Computational and Statistical Learning Theory

Problem set 3

Due: October 24th

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. **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 $h \in \mathcal{H}$:

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

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

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

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

2. **VC versus MDL :**
   Using the SRM rule above, for any countable hypothesis class $\mathcal{H}$, we can find a hypothesis with the generalization error arbitrary close to the generalization error of the best hypothesis

$h^* \in \mathcal{H}$. However, we saw earlier that the VC-dimension correctly captures the complexity of a hypothesis classes and no learning guarantee is possible for classes with infinite VC dimension. We will investigate why this is not a contradiction.

Let x be in 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) 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 $\hat{L}(h) = 0$ but $L(h) > 1 - \epsilon$.

(d) 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 $\text{SRM}_p(S) < 0.1$, 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.

(e) **(Optional)** Show how to construct a source distribution (as a function of $p(.)$ and m) such that there exists $h \in \mathcal{H}$ with $L(h) = 0$, but with probability at least 0.2 over a sample of size m, $L(\text{SRM}_p(S)) > 0.2$ (it is actually possible to get $L(\text{SRM}_p(S)) = 0.5$ with probability close to one).

3. **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 $h \in \mathcal{H}$:

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

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, $L(\text{SRM}_{\mathcal{H}}(S)) < L(h) + \epsilon$.

4. **PAC-Bayesian Theorem :**
The PAC-Bayesian Theorem which is a generalization of the cardinality bound and description length bound states that for any fixed prior distribution P on $\mathcal{H}$ and any $0 \leq \delta \leq 1$ the following statement holds with probability greater than $1 - \delta$ over S:

$$\forall Q \quad \text{KL}(\hat{L}_S(Q)||L(Q)) \leq \frac{\text{KL}(Q||P) + \log \frac{2m}{\delta}}{m - 1} \tag{2}$$

where Q is a distribution over $\mathcal{H}$, $\text{KL}(p||q) = p \ln(p/q) + (1 - p) \ln((1 - p)/(1 - q))$ is KL-divergence and we use $f(P)$ to denote $E_{x\sim P}[f(x)]$ for any probability distribution P.

In this problem we investigate the proof of PAC-Bayesian Theorem via two major steps:

(a) In this step, we aim to show that for any fix prior distribution P on $\mathcal{H}$:

$$(m-1)\mathrm{KL}(\hat{L}_S(Q)||L(Q)) \leq \mathrm{KL}(Q||P) + \ln E_{h\sim P}\left[e^{(m-1)\mathrm{KL}(\hat{L}_S(Q)||L(Q))}\right] \quad (3)$$

i. Prove the following inequality using Jensen's inequality and strong convexity of KL-divergence.

$$(m-1)\mathrm{KL}(\hat{L}_S(Q)||L(Q)) \leq E_{h\sim Q}[(m-1)\mathrm{KL}(\hat{L}_S(h)||L(h))] \quad (4)$$

ii. Show that for any function $f(x)$ and any distributions P and Q over the domain of x the following equation holds:

$$E_{x\sim Q}[f(x)] = \mathrm{KL}(Q||P) + E_{x\sim Q}\left[\ln \frac{dP(x)}{dQ(x)} e^{f(x)}\right] \quad (5)$$

iii. Show that the following inequality holds for any function $g(x)$ and any distributions P and Q over the domain of x:

$$E_{x\sim Q}\left[\ln \frac{dP(x)}{dQ(x)} g(x)\right] \leq \ln E_{x\sim P}[g(x)]. \quad (6)$$

iv. Use previous parts to conclude that inequality 3 holds.

(b) At this point, we want to an prove that for any fix probability distribution P on $\mathcal{H}$ and any $0 \leq \delta \leq 1$, the following upper bound holds with probability greater than $1-\delta$:

$$E_{h\sim P}\left[e^{(m-1)\mathrm{KL}(\hat{L}_S(Q)||L(Q))}\right] \leq \frac{2m}{\delta} \quad (7)$$

i. Prove that for any real valued random variable X satisfying $P(X \leq \epsilon) \leq e^{-mf(x)}$ where $f(x)$ is non-negative, the following inequality holds:

$$E[e^{(m-1)f(X)}] \leq m \quad (8)$$

Hint : First show that $P(e^{(m-1)f(x)} \geq \alpha) \leq \min(1, \alpha^{-m/(m-1)})$. Then Use the general fact that $E[Y] = \int_0^\infty P(Y \geq x)dx$ to prove inequality 8.

ii. Chernoff-Hoeffding bound states that for i.i.d random variables $X_1, \ldots, X_m$ from the interval $[0,1]$ if $\bar{X} = \sum_{i=1}^m (X_i/m)$ then for any $\epsilon \in [0,1]$ we have the following:

$$P(\bar{X} \leq \epsilon) \leq e^{-m\mathrm{KL}^+(\epsilon||E[X_i])} \quad (9)$$

where $\mathrm{KL}^+(p||q)$ is zero if $p \geq q$ and $\mathrm{KL}(p||q)$ otherwise. Prove the following inequality for a fixed $h \in \mathcal{H}$ using part (i) and Chernoff-Hoeffding bound.

$$E_{S\sim D^m}\left[e^{(m-1)\mathrm{KL}^+(\hat{L}_S(h)||L(h))}\right] \leq m \quad (10)$$

iii. The above inequality implies:

$$E_{S\sim D^m}\left[E_{h\sim\mathcal{P}}[e^{(m-1)\mathrm{KL}^+(\hat{L}_S(h)||L(h))}]\right] \leq m \tag{11}$$

Use Markov's inequality to prove that for any $\delta \in [0,1]$, we have the following inequality with probability greater than $1-(\delta/2)$ over S:

$$E_{h\sim\mathcal{P}}[e^{(m-1)\mathrm{KL}^+(\hat{L}_S(h)||L(h))}] \leq \frac{2m}{\delta} \tag{12}$$

iv. Write the inequality 12 also for the loss function $1-L$ and use the union bound to prove that for $\delta \in [0,1]$, the following inequality holds probability greater than $1-\delta$ over S:

$$E_{h\sim\mathcal{P}}[e^{(m-1)\mathrm{KL}(\hat{L}_S(h)||L(h))}] \leq \frac{2m}{\delta} \tag{13}$$

Research Problem :

- 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:

$$L(q) \leq \inf_{h\in\mathcal{H}} L(h) + \tilde{O}\left(\sqrt{d/\epsilon}\right) \tag{14}$$