

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 $I = \{[l_1, u_1], \dots, [l_k, u_k]\}$ be k disjoint intervals, such that $0 \leq l_1 \leq u_1 \leq l_2 \leq u_2 \leq \dots \leq u_k \leq 1$. For each such k disjoint intervals, define the corresponding hypothesis as

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

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

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

We note that $\mathcal{H}_k \subseteq \mathcal{H}_{k+1}$, since we can always take one of the intervals to be of length 0 by setting its endpoints to be equal. We are given a sample of size $m = \langle x_1, y_1 \rangle, \dots, \langle x_n, y_m \rangle$. Assume that the points are sorted, so that $0 \leq x_1 < x_2 < \dots < x_m \leq 1$.

Submission Guidelines:

- 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.
- Make sure to comment out / remove any code which halts the code execution, such as matplotlib popup.
- Your submission should include exactly two files: `assignment2.py`, `intervals.py`.

Explanation on intervals.py:

The file `intervals.py` includes a function that implements an ERM algorithm for \mathcal{H}_k . Given a sorted list $xs = [x_1, \dots, x_m]$, the respective labeling $ys = [y_1, \dots, y_m]$ 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 (d)-(f) you will need to use this function for large values of m . Execution in these cases could take time (more than 10 minutes for experiment), so plan ahead.

- (a) **(7 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 $P[y = 0|x] = 1 - P[y = 1|x]$.

Write a function that draws m pairs of (x, y) according to the distribution P . Use it to draw a sample of size $m = 100$ and create a plot:

- i. Plot the points and their label (have the y axis in range $-0.1, 1.1$ for clarity of presentation).
 - ii. Mark the lines $x = 0.2, 0.4, 0.6, 0.8$ clearly on the plot.
 - iii. Run the `find_best_interval` function on your sample with $k = 3$, and plot the intervals clearly.
- (b) **(7 points)** Note that here, we know the true distribution P , so for every given hypothesis $h \in \mathcal{H}_k$, we can calculate $error(h)$ precisely. What is the hypothesis with the smallest error?
- (c) **(7 points)** Write a function that, given a list of intervals, calculates the error of the respective hypothesis. Then, for $k = 3$, $m = 10, 15, 20, \dots, 100$, perform the following experiment $T = 100$ times: (i) Draw a sample of size m 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 m . Discuss the results. Do the empirical and true error decrease or increase in m ? Why?
- (d) **(7 points)** Draw a data set of $m = 1500$ samples. Find the best ERM hypothesis for $k = 1, 2, \dots, 20$, 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?
- (e) **(7 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 your results from question 3 in the theoretical part and the VC confidence bound to construct a penalty function.
 - Draw a data set of $m = 1500$ samples, run the experiment in (d) 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 (d)?
- (f) **(7 points)** Here we will use holdout-validation to search for a k that gives good test error. Draw a data set of $m = 1500$ samples and use 20% for a holdout-validation. Choose the best hypothesis based on 3 experiments. Discuss how close this gets you to finding the hypothesis with optimal true error.