Seminar 5

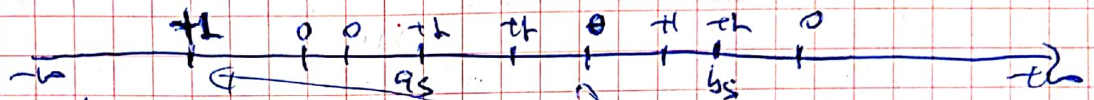
(1) $\mathcal{H}_{intervals} = \mathcal{H}^1_{rec} = \{ h_{a,b} : \mathbb{R} \to \mathbb{R},\ h_{a,b} = \mathbb{1}_{[a,b]},\ h_{a,b}(x) = \begin{cases} 1, x \in [a,b] \\ 0, otherwise \end{cases},\ a,b \in \mathbb{R} \}$

Consider a training set S of size m

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

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

Example:



The solution is h_{a_S, b_S} with loss $\frac{2}{9}$ · (there are 2 misclassified points out of 9)

Observations

1. We are in the agnostic case:
   - it might be the case that there is two labelling function but instead we are dealing with a distribution (same point might have different labels)
   - if there is a labelling function it might not be in $\mathcal{H}_{intervals}$

2. If all points are negative we should return an interval not containing any point in S.

3. If all point are positive should return an interval containing all points in S

We will first sort the training set S in ascending order of x's. So we obtain $S = \{ (x_{\pi(1)}, y_{\pi(1)}), (x_{\pi(2)}, y_{\pi(2)}), \dots (x_{\pi(m)}, y_{\pi(m)}) \}$

with $x_{\pi(1)} \le x_{\pi(2)} \le \dots \le x_{\pi(m)}$

As we are in the agnostic case we can have $x_{\pi(i)} = x_{\pi(i+1)}$ and

$$y_{\pi(i)} \neq y_{\pi(i+1)}$$

Consider the set Z containing the values of x with no repetition

$$Z = \{ z_1, z_2, \dots, z_n \}$$

$$z_1 = x_{\pi(1)} < z_2 < \dots < z_n = x_{\pi(m)},\ n \le m$$

If all initial x values are different than $z_1 = x_{\pi(1)}, \dots z_n = x_{\pi(m)}, n = m$

Idea of the implementation of $ERM_{\mathcal{H}}$

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

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

There are $m + (m-1) + (m-2) + \dots + 1 = \frac{m(m+1)}{2}$ such intervals.

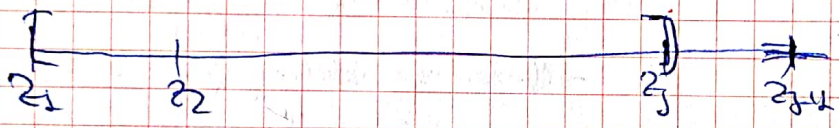
Determine the interval $Z^* = Z_{i^*,j^*}$ with the smallest empirical risk $Z_{i^*,j^*} = \underset{\substack{i=\overline{1,m} \\ j=\overline{i,m}}}{\operatorname{argmin}} \text{Loss}(Z_{ij})$

How to compute very fast $\text{Loss}(Z_{ij})$? Use dynamic programming!

$\text{loss}(Z_{ij}) = \frac{\#\text{negative points inside } Z_{ij} + \#\text{positive points outside } Z_{ij}}{m}$

Key observation: $\text{loss}(Z_{i,j+1})$ can be easily computed based on $\text{loss}(Z_{ij})$

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



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

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

General case (in the agnostic scenario)

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

Then: some of this points will have label $1 \Rightarrow p_{j+1}$

some of this point will have label $0 \Rightarrow n_{j+1}$ $\quad p_{j+1} + n_{j+1} = l$

In this case $\text{Loss}(Z_{i,j+1}) = \text{Loss}(Z_{ij}) - \frac{p_{j+1}}{m} + \frac{n_{j+1}}{m}$

because p_{j+1} points will be labeled correctly now

n_{j+1} points will be labeled incorrectly now

(if $l = 1$ we have $p_{j+1} + n_{j+1} = 1$ just have just one point labeled positive or negative)

Efficient implementation of the ERM_H rule for Thresholds

1. Sort S and obtain $x_{\pi(1)} \le x_{\pi(2)} \le \dots \le x_{\pi(m)}$. Build the set Z containing value x without repetition:
   $Z = \{z_1, z_2, \dots, z_n\}$, $z_1 = x_{\pi(1)} < z_2 < \dots < z_n = x_{\pi(m)}$
2. Check if all y_i, $i = \overline{1,m}$ have label 0. If so return (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 $p_j = \#\text{points } x_i = z_j \text{ with label } y_i = 1$
   $n_j = \#\text{points } x_i = z_j \text{ with label } y_i = 0$
4. $min\_error = \frac{m}{m} = 1$, $i^* = []$, $j^* = []$
   for $i = \overline{1,n}$
   for $j = \overline{i,n}$
   $z_{ij} = [z_i, z_j]$
   if $(j == i)$
   $Loss(z_{ij}) = \frac{P - p_j + n_j}{m}$
   else
   $Loss(z_{ij}) = Loss(z_{i,j-1}) + \frac{n_j - p_j}{m}$
   if $Loss(z_{ij}) < min\_error$
   $min\_error = Loss(z_{ij})$
   $i^* = i$
   $j^* = j$
5. Return i^*, j^*

Complexity: 1 sorting $O(m \log m)$
2 computing $P - O(m)$
3 computing $p_j, n_j - O(m)$
4 computing $Loss(z_{ij})$ = constant time $O(m^2)$

(2) $\mathcal{H} = \{h_{a,b,s} : \mathbb{R} \to \{-1,1\},\ a \le b,\ s \in \{-1,1\}\}$, where

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

a) realizable case

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

$$S = \{(x_1,y_1),(x_2,y_2),\dots,(x_m,y_m)\}$$

$$y_i = h_{a^*,b^*,s^*}(x_i).$$

We can have the following possibilities for examples appearing in S

+ + + + + (only positive examples)

− − − − − (only negative examples)

+ + + − − − + + +

− − − + + + − − −

+ + + + − − − −

− − − − − + + + +

Consider the following algorithm

Initialization: $a_+ = -\infty$, $a_- = -\infty$

$b_+ = +\infty$, $b_- = +\infty$

Compute $a_+ = \min_{\substack{i=1,m \\ y_i=+1}} x_i$, if there is no x_i with $y_i = +1$ then $a_+ = -\infty$

$b_+ = \max_{\substack{i=1,m \\ y_i=+1}} x_i$, // — // — $b_+ = +\infty$

$a_- = \min_{\substack{i=1,m \\ y_i=-1}} x_i$, — // // — $a_- = -\infty$

$b_- = \max_{\substack{i=1,m \\ y_i=-1}} x_i$, — // — $b_- = +\infty$

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

Else return $h_{a_+,b_+,+1}$.

5) agnostic case

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

$$\mathcal{H}^+_{\text{intervals}} = \{ h^+_{a,b} : \mathbb{R} \to \{-1,+1\},\ a \le b \}$$

$$h^+_{a,b}(x) = \begin{cases} 1, & x \in [a,b] \\ -1, & x \notin [a,b] \end{cases}$$

$$\mathcal{H}^-_{\text{intervals}} = \{ h^-_{a,b} : \mathbb{R} \to \{-1,1\},\ a \le b \}$$

$$h^-_{a,b}(x) = \begin{cases} -1, & x \in [a,b] \\ 1, & x \notin [a,b] \end{cases}$$

Use the algorithm in exercise 1 (efficient implementation of the ERM 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 risks.

(3) An algorithm with the following property:

$\exists \delta_0 \in (0,1)$ and $m_{\mathcal{H}} : (0,1) \to \mathbb{N}$ s.t. for every $\varepsilon \in (0,1)$ if $m \ge m_{\mathcal{H}}(\varepsilon)$ then for every D distribution it holds

$$P_{S \sim D^m}\left(L_D(A(S)) \le \min_{h \in \mathcal{H}} L_D(h) + \varepsilon \right) \ge 1 - \delta_0$$

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

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

$$k = \left\lceil \log \frac{\delta}{2} / \log \delta_0 \right\rceil$$

Def of Agnostic PAC: $\mathcal{H}$ is agnostic PAC if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A' with the following property: $\forall \varepsilon > 0$, $\forall \delta > 0$, $\forall D$ distribution over $Z = X \times \{0,1\}$ when running the algorithm A' on a learning set S of $m \ge m_{\mathcal{H}}(\varepsilon,\delta)$ examples sampled i.i.d. from D A' returns $h_S = A'(S)$ s.t.

$$P_{S \sim D^m}\left(L_D(h_S) \le \min_{h \in \mathcal{H}} L_D(h) + \varepsilon \right) \ge 1 - \delta$$

This is equivalent to:

$$P_{S\sim D^m}\left(L_D(h_S) > \min_{h\in\mathcal{H}} L_D(h) + \varepsilon\right) < \delta$$

$\min_{h\in\mathcal{H}} L_D(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$

Follow the indications.

Let $\varepsilon, \delta \in (0,1)$. Pick k "chunks" $S_1, S_2, \dots, S_k$ of size $m_{\mathcal{H}}(\frac{\varepsilon}{2})$

Use the property of the algorithm A given.

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

$$P_{S_i\sim D^{m_{\mathcal{H}}(\frac{\varepsilon}{2})}}\left(L_D(h_i) \le \min_{h\in\mathcal{H}} L_D(h) + \frac{\varepsilon}{2}\right) \ge 1-\delta_0$$

$$\Leftrightarrow \quad P_{S_0\sim D^{m_{\mathcal{H}}(\frac{\varepsilon}{2})}}\left(L_D(h_i) > \min_{h\in\mathcal{H}} L_D(h) + \frac{\varepsilon}{2}\right) < \delta_0.$$

(the probability of having a bad h_i).

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

$$P\left(L_D(h_1) > \min_{h\in\mathcal{H}} L_D(h) + \varepsilon/2 \text{ and } L_D(h_2) > \min_{h\in\mathcal{H}} L_D(h) + \frac{\varepsilon}{2} \text{ and } \dots\right) < (\delta_0)^k$$

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

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

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

Consider $\mathcal{H}' = \{h_1, h_2, \dots, h_k\}$

$\mathcal{H}'$ finite, apply Corollary (4.6)

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

$$P_{S_{k+1}\sim D^{m_{\mathcal{H}'}^{UC}(\varepsilon/2,\delta/2)}}\left(L_D(h_{k+1}) > \min_{h\in\mathcal{H}'} L_D(h) + \frac{\varepsilon}{2}\right) < \frac{\delta}{2}$$

Since $\left\lceil \frac{2\log(4k/\delta)}{\varepsilon^2} \right\rceil$ example

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

A: all h_i are bad : $L_D(h_i) > \min_{h \in \mathcal{H}} L_D(h) + \frac{\varepsilon}{2}$

B: h_{k+1} is bad : $L_D(h_{k+1}) > \min_{h \in \mathcal{H}'} L_D(h) + \frac{\varepsilon}{2}$

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

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

$$P_{(\underbrace{S_1, S_2 \ldots S_{i_1}}_{h_1 \text{ bad}} \ldots \underbrace{S_{i_k} \ldots S_m}_{\text{all}})} \sim D^m \left(L_D(h_{k+1}) > \min_{h \in \mathcal{H}} L_D(h) + \varepsilon \right) < \delta$$