AMAT583 (8432) Midterm II

Name .... Score: .../18

Problem 1. (3 points) What is the edit distance between

(a) XYZ and $AAAA$?

(b) ACB and $ZABZ$?

(c) $AABB$ and BA?

Explain your answers.

Solution:

(a) 4. An optimal edit sequence (out of several) is $XYZ \to XYZA \to XYAA \to XAAA \to AAAA$.

(b) 3. An optimal edit sequence (out of several) is $ACB \to AB \to ZAB \to ZABZ$.

(c) 3. An optimal edit sequence (out of several) is $AABB \to ABB \to BB \to BA$.

Problem 2. (3 points) Find the 2-means clustering of $\{0, 2, 3, 4, 5\}$. Justify your answer.

Solution: There are three realistic candidates: $C_1 = \{\{0\}, \{2, 3, 4, 5\}\}$, $C_2 = \{\{0, 2\}, \{3, 4, 5\}\}$ and $C_3 = \{\{0, 2, 3\}, \{4, 5\}\}$. ($\{\{0, 2, 3, 4\}, \{5\}\}$ does worse than C_1, since $\{0, 2, 3, 4\}$ has a higher cost than $\{1, 2, 3, 4\}$, which has the same cost as $\{2, 3, 4, 5\}$.) The means of C_1 are 0 and 3.5, which gives a cost of

$$(0-0)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 = 2 \cdot 1.5^2 + 2 \cdot 0.5^2 = 5.$$

The means of C_2 are 1 and 4, which gives a cost of

$$(0-1)^2 + (2-1)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 = 4.$$

The means of C_3 are $\frac{5}{3}$ and 4.5, which gives a cost of

$$(0-\frac{5}{3})^2 + (2-\frac{5}{3})^2 + (3-\frac{5}{3})^2 + (4-4.5)^2 + (5-4.5)^2 > (\frac{5}{3})^2 + (\frac{4}{3})^2 = \frac{41}{9} > 4.$$

Thus, C_2 has the lowest cost, so it is the 2-means clustering of X.

Problem 3. (4 points)

(a) Compute the Wasserstein distance between $(3, 1, 1)$ and $(1, 0, 4)$. (You can also view these as functions $f, g \colon \{1, 2, 3\} \to [0, \infty)$ given by $f(1) = 3, f(2) = f(3) = 1$ and $g(1) = 1, g(2) = 0, g(3) = 4$.)

(b) Write a transport plan realizing the distance in (a) on matrix form (if you have not already done so).

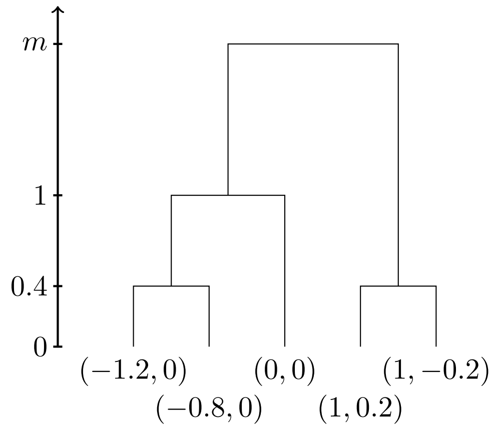


Figure 1: Solution to Problem 4.

Solution: (a) It is possible to go directly to the matrix form in (b). Another solution is the following: Move 2 elements from position 1 to position 3, and 1 element from position 2 to position 3. The cost of this is $2 \cdot 2 + 1 \cdot 1 = 5$, since we are moving two elements a distance of two, and one element a distance of one. There is no cheaper plan that lets us put 4 elements in total in position 3.

(b) Writing the transport plan from (a) on matrix form, we get $\begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$.

Problem 4. (3 points) Find the average linkage dendrogram of

$$X = \{(-1.2, 0), (-0.8, 0), (0, 0), (1, 0.2), (1, -0.2)\}$$

equipped with the Euclidean metric. You do not have to find the value at which the last two clusters merge; just label this merging point with "m".

Solution: The closest pairs of points are $\{(-1.2, 0), (-0.8, 0)\}$ and $\{(1, 0.2), (1, -0.2)\}$, and these clusters appear at 0.4. This gives $\{\{(-1.2, 0), (-0.8, 0)\}, \{(0, 0)\}, \{(1, 0.2), (1, -0.2)\}\}$, and we need to decide which cluster to merge with $\{(0, 0)\}$. We have

$$\delta(\{(-1.2, 0), (-0.8, 0)\}, \{(0, 0)\}) = \frac{1}{2}(1.2 + 0.8) = 1,$$

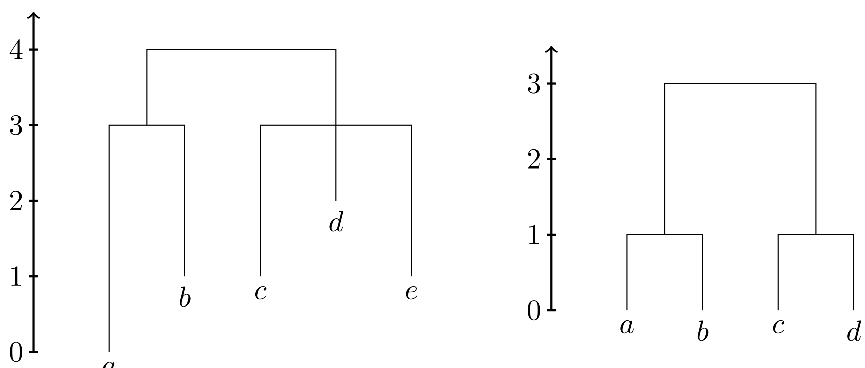
while $(1, 0.2)$ and $(1, -0.2)$ both have distance more than 1 to $(0, 0)$. Thus, $\{(-1.2, 0), (-0.8, 0)\}$ and $\{(0, 0)\}$ are next to merge at 1. Finally, the last two clusters merge at some point m. See Fig. 1.

Problem 5. (2 points) What is the barcode of the dendrogram in Fig. 2a?

Solution: The barcode is $\{[0, \infty), [1, 4), [1, 3), [1, 3), [2, 3)\}$.

Problem 6. (3 points) Find a set $X = \{a, b, c, d\} \subseteq \mathbb{R}$ equipped with the standard metric whose single linkage clustering dendrogram is the one in Fig. 2b.

Solution: There are many solutions. One is $a = 0$, $b = 1$, $c = 4$, $d = 5$, so $X = \{0, 1, 4, 5\}$.



(a) (b)

Figure 2