3.5, Q1:

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 increasing in both parameters.

Solution

3.5, Q2:

Let $\mathcal{X}$ be a discrete domain, and let $\mathcal{H}_{singleton} \equiv \{[z] : z \in \mathcal{X}\} \cup \{h\}$, where

$$[z] : \mathcal{X} \to \{0, 1\}; \quad [z](x) \equiv \begin{cases} 1 & x = z \\ 0 & \text{otherwise} \end{cases}$$
$$h : \mathcal{X} \to \{0, 1\}; \quad h^{-}(\_) \equiv 0$$

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 $\mathcal{H}_{singleton}$ in the realizable setup.

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

Solution, part (a)

Let $\mathcal{X}$ be the domain, let $f : \mathcal{X} \to \{0, 1\}$ be the underlying target function f that we are trying to approximate using $\mathcal{H}$.

We define the sample loss $L_S(h)$ as the number of elements in S that are mis-classified by h. More formally, $L_S(h) \equiv |\{(x, y) \in \S : h(x) \neq y\}|$.

The ERM algorithm must, given a particular sample set $S \in \mathcal{X}^n \sim \mathcal{D}^n$, provides a function $h_0 \in \mathcal{H} = ERM(S)$ which has minimum sample loss $L_S(h_0)$ across all functions in $\mathcal{H}$.

We can check over the classification of all the samples $s \in S$.

- If all samples $s \in S$ are classified as 0: we return h^{-} — this will always return 0.

- If some sample $s_1 \in S$ is classified as 1: notice that our hypothesis space $\mathcal{H}$ can only allow us to set *at most one sample to 1*. So, we can pick *any* sample s_1 to create our hypothesis function $h = [s_1]$, since that is the best we can do.

```
def hminus(_): return 0 # h-: sends all samples to 0
def indicator(z): return lambda x: 1 if x == z else 0 #indicator of z
def erm_sample(S):
    # all samples which have label 1
    one_samples = [y for (x, y) in S if y == 1]
    if len(one_samples) == 0: return hminus # send all samples to 0!
    else: # we will have at least on element in one_samples
        return indicator(ones_samples[0])
```

Solution, part (b)

4.5, Q1:

Prove that the following two statements are equivalent for any learning algorithm A, any probability distribution $\mathcal{D}$, and any loss function whose loss is in the range $[0, 1]$:

$$\forall \epsilon, \delta > 0, \exists M \equiv m(\epsilon, \delta), \forall m \geq M : \underset{S \sim D^m}{\mathbb{P}}[L_{\mathcal{D}}(A(S)) > \epsilon] < \delta.$$
$$\Updownarrow$$
$$\lim_{m \to \infty} \underset{S \sim D^m}{\mathbb{E}}[L_{\mathcal{D}}(A(S)) > \epsilon] = 0.$$

Solution

6.8, Q1:

For two hypothesis classes $\mathcal{H}, \mathcal{H}'$, if $\mathcal{H}' \subseteq \mathcal{H}$ then $\text{VCdim}(H') \leq \text{VCdim}(H)$.

Solution

Recall that the VC dimension of a given set family $\mathcal{H}$ is the size of the largest set C such that H shatters C. That is, the intersection of C with every element is H is equal to the powerset of C:

$$\text{VCdim}(\mathcal{H}) \equiv \max_C \{h \cap C : h \in \mathcal{H}\} = 2^C \qquad \text{We denote powerset of } C \text{ by } 2^C$$

Now, if a set family $\mathcal{H}'$ is a subset of another set family $\mathcal{H}$, and if $\mathcal{H}'$ shatters C, then:

$$\begin{aligned}
&\mathcal{H}' \text{ shatters C} \equiv \{h \cap C : h \in \mathcal{H}'\} = 2^C \qquad \text{Given, (1)} \\
&\{h \cap C : h \in \mathcal{H}'\} \subseteq \{h \cap C : h \in \mathcal{H}\} \qquad \text{Since } H' \subseteq H \\
&2^C \subseteq \{h \cap C : h \in \mathcal{H}\} \qquad \text{From (1)}
\end{aligned}$$

Hence, any set that can be shattered by H' can be shattered by H if $H' \subseteq H \implies \text{VCdim}(H') \leq \text{VCdim}(H)$.

On the other hand, clearly if H is larger than H', then H can shatter more. For example, let $H' = \phi \subsetneq H$. Then H' can only shatter the empty set, while H can in general shatter sets larger than the empty set. Hence, we have have strict inequality: $\text{VCdim}(\emptyset) < \text{VCdim}(H)$ for example.