Telecom Paris Introduction to Machine Learning (SI221/MICAS911)
Teacher: A. Tchamkerten

ASSIGNMENT 2 - SOLUTIONS

Exercise 1 (Finite classes are learnable)**.** In this exercise we will show that any finite class is learnable. More specifically we will establish the following result: Theorem: Let $\mathcal{H}$ be a finite set functions. Then, for any empirical risk minimization (ERM) function $\hat{f}_{S^m}$ we have

$$\mathbb{P}_{S^m}(\{S^m : \mathbb{P}_X(\hat{f}_{S^m}(X) \neq f(X)) > \varepsilon\}) \leq |\mathcal{H}|e^{-\varepsilon m}$$

for any data distribution and for any $f \in \mathcal{H}$.

1. Define the set of bad hypothesis as

$$\mathcal{H}_B = \{h \in \mathcal{H} : \mathbb{P}_X(h(X) \neq f(X)) > \varepsilon\})\}$$

and define the set of misleading samples as

$$\mathcal{M} = \{S^m : \exists h \in \mathcal{H}_B \text{ such that } L_{S^m}(h) = 0\}$$

Use realizability to show that (and this is the main step of the proof)

$$\mathbb{P}_{S^m}(\{S^m : \mathbb{P}_X(\hat{f}_{S^m}(X) \neq f(X)) \geq \varepsilon\}) \leq \mathbb{P}_{S^m}(\mathcal{M})$$

2. Argue that

$$\mathbb{P}_{S^m}(\mathcal{M}) \leq \sum_{h \in \mathcal{H}_B} \mathbb{P}_{S^m}(S^m : L_{S^m}(h) = 0)$$

3. Argue that

$$\mathbb{P}_{S^m}(S^m : L_{S^m}(h) = 0) \leq (1-\varepsilon)^m$$

4. Conclude the proof.

Solution. 1. Because of the realizability assumption ($f \in \mathcal{H}$) we have that there exists an $h \in \mathcal{H}$ such that $\hat{f}_{S^m} = h$ and $L_{S^m}(h) = 0$. Therefore

$$\begin{aligned}
&\mathbb{P}_{S^m}(\{S^m : \mathbb{P}_X(\hat{f}_{S^m}(X) \neq f(X)) > \varepsilon\}) \\
&= \mathbb{P}_{S^m}(\{S^m : \mathbb{P}_X(\hat{f}_{S^m}(X) \neq f(X)) > \varepsilon\} \cap \{S^m : \hat{f}_{S^m} = h, L_{S^m}(h) = 0 \text{ for some } h \in \mathcal{H}\}) \\
&= \mathbb{P}_{S^m}(\{S^m : \hat{f}_{S^m} = h, L_{S^m}(h) = 0 \text{ for some } h \in \mathcal{H}_B\}) \\
&\leq \mathbb{P}_{S^m}(L_{S^m}(h) = 0 \text{ for some } h \in \mathcal{H}_B)
\end{aligned}$$

2. Union bound

3. $\mathbb{P}_{S^m}(L_{S^m}(h) = 0) = \mathbb{P}_{S^m}(h(X_i) = f(X_i), 1 \leq i \leq m) = [\mathbb{P}_X(h(X) = f(X))]^m \leq (1-\varepsilon)^m$ where the upper bound holds for any $h \in \mathcal{H}_B$

4. Since $|\mathcal{H}_B| \leq |\mathcal{H}|$ and $(1-\varepsilon)^m \leq e^{-\varepsilon m}$ (which follows from the inequality $\ln x \leq x - 1$)

$$\mathbb{P}_{S^m}(\{S^m : \mathbb{P}_X(\hat{f}_{S^m}(X) \neq f(X)) \geq \varepsilon\}) \leq |\mathcal{H}_B|(1-\varepsilon)^m \leq |\mathcal{H}|e^{-m\varepsilon}$$

□

Exercise 2 (No Free Lunch Theorem)**.** In this exercise we show that if $\mathcal{H}$ is unrestricted, that is if $\mathcal{H}$ contains all the functions from $\mathcal{X}$ to $\mathcal{Y}$ then PAC learning $\mathcal{H}$ requires a "huge" number of samples, of the order of $|\mathcal{X}|$:
Theorem: Let $\mathcal{H}$ denote the set of all functions from $\mathcal{X}$ to $\mathcal{Y}$. Then the number of samples m needed to PAC learn $\mathcal{H}$ with accuracy $\varepsilon = 1/8$ and confidence $\delta = 1/8$, $m_{\mathcal{H}}(1/8, 1/8)$, is at least $|\mathcal{X}|/2$.

1. Suppose first that $|\mathcal{X}| < \infty$ and that $\mathcal{Y} = \{0, 1\}$. Let $\hat{f}_{S^m}(\cdot)$ be a predictor algorithm for the class $\mathcal{H}$. Let the test symbol $X \in \mathcal{X}$ and the training sequence $S^m = S_1, S_2 \ldots S_m \in \mathcal{X}^m$ be i.i.d. $\sim P$ for some distribution P. Justify the following inequalities:

$$\begin{aligned}
&\sup_{h\in\mathcal{H}} \mathbb{E}_{S^m}[P_X(\hat{f}_{S^m}(X) \neq h(X)] \\
&\quad \geq \mathbb{E}_h \mathbb{E}_{S^m}(P_X(\hat{f}_{S^m}(X) \neq h(X)) \quad \text{for any distribution over } h && (1)\\
&\quad = \mathbb{E}_h \mathbb{E}_{S^m} \mathbb{E}_X(\mathbb{1}\{\hat{f}_{S^m}(X) \neq h(X)\} && (2)\\
&\quad = \mathbb{P}_{h,S^m,X}(\hat{f}_{S^m}(X) \neq h(X)) && (3)\\
&\quad \geq \mathbb{P}_{h,S^m,X}(\hat{f}_{S^m}(X) \neq h(X))|X \notin \{S_1,\ldots,S_m\})\mathbb{P}_{S^m,X}(X \notin \{S_1,\ldots,S_m\}) && (4)
\end{aligned}$$

2. Suppose h is uniformly distributed over the set of functions from $\mathcal{X}$ to $\{0, 1\}$. Show that

$$\mathbb{P}_{h,S^m,X}(\hat{f}_{S^m}(X) \neq h(X)|X \notin \{S_1,\ldots,S_m\}) = 1/2.$$

3. Let P_X be the uniform distribution over $\mathcal{X}$. Show that

$$\mathbb{P}_{S^m,X}(X \notin \{S_1,\ldots,S_m\}) \geq \frac{|\mathcal{X}| - m}{|\mathcal{X}|}.$$

4. Argue that there exists h and a data distribution such that

$$\mathbb{P}_{S^m,X}(\hat{f}_{S^m}(X) \neq h(X)) \geq \frac{1}{4}$$

whenever $m \leq |\mathcal{X}|/2$.

5. Define events (for notational convenience)

$$\mathcal{E} = \{X, S^m : \hat{f}_{S^m}(X) \neq h(X)\}$$
$$\mathcal{E}_\varepsilon = \{S^m : \mathbb{P}_{S^m,X}(\hat{f}_{S^m}(X) \neq h(X)) \geq \varepsilon\}$$

Argue that there exists a function h and a data distribution such that

$$\frac{1}{4} \leq \mathbb{P}_{S^m,X}(\mathcal{E}) \leq Pr(S^m \in \mathcal{E}_\varepsilon) + \varepsilon.$$

6. Conclude that $m_{\mathcal{H}}(1/8, 1/8) \geq |\mathcal{X}|/2$.
7. Show that the previous bound also holds if $|\mathcal{Y}| > 2$.

Solution. 1. (1) follows from the fact that the supremum is at least as large as (any weighted) average.

2. Follows from the fact that for any value of X, $h(X)$ is uniformly distributed and independent of $\hat{f}_{S^m}(X)$.

3. Follows from the fact that over m trials X takes at most m different values.

4. Follows from 2.3. and the fact that if the inequality holds on average over h it also holds for at least one specific h.

5. Follows from

$$\mathbb{P}_{S^m,X}(\mathcal{E}) = \mathbb{P}_{S^m,X}(\mathcal{E}|S^m \in \mathcal{E}_\varepsilon)\mathbb{P}_S^m(S^m \in \mathcal{E}_\varepsilon) + \mathbb{P}_{S^m,X}(\mathcal{E}|S^m \notin \mathcal{E}_\varepsilon)\mathbb{P}_S^m(S^m \notin \mathcal{E}_\varepsilon)$$

6. Follows from 5. with $\varepsilon = 1/8$.

7. The only place where we used that $|\mathcal{Y}| = 2$ is in item 2. For the general case the $1/2$ should be replaced by $((|\mathcal{Y}| - 1)/|\mathcal{Y}|)$ which is always $\geq 1/2$. Hence, the result also holds for any $|\mathcal{Y}| \geq 2$.

□