# Esercises on Clustering

**Exercise.** Show that the $L_1$ (Euclidean) and Edit distances satisfy the four requirements for a metric space.

**Solution.** Recall that given a set $M$ with distance function $d(\cdot)$, $(M, d)$ is a *metric space* if the following conditions hold for any $x, y, x \in M$:

1. $d(x, y) \geq 0$;
2. $d(x, y) = 0$ if and only if $x = y$;
3. $d(x, y) = d(y, x)$; (symmetry)
4. $d(x, z) \leq d(x, y) + d(y, z)$; (triangle inequality)

$L_1$ **distance.** Recall that the $L_1$ distance between two points $x, y \in \Re^n$ is

$$d_{\mathrm{L1}}(x, y) = \sum_{i=1}^{n} |x_i - y_i|,$$

where the $x_i$'s and $y_i$'s denote the coordinates. We observe that for any three real numbers $a, b$ and $c$, we have $|a - b| \leq |a - c| + |c - b|$, which can be easily proved considering all possible relative positions of $a, b$ and $c$ on the line. Now we prove that the four conditions above are satisfied.

1. $d_{\mathrm{L1}}(x, y) \geq 0$. It follows since all terms of the summation are $\geq 0$.
2. $d_{\mathrm{L1}}(x, y) = 0$ if and only if $x = y$. It follows since, in order for $d_{\mathrm{L1}}(x, y)$ to be 0, all terms of the summation must be 0, hence $x_i = y_i$ for every $i$.
3. $d_{\mathrm{L1}}(x, y) = d_{\mathrm{L1}}(y, x)$. It follows since $|a - b| = |b - a|$ for any two reals $a$ and $b$.
4. $d(x, z) \leq d(x, y) + d(y, z)$. By the initial observation, we have that

$$\begin{aligned} d_{\mathrm{L1}}(x, z) &= \sum_{i=1}^{n} |x_i - z_i| \\ &\leq \sum_{i=1}^{n} (|x_i - y_i| + |y_i - z_i|) \\ &= (\sum_{i=1}^{n} |x_i - y_i|) + (\sum_{i=1}^{n} |y_i - z_i|) \\ &= d_{\mathrm{L1}}(x, y) + d_{\mathrm{L1}}(y, z) \end{aligned}$$

**Edit distance.** Recall that the edit distance is defined for strings and for any two strings $X$ e $Y$, over some alphabet, $d_{\text{edit}}(X,Y)$ is the minimum number of deletions or insertions that must be applied to transform $X$ into $Y$. Equivalently,

$$d_{\text{edit}}(X,Y) = |X| + |Y| - 2|\text{LCS}(X,Y)|,$$

where $\text{LCS}(X,Y)$ is the Longest Common Subsequence between $X$ and $Y$. Now we prove that the four conditions above are satisfied.

1. $d_{\text{edit}}(X,Y) \geq 0$. It follows since $|X|, |Y| \geq |\text{LCS}(X,Y)|$

2. $d_{\text{edit}}(X,Y) = 0$ if and only if $X = Y$. It follows since the relation $|X|, |Y| \geq \text{LCS}(X,Y)|$ implies that $d_{\text{edit}}(X,Y) = 0$ if and only if $|\text{LCS}(X,Y)| = |X| = |Y|$, i.e., $\text{LCS}(X,Y) = X = Y$.

3. $d_{\text{edit}}(X,Y) = d_{\text{edit}}(Y,X)$. It follows since the role of $X$ and $Y$ in the second definition is perfectly symmetrical.

4. $d_{\text{edit}}(X,Z) \leq d_{\text{edit}}(X,Y) + d_{\text{edit}}(Y,Z)$. Considering the first definition, we can transform $X$ into $Z$ by first transforming $X$ into $Y$, and then $Y$ into $Z$. Therefore, the minimum number of deletions or insertions required to transform $X$ into $Z$ cannot be larger than the minimum number of such operations required to transform $X$ into $Y$ plus the minimum number of such operations required to transform $Y$ into $Z$.

□

**Exercise.** For a given pointset $P$ from a metric space $(M,d)$, let $(C_1, C_2, \ldots, C_k; c_1, c_2, \ldots, c_k)$ be the clustering returned by the Farthest-First Traversal algorithm. Consider an arbitrary cluster $C_i$ and an arbitrary point $q \in C_i$, with $q \neq c_i$. Show that for every $1 \leq j1 \neq j2 \leq k$

$$d(c_{j1}, c_{j2}) \geq d(q, c_i).$$

**Solution.** By contradiction, suppose that $d(c_{j1}, c_{j2}) < d(q, c_i)$ for some pair of indices $j1, j2$. Let the indices of the centers denote the order in which they are discovered and assume, without loss of generality, that $j1 < j2$. Consider the iteration when $c_{j2}$ is discovered. By construction of the algorithm we have that for every point $p \in P$

$$\min_{j=1,\ldots,j2-1} d(c_{j2}, c_j) \geq \min_{j=1,\ldots,j2-1} d(p, c_j). \tag{1}$$

Moreover, since $q$ is assigned to $C_i$ we must have

$$d(q, c_i) \leq d(q, c_j) \quad \forall 1 \leq j \leq k.$$

By combining these observations and the initial hypothesis that $d(c_{j1}, c_{j2}) < d(q, c_i)$, we have

$$\min_{j=1,\ldots,j2-1} d(q, c_j) \geq d(q, c_i) > d(c_{j2}, c_{j1}) \geq \min_{j=1,\ldots,j2-1} d(c_{j2}, c_j),$$

hence,

$$\min_{j=1,\ldots,j2-1} d(c_{j2}, c_j) < \min_{j=1,\ldots,j2-1} d(q, c_j),$$

which contradicts Equantion 1. □

**Exercise.** For a pointset $P$ from a metric space $(M, d)$ and an integer $k > 1$, let $\Phi^{\text{opt}}_{\text{kcenter}}(k)$ be the minimum value of $\Phi_{\text{kcenter}}(\mathcal{C})$ over all possible $k$-clusterings $\mathcal{C}$ of $P$.

1. Show that $\Phi^{\text{opt}}_{\text{kcenter}}(k+1) \leq \Phi^{\text{opt}}_{\text{kcenter}}(k)$

2. Suppose that $\Phi^{\text{opt}}_{\text{kcenter}}(k)$ is attained by a clustering $(C_1, C_2, \ldots, C_k; c_1, c_2, \ldots, c_k)$. Show that for any point $p \in P$, with $p \in C_i$ for some $i$, we have that

$$C_i \subseteq \left\{q \,:\, d(p,q) \leq 2\Phi^{\text{opt}}_{\text{kcenter}}(k)\right\}$$

**Solution.**

1. Let $\mathcal{C} = (C_1, C_2, \ldots, C_k; c_1, c_2, \ldots, c_k)$ be the optimal $k$-clustering such that $\Phi_{\text{kcenter}}(\mathcal{C}) = \Phi^{\text{opt}}_{\text{kcenter}}(k)$. Consider an arbitrary point $c_{k+1} \in P$ with $c_{k+1} \neq c_i$, for $1 \leq i \leq k$, and let $\mathcal{C}'$ be the clustering induced by centers $c_1, c_2, \ldots, c_k, c_{k+1}$ (i.e., assigning each point of $P$ to the closest center). We have that for each $p \in P$

$$\min_{i=1,\ldots,k+1} d(p, c_i) \leq \min_{i=1,\ldots,k} d(p, c_i),$$

   which immediately implies $\Phi_{\text{kcenter}}(\mathcal{C}') \leq \Phi_{\text{kcenter}}(\mathcal{C})$. Hence,

$$\Phi^{\text{opt}}_{\text{kcenter}}(k+1) \leq \Phi_{\text{kcenter}}(\mathcal{C}') \leq \Phi_{\text{kcenter}}(\mathcal{C}) = \Phi^{\text{opt}}_{\text{kcenter}}(k).$$

2. By the triangle inequality, we have that for every $q \in C_i$

$$d(p,q) \leq d(p, c_i) + d(c_i, q) \leq 2\Phi^{\text{opt}}_{\text{kcenter}}(k).$$

□

**Exercise.** Specify in detail the map and reduce phases of each round of the MR-Farthest-First Traversal algorithm.

**Solution.** Let $P = \{p(0), p(1), \ldots, p(N-1)\}$ be the input set and assume that each point is initially represented as a key-value pair $(j, p(j))$, where the index $j$ is the key. Let $\ell$ be known to the algorithm. For each $p \in P$, the algorithm will output a pair $(c, p)$ where $c$ is the center of the cluster to which $p$ is assigned.

**Round 1**

- **Map phase.** Each input pair $(j, p(j))$ is mapped into the intermediate pair $((i_j, 0), p_j)$, where $i_j = (j \bmod \ell) + 1$. Note that the key $(i_j, 0)$ has two components: one is an index between 1 and $\ell$, while the other is a bit set to 0. The meaning of the second component will be clear in a moment.
- **Reduce phase.** For each $1 \leq i \leq \ell$, separately, gather all pairs with key $(i, 0)$, and denote the points corresponding to this set of pairs as $P_i$. Run the Farthest-First Traversal algorithm on $P_i$ and let $c_1^{(i)}, c_2^{(i)}, \ldots, c_k^{(i)}$ be the centers identified by the algorithm. Represent each $c_s^{(i)}$ through the pair $(-1, c_s^{(i)})$.

**Round 2**

- **Map phase.** Identity map.
- **Reduce phase.** Gather all pairs with key -1 and let $T$ denote the set of points corresponding to these pairs. Run the Farthest-First Traversal algorithm on $T$ and let $c_1, c_2, \ldots, c_k$ be the centers identified by the algorithm. For each $1 \leq i \leq k$ create the $\ell$ pairs: $((1,1), c_i), ((2,1), c_i), \ldots, ((\ell, 1), c_i)$. As before, the key has two components: one is an index between 1 and $\ell$, while the other is a bit set to 1. We assume that at this point all intermediate pairs $((i, 0), p)$ from Round 1 are still available.

**Round 3**

- **Map phase.** Identity map.
- **Reduce phase.** For each $1 \leq i \leq \ell$ separately, gather the $k$ pairs $((i, 1), c_s)$, with $1 \leq s \leq k$, and all intermediate pairs $((i, 0), p)$ from Round 1. For each $((i, 0), p)$, determine the center $c$ (among $c_1, c_2, \ldots, c_k$) closest to $p$, and output the pair $(c, p)$.

□

**Exercise.** Argue that when $k = o(N)$ the MR-Farthest-First Traversal algorithm requires local space $M_L = o(N)$ and aggregate space $M_A = O(N)$.

**Solution.** Fix $\ell = \lceil \sqrt{N/k} \rceil$. The first round partitions the points into $\ell$ subsets of size $N/\ell = O\left(\sqrt{Nk}\right)$ each, and runs the Farthest-First Traversal algorithm separately for each subset. Thus, $O\left(\sqrt{Nk}\right)$ local space is sufficient for the first round. The second round gathers the $k$ centers computed in each of the $\ell$ subsets into a set $T$ which, therefore, has size $k\ell = O\left(\sqrt{Nk}\right)$. It then runs the Farthest-First Traversal algorithm on $T$. Thus, $O\left(\sqrt{Nk}\right)$ local space is sufficient for the second round. The third round computes the final cluster assignment separately for each subset of the initial partition by gathering together each subset and the set of centers computed from $T$. Thus, $O\left(k+\sqrt{Nk}\right) = O\left(\sqrt{Nk}\right)$ local space is sufficient for the third round. We conclude that the local space required by the entire algorithm is $O\left(\sqrt{Nk}\right)$, and since we assumed $k = o(N)$ we have that $O\left(\sqrt{Nk}\right) = o(N)$. As for the aggregate space, the only replication performed is when a copy of the $k$ centers computed from $T$ is gathered together with each of the $\ell$ subsets of points of the initial partition. These copies contribute an additive factor $k\ell = O\left(\sqrt{Nk}\right) = o(N)$ to the aggregate space which, therefore, remains linear in $N$. □

**Exercise.** Let $P$ be a set of $N > 1$ points from $\Re^n$ and let $k > 1$ be an integer. We use $\Phi^{\rm opt}_{\rm kmeans}(k)$ to denote the optimal value of the objective function of the k-means clustering. Suppose that $A$ is a *randomized algorithm* that, if executed on $P$, returns a $k$-clustering $\mathcal{C}_A$ such that, with probability 1/2:

$$\Phi_{\rm kmeans}(\mathcal{C}_A) \leq (\ln k)\Phi^{\rm opt}_{\rm kmeans}(k),$$

where ln denotes the natural logarithm. Show how, by suitably using $A$ as a subroutine, the same clustering quality can be obtained with probability $\geq 1 - 1/N$.

**Hint:** Use the Chernoff bound that states that for a Binomial r.v. $Z$ with $E[Z] = \mu$, $\Pr(Z \leq (1-\delta)\mu) \leq 2^{-\mu\delta^2/2}$, for any $\delta \in (0,1)$.

**Solution.** Let $c > 1$ be a suitable constant that will be determined shortly. For $t = c\log_2 N$ times, run $A$ on $P$, where each run is independent of the others. Let $\mathcal{C}_i$ be the $k$-clustering computed in the $i$th run, for $1 \leq i \leq t$. Return the $k$-clustering $\mathcal{C}$ among the $\mathcal{C}_i$'s which minimizes the objective function $\Phi_{\rm kmeans}(\mathcal{C})$. For $1 \leq i \leq t$, define a Bernoulli r.v. $X_i$ that takes the value 1 if $\Phi_{\rm kmeans}(\mathcal{C}_i) \leq (\ln k)\Phi^{\rm opt}_{\rm kmeans}(k)$, and 0 otherwise. Let $Z = \sum_{i=1}^{t} X_i$, which is a Binomial r.v. with $\mu = E[Z] = t/2$. Clearly, the probability that the best $k$-clustering returned by $A$ in the $t$ runs has a value of the objective function $\leq (\ln k)\Phi^{\rm opt}_{\rm kmeans}(k)$ is the probability that $Z > 0$. We have that $\Pr(Z > 0) = 1 - \Pr(Z = 0)$ and

$$\begin{array}{rcl}
\Pr(Z=0) & \leq & \Pr(Z \leq \mu/2) \\
& \leq & 2^{-\mu/8} \quad \text{(by Chernoff bound)} \\
& = & 2^{-c\log_2 N/16} \quad \text{(since } \mu = c\log_2 N/2\text{)} \\
& = & \dfrac{1}{N^{c/16}}.
\end{array}$$

Thus, by setting $c = 16$, we have that $\Pr(Z = 0) \leq 1/N$, hence $\Pr(Z > 0) \geq 1 - 1/N$. □