Computational and Statistical Learning theory

Problem set 2

Due: January 31st
Email solutions to : karthik *at* ttic *dot* edu

Notation :
Input space : $\mathcal{X}$ Label space : $\mathcal{Y} = \{\pm 1\}$ Sample : $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$

Hypothesis Class : $\mathcal{H}$ Risk : $R(h) = \mathbb{E}\left[\mathbf{1}_{h(x) \neq y}\right]$ Empirical Risk : $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{m} \mathbf{1}_{h(x_i) \neq y_i}$

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:

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

   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. **Description-Length Based Structural Risk Minimization :**
In this problem we will consider more carefully an analysis of a slightly cleaner MDL-based SRM learning rule.

Recall that for any distribution p over hypotheses in $\mathcal{H}$ and any $\delta > 0$, with probability at least $1 - \delta$ over the sample $S := (x_1, y_1), \ldots, (x_n, y_n)$, for all $h \in \mathcal{H}$:

$$R(h) \leq \widehat{\hat{R}}(h) + \sqrt{\frac{\log \frac{1}{p(h)} + \log \frac{1}{\delta}}{2n}} \tag{3}$$

With the above in mind, for a prior distribution $p(\cdot)$, define the following learning rule:

$$\text{SRM}_p(S) = \tilde{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}(h) + \sqrt{\frac{\log \frac{1}{p(h)}}{2n}}$$

Prove that for any $h^\star \in \mathcal{H}$, any $\epsilon > 0$ and $\delta > 0$, with probability at least $1 - \delta$ over sample S of size :

$$m > \frac{\log \frac{1}{p(h^\star)} + 4 \log \frac{1}{\delta}}{\epsilon^2}$$

we will have $R(\text{SRM}_p(S)) \leq R(h^\star) + \epsilon$.
(Hint: you may find the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \leq \sqrt{2a + 2b}$ useful)

4. **VC versus MDL :**
Using the SRM rule above, for any countable hypothesis class $\mathcal{H}$, we can get arbitrary close to the generalization error of any $h \in \mathcal{H}$. But we saw that the VC-dimension correctly captures the complexity of a hypothesis classes, and that in some sense, no learning guarantee is possible for classes with infinite VC dimension. We will now try to understand why there is no contradiction here.

Let $\mathcal{X}$ be the interval $(0, 1]$. For any integer $r > 0$, consider the hypothesis class:

$$\mathcal{H}_r = \{\text{all binary functions of } \phi_r(x) = \lceil r \cdot x \rceil\}\,.$$

Our hypothesis class will be an infinite union of such classes. To make things a bit simpler, we consider only resolutions r that are integers power of two, that is:

$$\mathcal{H} = \bigcup_{q=1}^{\infty} \mathcal{H}_{2^q}.$$

(a) Show that $\mathcal{H}$ has infinite VC-dimension.

(b) Suggest either a binary description language for $\mathcal{H}$ or a distribution over it. It is OK if multiple descriptions refer to the same function, or if you prefer assigning probability to multiple functions that are actually the same one. But be sure that every hypothesis in $\mathcal{H}$ has a description or positive probability mass.

(c) Show that the VC dimension of $\mathcal{H}$ is in fact ∞.

(d) We first establish that the ERM is not appropriate here. Consider a source distribution which is uniform on $\mathcal{X}$ and for which:

$$y = \begin{cases} +1 & \text{if } x < 0.3473 \\ -1 & \text{otherwise} \end{cases}$$

Show that for any sample, and any $\epsilon > 0$, there exists a hypothesis $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ but $R(h) > 1 - \epsilon$.

(e) The ERM is not appropriate, but SRM$_p$ is. Calculate an explicit number n_0 (we are looking for an actual number here, not an expression), such that for with probability at least 0.99 over sample of size $n > n_0$, we will have SRM$_p(S) < 0.1$, where p is the prior (description language) you suggested above.

(Don't turn in) Think of a different description language or prior distribution for the same class $\mathcal{H}$ that would require a much smaller training set size to achieve a generalization error of 0.1 for the above source distribution. Then think of an example of a source distribution for which the new description language or prior distribution would require a much larger training set size then p to achieve low generalization error.

(f) For any prior distribution p, and any sample size $m > 0$, construct a source distribution such that there exists $h \in \mathcal{H}$ with $R(h) = 0$, but with probability at least 0.2 over a sample of size m, $R(\text{SRM}_p(S)) > 0.2$ (it is actually possible to get $R(\text{SRM}_p(S)) = 0.5$ with probability close to one).

5. **VC + MDL :**
MDL bounds are applicable for countable classes and VC bounds to possible uncountable classes with finite VC dimension. What about continuous classes with infinite VC dimensions? As an example consider the class of all polynomial threshold functions. Can we get learning guarantees for this class ?

Example : Consider input space $\mathcal{X} \subseteq \mathbb{R}$ and the hypothesis class

$$\mathcal{H}_{\text{poly}} = \{x \mapsto \text{sign}(f(x) < 0) : f \text{ is a polynomial function}\}.$$

This function class is uncountable and has infinite VC dimension (Eg. any binary function can be approximated by polynomial functions). However it is possible to get learning gaurantees, to do so we use the key observation that $\mathcal{H}_{\text{poly}} = \bigcup_{d=0}^{\infty} \mathcal{H}_{\text{poly}_d}$ where $\mathcal{H}_{\text{poly}_d}$ is the class of all polynomials of degree d. Note that by Problem 2(f) we have that VC dimension of $\mathcal{H}_d$ is bounded by $d + 1$.

We (i.e. you) shall prove learning guarantees for general hypothesis classes that can be written as countable union of classes with finite VC dimension. Consider:

$$\mathcal{H} = \bigcup_{d=1}^{\infty} \mathcal{H}_d$$

where $\mathcal{H}_d$ is a hypothesis class with VC dimension d.

(a) Prove a generalization error bound of the following form: For any $\delta > 0$, with probability at least $1 - \delta$ over sample of size m, for all $h \in \mathcal{H}$:

$$R(h) \leq \hat{R}(h) + \epsilon(m, \delta, d(h))$$

where:

$$d(h) = \min d \text{ s.t. } h \in \mathcal{H}_d$$

and for any δ and d, $\epsilon(m, \delta, h) \overset{m\to\infty}{\to} 0$. Be sure to specific $\epsilon(m, \delta, h)$ explicitly.

(b) Write down a learning rule $\text{SRM}_{\mathcal{H}}$ that guarantees that for any ϵ, δ and $h \in \mathcal{H}$, there exist $m(h)$ such that for any source distribution, with probability at least $1 - \delta$ over a sample of size m, $R(\text{SRM}_{\mathcal{H}}(S)) < R(h) + \epsilon$.

Challenge Problems :

1. **VC dimension of decision trees :**

   (a) 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.

   (b) 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

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

   (c) 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 $\mathcal{H}$.

2. **Refine PAC-Bayes**

   (a) Prove the refined PAC-Bayes bound: for any $\delta > 0$ and prior p, with probability at least $1 - \delta$ over the sample of size m, for any sample dependent distribution q over hypothesis,

   $$\text{KL}\left(R(q)||\hat{R}(q)\right) \leq \frac{\text{KL}\,(q\|p) + \log 2m + \log \frac{1}{\delta}}{m - 1}$$

   where recall that $R(q) = \mathbb{E}_{h\sim q}\left[R(h)\right]$ and $\hat{R}(q) = \mathbb{E}_{h\sim q}\left[\hat{R}(h)\right]$ are the risk and empirical risk of the randomized predictor defined by q.

(b) Show that close to $\hat{R}(q) = 0$ the above bound behaves as $1/m$ and far away from $\hat{R}(q) = 0$ it behaves as $1/\sqrt{m}$, matching our realizable and non-realizable bounds.

Research Problems :

1. Show how a VC-based learning guarantee can be obtained from the PAC-Bayes bound. That is, for any class with VC dimension d, describe a prior p and a learning rule that returns a distribution (randomized hypothesis) q, for which the PAC-Bayes bound guarantees $R(q) \leq \inf_{h \in \mathcal{H}} R(h) + \tilde{O}\left(\sqrt{d/\epsilon}\right)$.

2. 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.