ECE 6254, Spring 2024

Homework # 2

Due Sunday, February 4, at 11:59pm EST.

Suggested reading:

- *Learning from Data* (by Abu-Mostafa, Magdon-Ismail, Lin): Sections 2.1–2.2 (pages 39–62) contain a beautiful description of the VC generalization bound that closely mirrors what we did in class.
- "Introduction to Statistical Learning Thoery" by Bousquet, Boucheron, and Lugosi: If you read and liked the beginning of this paper, try to go back through this and re-read Sections 1-3 (first 13 pages) and then read Section 4, which provides another take on VC theory. You can download the paper on the course web page.
- Course notes on the growth function and VC bounds.

Problems:

1. Calculate the growth function $m_{\mathcal{H}}(n)$ for the following classes of classifiers:

   (a) The set of classifiers on $\mathbb{R}$ that can be written as either $h(x) = \text{sign}(x-a)$ for some $a \in \mathbb{R}$ or $h(x) = -\text{sign}(x-a)$ for some $a \in \mathbb{R}$, i.e., the set of both positive and negative rays.

   (b) The set of classifiers on $\mathbb{R}$ that can be written as either

   $$h(x) = \begin{cases} +1 & \text{for } x \in [a,b] \\ -1 & \text{otherwise} \end{cases}$$

   or

   $$h(x) = \begin{cases} -1 & \text{for } x \in [a,b] \\ +1 & \text{otherwise} \end{cases}$$

   for some $a, b \in \mathbb{R}$, i.e., the set of both positive and negative intervals.

   (c) The set of classifiers on $\mathbb{R}^n$ defined by the 1-nearest neighbor classification rule.

2. Consider classifiers defined by positive circles in $\mathbb{R}^2$, i.e., h such that for any $\boldsymbol{x} \in \mathbb{R}^2$,

   $$h(\boldsymbol{x}) = \begin{cases} +1 & \text{if } \|\boldsymbol{x} - \boldsymbol{c}\|_2 \le r \\ -1 & \text{otherwise,} \end{cases}$$

   for some $\boldsymbol{c} \in \mathbb{R}^2, r \in \mathbb{R}$.

   (a) Show that for this set of classifiers, $m_{\mathcal{H}}(3) = 8$.

(b) Argue that $m_{\mathcal{H}}(4) < 16$. [Hint: Review the approach we used in class for showing this same fact for linear classifiers in $\mathbb{R}^2$.]

3. Consider the set $\mathcal{H}$ of donut classifiers, functions on $\mathbb{R}^2$ of the form

$$h(\boldsymbol{x}) = \begin{cases} -1 \text{ if } \|\boldsymbol{x}\|_2^2 \leq a^2 \\ +1 \text{ if } a^2 < \|\boldsymbol{x}\|_2^2 \leq b^2 \\ -1 \text{ if } b^2 < \|\boldsymbol{x}\|_2^2 \end{cases} \tag{1}$$

for some $0 < a < b$.

(a) What is the growth function $m_{\mathcal{H}}(n)$ of this classifier?

(b) What is the VC dimension of $\mathcal{H}$?

4. Suppose that our input space $\mathcal{X} = \mathbb{R}$ and consider the hypothesis set

$$\mathcal{H} = \left\{ h : h(x) = \text{sign}\left(\sum_{i=0}^{d} c_i x^i \right) \quad \text{for some } c_0, \ldots c_d \in \mathbb{R} \right\}$$

In words, $\mathcal{H}$ is the set of classifiers obtained by evaluating some polynomial of degree d and comparing the result to a threshold. Prove that the VC dimension of $\mathcal{H}$ is exactly $d+1$ by showing that

(a) There are $d+1$ points which are shattered by $\mathcal{H}$. [Hint: Think about what happens when you define a polynomial by fixing the roots/zeros. What happens to the sign of this polynomial as you move across the intervals between the roots?]

(b) There are no $d+2$ points which are shattered by $\mathcal{H}$.

5. Suppose that the VC dimension of our hypothesis set $\mathcal{H}$ is $d_{\text{VC}} = 3$ (e.g., linear classifiers in $\mathbb{R}^2$) and that we have an algorithm for selecting some $h^\star \in \mathcal{H}$ based on a training sample of size n (i.e., we have n example input-output pairs to train on).

(a) Using the generalization bound given in class and derived in the handout, give a precise upper bound on $R(h^\star)$ that holds with probability at least 0.95 in the case where $n = 100$. Repeat for $n = 1\,000$ and $n = 10\,000$.

(b) Again using the generalization bound given in class, how large does n need to be to obtain a generalization bound of the form

$$R(h^\star) \leq \widehat{R}_n(h^\star) + 0.01$$

that holds with probability at least 0.95? How does this compare to the "rule of thumb" given in class?

6. (Optional) The VC dimension "usually" corresponds to the number of parameters or degrees of freedom in the hypothesis set. However, this is not *always* true. In this problem you will show a counter-example. Consider the hypothesis set for $x \in \mathbb{R}$ with

$$\mathcal{H} = \left\{ h : h(x) = (-1)^{\lfloor \alpha x \rfloor} \quad \text{for some } \alpha \in \mathbb{R} \right\},$$

where $\lfloor \cdot \rfloor$ is the floor function (i.e., the $\lfloor \alpha x \rfloor$ is the largest integer less than αx). Thus, even though this hypothesis set has a single parameter, it still has an infinite VC dimension! [Hint: Consider the inputs of $x_i = 10^i$ for $i = 1, \ldots, n$ and show how to choose α to realize an arbitrary dichotomy.]