TTIC 31250 : Introduction to Theory of Machine Learning

Spring 2020

•

Homework 3

•

Solutions by

Jinhong Du

dujinhong@uchicago.edu

12243476

Exercises

1. Think about what you would like to do for your project and propose it. Some ideas:

   - Read a paper from a recent COLT conference (say COLT 2013 through COLT 2017) and write a 4-5 page explanation of what it does. Papers can be found at: http://www.learningtheory.org/past-conferences-2/
   - Read a paper on learning theory from a recent related conference (ICML, NIPS) and write a 4-5 page explanation of what it does.
   - Think about a theoretical question, which could be modeling some machine learning setting, trying to give sufficient conditions for some approach to succeed, looking at a different model for how examples are selected or the kind of feedback the algorithm is given, etc. Write up your thoughts in 4-5 pages.
   - Conduct an experiment to compare different approaches to some problem. (Note: your approach doesnt have to turn out to be the best one!). Create a 4-5 page writeup explaining your experiment and findings.

   For this homework I just want a brief description, such as "I plan to read and explain the paper X from conference Y" or "I would like to think about how to theoretically model Z".

   I plan to read and explain the paper *Dynamic Local Regret for Non-Convex Online Forecasting* from conference NIPS 2019.

2. Consider the class $\mathcal{C}$ of axis-parallel rectangles in $\mathbb{R}^3$. Specifically, a legal target function is specified by three intervals $[x_1^{\min}, x_1^{\max}]$, $[x_2^{\min}, x_2^{\max}]$, and $[x_3^{\min}, x_3^{\max}]$, and classifies an example (x_1, x_2, x_3) as positive iff $x_1 \in [x_1^{\min}, x_1^{\max}]$, $x_2 \in [x_2^{\min}, x_2^{\max}]$, and $x_3 \in [x_3^{\min}, x_3^{\max}]$. Argue that $\mathcal{C}[m] = O(m^6)$.

   Proof. Let

$$\begin{aligned}\mathcal{S} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m) : \boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})^\top \in \mathbb{R}^3,\ y_i \in \{0,1\},\\ x_{i1} \neq x_{j1},\ x_{i2} \neq x_{j2},\ x_{i3} \neq x_{j3} \text{ for } i \neq j\}\end{aligned}$$

   with $|\mathcal{S}| = m$. For $(\boldsymbol{x}_i, y_i) \in \mathcal{S}$, we can choose $x_1^{\min}$ and $x_1^{\max}$ in $\binom{m+1}{2} - 1 = \frac{m(m+1)}{2} - 1$ ways. To see this, we first arange $x_{11}, \ldots, x_{m1}$ as $x_{(1),1} < \cdots < x_{(m),1}$. Then we divide $\mathbb{R}$ into $m+1$ sets except m points,

$$(-\infty, x_{(1),1}),\ (x_{(1),1}, x_{(2),1}),\ \ldots,\ (x_{(m),1}, \infty).$$

   So we can randomly throw $x_1^{\min}$ and $x_1^{\max}$ into these sets. There are $\binom{m+1}{2} - 1$ ways, since $x_1^{\max} > x_{(n),1}$ and $x_1^{\min} < x_{(1),1}$ have the same effect. Also, note that the break points will not effect the labeling. Analogously, there are also $\frac{m(m+1)}{2} - 1$ to choose $x_2^{\min}$ and $x_2^{\max}$, as well as $x_3^{\min}$ and $x_3^{\max}$. Combine the three dimensions, we can label $\mathcal{S}$ by at most totally

$$\left(\frac{m(m+1)}{2} - 1\right)^3 = O(m^6)$$

   different ways, i.e., $\mathcal{C}[m] = O(m^6)$. □

Problems

3. **VC-dimension of Two-Layer Networks.** Suppose that concept class $\mathcal{H}$ has VC-dimension d. Now suppose we create a 2-layer network by choosing k functions $h_1, h_2, \ldots, h_k$ from $\mathcal{H}$ and then running their output through some other Boolean function f. That is, given an input x, the network outputs $f(h_1(x), \ldots, h_k(x))$. For a given f, call the class of all such functions TWO-LAYER$_{f,k}(\mathcal{H})$. Show that TWO-LAYER$_{f,k}(\mathcal{H})$ has VC-dimension $O(kd\log(kd))$. Note that we are only asking for an upper bound here, not a lower bound.

   Hint: Suppose you have a set $\mathcal{S}$ of m data points. By Sauers lemma, we know there are at most $O(m^d)$ ways of labeling those points using functions in $\mathcal{C}$. Use that to get an upper bound on the number of ways of labeling those points using functions in TWO-LAYER$_{f,k}(\mathcal{C})$. Now select m so that this is less than 2^m which means the VC-dimension must be less than m.

Proof. Let $\mathcal{H}^k = \mathcal{H} \times \cdots \times \mathcal{H}$ be the cartesian product of k identical concept classes $\mathcal{H}$. For any sample $\mathcal{S}$ with $|\mathcal{S}| = m$, the ways to label points in $\mathcal{S}$ by $\mathcal{H}^k$, is exactly the product of the ways to label them by each $\mathcal{H}$. So

$$|\mathcal{H}^k[\mathcal{S}]| = |\mathcal{H}[\mathcal{S}]|^k \leq \mathcal{H}[m]^k,$$

by the definition of $\mathcal{H}[m]$.
Let $\mathcal{F}$ be the class of boolean function $f : \{0,1\}^k \mapsto \{0,1\}$, then for a given $f \in \mathcal{F}$, TWO-LAYER$_{f,k}(\mathcal{H}) = f \circ \mathcal{H}^k$. Since for any sample $\mathcal{S}$ with $|\mathcal{S}| = m$,

$$\begin{aligned}\text{TWO-LAYER}_{f,k}(\mathcal{H})[\mathcal{S}] &= \{f(g(x)) : x \in \mathcal{S}, g \in \mathcal{H}^k\} \\ &= \bigcup_{y \in \mathcal{H}^k} \{f(y)\},\end{aligned}$$

we have for a given $f \in \mathcal{F}$,

$$\begin{aligned}|\text{TWO-LAYER}_{f,k}(\mathcal{H})[\mathcal{S}]| &\leq \sum_{y \in \mathcal{H}^k} |\{f(y)\}| \\ &\leq 2|\mathcal{H}^k| \\ &\leq 2\mathcal{H}[m]^k.\end{aligned}$$

By Sauer's Lemma, $\mathcal{H}[m] \lesssim O(m^d)$. Thus,

$$\text{TWO-LAYER}_{f,k}(\mathcal{H})[m] \leq 2\mathcal{H}[m]^k \lesssim O(m^{kd}).$$

Now, let $\mathcal{S}_0$ be a set of size m that is shattered by TWO-LAYER$_{f,k}(\mathcal{H})$. Then

$$\text{TWO-LAYER}_{f,k}(\mathcal{H})[\mathcal{S}_0] = 2^m \lesssim O(m^{kd}),$$

i.e., $m \lesssim O(kd\log(m))$. Since $\log(m) \lesssim O(\log(kd\log(m))) \simeq O(\log(kd) + \log(\log(m)))$, i.e., $\log(m) \lesssim O(\log(kd))$, we have $m \lesssim O(kd\log(kd))$. As the VC-dimension of TWO-LAYER$_{f,k}(\mathcal{H})$ is the largest m such that there exists a sample $\mathcal{S}$ with $|\mathcal{S}| = m$ that can be shattered by TWO-LAYER$_{f,k}(\mathcal{H})$, we conclude that TWO-LAYER$_{f,k}(\mathcal{H})$ has VC-dimension $\lesssim O(kd\log(kd))$.

□

In problems 4-6, you will prove that the VC-dimension of the class $\mathcal{H}_n$ of halfspaces in n dimensions is $n+1$. ($\mathcal{H}_n$ is the set of functions $a_1x_1 + \ldots + a_nx_n \geq a_0$, where $a_0, \ldots, a_n$ are real-valued.) We will use the following definition: The convex hull of a set of points $\mathcal{S}$ is the set of all convex combinations of points in $\mathcal{S}$; this is the set of all points that can be written as $\sum_{x_i \in \mathcal{S}} \lambda_i x_i$, where each $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$. It is not hard to see that if a halfspace has all points from a set $\mathcal{S}$ on one side, then it must have the entire convex hull of $\mathcal{S}$ on that side as well.

4. **Lower Bound** Prove that VC-dim($\mathcal{H}_n$)$\geq n+1$ by presenting a set of $n+1$ points in n-dimensional space such that one can partition that set with halfspaces in all possible ways. (And, show how one can partition the set in any desired way.)

Proof. Let $\boldsymbol{x}_0 = \mathbf{0} \in \mathbb{R}^n$ and $\boldsymbol{x}_i = \boldsymbol{e}_i \in \mathbb{R}^n$ for $i = 1, \ldots, n$, where $\boldsymbol{e}_i$ is a unit vector with its ith entry being one. Let $\mathcal{X}_n = \{\boldsymbol{x}_i, i = 0, \ldots, n\}$.
If we want to label m different points $\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_m} \in \mathcal{X}_n$ as 1,

(a) If $\boldsymbol{x}_{i_j} \neq \mathbf{0}$ for all j, then we can choose $f(\boldsymbol{x}) = \sum_{j=1}^m x_j \geq \frac{1}{2}$.

(b) If $\boldsymbol{x}_{i_j} = \mathbf{0}$ for some j, then we can choose $f(\boldsymbol{x}) = -\sum_{j \notin \{i_1, \ldots, i_m\}} x_j \geq 0$.

Thus, $\boldsymbol{x}_{i_j}$ will have label 1 and $\boldsymbol{x}_k$ ($k \notin \{i_1, \ldots, i_m\}$) will have label 0. Also, this holds for $m = 1, \ldots, n$.
For the case $m = 0$ or $n+1$, we can choose $f(\boldsymbol{x}) \equiv 0 \geq 1$ and $f(\boldsymbol{x}) \equiv 0 \geq 0$, respectively.
Therefore, we can partition $\mathcal{X}_n$ with halfspaces in all possible ways. So VC-dim$(\mathcal{H}_n) \geq n+1$. □

5. **Upper Bound Part 1** The following is "Radon's Theorem", from the 1920's.

Theorem. *Let $\mathcal{S}$ be a set of $n+2$ points in n dimensions. Then $\mathcal{S}$ can be partitioned into two (disjoint) subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ whose convex hulls intersect.*

Show that Radon's Theorem implies that the VC-dimension of halfspaces is at most $n+1$. Conclude that VC-dim($\mathcal{H}_n$)$= n+1$.

Proof. Let conv($\mathcal{S}$) denote the convex hull of $\mathcal{S}$.
From Radon's Theorem, any sample $\mathcal{S}$ with $n+2$ points in n dimensions can be partitioned into two disjoint subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ whose convex hulls intersect. If there is a halfspace can separate $\mathcal{S}_1$ and $\mathcal{S}_2$, then it can also separate conv($\mathcal{S}_1$) and conv($\mathcal{S}_2$). To see this, without loss of generality, assume that points in $\mathcal{S}_1$ and $\mathcal{S}_2$ satisfy $\sum_{i=1}^n a_i x_i \geq a_0$ and $\sum_{i=1}^n a_i x_i < a_0$ respectively. Suppose that $\mathcal{S}_1 = \{s_1, \ldots, s_m\}$, then for all $s \in \text{conv}(\mathcal{S}_1)$, we have $s = \sum_{j=1}^m \lambda_j s_j$ for $\lambda_j \geq 0$ and $\sum_{j=1}^m \lambda_j = 1$. As $\sum_{i=1}^n a_i s_j \geq a_0$, we have

$$\sum_{i=1}^n a_i s = \sum_{i=1}^n a_i \sum_{j=1}^m \lambda_j s_j = \sum_{j=1}^m \lambda_j \sum_{i=1}^n a_i s_j \geq a_0,$$

which means that s is on the same side with $\mathcal{S}_1$. So conv($\mathcal{S}_1$) is on the same side of the halfspace with $\mathcal{S}_1$. Analogously, conv($\mathcal{S}_2$) is on the same side of the halfspace with $\mathcal{S}_2$. So conv($\mathcal{S}_1$) and conv($\mathcal{S}_2$) are separated by the halfspace.
However, since $\text{conv}(\mathcal{S}_1) \cap \text{conv}(\mathcal{S}_2) \neq \varnothing$, there is no way to separate conv($\mathcal{S}_1$) and conv($\mathcal{S}_2$) by a halfspace. Contradiction. So $\mathcal{S}$ with $|\mathcal{S}| = n+2$ cannot be shattered by $\mathcal{H}_n$.
Therefore, VC-dim$(\mathcal{H}_n) < n+2$. From Problem 4, we have VC-dim$(\mathcal{H}_n) = n+1$. □

6. **Upper Bound Part 2** Now we prove Radon's Theorem. We will need the following standard fact from linear algebra. If $x_1, \ldots, x_{n+1}$ are $n+1$ points in n-dimensional space, then they are linearly dependent. That is, there exist real values $\lambda_1, \ldots, \lambda_{n+1}$ not all zero such that $\lambda_1 x_1 + \ldots + \lambda_{n+1} x_{n+1} = 0$.

   You may now prove Radon's Theorem however you wish. However, as a suggested first step, prove the following. For any set of $n+2$ points $x_1, \ldots, x_{n+2}$ in n-dimensional space, there exist $\lambda_1, \ldots, \lambda_{n+2}$ not all zero such that $\sum_i \lambda_i x_i = 0$ and $\sum_i \lambda_i = 0$. (This is called *affine dependence*.) Now, think about the lambdas...

Proof. For any set of $n+2$ points $x_1, \ldots, x_{n+2}$, let $y_i = x_i$ for $i = 1, \ldots, n$ and $y_{n+1} = x_{n+1} + x_{n+2}$. Then there exists $\omega_1, \ldots, \omega_{n+1}$ not all zero such that $\sum_{i=1}^{n+1} \omega_i y_i = \sum_{i=1}^{n+2} \omega_i x_i = 0$, where $\omega_{n+2} = \omega_{n+1}$. That is, there exists $\omega_1, \ldots, \omega_{n+2}$ not all zero such that $\sum_{i=1}^{n+2} \omega_i x_i = 0$.
Without loss of generality, assume that $\omega_{n+2} \neq 0$.

(a) If $x_{n+2} = 0$, let $\omega_{n+2} = -\sum_{i=1}^{n+1} \omega_i$, then $\omega_1, \ldots, \omega_{n+2}$ not all zero, $\sum_{i=1}^{n+2} \omega_i x_i = 0$ and $\sum_{i=1}^{n+2} \omega_i = 0$.

(b) If $x_{n+2} \neq 0$, then

$$\sum_{i=1}^{n+1} \frac{\omega_i}{\omega_{n+2}} x_i + x_{n+2} = 0$$

and at least one of $\omega_1, \ldots, \omega_{n+1}$ is nonzero. Since there exists $\mu_1, \ldots, \mu_{n+1}$ not all zero such that $\sum_{i=1}^{n+1} \mu_i x_i = 0$, let $\mu_i' = \frac{\mu_i}{\sum_{j=1}^{n+1} \mu_j} \left(-\frac{1}{\omega_{n+2}} \sum_{j=1}^{n+1} \omega_j - 1 \right)$, then $\sum_{i=1}^{n+1} \mu_i' = -\frac{1}{\omega_{n+2}} \sum_{j=1}^{n+1} \omega_j - 1$. Notice that $\sum_{i=1}^{n+1} \mu_i' x_i = 0$, we have

$$\sum_{i=1}^{n+1} \left(\frac{\omega_i}{\omega_{n+2}} + \mu_i' \right) x_i + x_{n+2} = 0$$

and

$$\sum_{i=1}^{n+1} \left(\frac{\omega_i}{\omega_{n+2}} + \mu_i' \right) + 1 = 1 + \sum_{i=1}^{n+1} \frac{\omega_i}{\omega_{n+2}} + \sum_{i=1}^{n+1} \mu_i' = 0.$$

Thus, we have proved the following lemma,

Lemma 1. *For any set of $n+2$ points $x_1, \ldots, x_{n+2}$ in n-dimensional space, there exist $\lambda_1, \ldots, \lambda_{n+2}$ not all zero such that $\sum_i \lambda_i x_i = 0$ and $\sum_i \lambda_i = 0$.*

So, for any set of $n+2$ points in n-dimension space, there exists $\lambda_1, \ldots, \lambda_{n+2}$ not all zero such that $\sum_{i=1}^{n+2} \lambda_i x_i = \sum_{i=1}^{n+2} \lambda_i = 0$. Among the λ_i's, there must be $m > 0$ of them are positive. Without loss of generality, assume that $\lambda_1, \ldots, \lambda_m > 0$ and $\lambda_{m+1}, \ldots, \lambda_{n+2} \leq 0$. Then

$$\sum_{i=1}^{m} \lambda_i x_i = \sum_{j=m+1}^{n+2} (-\lambda_j) x_j.$$

Since $\sum_{i=1}^{n+2} \lambda_i = \sum_{i=1}^{m} \lambda_i + \sum_{j=m+1}^{n+2} \lambda_j = 0$, we have $\sum_{i=1}^{m} \lambda_i = \sum_{j=m+1}^{n+2} (-\lambda_j)$. Let $p_i = \frac{|\lambda_i|}{\sum_{j=1}^{m} \lambda_j}$ for $i = 1, \ldots, n+2$. Then

$$\sum_{i=1}^{m} p_i x_i = \sum_{j=m+1}^{n+2} p_j x_j,$$

$p_i \geq 0$ and $\sum_{i=1}^{m} p_i = \sum_{j=m+1}^{n+2} = 1$. Let $\mathcal{S}_1 = \{x_1, \ldots, x_m\}$ and $\mathcal{S}_2 = \{x_{m+1}, \ldots, x_{n+2}\}$. Then $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$, $\mathcal{S}_1 \cap \mathcal{S}_2 = \varnothing$ and $\sum_{i=1}^{m} p_i x_i \in \operatorname{conv}(\mathcal{S}_1) \cap \operatorname{conv}(\mathcal{S}_2)$, which implies that $\operatorname{conv}(\mathcal{S}_1) \cap \operatorname{conv}(\mathcal{S}_2) \neq \varnothing$. $\square$