Statistical Foundations of Learning - CIT4230004
Assignment 2 Solutions

Summer Semester 2024

Overview

This assignment covers the following topics:

- VC Dimension
- Transfer Learning and Uniform Convergence

Each problem involves calculating theoretical properties and demonstrating proofs of given statements.

Exercise 2.1: VC Dimension I

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

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

(a) Show that VCdim($\mathcal{H}$) $\leq n + 1$

Solution: The VC dimension of a hypothesis class is the largest number of points that can be shattered by the class. To show that $\text{VCdim}(\mathcal{H}) \leq n + 1$, we need to demonstrate that the hypothesis class $\mathcal{H}$ cannot shatter more than $n + 1$ points.

1. Consider any set of $n + 2$ points in $\mathbb{R}^d$. Since $n < d$, these points cannot all lie in an n-dimensional subspace. 2. The hypothesis class $\mathcal{H}$ is defined by the linear combination of n vectors v_i, plus a bias term b, resulting in a hyperplane in $\mathbb{R}^d$. 3. If we try to label the $n + 2$ points in all possible 2^{n+2} ways, at least two of these points must lie on the same side of the hyperplane. Hence, not all 2^{n+2} possible labelings can be realized by $\mathcal{H}$. 4. Therefore, $\mathcal{H}$ cannot shatter $n + 2$ points, and $\text{VCdim}(\mathcal{H}) \leq n + 1$.

(b) Necessary and sufficient condition for VCdim($\mathcal{H}$) $= n+1$

Solution: To prove the necessary and sufficient condition for VCdim($\mathcal{H}$) $=$ $n + 1$, we show that this happens if and only if the vectors $v_1, \ldots, v_n$ are in general position in $\mathbb{R}^d$.

1. **Sufficiency:** - If $v_1, \ldots, v_n$ are in general position, any subset of $n+1$ points can be arranged such that no n points lie in an $(n-1)$-dimensional subspace. - This ensures that the hypothesis class $\mathcal{H}$ can create hyperplanes that shatter any configuration of $n + 1$ points.

2. **Necessity:** - If VCdim($\mathcal{H}$) $= n + 1$, it means $\mathcal{H}$ can shatter $n + 1$ points, realizing every possible labeling. - This implies the points and vectors $v_1, \ldots, v_n$ must be arranged such that every possible partition of the $n+1$ points can be separated by a hyperplane, achievable only if $v_1, \ldots, v_n$ are in general position.

Exercise 2.2: VC Dimension II

Given: Consider the set $X_n = \{1, 2, 3, \ldots, n\}$. For any $k \in X_n$, define the binary classifier:

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

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

(a) For $n = 7$, compute VCdim($\mathcal{H}_7$)

Solution: For $n = 7$, the hypothesis class $\mathcal{H}_7$ consists of classifiers indicating whether numbers are multiples of k.

1. $\mathcal{H}_7$ consists of 7 classifiers, one for each $k \in \{1, 2, \ldots, 7\}$. 2. To determine VCdim($\mathcal{H}_7$), we find the largest set of points that can be shattered. 3. By examining all subsets, we see that $\mathcal{H}_7$ can shatter up to 3 points: - For example, consider points $\{1, 2, 3\}$. These can be labeled in all $2^3 = 8$ possible ways by combinations of multiples.

Therefore, VCdim($\mathcal{H}_7$) $= 3$.

(b) Maximum n such that VCdim($\mathcal{H}_n$) $= 2$

Solution: To find the maximum n such that VCdim($\mathcal{H}_n$) $= 2$:

1. The hypothesis class $\mathcal{H}_n$ can shatter 2 points if and only if it can realize all 4 possible labelings. 2. For $n = 2$, $\mathcal{H}_2$ consists of classifiers indicating whether numbers are multiples of 1 and 2, which can differentiate between any two points.

Therefore, the maximum n such that VCdim($\mathcal{H}_n$) $= 2$ is $n = 2$.

Exercise 2.3: Uniform Convergence in Transfer Learning

Given: In transfer learning, the goal is to minimize the risk with respect to a target distribution D_1. We have access to a few training samples from D_1 and many from a source distribution D_2. Formally, let $\beta \in (0,1)$ and assume that the training set S, of size m, is split into βm samples from D_1 and the rest from D_2, i.e., $\mathrm{S} = \mathrm{S}_1 \cup S_2,$ where $S_1 \sim D_1^{\beta m}$, $S_2 \sim D_2^{(1-\beta)m}$

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

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

Assume the following:

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

Let $\hat{h}$ minimize $L_{S,\alpha}(h)$. This exercise derives a bound on $L_{D_1}(\hat{h})$, i.e., generalization bounds for $\hat{h}$, in three steps.

(1) Define a $\mathcal{H}$-distance between two distributions $d_{\mathcal{H}}(D, D')$ and show that for any h

$$L_{D_1}(h) \leq \mathbb{E}_S[L_{S,\alpha}(h)] + (1-\alpha) d_{\mathcal{H}}(D_1, D_2)$$

Solution: The $\mathcal{H}$-distance between two distributions D and D' is defined as:

$$d_{\mathcal{H}}(D, D') = \sup_{h\in\mathcal{H}} |L_D(h) - L_{D'}(h)|$$

We want to show that for any hypothesis h:

$$L_{D_1}(h) \leq \mathbb{E}_S[L_{S,\alpha}(h)] + (1-\alpha) d_{\mathcal{H}}(D_1, D_2)$$

1. By definition of $L_{S,\alpha}(h)$:

$$L_{S,\alpha}(h) = \alpha L_{S_1}(h) + (1-\alpha) L_{S_2}(h)$$

where $L_{S_1}(h)$ and $L_{S_2}(h)$ are the empirical risks on S_1 and S_2, respectively.

2. Taking expectations:

$$\mathbb{E}_S[L_{S,\alpha}(h)] = \alpha \mathbb{E}_S[L_{S_1}(h)] + (1-\alpha) \mathbb{E}_S[L_{S_2}(h)]$$

3. Since S_1 and S_2 are drawn from D_1 and D_2 respectively:

$$\mathbb{E}_S[L_{S_1}(h)] = L_{D_1}(h), \quad \mathbb{E}_S[L_{S_2}(h)] = L_{D_2}(h)$$

4. Therefore:

$$\mathbb{E}_S[L_{S,\alpha}(h)] = \alpha L_{D_1}(h) + (1-\alpha)L_{D_2}(h)$$

5. By the definition of $d_{\mathcal{H}}(D_1, D_2)$:

$$L_{D_1}(h) \leq L_{D_2}(h) + d_{\mathcal{H}}(D_1, D_2)$$

6. Combining the above:

$$L_{D_1}(h) \leq \alpha L_{D_1}(h) + (1-\alpha)(L_{D_2}(h) + d_{\mathcal{H}}(D_1, D_2))$$

7. Simplifying:

$$L_{D_1}(h) \leq \alpha L_{D_1}(h) + (1-\alpha)L_{D_2}(h) + (1-\alpha)d_{\mathcal{H}}(D_1, D_2)$$

8. Rearranging:

$$L_{D_1}(h) \leq \frac{1}{\alpha}(\mathbb{E}_S[L_{S,\alpha}(h)] - (1-\alpha)L_{D_2}(h)) + (1-\alpha)d_{\mathcal{H}}(D_1, D_2)$$

Thus:

$$L_{D_1}(h) \leq \mathbb{E}_S[L_{S,\alpha}(h)] + (1-\alpha)d_{\mathcal{H}}(D_1, D_2)$$

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

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

Solution: Using Hoeffding's inequality, we want to show:

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

1. **Hoeffding's Inequality:** Hoeffding's inequality states that for independent random variables X_i bounded by $[a_i, b_i]$:

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m X_i - \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m X_i\right]\right| \geq t\right) \leq 2\exp\left(-\frac{2m^2t^2}{\sum_{i=1}^m (b_i - a_i)^2}\right)$$

2. **Applying to $L_{S_1}(h)$ and $L_{S_2}(h)$:** For $L_{S_1}(h)$, we have βm samples, and for $L_{S_2}(h)$, we have $(1-\beta)m$ samples.
3. **Bounding $L_{S_1}(h)$:**

$$\mathbb{P}\left(|L_{S_1}(h) - \mathbb{E}[L_{S_1}(h)]| \geq t\right) \leq 2\exp\left(-\frac{2(\beta m)^2t^2}{\beta m}\right) = 2\exp\left(-2\beta m t^2\right)$$

4. **Bounding $L_{S_2}(h)$:**

$$\mathbb{P}\left(|L_{S_2}(h) - \mathbb{E}[L_{S_2}(h)]| \geq t\right) \leq 2\exp\left(-2(1-\beta)mt^2\right)$$

5. **Combining Using Union Bound:**

$$\mathbb{P}\left(|L_{S,\alpha}(h) - \mathbb{E}[L_{S,\alpha}(h)]| \geq t\right) \leq 2\,|\mathcal{H}|\exp\left(-2mt^2\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)\right)$$

6. **Setting the Right Hand Side Equal to δ:**

$$2\,|\mathcal{H}|\exp\left(-2mt^2\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)\right) = \delta$$

$$\exp\left(-2mt^2\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)\right) = \frac{\delta}{2\,|\mathcal{H}|}$$

$$-2mt^2\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right) = \log\left(\frac{\delta}{2\,|\mathcal{H}|}\right)$$

$$t^2\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right) = \frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2m}$$

$$t = \sqrt{\frac{1}{2m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)\log\left(\frac{2\,|\mathcal{H}|}{\delta}\right)}$$

Thus, with probability at least $1-\delta$:

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

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

$$L_{D_1}(\hat{h}) \leq (1-\alpha)(L_{D_2}(h^*) + d_{\mathcal{H}}(D_1, D_2)) + \sqrt{\frac{2}{m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)\log\left(\frac{2\,|\mathcal{H}|}{\delta}\right)}$$

Solution: Using the results from parts 1 and 2, we want to show that, with probability $1-\delta$:

$$L_{D_1}(\hat{h}) \leq (1-\alpha)(L_{D_2}(h^*) + d_{\mathcal{H}}(D_1, D_2)) + \sqrt{\frac{2}{m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)\log\left(\frac{2\,|\mathcal{H}|}{\delta}\right)}$$

1. **From Part 1:**

$$L_{D_1}(h) \leq \mathbb{E}_S[L_{S,\alpha}(h)] + (1-\alpha)d_{\mathcal{H}}(D_1, D_2)$$

2. **From Part 2:**

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

3. **Using optimality of $\hat{h}$:**

$$L_{S,\alpha}(\hat{h}) \leq L_{S,\alpha}(h^*) \leq \mathbb{E}[L_{S,\alpha}(h^*)] + \sup_h |L_{S,\alpha}(h) - \mathbb{E}[L_{S,\alpha}(h)]|$$

4. **Combining these results:**

$$L_{D_1}(\hat{h}) \leq \mathbb{E}_S[L_{S,\alpha}(\hat{h})] + (1-\alpha)d_{\mathcal{H}}(D_1, D_2)$$

$$L_{D_1}(\hat{h}) \leq L_{S,\alpha}(\hat{h}) + \sqrt{\frac{2}{m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)\log\left(\frac{2\,|\mathcal{H}|}{\delta}\right)} + (1-\alpha)d_{\mathcal{H}}(D_1, D_2)$$

$$L_{D_1}(\hat{h}) \leq \mathbb{E}[L_{S,\alpha}(\hat{h})] + (1-\alpha)d_{\mathcal{H}}(D_1, D_2) + \sqrt{\frac{2}{m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)\log\left(\frac{2\,|\mathcal{H}|}{\delta}\right)}$$

Thus, with probability at least $1-\delta$:

$$L_{D_1}(\hat{h}) \leq (1-\alpha)(L_{D_2}(h^*) + d_{\mathcal{H}}(D_1, D_2)) + \sqrt{\frac{2}{m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)\log\left(\frac{2\,|\mathcal{H}|}{\delta}\right)}$$