Problem 1. [S-S] Problem 3.1

Monotonicity of Sample Complexity: Let $\mathcal{H}$ be a hypothesis class for a binary classification task. Suppose that $\mathcal{H}$ is PAC learnable and its sample complexity is given by $m_{\mathcal{H}}(\cdot,\cdot)$. Show that $m_{\mathcal{H}}$ is monotonically nonincreasing in each of its parameters. That is, show that given $\delta \in (0,1)$, and given $0 < \epsilon_1 \leq \epsilon_2 < 1$, we have that $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$. Similarly, show that given $\epsilon \in (0,1)$, and given $0 < \delta_1 \leq \delta_2 < 1$, we have that $m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$.

Solution. For fixed $\delta \in (0,1)$, suppose $0 < \epsilon_1 \leq \epsilon_2 \leq 1$. We need $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$. Given a training sequence of size $m \geq m_{\mathcal{H}}(\epsilon_1, \delta)$, we have that with probability at least $1-\delta$, we could learn a hypothesis h such that $L_{\mathcal{D},f}(h) \leq \epsilon_1 \leq \epsilon_2$. By the minimality of $m_{\mathcal{H}}(\epsilon_2, \delta)$, we get that $m_{\mathcal{H}}(\epsilon_2, \delta) \leq m_{\mathcal{H}}(\epsilon_1, \delta)$.

For fixed $\epsilon \in (0,1)$, suppose $0 < \delta_1 \leq \delta_2 < 1$. We need $m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$. Given a training sequence of size $m \geq m_{\mathcal{H}}(\epsilon, \delta_1)$, we have that with probability at least $1-\delta_1$, we could learn a hypothesis h such that $L_{\mathcal{D},f}(h) \leq \epsilon$. This also holds with $1-\delta_2$. By the minimality of $m_{\mathcal{H}}(\epsilon_2, \delta)$, we get that $m_{\mathcal{H}}(\epsilon_2, \delta) \leq m_{\mathcal{H}}(\epsilon_1, \delta)$. ■

Problem 2. [S-S] Problem 3.2

Let $\mathcal{X}$ be a discrete domain, and let $\mathcal{H}_{\text{Singleton}} = \{h_z : z \in \mathcal{X}\} \cup \{h^-\}$, where for each $z \in \mathcal{X}, h_z$ is the function defined by $h_z(x) = 1$ if $x = z$ and $h_z(x) = 0$ if $x \neq z$. h^- is simply the all-negative hypothesis, namely, $\forall x \in X, h^-(x) = 0$. The realizability assumption here implies that the true hypothesis f labels negatively all examples in the domain, perhaps except one.

1. Describe an algorithm that implements the ERM rule for learning $\mathcal{H}_{\text{Singleton}}$ in the realizable setup.

   Solution. Given a training set $\{(z_1, y_1), \ldots, (z_m, y_m)\}$, return h_{z_i} for the first (z_i, y_i) such that $y_i = 1$. If there is no such sample, return h^-. ■

2. Show that $\mathcal{H}_{\text{Singleton}}$ is PAC learnable. Provide an upper bound on the sample complexity.

   Solution. If $f = h^-$, our algorithm will clearly identify the correct hypothesis and have zero error. Hence, we may assume without loss of generality that $f = h_{z_0}$ for some $z_0 \in \mathcal{X}$. In this case, our algorithm will have nonzero error if and only if the training set S does not contain z_0, which will occur with probability

   $$\mathcal{D}^m(\{S \mid z_0 \notin S\}) = (1 - \mathcal{D}(\{z_0\}))^m$$

   in which case our algorithm will erroneously return h^- with error

   $$L_{(\mathcal{D},f)}(h^-) = \mathcal{D}(\{z \mid h^-(z) \neq f(z)\}) = \mathcal{D}(\{z_0\})$$

   Thus, assuming that $L_{(\mathcal{D},f)}(h^-) > \epsilon$ implies that $\mathcal{D}(\{z_0\}) > \epsilon$, and hence

   $$(1 - \mathcal{D}(\{z_0\}))^m < (1-\epsilon)^m \leq e^{-m\epsilon}$$

   Thus, $\mathcal{H}_{\text{Singleton}}$ is PAC learnable, and to satisfy thebounds it suffices to find an m such that $e^{-m\epsilon} \leq \delta$, i.e.

   $$m_{\mathcal{H}_{\text{Singleton}}}(\epsilon, \delta) \leq \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil$$

   ■

Problem 3. [S-S] Problem 3.5

Let $\mathcal{X}$ be a domain and let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$ be a sequence of distributions over $\mathcal{X}$. Let $\mathcal{H}$ be a finite class of binary classifiers over $\mathcal{X}$ and let $f \in \mathcal{H}$. Suppose we are getting a sample S of m examples, such that the instances are independent but are not identically distributed; the i th instance is sampled from $\mathcal{D}_i$ and then y_i is set to be $f(\mathbf{x}_i)$. Let $\bar{D}_m$ denote the average, that is, $\overline{\mathcal{D}}_m = (\mathcal{D}_1 + \cdots + \mathcal{D}_m)/m$.

Fix an accuracy parameter $\epsilon \in (0,1)$. Show that

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } L_{\left(\overline{\mathcal{D}}_m, f\right)}(h) > \epsilon \text{ and } L_{(S,f)}(h) = 0\right] \leq |\mathcal{H}| e^{-\epsilon m}.$$

Hint: Use the geometric-arithmetic mean inequality.

Solution. Fix some $h \in \mathcal{H}$ with $L_{\left(\overline{\mathcal{D}}_m, f\right)}(h) > \epsilon$. By definition,

$$\frac{\mathbb{P}_{X\sim\mathcal{D}_1}[h(X) = f(X)] + \ldots + \mathbb{P}_{X\sim\mathcal{D}_m}[h(X) = f(X))]}{m} < 1 - \epsilon.$$

We now bound the probability that h is consistent with S (i.e., that $L_S(h) = 0$) as follows:

$$\begin{aligned}
\mathbb{P}_{S\sim\prod_{i=1}^m \mathcal{D}_i}[L_S(h) = 0] &= \prod_{i=1}^m \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)] \\
&= \left(\left(\prod_{i=1}^m \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)]\right)^{\frac{1}{m}}\right)^m \\
&\leq \left(\frac{\sum_{i=1}^m \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)]}{m}\right)^m \\
&< (1-\epsilon)^m \\
&\leq e^{-\epsilon m}
\end{aligned}$$

The first inequality is the geometric-arithmetic mean inequality. Applying the union bound, we conclude that the probability that there exists some $h \in \mathcal{H}$ with $L_{\left(\overline{\mathcal{D}}_m, f\right)}(h) > \epsilon$, which is consistent with S is at most $|\mathcal{H}| \exp(-\epsilon m)$. ■

Problem 4. [S-S] Problem 3.6

Let $\mathcal{H}$ be a hypothesis class of binary classifiers. Show that if $\mathcal{H}$ is agnostic PAC learnable, then $\mathcal{H}$ is PAC learnable as well. Furthermore, if A is a successful agnostic PAC learner for $\mathcal{H}$, then A is also a successful PAC learner for $\mathcal{H}$.

Solution. Suppose that $\mathcal{H}$ is agnostic PAC learnable, and let A be a learning algorithm that learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\cdot,\cdot)$. We show that $\mathcal{H}$ is PAC learnable using A.

Let $\mathcal{D}, f$ be an (unknown) distribution over $\mathcal{X}$, and the target function respectively. We may assume w.l.o.g. that $\mathcal{D}$ is a joint distribution over $\mathcal{X} \times \{0,1\}$, where the conditional probability of y given x is determined deterministically by f. Since we assume realizability, we have $\inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) = 0$. Let $\epsilon, \delta \in (0,1)$. Then, for every positive integer $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, if we equip A with a training set S consisting of m i.i.d. instances which are labeled by f, then with probability at least $1 - \delta$ (over the choice of $S|_x$), it returns a hypothesis h with

$$\begin{aligned}
L_{\mathcal{D}}(h) &\leq \inf_{h'\in\mathcal{H}} L_{\mathcal{D}}\left(h'\right) + \epsilon \\
&= 0 + \epsilon \\
&= \epsilon.
\end{aligned}$$

■

Problem 5. [S-S] Problem 6.1

Show the following monotonicity property of VC-dimension: For every two hypothesis classes if $\mathcal{H}' \subseteq \mathcal{H}$ then $\text{VCdim}(\mathcal{H}') \leq \text{VCdim}(\mathcal{H})$.

Solution. Let $\mathcal{H}' \subseteq \mathcal{H}$ be two hypothesis classes for binary classification. Since $\mathcal{H}' \subseteq \mathcal{H}$, then for every $C = \{c_1, \ldots, c_m\} \subseteq \mathcal{X}$, we have $\mathcal{H}'_C \subseteq \mathcal{H}_C$. In particular, if C is shattered by $\mathcal{H}'$, then C is shattered by $\mathcal{H}$ as well. Thus, $\text{VCdim}(\mathcal{H}') \leq \text{VCdim}(\mathcal{H})$. ■

Problem 6. [S-S] Problem 6.4

We proved Sauer's lemma by proving that for every class $\mathcal{H}$ of finite VCdimension d, and every subset A of the domain,

$$|\mathcal{H}_A| \leq |\{B \subseteq A : \mathcal{H} \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{|A|}{i}$$

Show that there are cases in which the previous two inequalities are strict (namely, the $\leq$ can be replaced by $<$) and cases in which they can be replaced by equalities. Demonstrate all four combinations of $=$ and $<$.

Solution. Let $\mathcal{X} = \mathbb{R}^d$. We will demonstrate all the 4 combinations using hypothesis classes defined over $\mathcal{X} \times \{0, 1\}$. Remember that the empty set is always considered to be shattered.

- $(<, =)$: Let $d \geq 2$ and consider the class $\mathcal{H} = \left\{\mathbb{1}_{[\|x\|_2 \leq r]} : r \geq 0\right\}$ of concentric balls. The VC-dimension of this class is 1 . To see this, we first observe that if $\mathbf{x} \neq (0, \ldots, 0)$, then $\{\mathbf{x}\}$ is shattered. Second, if $\|\mathbf{x}_1\|_2 \leq \|\mathbf{x}_2\|_2$, then the labeling $y_1 = 0, y_2 = 1$ is not obtained by any hypothesis in $\mathcal{H}$. Let $A = \{\mathbf{e}_1, \mathbf{e}_2\}$, where $\mathbf{e}_1, \mathbf{e}_2$ are the first two elements of the standard basis of $\mathbb{R}^d$. Then, $\mathcal{H}_A = \{(0,0), (1,1)\}, \{B \subseteq A : \mathcal{H} \text{ shatters } B\} = \{\emptyset, \{\mathbf{e}_1\}, \{\mathbf{e}_2\}\}$, and $\sum_{i=0}^{d} \binom{|A|}{i} = 3$.

- $(=, <)$: Let $\mathcal{H}$ be the class of axis-aligned rectangles in $\mathbb{R}^2$. We have seen that the VC-dimension of $\mathcal{H}$ is 4 . Let $A = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, where $\mathbf{x}_1 = (0,0), \mathbf{x}_2 = (1,0), \mathbf{x}_3 = (2,0)$. All the labelings except $(1, 0, 1)$ are obtained. Thus, $|\mathcal{H}_A| = 7, |\{B \subseteq A : \mathcal{H} \text{ shatters } B\}| = 7$, and $\sum_{i=0}^{d} \binom{|A|}{i} = 8$.

- $(<, <)$: Let $d \geq 3$ and consider the class $\mathcal{H} = \left\{\text{sign}\langle w, x\rangle : w \in \mathbb{R}^d\right\}^2$ of homogenous halfspaces (see Chapter 9). We will prove in Theorem 9.2 that the VC-dimension of this class is d. However, here we will only rely on the fact that $\text{VCdim}(\mathcal{H}) \geq 3$. This fact follows by observing that the set $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is shattered. Let $A = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, where $\mathbf{x}_1 = \mathbf{e}_1, \mathbf{x}_2 = \mathbf{e}_2$, and $\mathbf{x}_3 = (1, 1, 0, \ldots, 0)$. Note that all the labelings except $(1, 1, -1)$ and $(-1, -1, 1)$ are obtained. It follows that $|\mathcal{H}_A| = 6, |\{B \subseteq A : \mathcal{H} \text{ shatters } B\}| = 7$, and $\sum_{i=0}^{d} \binom{|A|}{i} = 8$.

- $(=, =)$: Let $d = 1$, and consider the class $\mathcal{H} = \left\{\mathbb{1}_{[x \geq t]} : t \in \mathbb{R}\right\}$ of thresholds on the line. We have seen that every singleton is shattered by $\mathcal{H}$, and that every set of size at least 2 is not shattered by $\mathcal{H}$. Choose any finite set $A \subseteq \mathbb{R}$. Then each of the three terms in "Sauer's inequality" equals $|A| + 1$. ■