Problem Set 6

Instructions:

- This problem set covers the following topics.
  - Unit 12 – Sample Compression Schemes. You may cite without proof any claim that was proved in the lectures.
  - Unit 13 – Information-Theoretic Generalization Bounds. You may cite without proof any claim that was proved in the lectures.
  - Unit 14 – Online Learning. You may cite without proof any claim that was proved in the lectures or in Sections 8.1, 8.2.1, 8.2.2 and 8.2.3 in the MRT textbook.
- Before you start, make sure you are familiar with the course's Homework Policy.

1. Let $\mathcal{X}$ be a set, let $\mathcal{F}$ be the set of all functions $\mathcal{X} \to \{0,1\}$, let $\mathcal{H} \subseteq \mathcal{F}$, let I be a finite set. Recall the definitions we saw in class.

▶ **Notation 1.** For any $m \in \mathbb{N}$, let

$$S_{\mathcal{H}}(m) = \Big\{ \Big((x_1, y_1), \dots, (x_t, y_t) \Big) \in (\mathcal{X} \times \{0,1\})^t \;\Big|\; t \in \mathbb{N} \;\wedge\; t \leq m \;\wedge \exists h \in \mathcal{H} \; \forall i \in [t] : \; y_i = h(x_i) \Big\}$$

be the set of samples of length at most m that are consistent with $\mathcal{H}$. Furthermore, let $S_{\mathcal{H}}(\infty) = \cup_{m \in \mathbb{N}} S_{\mathcal{H}}(m)$. ⌟

▶ **Definition 2.** *Fix $m' \in \mathbb{N}$. A pair a functions*

$$c : \; S_{\mathcal{H}}(\infty) \to S_{\mathcal{H}}(m') \times I \qquad\qquad r : \; S_{\mathcal{H}}(m') \times I \to \mathcal{F}$$

is a (realizable) sample compression scheme for $\mathcal{H}$ of size $k \in \mathbb{N}$ if for any $S = \Big((x_1, y_1), \dots, (x_m, y_m)\Big) \in S_{\mathcal{H}}(\infty)$, the tuple $(S', i) = c(S)$ satisfies:

(i) *The entries of S' are a subset of the entries of S.*
(ii) *$f = r((S', i))$ labels S correctly. Namely for all $i \in [m]$, $f(x_i) = y_i$.*
(iii) *$m' + \log_2(|I|) \leq k$.*

Consider the following alternative definition.

▶ **Definition 3.** *Fix $m' \in \mathbb{N}$. A pair a functions*

$$c : \; (\mathcal{X} \times \{0,1\})^* \to (\mathcal{X} \times \{0,1\})^{m'} \times I \qquad\qquad r : \; (\mathcal{X} \times \{0,1\})^{m'} \times I \to \mathcal{F}$$

is a non-realizeable sample compression scheme for $\mathcal{H}$ of size $k \in \mathbb{N}$ if for any $S = \Big((x_1, y_1), \dots, (x_m, y_m)\Big) \in (\mathcal{X} \times \{0,1\})^*$, the tuple $(S', i) = c(S)$ satisfies:

(i) *The entries of S' are a subset of the entries of S.*
(ii) *The functions $f = r((S', i))$ satisfies that for all $h \in \mathcal{H}$, $L_S(f) \leq L_S(h)$.*
(iii) *$m' + \log_2(|I|) \leq k$.*

Prove that $\mathcal{H}$ has a realizable sample compression scheme of size k if and only if $\mathcal{H}$ has a non-realizable sample compression scheme of size k. [27 pts]

2. (a) Consider a generalization of Definition 2 called *lossy realizable sample compression with loss* ε, which is the same as Definition 2 except that Item (ii) is replaced by the requirement that $L_S(f) \leq \varepsilon$.[1] Consider the learning algorithm $A_{c,r}$ that for sample S outputs the hypothesis $f = r(c(S))$. Show that if (c, r) is a lossy realizable sample compression scheme for $\mathcal{H}$ of size k with loss $\varepsilon/2$, then $A_{c,r}$ PAC learns $\mathcal{H}$ with parameters ε and δ using $O\left(\frac{k\ln(k/\varepsilon)+\ln(1/\delta)}{\varepsilon^2}\right)$ samples. [13 pts]

(b) Consider a further generalization where instead of requiring $L_S(f) \leq \varepsilon$ for all S, we require that $\mathbb{P}_S\left[L_S(f) > \varepsilon\right] < \delta$ when S consists of any number of i.i.d. samples from a specific unknown distribution $\mathcal{D}$. Can you prove a similar PAC learning sample complexity bound given that (r, c) satisfies this definition for the specific distribution $\mathcal{D}$? [7 pts]

3. Recall that we saw that if a learning algorithm satisfies $I(S; h) \leq d \in \mathbb{N}$ and $m \geq \Omega\left(\frac{d}{\delta\varepsilon^2}\right)$ then

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[|L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon\right] \geq 1 - \delta, \tag{1}$$

where h denotes the hypothesis chosen by the algorithm and I denotes mutual information.

(i) Show that in the realizable case, if the algorithm is an ERM (namely, the equality $L_S(h) = 0$ always holds), then taking $m \geq \Omega\left(\frac{d}{\delta\varepsilon}\right)$ is sufficient to imply Eq. (1). You may assume that the algorithm is deterministic. Conclude that taking $\Omega\left(\frac{d}{\delta\varepsilon}\right)$ samples is sufficient to ensure that a deterministic ERM algorithm is a PAC learner. [13 pts]

(ii) Show that the sample complexity in (i) is tight for d and ε in the following sense. For fixed δ, show an example of a class that requires $\Omega\left(\frac{d}{\varepsilon}\right)$ samples for PAC learning. [17 pts]

Hint: You may use the fact that the fundamental theorem of learning is tight.

4. In this question we will see how to use bounds on the number of mistakes of an online learning algorithm to obtain generalization bounds in the batch (i.e., non-online) setting. Let $\mathcal{X}$ be a set, and let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0, 1\}$. Assume we have an online learning algorithm A that operates as follows. A starts with an initial hypothesis $h_1 \in \mathcal{H}$, and at each timestep $t \in [T]$, it: (i) receives x_t; (ii) predicts label $\hat{y}_t = h_t(x_t)$; (iii) receives y_t; (iv) pays loss $\ell(\hat{y}_t, y_y)$; (v) selects hypothesis h_{t+1}.

Fix $T \in \mathbb{N}$, let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{0, 1\}$, let $S = \Big((x_1, y_1), \ldots, (x_T, y_T)\Big) \sim \mathcal{D}^T$, and let ℓ be the 0-1 loss. Assume we execute the algorithm A above on the examples (x_t, y_t) sequentially for $t = 1, 2, \ldots, T$. Prove the following generalization bound. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\frac{1}{T}\sum_{t=1}^{T} L_{\mathcal{D}}(h_t) \leq \frac{1}{T}\sum_{t=1}^{T} \ell(h_t(x_t), y_t) + \sqrt{\frac{2\ln(1/\delta)}{T}}.$$

[23 pts]

Hint: You may use Azuma's inequality (Theorem D.7 in MRT). You essentially already proved Azuma's inequality as part of the proof of McDiarmid's inequality in Problem Set 4.

---

[1] Note that Definition 2 corresponds to the case $\varepsilon = 0$.