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.

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.

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 $m \geq m_{\mathcal{H}}(\epsilon)$, then, for every distribution $\mathcal{D}$, it holds that, with probability of at least $1 - \delta_0$, $L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$.

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

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

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 $\delta_0^k \leq \delta/2$. 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*

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

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

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