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. The following sections explain the concepts, the approach taken to solve each problem, and key points to remember.

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$(\mathcal{H}) \leq n + 1$

Concepts: - **VC Dimension:** Measures the capacity of a hypothesis class to shatter a set of points. A set of points is shattered if the hypothesis class can realize all possible labelings of those points. - **Linear Classifiers:** The given hypothesis class involves linear combinations of vectors, leading to hyperplanes in the feature space.

Approach: To show that VCdim$(\mathcal{H}) \leq n + 1$, we need to prove 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 $\mathcal{H}$ defines a hyperplane in $\mathbb{R}^d$ based on the linear combination of v_i's. 3. The arrangement of $n+2$ points in $\mathbb{R}^d$ implies that at least two points must lie on the same side of the hyperplane defined by $\mathcal{H}$. Therefore, not all 2^{n+2} possible labelings can be realized, proving that $\mathcal{H}$ cannot shatter $n + 2$ points. 4. Hence, VCdim$(\mathcal{H}) \leq n + 1$.

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

Concepts: - **General Position:** A set of vectors is in general position if no subset of n vectors lies in an $(n-1)$-dimensional subspace.

Approach: 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.

Key Points to Remember:

- VC dimension is a measure of the capacity of a hypothesis class. - To find the VC dimension, we need to determine the largest number of points that can be shattered by the hypothesis class. - General position is crucial for determining the exact VC dimension of linear classifiers.

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($\mathcal{H}_7$)

Concepts: - **Binary Classifiers:** Each classifier h_k indicates whether an element is a multiple of k.

Approach: To determine VCdim($\mathcal{H}_7$), we need to find the largest set of points that can be shattered by $\mathcal{H}_7$.

1. $\mathcal{H}_7$ consists of 7 classifiers, one for each $k \in \{1, 2, \ldots, 7\}$. 2. To find the VC dimension, we examine the subsets of $\{1, 2, \ldots, 7\}$ and determine the largest set that can be shattered. 3. We find that $\mathcal{H}_7$ can shatter up to 3 points, such as $\{1, 2, 3\}$, because we can label these points in all $2^3 = 8$ possible ways.

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

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

Approach: 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. This can differentiate between any two points in $\{1, 2\}$.

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

Key Points to Remember:

- The VC dimension of binary classifiers can be determined by analyzing the sets of points and the possible labelings that can be realized. - For simple hypothesis classes like multiples of integers, examining small cases can provide insight into the general behavior.

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.,

$$S = S_1 \cup S_2, \text{ where } S_1 \sim D_1^{\beta m}, S_2 \sim D_2^{(1-\beta)m}$$

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

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

Concepts: - **$\mathcal{H}$-distance:** Measures the maximum difference in the risk of any hypothesis in the class under two different distributions. - **Expectation of Weighted Empirical Risk:** The expected value of the weighted empirical risk over different samples from the source and target distributions.

Approach: 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:

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

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

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

Concepts: - **Hoeffding's Inequality:** Provides a bound on the probability that the sum of bounded independent random variables deviates from its expected value. - **Union Bound:** A technique to combine probabilities of multiple events.

Approach: 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 $L_{S_2}(h)$:**

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

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 $1-\delta$

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

Concepts: - **Optimality of $\hat{h}$:** The hypothesis $\hat{h}$ is chosen to minimize the weighted empirical risk $L_{S,\alpha}(h)$. - **Combining Bounds:** Using the bounds derived in parts 1 and 2 to obtain a final bound on the risk under the target distribution D_1.

Approach: 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 $1-\delta$:

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

Key Points to Remember:

- The $\mathcal{H}$-distance quantifies the maximum difference in risk between two distributions. - Hoeffding's inequality helps in deriving concentration bounds for empirical risk. - The union bound is used to combine probabilities across multiple hypotheses. - Generalization bounds combine empirical risk and distribution differences to bound the true risk.