Introduction to Machine Learning	**Fall Semester**

Homework 2: October 27, 2021

Due: November 10, 2021

Theory Questions

1. **(14 points) PAC in Expectation**. Consider learning in the realizable case. We say an hypothesis class $\mathcal{H}$ is ***PAC learnable in expectation*** if there exists a learning algorithm A and a function $N(a) : (0,1) \to \mathbb{N}$ such that $\forall a \in (0,1)$ and for any distribution P, given a sample set S, such that $|S| \geq N(a)$ it holds that,

$$\mathbb{E}\left[e_P(A(S))\right] \leq a$$

Show that $\mathcal{H}$ is PAC learnable *if and only if* $\mathcal{H}$ is PAC learnable in expectation (Hint: On one directions, use the law of total expectation. On the other direction, use Markov's inequality). You can assume that the loss function is bounded between 0 and 1.

2. **(14 points) Union Of Intervals.** Determine the VC-dimension of the subsets of the real line formed by the union of k intervals (see question 1 of the programming assignment for a formal definition of $\mathcal{H}$).

3. **(14 points) Axis-aligned Rectangles**. Fix $d \in \mathbb{N}$. Given two vectors $\mathbf{a} = (a_1, ..., a_d) \in \mathbb{R}^d$ and $\mathbf{b} = (b_1, ..., b_d) \in \mathbb{R}^d$, define

$$h_{\mathbf{a},\mathbf{b}}(\mathbf{x}) = \begin{cases} 1 & \forall i \in [d] : \quad a_i \leq x_i \leq b_i \\ 0 & \text{otherwise} \end{cases}.$$

Determine the VD-dimension of $\mathcal{H} = \{h_{\mathbf{a},\mathbf{b}} \mid \mathbf{a}, \mathbf{b} \in \mathbb{R}^d\}$.

4. **Growth function (18 points).** Fix two hypothesis classes $\mathcal{H}_1, \mathcal{H}_2$.

   (a) Show that for any n,

   $$\Pi_{\mathcal{H}_1 \cup \mathcal{H}_2}(n) \leq \Pi_{\mathcal{H}_1}(n) + \Pi_{\mathcal{H}_2}(n).$$

   (b) Denote $\mathcal{H}_1 \otimes \mathcal{H}_2 = \{x \mapsto h_1(x) \cdot h_2(x) \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$, where h_i return 0-1 labels ($h_i(x) \in \{0,1\}$). Show that for any n,

   $$\Pi_{\mathcal{H}_1 \otimes \mathcal{H}_2}(n) \leq \Pi_{\mathcal{H}_1}(n) \cdot \Pi_{\mathcal{H}_2}(n).$$

Programming Assignment

1. **Union Of Intervals.** In this question, we will study the hypothesis class of a finite union of disjoint intervals, and the properties of the ERM algorithm for this class.

   To review, let the sample space be $\mathcal{X} = [0, 1]$ and assume we study a binary classification problem, i.e. $\mathcal{Y} = \{0, 1\}$. We will try to learn using an hypothesis class that consists of k intervals. More explicitly, let $I = \{[l_1, u_1], \ldots, [l_k, u_k]\}$ be k disjoint intervals, such that $0 \leq l_1 \leq u_1 \leq l_2 \leq u_2 \leq \ldots \leq u_k \leq 1$. For each such k disjoint intervals, define the corresponding hypothesis as

   $$h_I(x) = \begin{cases} 1 & \text{if } x \in [l_1, u_1], \ldots, [l_k, u_k] \\ 0 & \text{otherwise} \end{cases}$$

   Finally, define $\mathcal{H}_k$ as the hypothesis class that consists of all hypotheses that correspond to k disjoint intervals:

   $$\mathcal{H}_k = \{h_I | I = \{[l_1, u_1], \ldots, [l_k, u_k]\},\ 0 \leq l_1 \leq u_1 < l_2 \leq u_2 < \ldots \leq u_k \leq 1\}$$

   <u>Submission Guidelines:</u>

   - Download the files `skeleton.py` and `intervals.py` from Moodle. You should implement only the missing code in `skeleton.py`, as specified in the following questions. In every method description, you will find specific details on its input and return values.
   - Your code should be written with python 3.
   - Your submission should include exactly two files: `assignment2.py`, `intervals.py`.

   <u>Explanation on `intervals.py`:</u>

   The file `intervals.py` includes a function that implements an ERM algorithm for $\mathcal{H}_k$. Given a sorted list xs $= [x_1, \ldots, x_n]$, the respective labeling ys $= [y_1, \ldots, y_n]$ and k, the given function `find_best_interval` returns a list of up to k intervals and their error count on the given sample. These intervals have the smallest empirical error count possible from all choices of k intervals or less.

   Note that in sections (c)-(e) you will need to use this function for large values of n. Execution in these cases could take time (more than 10 minutes for experiment), so plan ahead.

   (a) **(8 points)** Assume that the true distribution $P[x, y] = P[y|x] \cdot P[x]$ is: x is distributed uniformly on the interval $[0, 1]$, and

   $$P[y = 1|x] = \begin{cases} 0.8 & \text{if } x \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1] \\ 0.1 & \text{if } x \in (0.2, 0.4) \cup (0.6, 0.8) \end{cases}$$

   and $P[y = 0|x] = 1 - P[y = 1|x]$. Since we know the true distribution P, we can calculate $e_P(h)$ precisely for any hypothesis $h \in \mathcal{H}_k$. What is the hypothesis in $\mathcal{H}_{10}$ with the smallest error (i.e., $\arg\min_{h \in \mathcal{H}_{10}} e_P(h)$)?

(b) **(8 points)** Write a function that, given a list of intervals I, calculates the true error $e_P(h_I)$. Then, for $k = 3$, $n = 10, 15, 20, \ldots, 100$, perform the following experiment $T = 100$ times: (i) Draw a sample of size n and run the ERM algorithm on it; (ii) Calculate the empirical error for the returned hypothesis; (iii) Calculate the true error for the returned hypothesis. Plot the average empirical and true errors, averaged across the T runs, as a function of n. Discuss the results. Do the empirical and true error decrease or increase in n? Why?

(c) **(8 points)** Draw a data set of $n = 1500$ samples. Find the best ERM hypothesis for $k = 1, 2, \ldots, 10$, and plot the empirical and true errors as a function of k. How does the error behave? Define k^* to be the k with the smallest empirical error for ERM. Does this mean the hypothesis with k^* intervals is a good choice?

(d) **(8 points)** Now we will use the principle of structural risk minimization (SRM), to search for a k that gives good test error. Let $\delta = 0.1$:

- Use to following penalty function:

$$2\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}_k) + \ln \frac{2}{\delta}}{n}}$$

- Draw a data set of $m = 1500$ samples, run the experiment in (c) again, but now plot two additional lines as a function of k: 1) the penalty for the best ERM hypothesis and 2) the sum of penalty and empirical error.
- What is the best value for k in each case? is it better than the one you chose in (c)?

(e) **(8 points)** Here we will use holdout-validation to search for a $k \in \{1, ..., 10\}$ that gives good test error. Draw a data set of $n = 1500$ samples and use 20% for a holdout-validation. Choose the best hypothesis and discuss how close this gets you to finding the hypothesis with optimal true error.