COMS 4995-1 S20 Homework 2 (due March 4, 2020)

Instructions

Submit your write-up on Gradescope as a neatly typeset PDF document by 11:00 PM of the due date. Please use TEX, LATEX, or a similar system.

On Gradescope, be sure to select the pages containing your answer for each problem. More details can be found on the Gradescope Student Workflow help page.

(If you don't select pages containing your answer to a problem, you'll receive a zero for that problem.)

Also, please make sure that your "Student ID #" on Gradescope is set to your UNI, using only lowercase letters and numbers (e.g., abc1234).

Finally, please make sure **your name and your UNI** appear prominently on the first page of your write-up.

Problem 1

Let $\mathsf{Halfspaces}_d := \{h_{\boldsymbol{w},b} : \boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$, the class of half-space functions in $\mathbb{R}^d$, where

$$h_{\boldsymbol{w},b}(\boldsymbol{x}) = \mathbb{1}_{\{\boldsymbol{x}^\mathsf{T}\boldsymbol{w} \leq b\}} \quad (\forall \boldsymbol{x} \in \mathbb{R}^d).$$

In the first part of this problem, you will show that for any n points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^2$, the set of effective behaviors of $\mathcal{H}$ on these points,

$$\mathsf{Halfspaces}_2(\boldsymbol{x}_{1:n}) = \{(h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_n)) : h \in \mathsf{Halfspaces}_2\},$$

has cardinality $O(n^2)$. Note that $\mathsf{Halfspaces}_d$ has VC dimension $d+1$, so Sauer's lemma only guarantees

$$|\mathsf{Halfspaces}_2(\boldsymbol{x}_{1:n})| \leq \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \binom{n}{3} = \Theta(n^3).$$

So this will show that Sauer's lemma is not tight for all hypothesis classes of a given VC dimension.

(a) Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n arbitrary points in $\mathbb{R}^2$. First, prove that there are at most $2(n-1)$ "behaviors" $\boldsymbol{a} = (a_1, \ldots, a_n)$ realized by half-space functions $h_{\boldsymbol{w},b}$ such that $\boldsymbol{x}_1^\mathsf{T}\boldsymbol{w} = b$.
Hint: Consider lines that pass through $\boldsymbol{x}_1$ and $\boldsymbol{x}_i$ for $i = 2, \ldots, n$. Then consider lines that pass through $\boldsymbol{x}_1$ and the angle between two "adjacent" lines of the previous type. What are the different behaviors that these lines determine?

(b) Use the result from Part (a) to prove that $|\mathsf{Halfspaces}_2(\boldsymbol{x}_{1:n})| \leq 4n(n-1)+1$ for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^2$.

Let $\mathsf{Balls}_d := \{g_{\boldsymbol{\mu},r} : \boldsymbol{\mu} \in \mathbb{R}^d, r > 0\}$ be the class of (Euclidean) ball functions in $\mathbb{R}^d$, where

$$g_{\boldsymbol{\mu},r}(\boldsymbol{x}) = \mathbb{1}_{\{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2 \leq r\}} \quad (\forall \boldsymbol{x} \in \mathbb{R}^d).$$

In the remainder of this problem, you will obtain bounds on the VC dimension of Balls_d.

(c) Let $\boldsymbol{\varphi}\colon \mathbb{R}^d \to \mathbb{R}^{d+1}$ be a feature map defined by $\boldsymbol{\varphi}(\boldsymbol{x}) := (\boldsymbol{x}, \|\boldsymbol{x}\|_2^2)$. Prove that if $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ are shattered by Balls_d, then $\boldsymbol{\varphi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\varphi}(\boldsymbol{x}_n)$ are shattered by $\mathsf{Halfspaces}_{d+1}$. Use this fact to deduce a bound on the VC dimension of Balls_d in terms of d.

(d) Prove that the VC dimension of Balls_d is at least $d+1$.

Problem 2

(a) Consider two hypothesis classes on the same input space, $\mathcal{H}_1$ and $\mathcal{H}_2$, where $\mathcal{H}_i$ has VC dimension $d_i \neq 0$. Prove that the VC dimension of $\mathcal{H}_1 \cup \mathcal{H}_2$ is at most an absolute constant times $\max\{d_1, d_2\}$.

(b) Prove that, for any non-negative integers d and n with $n \geq d$, there exists an input domain $\mathcal{X}$ and a hypothesis class $\mathcal{H}$ defined on $\mathcal{X}$, such that the following two properties hold.

(i) $\mathcal{H}$ has VC dimension d.
(ii) There exists $x_1, \ldots, x_n \in \mathcal{X}$ such that

$$|\mathcal{H}(x_{1:n})| = \sum_{k=0}^{d} \binom{n}{k}.$$

Problem 3

Let P be a probability distribution on $\mathcal{X} \times \{0,1\}$, and let $(\mathcal{H}_k)_{k\geq 1}$ be an infinite sequence of hypothesis classes on $\mathcal{X}$, where $\mathcal{H}_k$ has VC dimension d_k, and $d_1 < d_2 < \cdots$. Using the same "δ-splitting" trick as in the proof of the Occam's razor bound, one can prove that for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the random choice of an iid sample of size n from P,

$$|\operatorname{err}(h) - \widehat{\operatorname{err}}(h)| \leq \mathsf{bound}(k,n,\delta) \quad \text{(for all } k \geq 1 \text{ and all } h \in \mathcal{H}_k)$$

where

$$\mathsf{bound}(k,n,\delta) := C\sqrt{\frac{d_k \log(n) + \log(k) + \log(1/\delta)}{n}}.$$

Above, C is a universal constant. (Before working on this problem, make sure you see why this above assertion is true.)

Consider the following strategy for choosing $\hat{h} \in \bigcup_{k\geq 1} \mathcal{H}_k$. Let $\hat{h}_k := \arg\min_{h\in\mathcal{H}_k} \widehat{\operatorname{err}}(h)$ for each $k \geq 1$. Then, let $\hat{k} := \arg\min_{k\geq 1} \widehat{\operatorname{err}}(\hat{h}_k) + \mathsf{bound}(k,n,\delta)$. Finally, let $\hat{h} := \hat{h}_{\hat{k}}$.

(For simplicity, you can assume that minimizers exist so that "arg min" makes sense, and further that if there are ever multiple minimizers, then one can be selected arbitrarily.)

(a) The strategy above is simple to write down, but the "$\arg\min_{k\geq 1}$" should give some pause. Briefly explain how, in principle, the strategy can be executed in finite time. (Assume, for each k, that you have an algorithm for computing $\hat{h}_k$.)

(b) Prove that with probability at least $1-\delta$,

$$\operatorname{err}(\hat{h}) \leq \inf_{k\geq 1} \operatorname{err}(h_k^*) + 2\,\mathsf{bound}(k,n,\delta)$$

where $h_k^* := \arg\min_{h\in\mathcal{H}_k} \operatorname{err}(h)$ for each $k \geq 1$.

Problem 4

Recall that for any $A \subseteq \mathbb{R}^n$, we define

$$\mathrm{Rad}_n(A) := \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\boldsymbol{a} \in A} |\langle \boldsymbol{\sigma}, \boldsymbol{a} \rangle_n|$$

where $\boldsymbol{\sigma}$ is a random vector distributed uniformly in $\{\pm 1\}^n$ (i.e., the coordinates of $\boldsymbol{\sigma}$ are iid Rademacher random variables), and $\langle \cdot, \cdot \rangle_n$ is the *normalized* inner product

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_n := \frac{1}{n} \sum_{i=1}^{n} u_i v_i.$$

For any $A \subseteq \mathbb{R}^n$, define

$$\mathrm{Gauss}_n(A) := \mathbb{E}_{\boldsymbol{g}} \sup_{\boldsymbol{a} \in A} |\langle \boldsymbol{g}, \boldsymbol{a} \rangle_n|$$

where $\boldsymbol{g} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. This is the same as $\mathrm{Rad}_n(A)$, except with the Rademacher random vector $\boldsymbol{\sigma}$ replaced by the standard Gaussian random vector $\boldsymbol{g}$.

(a) Let A be an arbitrary subset of $\mathbb{R}^n$, and let $\boldsymbol{v} \in [-1, +1]^n$ be an arbitrary fixed vector. Define

$$A + \boldsymbol{v} := \{\boldsymbol{a} + \boldsymbol{v} : \boldsymbol{a} \in A\}.$$

Prove that $|\mathrm{Rad}_n(A) - \mathrm{Rad}_n(A + \boldsymbol{v})| \leq \sqrt{\frac{2 \ln 2}{n}}$.

(b) Let A and B be arbitrary subsets of $\{0, 1\}^n$. Define

$$A \odot B := \{\boldsymbol{a} \odot \boldsymbol{b} : \boldsymbol{a} \in A, \boldsymbol{b} \in B\},$$

where $\boldsymbol{u} \odot \boldsymbol{v}$ denotes the element-wise product of $\boldsymbol{u}$ and $\boldsymbol{v}$, and

$$A + B := \{\boldsymbol{a} + \boldsymbol{b} : \boldsymbol{a} \in A, \boldsymbol{b} \in B\}.$$

Prove that $\mathrm{Rad}_n(A \odot B) \leq 2\,\mathrm{Rad}_n(A + B) \leq 2(\mathrm{Rad}_n(A) + \mathrm{Rad}_n(B))$.
Hint: Try to use the contraction lemma for the first inequality.

(c) Compute $\mathrm{Gauss}_n(S^{n-1})$, where S^{n-1} is the unit sphere in $\mathbb{R}^n$ (i.e., the set of all unit vectors). (It is fine to just give the asymptotic dependence on n as $n \to \infty$.)

(d) Prove that for any $A \subseteq \mathbb{R}^n$,

$$\mathrm{Rad}_n(A) \leq \sqrt{\frac{\pi}{2}}\,\mathrm{Gauss}_n(A) \leq O(\sqrt{\log n})\,\mathrm{Rad}_n(A).$$

Problem 5

Below, k and d denote fixed positive integers, and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are iid random vectors taking values in B^d (the unit Euclidean norm ball).

(a) Consider the class of functions on $\mathbb{R}^d$ of the form

$$\mathcal{F}_1 := \left\{ \boldsymbol{x} \mapsto \mathrm{ReLU}(\boldsymbol{x}^\mathsf{T} \boldsymbol{b}) : \boldsymbol{b} \in B^d \right\},$$

where $B^d := \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_2 \leq 1\}$ is the unit ball in $\mathbb{R}^d$, and $\mathrm{ReLU} \colon \mathbb{R} \to \mathbb{R}$ is the function given by $\mathrm{ReLU}(z) := \max\{0, z\}$ for all $z \in \mathbb{R}$. Derive a bound on $\mathbb{E}\,\mathrm{Rad}_n(\mathcal{F}_1(\boldsymbol{X}_{1:n}))$.

(b) Consider the class of functions on $\mathbb{R}^d$ of the form

$$\mathcal{F}_2 := \left\{ \boldsymbol{x} \mapsto \sum_{i=1}^k a_i \,\mathrm{ReLU}(\boldsymbol{x}^\mathsf{T} \boldsymbol{b}_i) : a_1, \ldots, a_k \in \mathbb{R}, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_k \in B^d, \sum_{i=1}^k |a_i| \leq 1 \right\}.$$

Derive a bound on $\mathbb{E}\,\mathrm{Rad}_n(\mathcal{F}_2(\boldsymbol{X}_{1:n}))$.

(c) Consider the class of functions on $\mathbb{R}^d$ of the form

$$\mathcal{F}_3 := \left\{ \boldsymbol{x} \mapsto \sum_{i=1}^k a_i \,\mathrm{sign}(\boldsymbol{x}^\mathsf{T} \boldsymbol{b}_i) : a_1, \ldots, a_k \in \mathbb{R}, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_k \in \mathbb{R}^d, \sum_{i=1}^k |a_i| \leq 1 \right\}.$$

Derive a bound on $\mathbb{E}\,\mathrm{Rad}_n(\mathcal{F}_3(\boldsymbol{X}_{1:n}))$.