Homework 8

Due by the start of class on Tuesday, April 28. (Submissions will be through Gradescope.) Late homeworks are not accepted (unless an extension has been prearranged) so please turn in whatever you have completed by the due date. Unless otherwise specified, you may assume that all inputs are given in *general position*.

Problem 1. In this problem, we consider a common clustering technique. We are given a set $P = \{p_1, \ldots, p_n\}$ of points in $\mathbb{R}^d$ and an integer k, where $1 \leq k \leq n$. The objective is to compute a subset $C \subseteq P$ of size k, called *cluster centers*, in order to minimize the maximum distance from every point of P to its closest cluster center.[1] This can be expressed more formally as follows. Given two discrete sets C and P in $\mathbb{R}^d$, define the max-min distance

$$\text{dist}(P, C) \;=\; \max_{p \in P} \min_{c \in C} \|p - c\|,$$

and the objective function is

$$r^*(P, k) \;=\; \min_{\substack{C \subseteq P \\ |C| = k}} \text{dist}(P, C).$$

The set of centers C^* that realizes the minimum distance $r^*(P, k)$ is desired set of cluster centers (see Fig. 1).

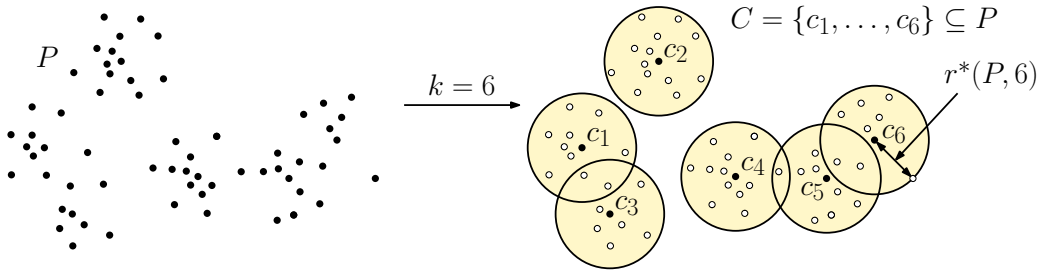

Figure 1: Clustering.

The objective of this problem is to compute an ε-coreset for this clustering problem. In particular, our objective is to compute a subset $R \subseteq P$ such that for any $0 < \varepsilon < 1$,

$$(1 - \varepsilon) r^*(P, k) \;\leq\; r^*(R, k) \;\leq\; (1 + \varepsilon) r^*(P, k).$$

Present an algorithm which given P, k, and ε, computes an ε-coreset for the clustering problem of size $O((1/\varepsilon)^d k)$. Justify your algorithm's correctness and derive its running time.

[1] As an example application, imagine that the point set P represents the residents of some city that has sufficient funds to build k urgent-care centers. The objective is to build k centers to minimimize the maximum distance that anyone needs to travel to get to the closest one.

Hint: A running time of $O(n \log n + (1/\varepsilon)^d k)$ is achievable, under the assumption that d is a constant. The construction involves building a square grid of an appropriate size and selecting an arbitrary point from each nonempty cell of the grid.

To assist you, you may assume that you have access to an algorithm that runs in time $O(n \log n)$ that computes a weak approximation in the form of a subset C' of size k such that

$$r^*(P,k) \;\leq\; \text{dist}(P,C') \;\leq\; 2r^*(P,k).$$

(Computing this weak approximation is an interesting problem in itself, but we will just assume we can access it like a black box.)

Problem 2. The objective of this problem is to investigate the *VC-dimension* of some range spaces. Recall that a *range space* Σ is a pair $(X, \mathcal{R})$, where X is a (finite or infinite) set, called *points*, and $\mathcal{R}$ is a (finite or infinite) family of subsets of X, called *ranges*.

For each of the following range spaces, derive its VC-dimension and prove your result. Throughout, you may assume that points are in general position.

(a) $\Sigma = (\mathbb{R}^2, \mathcal{U})$, where $\mathcal{U}$ consists of all orthogonal U-shaped ranges. An *orthogonal U-shaped range* is defined by three numbers (x_0, x_1, y_0). It is the region bounded between the two vertical lines $x_0 \leq x \leq x_1$ and above a horizontal line $y \geq y_0$ (see Fig. 2(a)).

(b) $\Sigma = (\mathbb{R}^2, \mathcal{V})$, where $\mathcal{V}$ consists of all orthogonal V-shaped ranges. An *orthogonal V-shaped range* is defined by four numbers (x_0, x_1, a, b). It is the region bounded between the two vertical lines $x_0 \leq x \leq x_1$ and above a line $y \geq ax + b$ (see Fig. 2(b)).

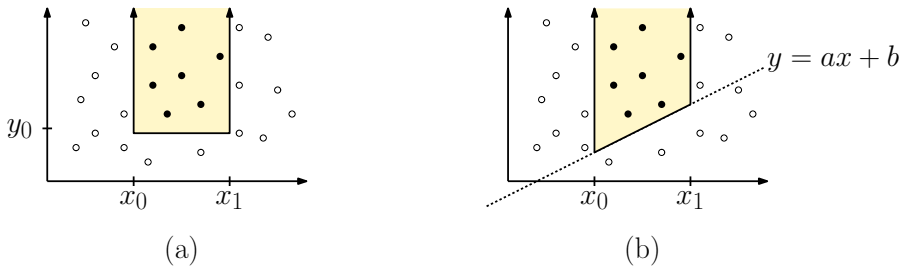

Figure 2: VC-Dimension of some range spaces.

Hint: As a template for your solution, here is an example of what a typical proof of VC dimensions looks like. In order to show that the VC-dimension is k, you need to give an example of a k-element subset that is shattered and prove that no set of size $k+1$ can be shattered.

Example: Consider the range space $\Sigma = (\mathbb{R}^2, H)$ where H consists of all closed horizontal halfspaces, that is, halfplanes of the form $y \geq y_0$ or $y \leq y_0$. We claim that $\text{VC}(\Sigma) = 2$.

$\text{VC}(\Sigma) \geq 2$: Consider the points $a = (0, -1)$ and $b = (0, 1)$. The ranges $y \geq 2$, $y \geq 0$, $y \leq 0$ and $y \leq 2$ generate the subsets $\{\emptyset, \{a\}, \{b\}, \{a, b\}\}$, respectively. Therefore, there is a set of size two that is shattered.

$\mathrm{VC}(\Sigma) < 3$: Consider any three element set $\{a, b, c\}$ in the plane. Let us assume that these points have been given in increasing order of their y-coordinates. Observe that any horizontal halfplane that contains b, must either contain a or c. Therefore, no 3-element point set can be shattered.

Problem 3. In this problem we will explore an idea for constructing a *weak ε-net* for a set of P points in the plane. By a "weak" ε-net, we mean a set of points that satisfies the standard definition of an ε-net, but it can be formed from *any* set of points, not just the points of P. We'll give the broad outline, and you will fill in the details.

You are given n points P in $\mathbb{R}^2$. Let us make the general-position assumption that there are no duplicate x- or y-coordinates. We construct a set $S \subset \mathbb{R}^2$ as follows. We first compute an integer $k \geq 1$, and let $m = \lfloor n/k \rfloor$. Next, we sort the points of S in increasing order according to their x-coordinates, and let $\langle x_1, \ldots, x_n \rangle$ denote the resulting sorted sequence of coordinates. We take every kth element from this sorted sequence:

$$X \;=\; \{x_k, x_{2k}, x_{3k}, \ldots, x_{mk}\}.$$

We repeat the same process for the y-coordinates, by first sorting them in increasing order as $\langle y_1, \ldots, y_n \rangle$, and setting

$$Y \;=\; \{y_k, y_{2k}, y_{3k}, \ldots, y_{mk}\}.$$

(Note that we use the same value of k and m in defining both X and Y.) Finally, we set $S = X \times Y$, that is, for $1 \leq i, j \leq m$, we include the point (x_i, y_j) into S. Clearly, S has m^2 elements. Observe that while they are based on the coordinates of the points of P, the points of S need not belong to P.

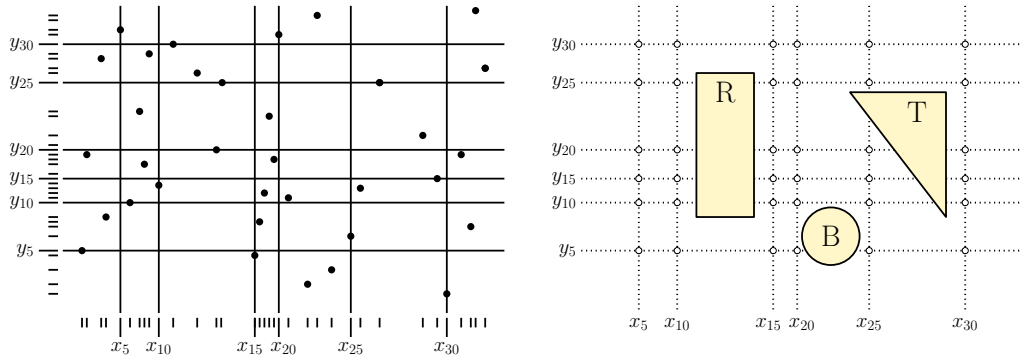

Figure 3: Weak ε-net construction (for $k = 5$).

(a) For each of the following range spaces, answer the following question. If a range from the set contains no elements of S, what is the maximum number of elements of P that it might contain? (That is, if $Q \cap S = \emptyset$ then how large can $|Q \cap P|$ be?) In each case justify your answer. (Express your answer precisely, not as an asymptotic function.)

(R) Axis-aligned rectangles (of any width and height).

(T) Axis-aligned right triangles. This is defined to be any right triangle such that the legs (i.e., the sides incident to the 90° angle) are parallel to the x- and y-axes. The hypotenuse can have any slope.

(B) Euclidean balls (of any radius).

(b) Suppose you are given a parameter $\varepsilon > 0$. Based on your answers to (R), (T), and (B) above, what value should we set k to (as a function of n and ε) in the above construction so that the resulting set S is an ε-net (in the weak sense) for P. We want k to be as large as possible, so that the resulting ε-net is as small as possible.

(c) Suppose that further, we would like a weak ε-sample. For each of the range spaces, (R), (T), and (B) above, is there any value of k such that the resulting set S is an ε-sample of P, where the size of S is a function of ε (but not n)? If so, give this value and justify its correctness. If not, explain why the resulting set S's size must depend on n.

Problem 4. (Optional–Ungraded) The objective of this problem is to investigate the *VC-dimension* of some more range spaces. For each of the following range spaces, derive its VC-dimension and prove your result. Throughout, you may assume that points are in general position.

In the following parts, let τ denote the triangle whose vertices are $(0, 0)$, $(1, 0)$, and $(0, 1)$ (see Fig. 4(a)).

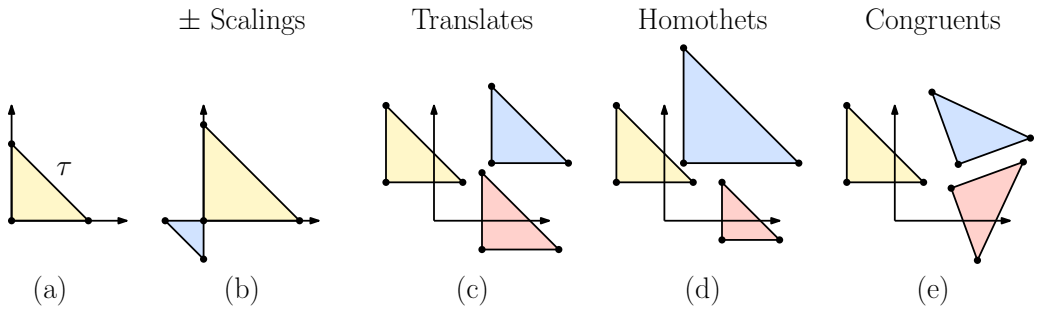

Figure 4: Problem 1: VC-dimension.

(a) A *uniform ±-scaling* of τ is any triangle with vertices $(0, 0)$, $(\alpha, 0)$ and $(0, \alpha)$ for any $\alpha \in \mathbb{R}$, possibly negative (see Fig. 4(b)). What is the VC-dimension of $\Sigma_1 = (\mathbb{R}^2, \mathcal{S})$, where $\mathcal{S}$ is the set of all uniform ±-scalings of τ?

(b) What is the VC-dimension of $\Sigma_2 = (\mathbb{R}^2, \mathcal{T})$, where $\mathcal{T}$ is the set of all translates of τ (see Fig. 4(c))?

(c) A shape τ' is a *homothet* of τ if τ' can be formed by performing a uniform scaling of τ by a any positive scale factor followed by any translation (see Fig. 4(d)). What is the VC-dimension of $\Sigma_3 = (\mathbb{R}^2, \mathcal{H})$, where $\mathcal{H}$ is the set of all homothets of τ?

Note: Challenge problems are not graded as part of the homework. The grades are recorded separately. After final grades have been computed, I may "bump-up" a grade that is slightly below a cutoff threshold based on these extra points. (But there is no formal rule for this.)

Challenge Problem: Answer Problem 4 for the range space of *congruent copies* of τ. A shape τ' is *congruent* to τ if τ' can be formed by performing a rotation, translation, and possible reflection of τ, but no scaling (see Fig. 4(e)). What is the VC-dimension of $\Sigma_4 = (\mathbb{R}^2, \mathcal{C})$, where $\mathcal{C}$ is the set of all shapes that are congruent to τ? (**Hint:** If you cannot obtain the exact dimension, I would be satisfied with upper and/or lower bounds.)