Assignment 2

Due: 30.05.2024, 23:59

Points: 15

The solutions have to be handed in via Moodle. We do not accept late submissions.

We would recommend using LaTeX for writing your submission but also accept handwritten solutions, but please note that if we can not read or understand it, we cannot grade it.

To get full points, always provide the steps in your derivation/proofs and make clear when/how you use known results, for example, from the lecture (e.g. already proven concentration inequalities).

Exercise 2.1: VC Dimension I

Let $v_1, \ldots, v_n \in \mathbb{R}^d$ for some $n < d$. Define the hypothesis class

$$\mathcal{H} = \left\{ x \mapsto sign\left(\sum_{i=1}^{n} \alpha_i \langle v_i, x \rangle + b \right) \mid \alpha_1, \ldots, \alpha_n, b \in \mathbb{R} \right\}$$

1. Show that $\text{VCdim}(\mathcal{H}) \leq n + 1$
2. Prove a necessary and sufficient condition on $v_1, \ldots, v_n$ such that $\text{VCdim}(\mathcal{H}) = n+1$.

(3 + 2 = 5 points)

Exercise 2.2: VC Dimension II

Consider the set $\mathcal{X}_n = \{1, 2, 3, \ldots, n\}$. For any $k \in \mathcal{X}_n$, define the binary classifier

$$h_k : \mathcal{X}_n \to \{0, 1\},\ h_k(x) = \begin{cases} 1 & \text{if } x \text{ is a multiple of } k \\ 0 & \text{otherwise} \end{cases}$$

Let $\mathcal{H}_n = \{h_k \ : \ k \in \mathcal{X}_n\}$ be the hypothesis class of all binary classifiers of above form.

1. For $n = 7$, compute $\text{VCdim}(H_7)$. **Hint:** There's a tight upper bound based on $|\mathcal{H}_7|$.
2. What is the maximum value of n such that $\text{VCdim}(\mathcal{H}_n) = 2$?

(2 + 2 = 4 points)

Exercise 2.3: Uniform Convergence in Transfer Learning

In transfer learning, the goal is to minimise the risk with respect to a target distribution $\mathcal{D}_1$, that is, $\min_{h\in\mathcal{H}} L_{\mathcal{D}_1}(h)$.

However, we have access to few training samples from $\mathcal{D}_1$ and many training samples from a source distribution $\mathcal{D}_2$. Formally let $\beta \in (0,1)$ and assume that the training set S, of size m, is split into βm samples from $\mathcal{D}_1$ and rest from $\mathcal{D}_2$, that is, $S = S_1 \cup S_2$, where $S_1 \sim \mathcal{D}_1^{\beta m}, S_2 \sim \mathcal{D}_2^{(1-\beta)m}$.

We aim to minimise a weighted empirical risk. For $\alpha \in (0,1)$, define the weighted empirical risk of classifier h as

$$L_{S,\alpha}(h) \;=\; \alpha L_{S_1}(h) + (1-\alpha) L_{S_2}(h) \;=\; \frac{\alpha}{\beta m} \sum_{(x,y)\in S_1} \mathbf{1}\{h(x) \neq y\} + \frac{1-\alpha}{(1-\beta)m} \sum_{(x,y)\in S_2} \mathbf{1}\{h(x) \neq y\}$$

You may assume the following:

- $\mathcal{H}$ has a finite number of hypotheses.
- There is a target predictor $h^* \in \mathcal{H}$ such that $L_{D_1}(h^*) = 0$ (equivalently, $\mathcal{D}_1$ is realisable).

Let $\widehat{h}$ minimise $L_{S,\alpha}(h)$. This exercise derives a bound on $L_{\mathcal{D}_1}(\widehat{h})$, i.e. generalisation bounds for $\widehat{h}$, in three steps.

1. Define a $\mathcal{H}$-distance between two distributions $d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = \sup_{h\in\mathcal{H}} \big|L_{\mathcal{D}}(h) - L_{\mathcal{D}'}(h)\big|$. Show that for any h,

$$L_{D_1}(h) \;\leq\; \mathbb{E}_S[L_{S,\alpha}(h)] + (1-\alpha) d_{\mathcal{H}}(\mathcal{D}_1, \mathcal{D}_2).$$

2. Use Hoeffding's inequality and a union bound to show that, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{H}} \big|L_{S,\alpha}(h) - \mathbb{E}_S[L_{S,\alpha}(h)]\big| \leq \sqrt{\frac{1}{2m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{(1-\beta)}\right) \log\left(\frac{2|\mathcal{H}|}{\delta}\right)}.$$

3. Use the bounds from previous parts, and optimality of $\widehat{h}$ to conclude that, with probability $1-\delta$,

$$L_{\mathcal{D}_1}(\widehat{h}) \;\leq\; (1-\alpha)\big(L_{\mathcal{D}_2}(h^*) + d_{\mathcal{H}}(\mathcal{D}_1, \mathcal{D}_2)\big) + \sqrt{\frac{2}{m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{(1-\beta)}\right) \log\left(\frac{2|\mathcal{H}|}{\delta}\right)}$$

(1 + 3 + 2 = 6 points)