Homework 7

Due April 3 at 11pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. Let $\mathcal{R}$ be the set of rectangles in $\mathbb{R}^2$ with sides aligned with the x and y axis. Consider the set $\mathcal{F}$ of functions on $\mathbb{R}^2$ that are equal to 1 on some rectangle $R \in \mathcal{R}$ and 0 everywhere else. What is the VC dimension of $\mathcal{F}$? (Your answer can contain drawings.) A rectangle $R \in \mathcal{R}$ is of the form $[a, b] \times [c, d]$ for some real numbers $a \leq b$ and $c \leq d$.

2. For $i = 1, \ldots, k$, consider a set $\mathcal{F}_i$ of functions from $\mathcal{X}$ to $\{-1, +1\}$, with $\mathrm{VC}(\mathcal{F}_i) \leq D$. Let $\mathcal{H}$ be the set of functions that can be written in the form, for some $f_1 \in \mathcal{F}_1, \ldots, f_k \in \mathcal{F}_k$,

$$h(x) = \begin{cases} 1 & \text{if } f_1(x) = \cdots = f_k(x) = 1 \\ -1 & \text{otherwise.} \end{cases}$$

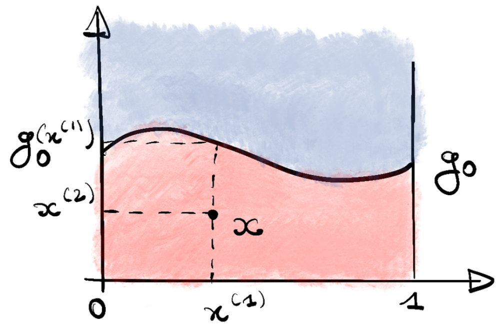

Figure 1: If the point $\mathbf{x}$ is "below" the graph of g_0, then $\mathbf{y} = +1$. Otherwise, $\mathbf{y} = -1$.

(a) Show that for every $x_1, \ldots, x_n \in \mathcal{X}$,

$$\log(\mathcal{N}_{\mathcal{H}}(x_1, \ldots, x_n)) \leq \sum_{i=1}^{k} \log(\mathcal{N}_{\mathcal{F}_i}(x_1, \ldots, x_n)).$$

(Hint: for $h \in \mathcal{H}$ (with corresponding functions $f_i \in \mathcal{F}_i$), we have $(h(x_1), \ldots, h(x_n)) = (\max_i f_i(x_1), \ldots, \max_i f_i(x_n))$.)

(b) Use Sauer's lemma to show that there exists an absolute constant C (not depending on D or k) such that

$$\mathrm{VC}(\mathcal{H}) \leq CDk \log(k).$$

The optimal value of C is not needed. (Hint: take $n = CDk\log(k)$ and use Sauer's lemma to show that $\mathcal{N}_{\mathcal{H}}(x_1, \ldots, x_n) < 2^n$. Hint 2: you may also use that the function $x \mapsto x \log(en/x)$ is increasing on $[0, n]$ as long as $n \geq 2$.)

3. Let $k \geq 1$ be an integer. Let $g_0 : [0,1] \to [0,1]$ be a $\mathcal{C}^k$ function with $|g_0^{(i)}(x)| \leq R$ for all $x \in [0,1]$ and $0 \leq i \leq k$. Let $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ be a uniform random variable on $[0,1]^2$. Let $\mathbf{y} = 1$ if $\mathbf{x}^{(2)} \leq g_0(\mathbf{x}^{(1)})$ and $\mathbf{y} = -1$ otherwise (see Figure 2). We let P be the law of $(\mathbf{x}, \mathbf{y})$ and consider a sample of n independent observations $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})$ of law P. We consider the ℓ_{01} loss: the P-risk of a predictor $f : \mathcal{X} \to \{-1, +1\}$ is given by $\mathbb{E}_P[\mathbf{1}\{f(\mathbf{x}) \neq \mathbf{y}\}] = P(f(\mathbf{x}) \neq \mathbf{y})$.

(a) What is the Bayes predictor $f_P^\star$? What is the Bayes risk $\mathcal{R}_P(f_P^\star)$?

(b) Let $\mathcal{G}$ be the set of $\mathcal{C}^k$ functions g with $|g^{(i)}(x)| \leq R$ for all $x \in [0,1]$ and $0 \leq i \leq k$. We let $\mathcal{F}$ be the set of classifiers f that can be written as

$$f(x) = \begin{cases} 1 & \text{if } x^{(2)} \leq g(x^{(1)}) \\ -1 & \text{otherwise,} \end{cases}$$

where $g \in \mathcal{G}$. Consider the empirical risk minimizer $\hat{f}_{\mathcal{F}}$. What is the approximation error of $\hat{f}_{\mathcal{F}}$? What is $\mathcal{R}_n(\hat{f}_{\mathcal{F}})$ equal to? What is the performance of this predictor?

(c) Let $\mathcal{G}_{L,k}$ be the set of piecewise polynomial functions: a function $g \in \mathcal{G}_{L,k}$ is of the form

$$g(x) = \sum_{i=0}^{k-1} a_{i,l}(x - x_l)^i \tag{1}$$

for $x \in [l/L, (l+1)/L]$, $x_l = (l + 1/2)/L$ and $l = 0, \dots, (L-1)$. We then let $\mathcal{F}_{l,k}$ be the set of classifiers f that can be written as

$$f(x) = \begin{cases} 1 & \text{if } x^{(2)} \leq g(x^{(1)}) \\ -1 & \text{otherwise,} \end{cases}$$

where $g \in \mathcal{G}_{l,k}$. By writing a Taylor expansion (with integral form of the remainder) of g_0 around each x_l, give a bound on the approximation error $\inf_{f \in \mathcal{F}_{L,k}} \mathcal{R}_P(f) - \mathcal{R}_P(f_P^\star)$.

(d) What is the VC dimension of $\mathcal{F}_{1,1}$? What is the VC dimension of $\mathcal{F}_{1,k}$? What is the VC dimension of $\mathcal{F}_{L,1}$? What is the VC dimension of $\mathcal{F}_{L,k}$? You do not have to prove your answers in this question.

(e) Consider the empirical risk minimizer $\hat{f}_{\mathcal{F}_{L,k}}$. Using the previous question, find a bound on the expected estimation error.

(f) Using the previous questions, give a bound on the expected excess of risk $\mathbb{E}[\mathcal{R}_P(\hat{f}_{\mathcal{F}_{L,k}}) - \mathcal{R}_P(f_P^\star)]$. How should L be chosen to (approximately) minimize this expression?