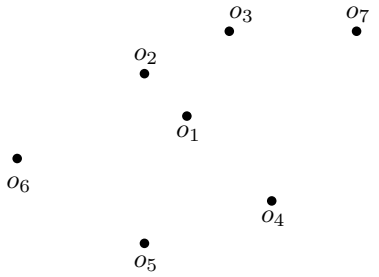
INFS 4205/7205: Exercise Set 8

Prepared by Yufei Tao and Junhao Gan

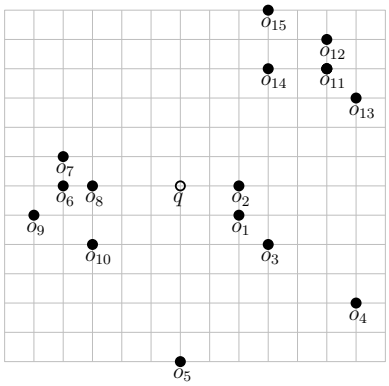
Problem 1. Consider the cluster as shown in the figure below with point o_1 as the centroid (i.e., all 7 points in the same cluster).



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.

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



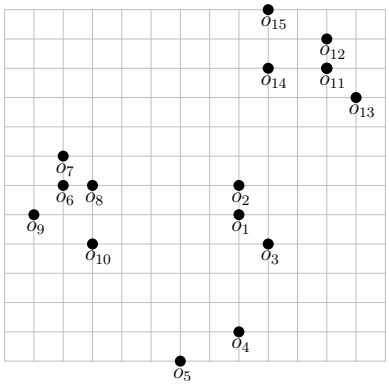
There are three clusters with centroids o_1, o_6, and o_{11}, respectively. The sorted lists of the three clusters are:

- 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.

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.

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?

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.

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:

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

Then define

$$penalty(S) \quad = \quad \min_{p \in P} mindist_S(p).$$

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