Homework 3 (Graded): due Monday, March 18 2024
CS-526 Learning Theory

Problems from *Understanding Machine Learning: From Theory to Algorithms* by Shai Shalev-Shwartz and Shai Ben-David:

1. Exercise 5 of Chapter 3.
2. Exercise 2 of Chapter 5. Note that here we expect just a qualitative answer, without any computations.
3. Exercise 3 of Chapter 5.
4. Exercise 2 of Chapter 6.
5. Exercise 8 of Chapter 6.

Problem 6. Stable Learning.

Let $\mathcal{X}$ be a domain set and $\mathcal{Y}$ be a set of labels. Let $\mathcal{F}$ be a set of possible labelling functions, $\mathcal{F} \subset \{f | f : \mathcal{X} \to \mathcal{Y}\}$.

Definition: We say that A is a *stable learner* for $\mathcal{F}$ using the hypothesis class $\mathcal{H}$, if for any labeling function $f \in \mathcal{F}$ and for all $m \geq 1$, when given as input the set of samples $S = \{(x_1, f(x_1)), \cdots, (x_m, f(x_m))\}$ where $x_i \in \mathcal{X}$, A outputs $h_S \in \mathcal{H}$ such that $h_S(x_i) = f(x_i)$ for $1 \leq i \leq m$.

Remark: Question 1 is about the proof of a statement and question 2 is an application. You can answer question 2 even if you do not prove the statement in question 1.

1. Let $\mathcal{F}$ be a labelling class and $\mathcal{H}$ a finite hypothesis class which are not necessarily equal. We suppose there exists a stable learner A for $\mathcal{F}$ using $\mathcal{H}$. Prove the following statement:

   For all $f \in \mathcal{F}$ and all distributions $\mathcal{D}$ over $\mathcal{X}$ and all $\epsilon, \delta \in (0, 1)$, if A is given a set of samples $S = \{(x_i, f(x_i))\}_{i=1}^m$ with $x_i \sim \mathcal{D}$ and size m such that

   $$m \geq \frac{1}{\epsilon}\left(\log |\mathcal{H}| + \log \frac{1}{\delta}\right),$$

   then with probability at least $1 - \delta$ the learner A outputs a hypothesis $h_S \in \mathcal{H}$ that satisfies

   $$P_{x \sim \mathcal{D}}\big[h_S(x) \neq f(x)\big] \leq \epsilon$$

   <u>Hint</u>: *Fix the labeling function. Then, define a notion of "bad" hypotheses, and use union bound.*

Now, we consider the problem of learning conjunctions. Let $\mathcal{X} = \{0,1\}^n$. Let $\mathcal{F} = \text{CONJUNCTIONS}_n$ denote the class of conjunctions over the n boolean variables $z_1, \ldots, z_n$. A *literal* is either a boolean variable z_i or its negation $\bar{z}_i$. A conjunction is simply an 'and' ($\wedge$) of literals. An example conjunction φ with $n = 10$ is

$$\varphi(z_1, \ldots, z_{10}) = z_1 \wedge \bar{z}_3 \wedge \bar{z}_8 \wedge z_9$$

We want to learn a target conjunction $\phi^* \in \text{CONJUNCTIONS}_n$ from a sampling set $S = \{(x_i, \phi^*(x_i))\}_{i=1}^m$, and the hypothesis class is $\mathcal{H} = \text{CONJUNCTIONS}_n$. So here each sample x_i is a binary vector $(x_{i,1}, \cdots, x_{i,10})$ assigned to $(z_1, \ldots, z_{10})$. The corresponding label $\phi^*(x_i)$ equals 0 or 1.

2. Consider the following algorithm for learning conjunctions:

    1. Set $h = z_1 \wedge \bar{z}_1 \wedge z_2 \wedge \bar{z}_2 \wedge \cdots \wedge z_n \wedge \bar{z}_n$.
    2. For $i = 1, \ldots, m$:
    3.   If $\phi^*(x_i) == 1$: (Ignore samples with 0 label)
    4.   For $j = 1, \ldots, n$:
    5.   If $x_{i,j} == 0$: (j-th bit of x_i)
    6.   Drop z_j from h.
    7.   Else:
    8.   Drop $\bar{z}_j$ from h.
    9. Output h.

    (a) Apply the algorithm to the sample set $S = \{(0001, 0), (0111, 0), (1001, 1), (1011, 0)\}$, and determine the output. Check that the algorithm has outputed a *stable* hypothesis.

    (b) Suppose now that the algorithm is indeed a stable learner. Given (ϵ, δ) how many samples are needed to have:

    $$P_{x \sim \mathcal{D}}\big[h_S(x) \neq f(x)\big] \leq \epsilon \qquad \text{with probability at least } 1 - \delta$$

    for any distribution $\mathcal{D}$, and set S ?