Advanced Machine Learning Seminar 5

Exercise 1 (exercise 8.1 in the book)

Let $\mathcal{H}$ be the class of intervals on the line (formally equivalent to axis aligned rectangles in dimension $n = 1$). Propose an implementation of the $\mathrm{ERM}_{\mathcal{H}}$ learning rule (in the agnostic case) that given a training set of size m, runs in time $\mathcal{O}(m^2)$. Hint: Use dynamic programming.

Solution.

$$\mathcal{H}_{\text{intervals}} = \mathcal{H}^1_{rec} = \left\{ h_{a,b} \colon \mathbb{R} \to \mathbb{R},\ h_{a,b} = \mathbb{1}_{[a,b]},\ h_{a,b}(x) = \begin{cases} 1 & x \in [a,b] \\ 0 & \text{otherwise} \end{cases},\ a, b \in \mathbb{R} \right\}$$

Consider a training set S of size m:

$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \mid x_i \in \mathbb{R},\ y_i \in \{0,1\}, i = \overline{1,m}\}$

Propose an implementation of the $\mathrm{ERM}_{\mathcal{H}}$ learning rule in the agnostic case that runs in $\mathcal{O}(m^2) \Leftrightarrow$ find a hypothesis h_{a_S,b_S} with the smallest empirical risk.

Example:

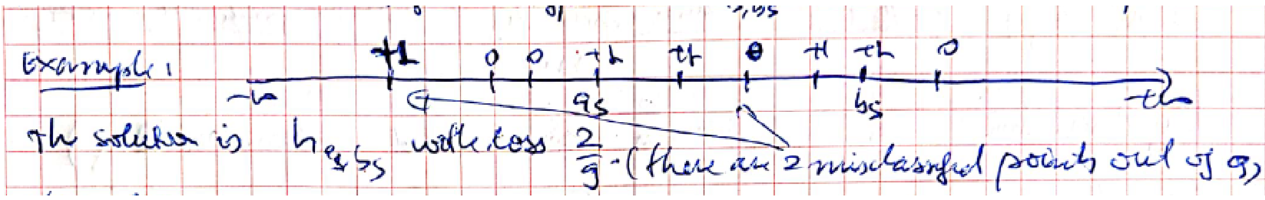

Figure 1: Example for agnostic case: 9 points scattered on the real line with some labels (5 positives and 4 negatives).

The solution for the example in Figure 2 is h_{a_S,b_S} with loss $\dfrac{2}{9}$ (there are 2 misclassified points out of 9).

Observations

1. We are in the agnostic case:
   - it might be the case that there is no labeling function but instead we are dealing with a distribution (same point might have different labels);
   - if there is a labeling function, it might not be in $\mathcal{H}_{\text{intervals}}$
2. If all points are negative, we should return an interval not containing any point in S
3. If all points are positive, we should return an interval containing all points in S

We will first sort the training set S in ascending order of $x's$.
We obtain $S = \{(x_{\sigma(1)}, y_{\sigma(1)}), (x_{\sigma(2)}, y_{\sigma(2)}), \ldots, (x_{\sigma(m)}, y_{\sigma(m)})\}$ with $x_{\sigma(1)} \le x_{\sigma(2)} \le \cdots \le x_{\sigma(m)}$.
As we are in the agnostic case, we can have $x_{\sigma(i)} = x_{\sigma(i+1)}$ and $y_{\sigma(i)} \neq y_{\sigma(i+1)}$.
Consider the set Z containing the values of x' with no repetition:

$$Z = \{z_1, z_2, \ldots, z_n\}$$
$$z_1 = x_{\sigma(1)} < z_2 < \cdots < z_n = x_{\sigma(m)} \quad n \le m$$

If all initial x values are different, then $z_1 = x_{\sigma(1)}, \ldots, z_n = x_{\sigma(m)}, n = m$.

Idea of the implementation of $\mathrm{ERM}_{\mathcal{H}}$

1. If all $y_i = 0$, return an interval not containing any point x: $[z_1 - 2, z_1 - 1]$.

2. Consider all possible intervals $Z_{i,j} = [z_i, z_j] \qquad i = \overline{1,n}, j = \overline{i,n}$

There are $n + (n-1) + (n-2) + \cdots + 1 = \dfrac{n(n+1)}{2}$ such intervals.

Determine the interval $Z^* = Z_{i^*,j^*}$ with the smallest empirical risk. $Z_{i^*,j^*} = \underset{\substack{i=\overline{1,n} \\ j=\overline{i,m}}}{\operatorname{argmin}} \operatorname{Loss}(Z_{i,j})$

How to compute very fast $\operatorname{Loss}(Z_{i,j})$? Use dynamic programming!

$\operatorname{Loss}(Z_{i,j}) = \dfrac{\#\text{ negative points inside } Z_{i,j} + \#\text{ positive points outside } Z_{i,j}}{m}$

Key observation: $\operatorname{Loss}(Z_{i,j+1})$ can be computed based on $\operatorname{Loss}(Z_{i,j})$.

Simple case: there is just one point (x_k, y_k) in the training set S such that $x_k = z_{j+1}$.

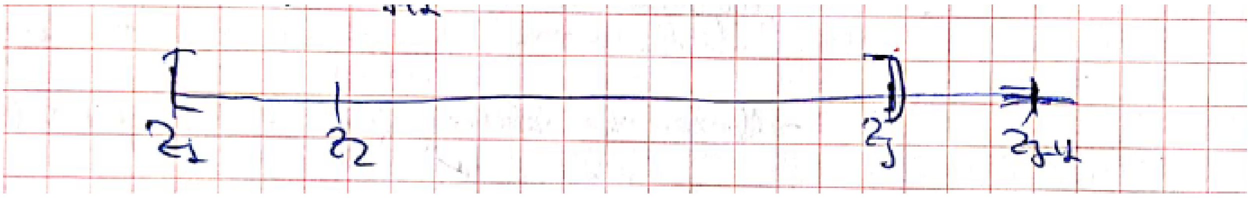

Figure 2: Sorted values $z_1, z_2, \ldots, z_{j+1}$.

$$\text{If } y_k = +1 \text{ then } \operatorname{Loss}(Z_{i,j+1}) = \operatorname{Loss}(Z_{i,j}) - \frac{1}{m} \quad \text{(the loss decreases)}$$

$$\text{If } y_k = 0 \quad \text{then } \operatorname{Loss}(Z_{i,j+1}) = \operatorname{Loss}(Z_{i,j}) + \frac{1}{m} \quad \text{(the loss increases)}$$

General case (in the agnostic scenario)

We have multiple points $x_{k_1}, x_{k_2}, \ldots, x_{k_l} = z_{j+1}$ (l points)

Then: some of the points will have label $1 = p_{j+1}$ some of the points will have label $0 = n_{j+1}$ $p_{j+1} + n_{j+1} = l$

In this case we have that:

$$\operatorname{Loss}(Z_{i,j+1}) = \operatorname{Loss}(Z_{i,j}) - \frac{p_{j+1}}{m} + \frac{n_{j+1}}{m}$$

as p_{j+1} points will be labeled correctly now and n_{j+1} points will be labeled incorrectly now (if $l = 1$, we have $p_{j+1} + n_{j+1} = 1$, so we have just one point labeled positive or negative)

Efficient implementation of the $\text{ERM}_{\mathcal{H}}$ rule for $\mathcal{H}_{\text{intervals}}$

1. Sort S and obtain $x_{\sigma(1)} \le x_{\sigma(2)} \le \cdots \le x_{\sigma(m)}$. Build set Z containing value x without repetition: $Z = \{z_1, z_2, \ldots, z_n\}$, $z_1 = x_{\sigma(1)} < z_2 < \cdots < z_n = x_{\sigma(m)}$

2. Check if all y_i $i = \overline{1,m}$ have value 0. If so, return h_{a_S,b_S}, where $a_S = z_1 - 2$, $b_S = z_1 - 1$. Compute $P = \sum_{i=1}^{m} y_i$ (# positive examples)

3. For $j = \overline{1,n}$

   compute values $\begin{aligned} p_j &= \#\text{ points } x_i = z_j \text{ with label } y_i = 1 \\ n_j &= \#\text{ points } x_i = z_j \text{ with label } y_i = 0 \end{aligned}$

4. min_error $= \frac{m}{m} = 1$, $i^* = []$, $j^* = []$

   for $i = \overline{1,m}$
     for $j = \overline{i,n}$
       $Z_{i,j} = [z_i, z_j]$
       if $(j == i)$
         $\operatorname{Loss}(Z_{i,j}) = \dfrac{P - p_j + n_j}{m}$
       else
         $\operatorname{Loss}(Z_{i,j}) = \operatorname{Loss}(Z_{i,j-1}) + \dfrac{n_j - p_j}{m}$
       if $\operatorname{Loss}(Z_{i,j}) <$ min_error
         min_error $= \operatorname{Loss}(Z_{i,j})$
         $i^* = i$
         $j^* = j$

5. Return i^*, j^*

Complexity:

1. sorting $\mathcal{O}(m \cdot \log m)$
2. computing P – $\mathcal{O}(m)$
3. computing p_j, n_j – $\mathcal{O}(m)$
4. $\text{Loss}(Z_{i,j})$ = constant time

Total: $\mathcal{O}(m^2)$ □

Exercise 2 Let $\mathcal{X} = \mathbf{R}$ and consider $\mathcal{H}$ the class of 3-piece classifiers (signed intervals):

$$\mathcal{H} = \{h_{a,b,s} \colon \mathbf{R} \to \{-1,1\},\ a \leq b,\ s \in \{-1,+1\}\}$$

$$\text{where } h_{a,b,s}(x) = \begin{cases} s & \text{if } x \in [a,b] \\ -s & \text{if } x \notin [a,b] \end{cases}$$

Give an efficient ERM algorithm for class $\mathcal{H}$ and compute its complexity for each of the following cases:

a. realizable case.

b. agnostic case.

Solution. a. realizable case

There exists a function $h_{a^*,b^*,s^*} \in \mathcal{H}$ that labels the training points

$$S = \{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\} \qquad y_i = h_{a^*,b^*,s^*}(x_i)$$

We can have the following possibilities for examples appearing in S:

$$\begin{array}{ll} +++++++++ & \text{(only positive examples)} \\ --------- & \text{(only negative examples)} \\ +++---+++ & \\ ---+++--- & \\ ++++----- & \\ -----++++ & \end{array}$$

Consider the following algorithm

$$\begin{array}{llll} \text{Initialization:} & a_+ = -\infty & a_- = -\infty & \\ & b_+ = +\infty & b_- = +\infty & \\ \text{Compute } & a_+ = \min\limits_{\substack{i=\overline{1,m} \\ y_i=+1}} x_i & & \text{if there is no } x_i \text{ with } y_i = +1 \text{, then } a_+ = -\infty \\ & b_+ = \max\limits_{\substack{i=\overline{1,m} \\ y_i=+1}} x_i & & \text{if there is no } x_i \text{ with } y_i = +1 \text{, then } b_+ = +\infty \\ & a_- = \min\limits_{\substack{i=\overline{1,m} \\ y_i=-1}} x_i & & \text{if there is no } x_i \text{ with } y_i = -1 \text{, then } a_- = -\infty \\ & b_- = \max\limits_{\substack{i=\overline{1,m} \\ y_i=-1}} x_i & & \text{if there is no } x_i \text{ with } y_i = -1 \text{, then } b_- = +\infty \end{array}$$

If $a_+ < a_-$ return $h_{a_-,b_-,-1}$

else return $h_{a+,b+,+1}$

b. agnostic case

Can think of $\mathcal{H}_{\text{signedintervals}} = \mathcal{H}^+_{\text{intervals}} \cup \mathcal{H}^-_{\text{intervals}}$

$$\mathcal{H}^{+}_{\text{intervals}} = \left\{ h^{+}_{a,b} \colon \mathbb{R} \to \{-1,1\},\ a \le b,\ h^{+}_{a,b}(x) = \begin{cases} 1 & x \in [a,b] \\ -1 & x \notin [a,b] \end{cases} \right\}$$

$$\mathcal{H}^{-}_{\text{intervals}} = \left\{ h^{-}_{a,b} \colon \mathbb{R} \to \{-1,1\},\ a \le b,\ h^{-}_{a,b}(x) = \begin{cases} -1 & x \in [a,b] \\ 1 & x \notin [a,b] \end{cases} \right\}$$

Use the algorithm in exercise 1 (efficient implementation of the $\text{ERM}_{\mathcal{H}}$ rule) and run it for $\mathcal{H}^{+}_{\text{intervals}}$ and $\mathcal{H}^{-}_{\text{intervals}}$.

Obtain the hypotheses $h^{+}_{a^*,\,b^*}$ and $h^{-}_{c^*,\,d^*}$.

Choose the one with the minimum empirical risk. □

Exercise 3 (exercise 10.1 in the book)

Boosting the Confidence: Let A be an algorithm that guarantees the following: There exist some constant $\delta_0 \in (0,1)$ and a function $m_{\mathcal{H}} \colon (0,1) \to \mathbb{N}$ such that, for every $\epsilon \in (0,1)$, if $m \ge m_{\mathcal{H}}(\epsilon)$, then, for every distribution $\mathcal{D}$, it holds that, with probability of at least $1-\delta_0$, $L_{\mathcal{D}}(A(S)) \le \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$.

Suggest a procedure that relies on A and learns $\mathcal{H}$ in the usual agnostic PAC learning model and has a sample complexity of

$$m_{\mathcal{H}}(\epsilon,\delta) \le k\, m_{\mathcal{H}}(\epsilon/2) + \left\lceil \frac{2\log(4k/\delta)}{\epsilon^2} \right\rceil$$

where

$$k = \lceil \log(\delta/2)/\log(\delta_0) \rceil$$

Hint: Divide tha data into $k+1$ chunks, where each of the first k chunks is of size $m_{\mathcal{H}}(\epsilon/2)$ examples. Train the first k chunks using A. Argue that the probability that for all these chunks we have $L_{\mathcal{D}}(A(S)) > \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$ is at most $\delta_0^k \le \delta/2$. Finally, use the last chunk to choose from the k hypotheses that A generated from the k chunks (by relying on Corollary 4.6).

Corollary 4.6. *Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, ane let $\ell \colon \mathcal{H} \times Z \to [0,1]$ be a loss function. Then, $\mathcal{H}$ enjoys the uniform convergence property with sample complexity*

$$m^{UC}_{\mathcal{H}}(\epsilon,\delta) \le \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$$

Furthermore, the class is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m_{\mathcal{H}}(\epsilon,\delta) \le m^{UC}_{\mathcal{H}}(\epsilon/2,\delta) \le \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

Solution. A algorithm with the following property: $\exists\, \delta_0 \in (0,1)$ and $m_{\mathcal{H}} \colon (0,1) \to \mathbb{N}$ such that for every $\epsilon \in (0,1)$ if $m \ge m_{\mathcal{H}}(\epsilon)$ then for every distribution $\mathcal{D}$ it holds

$$\underset{S\sim\mathcal{D}^m}{P}\left(L_{\mathcal{D}}(A(S)) \le \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \right) \ge 1 - \delta_0$$

Suggest a procedure based on algorithm A that learns $\mathcal{H}$ in the agnostic PAC setting and has a sample complexity of

$$m_{\mathcal{H}}(\epsilon,\delta) \le k * m_{\mathcal{H}}(\epsilon/2) + \left\lceil \frac{2\log(4k/\delta)}{\epsilon^2} \right\rceil \qquad \text{where } k = \left\lceil \frac{\log\delta/2}{\log\delta_0} \right\rceil$$

Definition of agnostic PAC: $\mathcal{H}$ ia agnostic PAC if there exists a function $m_{\mathcal{H}} \colon (0,1)^2 \to \mathbb{N}$ and a learning algorithm A' with the following property: $\forall\,\epsilon > 0$, $\forall\,\delta > 0$, $\forall\,\mathcal{D}$ distribution function over $Z = \mathcal{X} \times \{0,1\}$ when we run the algorithm A' on a training set S of $m \ge m_{\mathcal{H}}(\epsilon,\delta)$ examples sampled i.i.d. from $\mathcal{D}$, A' returns $h_S = A'(S)$ such that

$$\underset{S\sim\mathcal{D}^m}{P}\left(L_{\mathcal{D}}(h_S) \le \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \right) \ge 1 - \delta$$

This is equivalent to:

$$\underset{S\sim\mathcal{D}^m}{P}\left(L_{\mathcal{D}}(h_S) > \underbrace{\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon}_{\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)+\frac{\epsilon}{2}+\frac{\epsilon}{2}}\right) < \delta$$

Follow the indications.

Let ϵ, $\delta \in (0,1)$. Pick k "chunks" $S_1, S_2, \ldots, S_k$ of size $m_{\mathcal{H}}(\frac{\epsilon}{2})$. Use the property of the algorithm A given.

$$\forall i = \overline{1,k} \qquad A(S_i) = h_i$$

$$\underset{S_i\sim\mathcal{D}^{m_{\mathcal{H}}(\frac{\epsilon}{2})}}{P}\left(L_{\mathcal{D}}(h_i) \le \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h_i) + \frac{\epsilon}{2}\right) \ge 1-\delta_0$$

$$\Leftrightarrow \underset{S_i\sim\mathcal{D}^{m_{\mathcal{H}}(\frac{\epsilon}{2})}}{P}\left(L_{\mathcal{D}}(h_i) > \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h_i) + \frac{\epsilon}{2}\right) < \delta_0 \quad \text{(the probability of having a bad } h_i\text{)}$$

The probability that all $h_i,\ i=\overline{1,k}$ are bad is given by:

$$P\left(L_{\mathcal{D}}(h_1) > \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2} \text{ and } L_{\mathcal{D}}(h_2) > \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2} \text{ and } \ldots\right) < (\delta_0)^k$$

Find k such that $\delta_0^k < \delta/2$

$$\Leftrightarrow k \cdot \ln\delta_0 < \ln\frac{\delta}{2} \,\Big|\, :\ \ln\delta_0$$

$$k \ge \left\lceil \frac{\ln\delta - \ln 2}{\ln\delta_0} \right\rceil$$

Consider $\mathcal{H}' = \{h_1, h_2, \ldots, h_k\}$. $\mathcal{H}'$ finite, apply Corrolary (4.6).

If $m \ge m_{\mathcal{H}}^{UC}(\epsilon/2, \delta/2) \le \left\lceil \frac{2\log(4k/\delta)}{\epsilon^2} \right\rceil$ we have that

$$\underset{S_{k+1}\sim\mathcal{D}^{m_{\mathcal{H}}^{UC}(\epsilon/2,\delta/2)}}{P}\left(L_{\mathcal{D}}(h_{k+1}) > \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2}\right) < \frac{\delta}{2}$$

S_{k+1} has $\left\lceil \frac{2\log(4k/\delta)}{\epsilon^2} \right\rceil$ examples.

So: $L_{\mathcal{D}}(h_{k+1}) > \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$ if either we have

$$\text{A: all } h_i \text{ are bad:} \quad L_{\mathcal{D}}(h_i) > \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2}$$

$$\text{B: } h_{k+1} \text{ is bad:} \quad L_{\mathcal{D}}(h_{k+1}) > \min_{h\in\mathcal{H}'} L_{\mathcal{D}}(h) + \frac{\epsilon}{2}$$

$P(A \cup B) \le P(A) \cup P(B) = \frac{\delta}{2} + \frac{\delta}{2} = \delta.$

So, take $m = k \cdot m_{\mathcal{H}}(\frac{\epsilon}{2}) + \left\lceil \frac{2\log(4k/\delta)}{\epsilon^2} \right\rceil$, $k = \left\lceil \frac{\ln\delta - \ln 2}{\ln\delta_0} \right\rceil$

$$\underset{(\underbrace{S_1, S_2, \ldots, S_k}_{h_1,h_2,\ldots,h_k}, \underset{h_{k+1}}{\overset{S_{k+1}}{\downarrow}})}{P}\left(L_{\mathcal{D}}(h_{k+1}) > \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\right) < \delta \quad \checkmark$$

$\square$