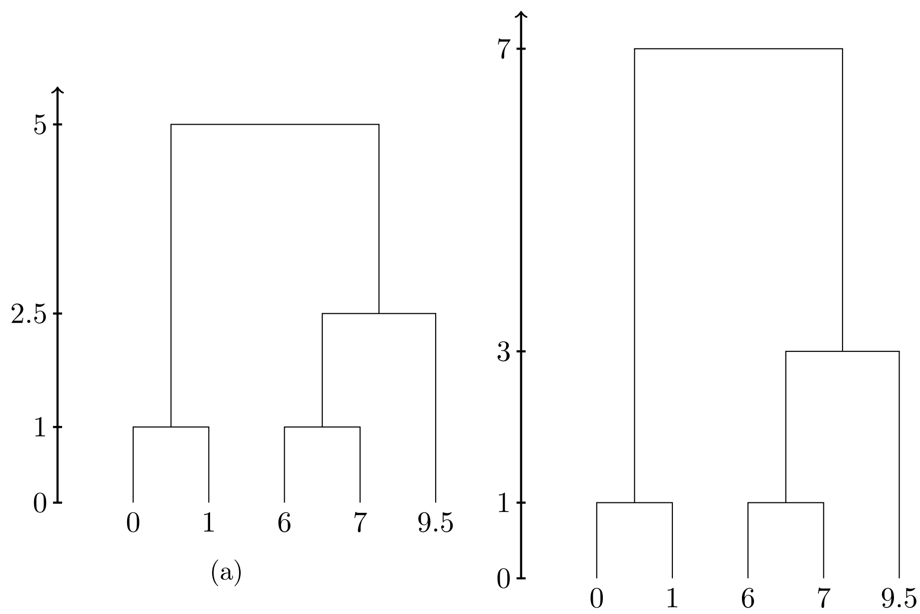
Solution: (a) At height 1, 0 and 1 merge, and so do 6 and 7. Next, $\{6, 7\}$ and $\{9.5\}$ merge at 2.5, and finally, $\{0, 1\}$ and $\{6, 7, 9.5\}$ merge at 5. See Fig. 1a.

(b) At height 1, 0 and 1 merge, and so do 6 and 7. The next two clusters to merge are $\{6, 7\}$ and $\{9.5\}$. The merging point is $\frac{1}{2}(3.5 + 2.5) = 3$. Finally, $\{0, 1\}$ and $\{6, 7, 9.5\}$ merge, which happens at $\frac{1}{2 \cdot 3}(6 + 5 + 7 + 6 + 9.5 + 8.5) = 7$. See Fig. 1b.

Problem 5. Draw a dendrogram that has the barcode $\{[1, \infty), [1, 3), [1, 3), [1, 3), [2, 4)\}$.

Solution: See Fig. 2. (It is important to not glue $[1, 3)$ to $[2, 4)$, as this would give intervals $[1, 4)$ and $[2, 3)$ in the barcode instead of $[1, 3)$ and $[2, 4)$.)

[1] This can be thought of as "move two blocks from position 1 to position 3, and one block from position 2 to position 3".



(a)

(b)

Figure 1: Solution to Problem 4.

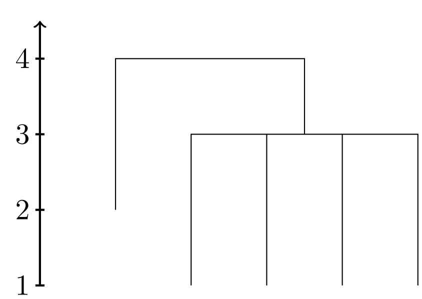


Figure 2: Solution to Problem 5.