Machine Learning Theory (CSC 482A/581B)

Problem Set 2 **Due on Tuesday, February 26th, 7pm**

Instructions:

- You must write up your solutions individually.
- You may have high-level discussions with 1 other student registered in the course. If you discuss problems with another student, include at the top of your submission: their name, V#, and the problems discussed.
- Your must type up your solutions and are encouraged to use LaTeX to do this. For any problems where you only have a partial solution, be clear about any parts of your solution for which you have low confidence.
- Please submit your solutions via conneX by the due date of Tuesday, February 26th, 7pm. This is a hard deadline.

Questions:

1. Let $\mathcal{X} = \mathbb{R}^2$ and take $\mathcal{F}$ to be the set of all convex polygons; the classifier corresponding to a convex polygon labels as positive all points inside the polygon (including the boundary) and labels all other points as negative. Prove that $\mathrm{VCdim}(\mathcal{F}) = \infty$.

2. Let $\mathcal{F}$ be the class of linear separators in d dimensions, so that $\mathcal{F} = \left\{ f_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R} \right\}$ with $f_{w,b}(x) = \mathbf{1}\left[\langle w, x \rangle + b \geq 0\right]$.
   (a) Prove that $\mathrm{VCdim}(\mathcal{F}) \geq d + 1$.
   (b) Radon's Theorem states that any set of $d + 2$ points in $\mathbb{R}^d$ can be partitioned into two sets A and B such that the convex hulls of A and B intersect. Using Radon's Theorem, prove that $\mathrm{VCdim}(\mathcal{F}) \leq d + 1$ (and hence, combined with part (a), we may conclude that $\mathrm{VCdim}(\mathcal{F}) = d + 1$).
   (c) Next, prove Radon's Theorem. Any valid proof is allowed. Here is the start of one potential proof. Recall from linear algebra that any $d + 1$ points $x_1, \ldots, x_{d+1} \in \mathbb{R}^d$ must be linearly dependent, i.e., there exists a vector $\lambda \in \mathbb{R}^{d+1}$ not equal to the zero vector such that

$$\sum_{j=1}^{d+1} \lambda_j x_j = 0.$$

   The hint is to first prove that any set of $d + 2$ points $x_1, \ldots, x_{d+2} \in \mathbb{R}^d$ must be affine dependent, meaning that there exists a vector $\lambda \in \mathbb{R}^{d+2}$ not equal to the zero vector such that

$$\sum_{j=1}^{d+2} \lambda_j x_j = 0 \qquad \text{and} \qquad \sum_{j=1}^{d+2} \lambda_j = 0.$$

3. Suppose that P is a probability distribution over $\mathbb{R}^d$, and let the training sample $X_1, \ldots, X_n, X_{n+1}$ be i.i.d. samples with distribution P. We say that $(X_1, \ldots, X_n)$ is *s-sparse* if

$$\|X_j\|_0 \le s \qquad \text{for all } j \in [n],$$

where, for any vector x, the ℓ_0 "norm" $\|x\|_0$ is defined as the number of non-zero components in x.

Let $\hat{s}$ be the minimum value of $s \in \{0, 1, \ldots, d\}$ such that $(X_1, \ldots, X_n)$ is s-sparse.

(a) Derive an upper bound (which holds with high probability over $X_1, \ldots, X_n$) on the probability that X_{n+1} is $\hat{s}$-sparse. Specifically, your bound should be of the form:

With probability at least $1 - \delta$, $\Pr(\|X_{n+1}\|_0 > \hat{s}) = O\left(\frac{\log \frac{1}{\delta}}{n}\right)$.

The bound can also depend on the dimension, but the rate with respect to n cannot be worse than $O\left(\frac{1}{n}\right)$ (so $O\left(\frac{\log n}{n}\right)$ is not allowed).

(b) If your upper bound from part (a) depended on the dimension d, it degrades severely as $d \to \infty$. Derive an upper bound that is dimension-free. Unlike part (a), the rate with respect to n now can be $O\left(\frac{\log n}{n}\right)$.

4. Suppose that P is a probability distribution over the unit Euclidean ball in $\mathbb{R}^d$, and let $X_1, \ldots, X_n$ be i.i.d. samples with distribution P.

Using tools from class, prove that the average distance (considering all pairs) between n points is tightly concentrated around its expectation. That is, show that

$$\frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \|X_i - X_j\|_2$$

is tightly concentrated around

$$\mathsf{E}_{X,Y \sim P} \|X - Y\|_2.$$

Specifically, you should show that with probability at least $1 - \delta$,

$$\left| \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \|X_i - X_j\|_2 - \mathsf{E}_{X,Y \sim P} \|X - Y\|_2 \right| = O\left(\sqrt{\frac{\log \frac{1}{\delta}}{n}}\right).$$