TFML HW-3

SIDDHANT CHAUDHARY
BMC201953

(1) Problem 7.4 of the book. Let $\mathcal{H}$ be some hypothesis class, and for $h \in \mathcal{H}$ suppose $|h|$ denotes the description length of h. Because we are using the MDL paradigm, we assume that the description language for the class $\mathcal{H}$ is *prefix-free*. Clearly, this implies that $\mathcal{H}$ is countable; this is because the description function $d : \mathcal{H} \to \Sigma^*$ must be injective. Since Σ^* is countable, we immediately see that $\mathcal{H}$ is countable. So, we will now assume that $\mathcal{H} = \bigcup_{n\in\mathbb{N}} \{h_n\}$. Also, the weight of h_n is just $\frac{1}{2^{|h_n|}}$ (and as seen in class, these weights add up to atmost 1 by *Kraft Inequality*).

For a sample set S of size m, let

$$h_S \in \underset{h\in\mathcal{H}}{\operatorname{argmin}} \left[L_S(h) + \sqrt{\frac{|h| + \log(2/\delta)}{2m}} \right]$$

For any $B > 0$, let

$$\mathcal{H}_B = \{h \in \mathcal{H} : |h| \leq B\}$$

and we also define

$$h_B^* = \underset{h\in\mathcal{H}_B}{\operatorname{argmin}} L_{\mathcal{D}}(h)$$

We will show that

$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h_B^*) \leq 2\sqrt{\frac{B + \log(2/\delta)}{2m}} \tag{0.1}$$

To prove this, we will use the following fact that was proven in class: if $\mathcal{H} = \bigcup \mathcal{H}_n$, then with probability atleast $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$, the following bound holds (simultaneously) for all $n \in \mathbb{N}$ and $h \in \mathcal{H}_n$ (this is **Theorem 7.4** of the book).

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, w(n) \cdot \delta)$$

In the MDL paradigm, the function ϵ_n was the following.

$$\epsilon_n(m, \delta) = \sqrt{\frac{\log(2/\delta)}{2m}}$$

As we claimed above, if $h \in \mathcal{H}_n$, then $w(n) = \frac{1}{2^{|h|}}$. So, we see that with probability atleast $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$, the following holds for all n and $h \in \mathcal{H}_n$.

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n\left(m, \frac{1}{2^{|h|}}\delta\right) = \sqrt{\frac{|h| + \log(2/\delta)}{2m}} \tag{0.2}$$

Date: 18 December, 2021.

So, for every $B > 0$, with probability of atleast $1-\delta$ over the choice of $S \sim \mathcal{D}^m$, we have the following.

$$\begin{aligned}
L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \sqrt{\frac{|h_S| + \log(2/\delta)}{2m}} && \text{(By inequality (0.2))} \\
&\leq L_S(h_B^*) + \sqrt{\frac{|h_B^*| + \log(2/\delta)}{2m}} && \text{(By defn. of } h_S) \\
&\leq L_{\mathcal{D}}(h_B^*) + 2\sqrt{\frac{|h_B^*| + \log(2/\delta)}{2m}} && \text{(Again by inequality (0.2))} \\
&\leq L_{\mathcal{D}}(h_B^*) + 2\sqrt{\frac{B + \log(2/\delta)}{2m}} && (|h_B^*| \leq B)
\end{aligned}$$

and hence it follows that

$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h_B^*) \leq 2\sqrt{\frac{B + \log(2/\delta)}{2m}}$$

and this proves our claim (0.1), completing the solution.

(2) Problem 7.5 of the book. Here we will solve all five parts of this problem.

1. Let A be a nonuniform learner for a class $\mathcal{H}$. For each $n \in \mathbb{N}$, we define

$$\mathcal{H}_n^A := \left\{h \in \mathcal{H} : m^{\text{NUL}}(0.1, 0.1, h) \leq n\right\}$$

We will show that each class $\mathcal{H}_n^A$ has finite VC dimension. Note that $0.1 < 1/8$ and $0.1 < 1/7$. Clearly, the definition of $\mathcal{H}_n^A$ implies that with probability of atleast $1 - 0.1 = 0.9 > 1 - \frac{1}{7}$ over the choice of $S \sim \mathcal{D}^n$, it is true that

$$L_{\mathcal{D}}(A(S)) \leq \operatorname*{argmin}_{h \in \mathcal{H}_n^A} L_{\mathcal{D}}(h) + 0.1 < \operatorname*{argmin}_{h \in \mathcal{H}_n^A} L_{\mathcal{D}}(h) + 1/8$$

In particular, if $\mathcal{D}$ is a distribution satisfying the realizability assumption w.r.t $\mathcal{H}_n^A$, we have that with probability of atleast $1 - \frac{1}{7}$, it is true that

$$L_{\mathcal{D}}(A(S)) < \frac{1}{8}$$

But this clearly implies that the VC dimension of $\mathcal{H}_n^A$ is finite; otherwise, by the **No Free Lunch Theorem** (**Theorem 5.1** of the book) there will be some distribution $\mathcal{D}$ satisfying the realizability assumption w.r.t $\mathcal{H}_n^A$ for which, with probability $\geq \frac{1}{7}$, it will be the case that

$$L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}$$

and that will be a contradiction to what we've seen above. So, it follows that $\text{VCdim}(\mathcal{H}_n^A) < \infty$.

2. Suppose a class $\mathcal{H}$ is nonuniformly learnable. From part **1.**, we see that each class $\mathcal{H}_n^A$ has finite VC dimension. Also, it is clear that

$$\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n^A$$

and this proves this part.

3. Let $\mathcal{H}$ be a class that shatters an infinite set. Let $\{\mathcal{H}_n\}$ be a sequence of classes such that $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$. We show that there is some n for which $\text{VCdim}(\mathcal{H}_n) = \infty$.

Suppose the infinite set shattered by $\mathcal{H}$ is K. In addition, suppose $\{\mathcal{H}_n\}$ is a sequence of classes each having a finite VC dimension. We claim that $\mathcal{H} \setminus \bigcup_{n\in\mathbb{N}} \mathcal{H}_n$ is non-empty, and clearly that will prove our claim. We will define a sequence of sets $\{K_n\}$ as follows.

(1) Define K_1 to be any finite subset of K such that $|K_1| > \text{VCdim}(\mathcal{H}_1)$.
(2) Suppose sets $K_1, ..., K_i$ have been defined, each of them being finite and satisfying $|K_j| > \text{VCdim}(\mathcal{H}_j)$ for each $1 \le i \le j$. Consider $K \setminus \bigcup_{j=1}^{i} K_j$; this is still an infinite set. So, choose any finite subset K_{i+1} of $K \setminus \bigcup_{j=1}^{i} K_j$ such that $|K_{i+1}| > \text{VCdim}(\mathcal{H}_{i+1})$.
(3) Choosing the K_is as above ensures that all the K_is are mutually disjoint sets.

Now, take any $n \in \mathbb{N}$, and consider the set K_n. Because $\mathcal{H}_n$ satisfies $\text{VCdim}(\mathcal{H}_n) < |K_n|$, we see that there is some function $f_n : K_n \to \{0, 1\}$ such that no $h \in \mathcal{H}_n$ agrees with f_n on the domain K_n. Since n was arbitrary, we have thus obtained a sequence of functions $\{f_n\}$ such that for any n, there is no hypothesis in $\mathcal{H}_n$ which agrees with f_n on K_n.

Now, consider the disjoint union $\bigsqcup_{n=1}^{\infty} K_n$. It is clear that this disjoint union is a subset of K. Now, consider the function $f' : \bigsqcup_{n=1}^{\infty} K_n \to \{0, 1\}$ defined as follows.

$$f'(x) = f_n(x), \quad \text{if } x \in K_n$$

f' is well defined because of the disjoint union. Now, because K is shattered by $\mathcal{H}$, there is some $f \in \mathcal{H}$ that agrees with f' on $\bigsqcup_{n=1}^{\infty} K_n$. But by our construction, no hypothesis in $\mathcal{H}_n$ for any $n \in \mathbb{N}$ can agree with f'; hence, it follows that $f \in \mathcal{H} \setminus \bigcup_{n=1}^{\infty} \mathcal{H}_n$, and this proves our claim, and also completes the proof.

4. We will now construct a class $\mathcal{H}_1$ of functions from the unit interval $[0, 1]$ to $\{0, 1\}$ that is nonuniformly learnable but not PAC learnable. So, let our domain be $\mathcal{X} = [0, 1]$. For each $n \in \mathbb{N}$, let $\mathcal{H}_n$ denote the class of unions of atmost n intervals; in particular, the class $\mathcal{H}_n$ contains all indicator functions of subsets of $[0, 1]$ which are unions of atmost n closed intervals in $[0, 1]$, i.e

$$\mathcal{H}_n = \{h_{a_1,b_1,...,a_n,b_n} : a_i \le b_i \forall\ i \in [n]\}$$

where

$$h_{a_1,b_1,...,a_n,b_n}(x) = \bigwedge_{i=1}^{n} \mathbf{1}_{x\in[a_i,b_i]}$$

We claim that $\text{VCdim}(\mathcal{H}_n) = 2n$, and let us now prove this. Suppose $z_1 < z_2 < \cdots < z_{2n}$ is any set of points in $[0, 1]$, and consider any labelling of these points. Suppose the labelling is $\{y_1, y_2, ..., y_{2n}\}$. The idea is two group the points into pairs of two; so, we consider the groups $\{z_1, z_2\}, \{z_3, z_4\}, ..., \{z_{2n-1}, z_{2n}\}$. For each group, we use a closed interval to shatter that group. For example, if $\{z_1, z_2\}$ have labels $\{0, 0\}$, we take $[a_1, b_1]$ to be any interval such that $b_1 < z_1$. If the labels are $\{1, 1\}$, we take the interval $[a_1, b_1] = [z_1, z_2]$. Like this, using one interval for each group, we can get the desired labelling. So, it follows that $\text{VCdim}(\mathcal{H}_n) \ge 2n$.

Next, suppose $z_1 < z_2 < \cdots < z_{2n} < z_{2n+1}$ is a set of $2n + 1$ points in $[0, 1]$. We claim that the labelling $\{1, 0, 1, 0, ..., 0, 1\}$ cannot be obtained using the class $\mathcal{H}_n$. Note that to obtain such a labelling, the intervals have to be disjoint. Moreover, the interval $[a_i, b_i]$ will have to contain the point z_{2i-1}, and it cannot contain the points z_{2i-2} or z_{2i}. So then it follows that the last point can never be contained in any of the n intervals; so, this labelling cannot be attained. This shows that $\text{VCdim}(\mathcal{H}_n) = 2n$.

Now, let $\mathcal{H}_1 = \bigcup_{n\in\mathbb{N}} \mathcal{H}_n$. It is clear that $\mathrm{VCdim}(\mathcal{H}_1) = \infty$, and hence $\mathcal{H}_1$ is not PAC learnable. But clearly, because each $\mathcal{H}_n$ has finite VC dimension, we know that $\mathcal{H}_1$ is nonuniformly learnable by a theorem proved in class (namely a hypothesis class is nonuniformly learnable if and only if it can be written as a countable union of hypothesis classes satisfying the uniform convergence property). This completes our construction.

5. Let $\mathcal{H}_2$ be the class of all functions from $[0, 1]$ to $\{0, 1\}$. Clearly, the set $[0, 1]$ is shattered by $\mathcal{H}_2$, and also we know that the set $[0, 1]$ is infinite. So, it must be true that $\mathcal{H}_2$ is not nonuniformly learnable; if $\mathcal{H}_2$ were nonuniformly learnable, then by part **2.** of this problem, we can write $\mathcal{H}_2$ as a union of countably many classes of finite VC dimension. But by part **3.** of this problem, because $\mathcal{H}_2$ shatters an infinite set, some class in this countable union must have infinite VC dimension, which is a contradiction. So, it follows that $\mathcal{H}_2$ is not nonuniformly learnable.

(3) Problem 11.1 of the book. Suppose the the labels are chosen at random according to $\mathbf{P}[y = 1] = \mathbf{P}[y = 0] = 1/2$. Let A be a learning algorithm as given in the problem statement, i.e A returns the constant predictor $h(\boldsymbol{x}) = 1$ if the parity of the labels on the training set is 1 and otherwise the algorithm outputs the constant predictor $h(\boldsymbol{x}) = 0$.

Now suppose S is any training set. So, $A(S)$ will be a constant hypothesis. Now because $A(S)$ is a constant function, we see that

$$L_{\mathcal{D}}(A(S)) = \frac{1}{2}$$

This is simply because for point, $A(S)$ will correctly classify it with probability $1/2$, because the labels are generated using the uniform distribution on two objects. So, it follows that no matter what the set S is, the true error of the output $A(S)$ will always be $1/2$.

Next, we will deal with two cases on the nature of the training set S.

(1) In the first case, suppose that the parity of the labels in S is 1. Fix any singleton subset $\{(\boldsymbol{x}_0, y_0)\} \subset S$ (in other words, this denotes the leave-out set during the 1-fold cross validation step). We have the following two subcases.
   (a) In the first case, $y_0 = 0$, i.e the parity of the labels in $S \setminus \{(\boldsymbol{x}_0, y_0)\}$ is 1. In this case, when the algorithm A is trained on the set $S \setminus \{(\boldsymbol{x}_0, y_0)\}$, it returns the constant hypothesis $A(S)(\boldsymbol{x}) = 1$. In this case, the leave-one-out estimate of $A(S)$ is simply 1 (because it makes an error on the point $\boldsymbol{x}_0$).
   (b) In the second case, we have $y_0 = 1$, i.e the parity of the labels in $S \setminus \{(\boldsymbol{x}_0, y_0)\}$ is 0. In this case, the algorithm A returns the hypothesis $A(S)(\boldsymbol{x}) = 0$ when trained on $S \setminus \{(\boldsymbol{x}_0, y_0)\}$. Again, it follows that the leave-one-out estimate of $A(S)$ in this case is simply 1 (because $A(S)$ makes an error on the point $\boldsymbol{x}_0$).

   Now taking the average over all possible singleton subsets $\{(\boldsymbol{x}_0, y_0)\}$ of S, we see that the estimate of the error of $A(S)$ using leave-one-out validation is 1.
(2) In the second case, we the parity of the labels in S is 0. The same exact analysis as above can be repeated, and in this case too, the estimate of the error of $A(S)$ using leave-one-out validation is 1 again.

So in any case, the leave-one-out error estimate of $A(S)$ is 1. So, it follows that the difference between the error estimate of $A(S)$ and the true error of $A(S)$ is $1 - \frac{1}{2} = \frac{1}{2}$, and this completes the solution of the problem.

(5) Problem 12.2 of the book. Let $\mathcal{H} = \mathcal{X} = \left\{\boldsymbol{x} \in \mathbf{R}^d : ||\boldsymbol{x}|| \leq B\right\}$, where $B > 0$ is some real constant. Let $\mathcal{Y} = \{\pm 1\}$. Let the loss function ℓ be defined as follows.

$$\ell(\boldsymbol{w}, (\boldsymbol{x}, y)) = \log(1 + e^{-y\boldsymbol{w}^T\boldsymbol{x}})$$

We will now show that the resultant learning problem is convex-Lipschitz-bounded and convex-smooth-bounded.

First, let us show that the learning problem is indeed a convex learning problem. To do that, define the function $g : \mathbf{R} \to \mathbf{R}$ as follows.

$$g(z) = \log(1 + e^{-z})$$

We claim that g is a convex function. Note that g is also differentiable on $\mathbf{R}$, with the derivative of g being

$$g'(z) = \frac{-e^{-z}}{1 + e^{-z}} = \frac{-1}{1 + e^{z}}$$

Also, g is infact twice differentiable, and the second derivative of g is the following.

$$g''(z) = \frac{e^{z}}{(1 + e^{z})^2}$$

So, we see that g'' is positive everywhere on $\mathbf{R}$. Hence, by the second derivative test for convexity, it follows that g is convex.

Now, fix the data point $(\boldsymbol{x}, y)$, and consider the loss function ℓ as a function of $\boldsymbol{w}$. We can write

$$\ell(\boldsymbol{w}, (\boldsymbol{x}, y)) = g(y\boldsymbol{w}^T\boldsymbol{x})$$

Then, by a theorem proved in class (which is **Claim 12.4** of the book), we conclude that ℓ is a convex function of $\boldsymbol{w}$. So, this learning problem is really a convex learning problem.

Now, for this fixed data point $(\boldsymbol{x}, y)$, define a function $h : \mathbf{R}^d \to \mathbf{R}$ as follows.

$$h(\boldsymbol{w}) = y\boldsymbol{w}^T\boldsymbol{x}$$

Observe that

$$\nabla_{\boldsymbol{w}} h(\boldsymbol{w}) = y\boldsymbol{x}$$

and hence

$$||\nabla h(\boldsymbol{w})|| = ||\boldsymbol{x}|| \leq B$$

and hence it follows that h is B-Lipschitz (see problem **(7)** of this homework, which is solved later in this document).

Showing Convex-Lipschitz-Boundedness. Note that the hypothesis class $\mathcal{H}$ is bounded by B (by definition), and it is also a convex domain. Next, the function g that we defined above is 1-Lipschitz. This is easy to see because for all $z \in \mathbf{R}$,

$$|g'(z)| = \left|\frac{1}{1 + e^{z}}\right| \leq 1$$

Now, note that

$$\ell(\boldsymbol{w}, (\boldsymbol{x}, y)) = g(h(\boldsymbol{w}))$$

and it follows that ℓ is $1 \cdot B = B$-Lipschitz (by a theorem on the Lipschitzness of a composition of Lipschitz functions). So, it follows that this problem is Convex-Lipschitz-Bounded with parameters B, B.

Showing Convex-Smooth-Boundedness. Again, as above, the hypothesis class $\mathcal{H}$ is bounded by B, and it is also a convex domain. We claim that the function g defined above is $\frac{1}{4}$-smooth. To see this, observe that

$$\begin{aligned} |g''(z)| &= \left|\frac{e^z}{(1+e^z)^2}\right| \\ &= \left|\frac{e^z}{1+2e^z+(e^z)^2}\right| \\ &= \left|\frac{1}{2+e^{-z}+e^z}\right| \\ &\leq \frac{1}{2+2} \\ &= \frac{1}{4} \end{aligned}$$

where above we have used the inequality $e^z + e^{-z} \geq 2$ for all $z \in \mathbf{R}$ (which is a simple implication of the AM-GM inequality). Hence, it follows that g is $\frac{1}{4}$-smooth (by problem **(7)** of this assignment, which is solved later in this document).

Now, again, we know that

$$\ell(\boldsymbol{w}, (\boldsymbol{x}, y)) = g(y\boldsymbol{w}^T\boldsymbol{x})$$

and hence we see that ℓ is $\frac{1}{4} \cdot ||\boldsymbol{x}||^2 = \frac{B^2}{4}$ smooth (this was also mentioned in class, and this is **Claim 12.9** of the book). So, it follows that this problem is a Convex-Smooth-Bounded problem with parameters $\frac{B^2}{4}, B$. This completes the solution to the problem.

(6). Consider the set of $n \times n$ matrices of rank k, where $1 \leq k \leq n$. We will show that this set is *not* convex. The proof is quite simple. Suppose M is a rank k matrix. Clearly, $-M$ is a rank k matrix as well. However, observe that

$$\frac{1}{2}M + \frac{1}{2}(-M) = 0$$

does not have rank k (it has rank 0, and $1 \leq k$). So, this set is not convex.

(7). Let $f : U \to \mathbf{R}$ be a differentiable function, where $U \subseteq \mathbf{R}^n$ is an open convex set. Suppose that $||\nabla f(\boldsymbol{x})||_2 \leq G$ for all $\boldsymbol{x} \in U$. We will show that f is Lipschitz with Lipschitz constant G. So, suppose $\boldsymbol{x}, \boldsymbol{y}$ are any two points in U. Define the map $\gamma : [0,1] \to \mathbf{R}^n$ as follows.

$$\gamma(t) = \boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})$$

Since U is convex, $\gamma(t) \in U$ for all $t \in [0,1]$. Now, let $g : [0,1] \to \mathbf{R}$ be the composition $f \circ \gamma$. Clearly, because both γ and f are differentiable, g is also differentiable on $(0,1)$. Moreover,

$$g'(t) = \nabla f(\gamma(t))^T \gamma'(t) = \nabla f(\gamma(t))^T(\boldsymbol{y} - \boldsymbol{x})$$

So, observe that for all $t \in (0,1)$, we have that

$$|g'(t)| = |\nabla f(\gamma(t))^T(\boldsymbol{y} - \boldsymbol{x})| \leq ||\nabla f(\gamma(t))||_2 \cdot ||\boldsymbol{y} - \boldsymbol{x}||_2 \leq G\,||\boldsymbol{y} - \boldsymbol{x}||_2$$

Also, by the **Mean Value Theorem**, we know that

$$g(1) - g(0) = g'(t)$$

for some $t \in (0,1)$. This means that for some $t \in (0,1)$,

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = g'(t)$$

and hence

$$|f(\boldsymbol{y}) - f(\boldsymbol{x})| \leq G\, ||\boldsymbol{y} - \boldsymbol{x}||_2$$

Since $\boldsymbol{x}, \boldsymbol{y} \in U$ were arbitrary, this shows that f is indeed G-Lipschitz. Infact, the exact same proof holds for functions $f : U \to \mathbf{R}^m$ as well, because there is a version of the **Mean Value Theorem** for differentiable functions $[a,b] \to \mathbf{R}^m$.

(8) Problem 12.4 of the book. Given below are the solutions to the two parts of the problem.

(a) Fix a turing machine T. First, suppose T halts on the input 0. Then, we see that for $h \in [0,1]$,

$$\begin{aligned} \ell(h,T) &= h\ell(0,T) + (1-h)\ell(1,T) \\ &= h \end{aligned}$$

In the second case, suppose T does not halt on input 1. Then, we see that for $h \in [0,1]$, we have

$$\begin{aligned} \ell(h,T) &= h\ell(0,T) + (1-h)\ell(1,T) \\ &= (1-h) \end{aligned}$$

In either case, $\ell(h,T)$ is a linear function over $\mathcal{H}$, and hence it is convex. Moreover, the derivative of $\ell(h,T)$ is always bounded above by 1 (easy to see from the above two formulae), and hence ℓ is 1-Lipshitz. Finally, $\mathcal{H} = [0,1]$, and it is trivially bounded. So, this problem is a Convex-Lipschitz-Bounded problem.

(b) Couldn't do this.