Computational and Statistical Learning Theory

Problem set 2

Due: October 17th

Please send your solutions to `learning-submissions@ttic.edu`

Notation :

Input space : $\mathcal{X}$	Label space : $\mathcal{Y} = \{\pm 1\}$	Sample : $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathcal{Y}$
Hypothesis Class : $\mathcal{H}$	Risk : $L(h) = \mathbb{E}\left[\mathbf{1}_{h(x) \neq y}\right]$	Empirical Risk : $\hat{L}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{h(x_i) \neq y_i}$

1. **Shatter Lemma :**
Given a set $S = \{x_1, \ldots, x_m\}$ let $\mathcal{H}_{x_1,\ldots,x_m} = \{(h(x_1), \ldots, h(x_m)) \in \{\pm 1\}^m : h \in \mathcal{H}\}$. Recall that we say that such a set is *shattered* by $\mathcal{H}$ if $|\mathcal{H}_{x_1,\ldots,x_m}| = 2^m$, and that the VC dimension of $\mathcal{H}$ is the size of he largest sample that can be shattered. Also recall that the *growth function* of the hypothesis class $\mathcal{H}$ is given by:

$$\Pi_{\mathcal{H}}(m) = \sup_{x_1,\ldots,x_m} |\mathcal{H}_{x_1,\ldots,x_m}| .$$

That is, we can also define the VC dimension as the largest m for which $\Pi_{\mathcal{H}}(m) = 2^m$.

The aim of this exercise is to prove the "Shatter Lemma": if $\mathcal{H}$ has VC dimension d, then for any m,

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}. \tag{1}$$

In order to prove (1), we will actually prove the following statement: for any set $S = \{x_1, \ldots, x_m\}$:

$$|\mathcal{H}_S| \leq |\{B \subset S : B \text{ is shattered by } \mathcal{H}\}| \tag{2}$$

That is, the number of possible labeling of a S is bounded by the number of different subsets of S that can be shattered.

We (i.e. you) will prove (2) by induction.

(a) Establish that (2) holds for $S = \emptyset$ (the empty set).

(b) For any set S and any point $x' \notin S$, assume (2) holds for S and for any hypothesis class, and prove that (2) holds for $S' = S \cup \{x'\}$ and any hypothesis class. To this end, for any hypothesis class $\mathcal{H}$, write $\mathcal{H} = \mathcal{H}^- \cup \mathcal{H}^+$ where:

$$\mathcal{H}^+ = \{h \in \mathcal{H} : h(x') = +1\}$$
$$\mathcal{H}^- = \{h \in \mathcal{H} : h(x') = -1\}$$

i. Prove that $|\mathcal{H}_{S'}| = \left|\mathcal{H}_S^+\right| + \left|\mathcal{H}_S^-\right|$.

ii. Prove that (2) holds for S and $\mathcal{H}$ by applying (2) to each of the two terms on the right-hand-side above.

We can now conclude that (2) holds for any (finite) S and any $\mathcal{H}$.

(c) Use (2) to establish (1).

(d) For $d \leq n$, prove that $\sum_{i=0}^{d} \binom{m}{i} \leq m^d$. **Optional:** Prove the tighter bound: $\sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d$

2. **VC Dimension :**

(a) Consider the hypothesis class $\mathcal{H}_\bullet$ of positve circles in $\mathbb{R}^2$. That is set of all hypothesis that are positive inside some circle and negative outside. Calculate the VC dimension of this class, and show that this is the exact value of the VC dimension.

(b) Consider the hypothesis class $\mathcal{H}_\circ$ of both positive and negative circles in $\mathbb{R}^2$. That is set of all hypothesis that are positive inside some circle and negative outside and all hypothesis that are negative inside that circle and positive outside. Show how to shatter 4 points using this class and establish a lower bound of 4 on the VC dimension of the class.

We now consider the VC dimension of the class $\mathcal{H}_d$ of linear separators in $\mathbb{R}^d$:

$$\mathcal{H}_d = \left\{x \mapsto \text{sign}(w^\top x + b)\middle| w \in \mathbb{R}^d, b \in \mathbb{R}\right\}$$

(c) Consider the set of $d+1$ points that include origin and the d bases e_i (ie. 1 on ith co-ordinate and 0 elsewhere). Show that the points can be shattered by $\mathcal{H}_d$.

(d) Prove that no set of $d+2$ points can be shattered by $\mathcal{H}_d$.
(Hint : Use Radon's theorem which states that any set of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two disjoint sets whose convex hulls intersect.)

From this we conclude that the VC dimension of $\mathcal{H}_d$ is exactly $d+1$.

(e) Prove that for any $\mathcal{H}_1 \subseteq \mathcal{H}_2$, the VC dimension of $\mathcal{H}_1$ is not larger than that of $\mathcal{H}_2$.

(f) Use the above to prove that if for some hypothesis class $\mathcal{H}$, there exists a feature map $\phi : \mathcal{X} \mapsto \mathbb{R}^d$ such that any hypothesis $h \in \mathcal{H}$ can be written as

$$h(x) = \text{sign}\left(\sum_{i=1}^{d} w_i \phi_i(x) + b\right)$$

for some $w \in \mathbb{R}^d$ and some $b \in \mathbb{R}$, then VC dimension of $\mathcal{H}$ is at most $d+1$.

(g) Use this to obtain a tight upper bound on the VC-dimension of $\mathcal{H}_\circ$ and conclude that the VC-dimension of this class is indeed four. Note that the bound you can get on $\mathcal{H}_\bullet$ is not tight.

3. **Hoeffding Bounds :**

We will now see how to use systematization to obtain a learning guarantee that depends on the growth function, and hence on the VC dimension.

(a) For any sequence of $2m$ points $S = (z_1, ...z_m, z'_1, ..., z'_m)$, consider m i.i.d. uniform random signs s_1,. . . ,s_m which define the samples S_1, S_2 in the following way: for each $i = 1..m$, if $s_i = 1$ then $z_i \in S_1$ and $z'_i \in S_2$, otherwise (if $s_i = -1$) then $z_i \in S_2$ and $z'_i \in S_1$; i.e. the variables s_1,. . . ,s_m specify how to "deal" the $2m$ points into the two sets S_1 and S_2. Now, for any sequence S of $2m$ points, and any hypothesis h, with $l(h, z) \in \{0, 1\}$, prove that with probability $\geq 1 - \delta$ over the separation to S_1,S_2:

$$|L_{S_1}(h) - L_{S_2}(h)| \leq \sqrt{f(\delta)/m} \tag{3}$$

(Hint: write $L_{S_1}(h) - L_{S_2}(h) = \frac{1}{m}\sum_{i=1}^{m}(-1)^{s_i}(l(h, z'_i) - l(h, z_i)))$)

(b) For any sequence S of $2m$ points as above, prove that with probability $\geq 1 - \delta$ over the separation to S_1,S_2, for every $h \in \mathcal{H}$:

$$|L_{S_1}(h) - L_{S_2}(h)| \leq \sqrt{f(\delta, \Pi_{\mathcal{H}})/m)} \tag{4}$$

and conclude that the same inequality holds with probability $\geq 1 - \delta$ over $S_1, S_2 \sim$ i.i.d.D^m, and every $h \in \mathcal{H}$.

(c) Recall the symmetrization lemma:

$$P_{S \sim D^m}\left(\exists_{h \in \mathcal{H}} |L(h) - L_S(h)| > 2\epsilon\right) \leq 2P_{S,S' \sim D^m}\left(\exists_{h \in \mathcal{H}} |L_S(h) - L_{S'}(h)| > \epsilon\right) \tag{5}$$

Use this, and part (c) above, to prove that, with probability $\geq 1 - \delta$ over $S \sim D^m$, for all $h \in \mathcal{H}$:

$$|L_S(h) - L(h)| < \sqrt{f(\delta, \Pi_{\mathcal{H}})/m} \tag{6}$$

and conclude that if VC-dim($\mathcal{H}$) $\leq D$ then:

$$L(\hat{h}) \leq L(h^*) + \sqrt{f(\delta, D\log(2em/D))/m} \tag{7}$$

(d) Conclude that $m = O(D\log(1/\epsilon)/\epsilon^2)$ samples are enough to ensure that with probability $\geq 1 - \delta$, $L(\hat{h}) < L(h^*) + \epsilon$. Write down an explicit bound (without big-O notation, though the constants need not be the tightest possible).

(Hint: start with the expression for m, plug it into the r.h.s. of part (c) above, and verify that the r.h.s is less than ϵ.)

4. **Bernstein Bounds :**

   (a) Use Bernstein inequality to prove that for any $0 \leq \delta \leq 1$ and any $h \in \mathcal{H}$ with probability greater than $1-\delta$, the following bound holds where m is the number of samples. Specify functions $f(\delta)$ and $g(\delta)$.

$$|\hat{L}(h) - L(h)| \leq \frac{f(\delta)}{m} + \sqrt{\frac{g(\delta)L(h)}{m}} \tag{8}$$

   (b) i. Use the union bound to prove that with probability greater than $1-\delta$, the following bound holds for all $h \in \mathcal{H}$. Find $f(\delta, |\mathcal{H}|)$ and $g(\delta, |\mathcal{H})|$.

$$|\hat{L}(h) - L(h)| \leq \frac{f(\delta, |\mathcal{H}|)}{m} + \sqrt{\frac{g(\delta, |\mathcal{H}|)L(h)}{m}} \tag{9}$$

   ii. Use the bound on the previous part to prove that for any $0 \leq \delta \leq 1$ with probability greater than $1-\delta$. Specify functions $f(\delta, |\mathcal{H}|)$, $g_1(\delta, |\mathcal{H})|$ and $g_2(\delta, |\mathcal{H})|$.

$$L(\hat{h}) \leq L(h^*) + \frac{f(\delta, |\mathcal{H}|)}{m} + \sqrt{\frac{g_1(\delta, |\mathcal{H}|)L(\hat{h})}{m}} + \sqrt{\frac{g_2(\delta, |\mathcal{H}|)L(h^*)}{m}} \tag{10}$$

   iii. Solve the above inequality for $L(\hat{h})$ to find the following bound. Find $f(\delta, |\mathcal{H}|)$ and $g(\delta, |\mathcal{H})|$.

$$L(\hat{h}) \leq L(h^*) + \frac{f(\delta, |\mathcal{H}|)}{m} + \sqrt{\frac{g(\delta, |\mathcal{H}|)L(h^*)}{m}} \tag{11}$$

   (c) Find a lower bound on the number of samples for the inequality 11 to be hold with probability greater than $1-\delta$.

   (d) Prove that with probability greater than $1-\delta$ the following bound holds for any $a > 0$ and all $h \in \mathcal{H}$.

$$L(h) \leq (1+a)L(h^*) + (1+1/a)f(\delta, \mathcal{H})/m \tag{12}$$

   This means that as long as we want a constant factor approximation of $L(h^*)$, e.g. we want error that is $1.1L(h^*)$, we get a "rate" of $1/m$. (Hint: write this as $L(h) \leq \inf_a (1+a)L(h^*) + (1+1/a)f(\delta, \mathcal{H})/m$ and optimize over a.)

Challenge Problems

- VC bounds :

  1. Combine ideas from Problems 3 and 4 to show that for any class $\mathcal{H}$ with VC-dim$(\mathcal{H}) \leq D$, with probability $\geq 1-\delta$ over $S \sim D^m$:

$$L(\hat{h}) < L(h^*) + O\Bigg(\frac{Dlog(m/D) + \log(1/\delta)}{m} + \sqrt{\frac{(D\log(m/D) + \log(1/\delta))L(h^*)}{m}}\Bigg) \tag{13}$$

2. Show that this means that to get $L(\hat{h}) < L(h^*) + \epsilon$, we need

$$m = O\left(\frac{d \log(1/\epsilon\delta)}{\epsilon} . \frac{L^* + \epsilon}{\epsilon}\right) \tag{14}$$

- VC dimension of decision trees :

1. Prove a learning guarantee for decision trees of size k (i.e. having at most k leaves) over an input space of n binary variables, where each decision is over a single binary variable.
2. For the input space $\mathcal{X} = \mathbb{R}^d$, provide an upper bound (that is as tight as possible) on the VC dimension of the class of stumps

$$\mathcal{H} = \{x \mapsto \text{sign}\,(ax_i - b) : i \in [d], b \in \mathbb{R}, a \in \pm 1\}$$

3. For the input space $\mathcal{X} = \mathbb{R}^d$, provide an upper bound (that is as tight as possible) on the VC dimension of the class of decision trees of size k where each decision is based on a stump from $\mathcal{H}$.

Research Problem :

- Show that any hypothesis class with VC-dimension d has a compression scheme of size d. There is a 600 dollar prize on this problem.