AMAT583 (8433) Midterm II

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

Problem 1. (3 points) Over the alphabet $\{a, b\}$, find all the words with edit distance exactly 1 from the word aa.

Solution: $a, ab, ba, aaa, baa, aba, aab$.

Problem 2. (3 points) Compute the Wasserstein distance between $(1, 0, 2, 0, 1)$ and $(2, 0, 0, 0, 2)$. (You can also view these as functions $f, g\colon \{1, 2, 3, 4, 5\} \to [0, \infty)$ given by $f(1) = 1, f(2) = 0, f(3) = 2, f(4) = 0, f(5) = 1$ and $g(1) = 2, g(2) = g(3) = g(4) = 0, g(5) = 2$.)

Solution: We send 1 from the third position to the first position, and 1 from the third position to the fifth position. We are twice moving a weight of 1 a distance of 2, which gives a cost of $2 \cdot 1 \cdot 2 = 4$. This is the cheapest transport plan, so the Wasserstein distance is 4.

(We can also write the transport plan on matrix form: $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$ We have two 1s with distance 2 two the diagonal, which gives a cost of $2 \cdot 1 \cdot 2 = 4$.)

Problem 3. (5 points) Let $X = \{0, 1, 4, 5, 9\}$ 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 1, 0 merges with 1 and 4 with 5. At 3, $\{0, 1\}$ merges with $\{4, 5\}$ since $d(1, 4) = 3$, and at 4, $\{0, 1, 4, 5\}$ merges with $\{9\}$. See Fig. 1a.

(b) At 1, 0 merges with 1 and 4 with 5. We get three clusters $C_1 = \{0, 1\}$, $C_2 = \{4, 5\}$ and $C_3 = \{9\}$.

$$\begin{aligned} \delta(C_1, C_2) &= \frac{1}{2 \cdot 2}(4 + 3 + 5 + 4) = 4 \\ \delta(C_2, C_3) &= \frac{1}{2 \cdot 1}(5 + 4) = 4.5. \end{aligned}$$

Thus, we merge C_1 and C_2 at 4. Finally, we merge $\{0, 1, 4, 5\}$ and $\{9\}$ at

$$\delta(\{0, 1, 4, 5\}, \{9\}) = \frac{1}{4 \cdot 1}(9 + 8 + 5 + 4) = 6.5.$$

See Fig. 1b.

Problem 4. (3 points) Suppose $C = \{C_1, C_2, C_3\}$ is the 3-means clustering of a set $X \subseteq \mathbb{R}^2$, where the mean of C_1 is $\mu_1 = (0, 0)$, the mean of C_2 is $\mu_2 = (1, 1)$ and the mean of C_3 is $\mu_3 = (2, 0)$. Suppose $(1, 2), (0.7, 0) \in X$. To which cluster C_i does $(1, 2)$ belong? To which cluster C_i does $(0.7, 0)$ belong? Justify your answer.

Solution: We know that if a point $x \in X$ is closer to μ_i than any other μ_j, then $x \in C_i$. The μ_i closest to $(1, 2)$ is $\mu_2 = (1, 1)$, and the μ_i closest to $(0.7, 0)$ is $\mu_1 = (0, 0)$. Thus, $(1, 2)$ belongs to C_2, and $(0.7, 0)$ belongs to C_1.

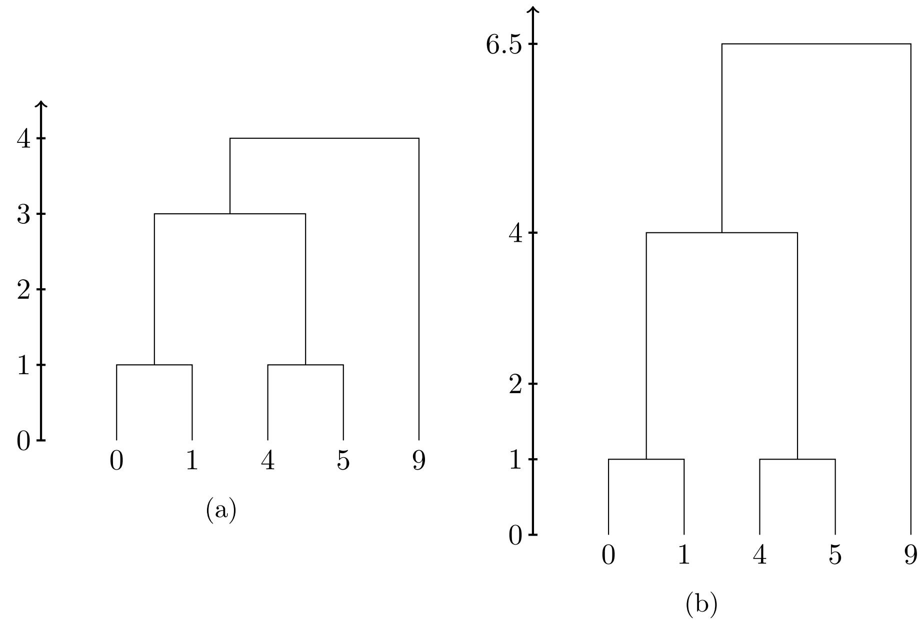

Figure 1: Solution to Problem 3.

Problem 5. (4 points)

(a) Find the barcode of the dendrogram in Fig. 2.

(b) For each $h \in [0, \infty)$, one can read off a partition of a subset of $X = \{a, b, c, d, e\}$ from the dendrogram in Fig. 2. Write these partitions for $h = 0, 1, 2, 3, 4$.

Solution: (a) $\{[0, \infty), [0, 2), [0, 2), [1, 4), [2, 3)\}$
(b)

h=0: $\{\{c\}, \{d\}, \{e\}\}$

h=1: $\{\{a\}, \{c\}, \{d\}, \{e\}\}$

h=2: $\{\{a\}, \{b\}, \{c, d, e\}\}$

h=3: $\{\{a, b\}, \{c, d, e\}\}$

h=4: $\{\{a, b, c, d, e\}\}$

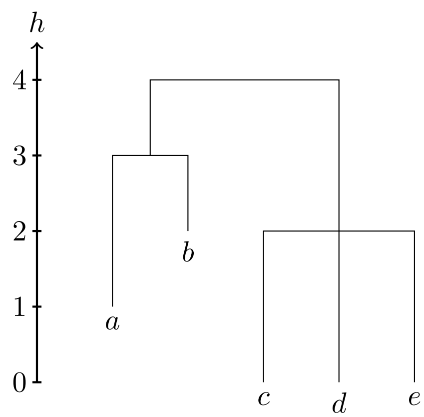

Figure 2