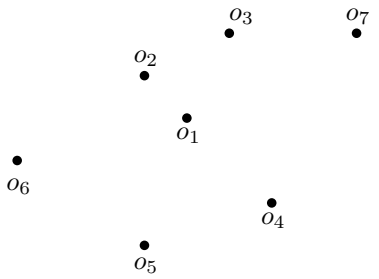
# INFS 4205/7205: Exercise Set 8

Prepared by Yufei Tao and Junhao Gan

**Problem 1.** Consider the cluster as shown in the figure below with point $o_1$ as the centroid (i.e., all 7 points in the same cluster).



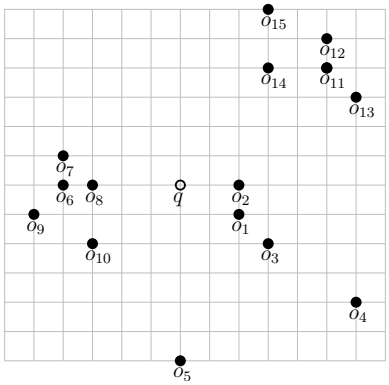

The list of points in ascending order of their distances to $o_1$ is $o_1, o_2, o_3, o_4, o_5, o_6, o_7$. Answer the following questions:

- Consider the nearest neighbor (NN) query with point $q_1$. Suppose that we have scanned the list up to point $o_4$. According to Pruning Rule 1, what is the lower bound we can derive on $\|q_1, p_{aft}\|$ for any point $p_{aft}$ that ranks after $o_4$ in the sorted list?
- Suppose that we have another NN query with query point $q_2$ (not shown in the figure). According to Pruning Rule 2, how far must $q_2$ be from $o_1$ so that the query can prune away the entire cluster? You can assume that the current NN of $q_2$ has distance 1 to $q_2$.

**Solution.** For the first bullet, $\|q_1, p_{aft}\| \geq \|o_4, o_1\| - \|o_1, q_1\|$.

Let $p_{nn}$ be the current nearst point to $q_2$ so far. The entire cluster can be pruned if $\|q_2, o_1\| > \|o_1, o_7\| + 1$.

**Problem 2.** Consider the dataset as shown in the figure below.




There are three clusters with centroids $o_1$, $o_6$, and $o_{11}$, respectively. The sorted lists of the three clusters are:

- $o_1, o_2, o_3, o_4, o_5$

- $o_6, o_7, o_8, o_9, o_{10}$
- $o_{11}, o_{12}, o_{13}, o_{14}, o_{15}$

Consider the NN query with point $q$. What are the points that are scanned by the query in each cluster? And in what order are the points scanned? You need to provide explanation for your answer.

**Solution.** The query scans $\{o_1, o_2, o_3, o_4\}$ from the first cluster (in this order), $\{o_6, o_7, o_8, o_9, o_{10}\}$ from the second, and no points from the third.

Since $o_1$ is the closest cluster centroid to $q$, the points in the first cluster are scanned first. When $o_4$ is scanned, $o_2$ is the current NN of $q$. As $\|o_1, o_4\| - \|o_1, q\| = 5 - 2\sqrt{2} > 2 = \|o_2, q\|$, the remaining points of the cluster are pruned.

The cluster with centroid $o_6$ is the second one scanned. As $\|o_6, q\| - \|o_6, o_{10}\| = 4 - \sqrt{5} < 2 = \|o_2, q\|$, Pruning Rule 2 is ineffective; and all the points in this cluster need to be scanned. The NN of $q$, however, remains as $o_2$.

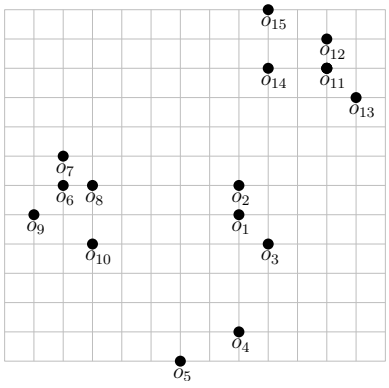
For the third cluster, as $\|o_{11}, q\| - \|o_{11}, o_{15}\| = \sqrt{41} - 2\sqrt{2} > 2 = \|o_2, q\|$, the entire cluster is pruned.

**Problem 3.** Suppose that we have created an index (as described in the class) on a set $P$ of $n$ points. There are $k$ clusters, each of which is covered by a circle centered at the cluster's centroid, and having a radius at most $r$. We know that any two of the $k$ centroids have distance greater than $4r$ from each other. Consider an NN query whose query point falls in one of these circles. Prove that our query algorithm will use Pruning Rule 2 to eliminate $k - 1$ clusters directly.

**Proof.** Without loss of generality, let $c_1$ be the centroid of the circle covering $q$, and let $c_2, \cdots, c_k$ be the other $k - 1$ centroids. The cluster of $c_1$ is the first scanned by our query algorithm. Let $p_{nn}$ be the NN of $q$ in this cluster.

It must hold that $\|q, p_{nn}\| \leq \|q, c_1\| + \|p_{nn}, c_1\| \leq r + r = 2r$. Now, for any $i \in [2, k]$, consider the cluster with centroid $c_i$. Denote by $p_i$ the farthest point from $c_i$ in this cluster. Since $\|q, c_i\| \geq \|c_i, c_1\| - \|q, c_1\| > 4r - r = 3r$, we know $\|q, c_i\| - \|c_i, p_i\| > 3r - r = 2r \geq \|q, p_{nn}\|$. Thus, according to Pruning Rule 2, the cluster can be pruned.

**Problem 4.** Consider the set of points shown in the figure below:




Suppose that we run the $k$-center algorithm with $k = 4$, and that $o_1$ is the first centroid chosen. What are the other 3 centroids output by the algorithm? Also, what are the resulting 4 clusters?

**Solution.** The centroids are $o_1, o_{15}, o_9, o_5$. The corresponding 4 clusters are: $\{o_1, o_2, o_3\}$, $\{o_{11}, o_{12}, o_{13}, o_{14}, o_{15}\}$, $\{o_6, o_7, o_8, o_9, o_{10}\}$ and $\{o_4, o_5\}$.

**Problem 5.** Describe a way to implement the $k$-center algorithm in $O(nk)$ time, where $n$ is the number of points in the dataset.

**Solution.**

- For each point $p \in P$, we will maintain its distance $t(p)$ to the nearest centroid found so far. At the beginning, $t(p) = \infty$.
- Pick an arbitrary point as the first centroid $c_1$. Update $t(p)$ to $\|p, c_1\|$ for all the points $p \in P$.
- For $i = 2$ to $k$, do the following:
  - Pick as the next centroid the point $p$ with the largest $t(p)$. Denote this point as $c_i$.
  - Update $t(p) = \min\{t(p), \|p, c_i\|\}$ for all the points $p \in P$.

Clearly, the algorithm spends $O(n)$ time finding a new centroid. Therefore, the overall running time is $O(nk)$.

**Problem 6.** In the class, we proved that the $k$-center algorithm is 2-approximate when the centroids must be selected from the dataset. In this exercise, we will prove that the same is true even if the centroids are arbitrary points in the data space.

For this purpose, let us rephrase the $k$-center problem as follows. Let $P$ be a set of points in $\mathbb{R}^d$. Let $S$ be a set of $k$ arbitrary points in $\mathbb{R}^d$. For each point $p \in P$, define:

$$\mathit{mindist}_S(p) \quad = \quad \min_{c \in S} \|c, p\|.$$

Then define

$$\mathit{penalty}(S) \quad = \quad \min_{p \in P} \mathit{mindist}_S(p).$$

The goal is to find a set $S$ with the smallest $\mathit{penalty}(S)$. Prove that the $k$-center algorithm is 2-approximate even for this (more general) problem.

**Solution.** Precisely the same proof presented in the class suffices—noticing that every sentence there is still correct even on the more general problem.