Computational and Statistical Learning Theory

Problem set 1

Due: Monday, October 10th

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

Notation:

- Input space: $\mathcal{X}$
- Label space: $\mathcal{Y} = \{\pm 1\}$
- Sample: $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X}$
- Hypothesis Class: $\mathcal{H}$
- Risk: $L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}_{h(x)\neq y}\right]$
- Empirical Risk: $L_S(h) = \frac{1}{m}\sum_{(x,y)\in S}\mathbf{1}_{h(x)\neq y}$

1. **Binomial Tail Bounds:**
   $\{0,1\}$-valued random variables $X_1, \ldots, X_n$ are drawn independently each from Bernoulli distribution with parameter $p = 0.1$. Define $P_n := \mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i \leq 0.2)$.

   (a) For $n = 1$ to 30 calculate and plot the below in the same plot (see [1, section 6.1] for definition of Hoeffding and Bernstein inequalities):

      i. Exact value of P_n (binomial distribution).
      ii. Normal approximation for P_n.
      iii. Hoeffding inequality bound on P_n.
      iv. Bernstein inequality bound on P_n.
      v. (Optional) Bound using relative entropy (see [2]).

   (b) For $n = 30$ to 300 calculate and plot the below in the same plot :

      i. Normal approximation for P_n.
      ii. Hoeffding inequality bound on P_n.
      iii. Bernstein inequality bound on P_n.
      iv. (Optional) Bound using relative entropy.

   (c) Try above bounds under the following settings:

      i. $p = 0.4$ and $P_n := \mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i \leq 0.45)$.
      ii. $p = 0.01$ and $P_n := \mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i \leq 0.02)$.

2. **Improve bound in realizable case:**
For the agnostic case we saw that $\forall^{\delta}_{S\sim\mathcal{D}^m}, \forall_{h\in\mathcal{H}}$, we have

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2m}},$$

in this problem we will consider the realizable case and show how to get a dependence on $1/m$ instead of $1/\sqrt{m}$.

   (a) Prove
   $$\forall h, \quad \mathbb{P}(L_S(h) = 0) \leq e^{-L_{\mathcal{D}}(h)\cdot m}.$$
   (Hint : for $t > 1$, $(1 - 1/t)^t \leq 1/e$).

   (b) Use above bound and a union bound to prove
   $$\forall^{\delta}_{S\sim\mathcal{D}^m}, \forall_{h\in\mathcal{H},\text{s.t.}L_S(h)=0}, \quad L_{\mathcal{D}}(h) \leq \frac{\log|\mathcal{H}| + \log(1/\delta)}{m}$$

   (c) Use this to obtain a PAC learning guarantee for the realizable case (i.e. assume $\exists_{h\in\mathcal{H}} L_{\mathcal{D}}(h) = 0$), wrote down the corresponding sample complexity.

   (d) (Bonus) Prove a more general form that interpolates the realizable and agnostic case using Bernstein inequality, in the following form, specify $f(\delta, |\mathcal{H}|)$ and $g(\delta, |\mathcal{H}|)$.
   $$L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{D}}(h^*) + \frac{f(\delta, |\mathcal{H}|)}{m} + \sqrt{\frac{g(\delta, |\mathcal{H}|)L_{\mathcal{D}}(h^*)}{m}}. \tag{1}$$

3. **VC Dimension of Sparse Linear Classifiers Using Concept Class Union Arguments:**
Fix an instance space $\mathcal{X}$. Let $\{\mathcal{H}_i\}$ be a family of concept classes over $\mathcal{X}$, where i ranges from 1 to some integer r. Define the concept $\mathcal{H}$ to be the union $\cup_{i=1\dots r}\mathcal{H}_i$.

   (a) Give an upper bound for the growth function $\Gamma_m(\mathcal{H})$ in terms of the growth functions $\Gamma_m(\mathcal{H}_i)$ for $i = 1 \dots r$.

   (b) If the VC dimension of all $\mathcal{H}_i$ are bounded by D, i.e. $\text{VCdim}(\mathcal{H}_i) \leq D$, give a bound on the growth function $\Gamma_m(\mathcal{H})$ in terms of m, r and D.
   Hint: Use Sauer's lemma.

   (c) From (b) conclude that $\text{VCdim}(\mathcal{H}) = O(\max(D, \log(r) + D\log(\log(r)/D)))$. Give an exact bound, without the $O(\cdot)$ notation.

For the remainder or the question, consider a feature mapping $\phi : \mathcal{X} \mapsto \mathbb{R}^d$. For a vector $w \in \mathbb{R}^d$, we say that w is k-sparse if the number of coordinates j for which $w(j) \neq 0$ is at most k. We Define the following concept class:

$$\mathcal{H}^{(k)} := \{h_w(x) = \text{sign}\left(\langle w, \phi(x)\rangle\right) | \text{ s.t. } w \text{ is } k\text{-sparse}\}.$$

   (d) Using (c) above, show that the VC dimension of $\mathcal{H}^{(k)}$ is at most $O(k\log(d/k))$.

(e) Show how to shatter a subset of size $\Omega(\log d)$ with respect to $\mathcal{H}^{(1)}$, establishing tight upper and lower bounds on the VC dimension of $\mathcal{H}^{(1)}$.

(f) (Bonus) Show how to shatter a subset of size $\Omega(k\log(d/k))$ with respect to $\mathcal{H}^{(k)}$, establishing tight upper and lower bounds on the VC dimension of $\mathcal{H}^{(k)}$.

4. **Tutorial walkthrough on deriving the VC Bound (will not be graded):**
Given a set $C = \{x_1, \ldots, x_m\}$ 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 the largest sample set that can be shattered. Also recall that the *growth function* of the hypothesis class $\mathcal{H}$ is given by:

$$\Gamma_{\mathcal{H}}(m) = \sup_{x_1,\ldots,x_m} |\mathcal{H}_{x_1,\ldots,x_m}| \,. \tag{2}$$

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

In this problem, we will see how to obtain a uniform convergence guarantee, bounding the differences between empirical and expected errors, in terms of the growth function. We already know how uniform convergence ensures learning guarantees for ERM, and so we can obtain learning guarantees in terms of the growth function, and thus using 2 also in terms of the VC dimension.

The growth function bounds the number of behaviors of the hypothesis class on a finite number of points, while the expected error depends on the behavior on all the points. To overcome this, we will first bound the difference between two the empirical error on two different samples, and then show that the difference between the expected and empirical errors cannot be much larger than the difference between two different empirical errors. This technique is called "systematization": we are introducing a "ghost sample" of another m points sampled i.i.d. from the source distribution to stand in for the expected error and make the problem symmetric.

We will use the following Sauer's Lemma: if $\mathcal{H}$ has VC dimension d, then for any m,

$$\Gamma_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d \tag{3}$$

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 $S = (z_1, \ldots z_m, z'_1, \ldots, z'_m)$, consider m i.i.d. uniform random signs $s_1,\ldots,s_m$ which define the samples S_1, S_2 in the following way: for each $i = 1 \ldots m$, 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 $s_1,\ldots,s_m$ 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 $\ell(h, z) \in \{0, 1\}$, prove that with probability $\geq 1 - \delta$ over the separation to S_1, S_2:

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

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

(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 $h \in \mathcal{H}$:

$$|L_{S_1}(h) - L_{S_2}(h)| \leq \sqrt{f(\delta, \Gamma_{\mathcal{H}}(2m))/m)} \tag{5}$$

and conclude that for any hypothesis class $\mathcal{H}$, any source distribution $\mathcal{D}$ and any sample size m, with probability at least $1-\delta$ over $S_1, S_2 \overset{\text{i.i.d}}{\sim} \mathcal{D}^m$ (i.e. each sample is drawn independently from $\mathcal{D}^m$), for all $h \in H$,

$$|L_{S_1}(h) - L_{S_2}(h)| \leq \sqrt{f(\delta, \Gamma_{\mathcal{H}}(2m))/m)} \tag{6}$$

(c) Prove the symmetrization lemma; i.e. prove that for any hypothesis class $\mathcal{H}$, any source distribution $\mathcal{D}$, any $\epsilon > 0$ and any sample size m such that $m \geq \frac{1}{2\epsilon^2}$:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\exists_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)| > 2\epsilon\right) \leq 2\mathbb{P}_{S,S'\sim\mathcal{D}^m}\left(\exists_{h\in\mathcal{H}}|L_S(h) - L_{S'}(h)| > \epsilon\right) \tag{7}$$

Hint: Prove the above inequality is two steps:

i. Show that for any S and S', the following inequality holds:

$$\mathbf{1}_{\exists_{h\in\mathcal{H}}|L_{\mathcal{D}}(h)-L_S(h)|>2\epsilon}\mathbf{1}_{\exists_{h\in\mathcal{H}}|L_{\mathcal{D}}(h)-L_{S'}(h)|<\epsilon} \leq \mathbf{1}_{\exists_{h\in\mathcal{H}}|L_S(h)-L_{S'}(h)|>\epsilon}$$

ii. Take the expectation with respect to S' and use Chebychevs inequality to bound $\mathbb{P}(\exists_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_{S'}(h)| < \epsilon)$ and then take the expectation with respect to S. Chebychev's inequality bounds the probability of deviation from the expected value:

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq t\right) \leq \frac{\text{Var}(X)}{t^2}$$

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

$$|L_S(h) - L_{\mathcal{D}}(h)| < \sqrt{f(\delta, \Gamma_{\mathcal{H}}(2m))/m} \tag{8}$$

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

$$L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{D}}(h^*) + \sqrt{f(\delta, D\log(2em/D))/m} \tag{9}$$

(e) Conclude that $m = O(D\log(1/\epsilon)/\epsilon^2)$ samples are enough to ensure that with probability $\geq 1-\delta$, $L_{\mathcal{D}}(\hat{h}) < L_{\mathcal{D}}(h^*) + \epsilon$. 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 (d) above, and verify that the r.h.s is less than ϵ.

(f) (Optional) Show how to use Bernstein's inequality instead of Hoefding's to obtain the bound.

References

[1] O. Bousquet, S. Boucheron, and G. Lugosi. *Introduction to statistical learning theory*. Advanced Lectures on Machine Learning, pp. 169-207. Springer Berlin Heidelberg, 2004.

[2] J. Langford. *Tutorial on practical prediction theory for classification*. Journal of machine learning research, pp. 273-306, 2005