Assignment 2:
Komal Wavhal
Student Id -20034443

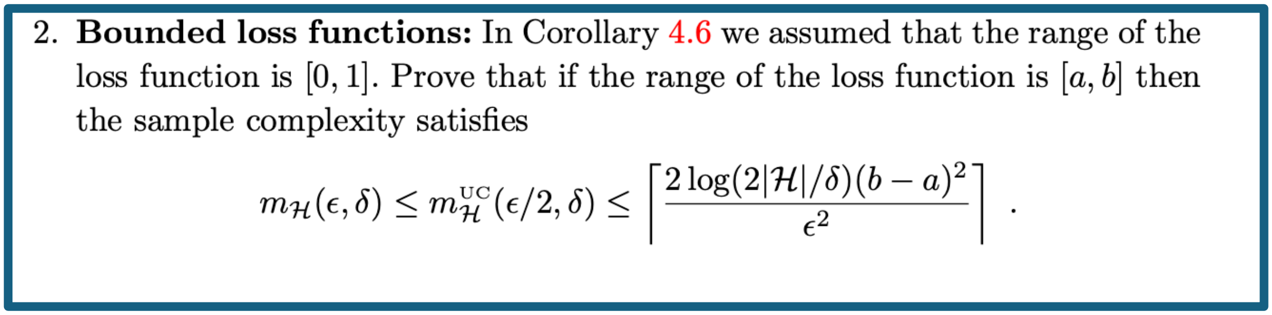
Q1. Exercise 2 of the UML book (the problem starting with "bounded loss function" in boldface).

Q2. Read Section 6.1 through 6.4

Submit your solution to Exercise 2; summarize the intuition you gained from reading chapter 6.

Table of Contents

Q1. Exercise 2 of the UML book (the problem starting with "bounded loss function" in boldface).



2. **Bounded loss functions:** In Corollary 4.6 we assumed that the range of the loss function is $[0,1]$. Prove that if the range of the loss function is $[a,b]$ then the sample complexity satisfies

$$m_{\mathcal{H}}(\epsilon,\delta) \le m_{\mathcal{H}}^{\text{UC}}(\epsilon/2,\delta) \le \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)(b-a)^2}{\epsilon^2} \right\rceil .$$

Solution:

Bounded Loss Functions Proof

Step-by-Step Proof

Step 1: Understanding Sample Complexity

The uniform convergence sample complexity bound describes how many samples are needed to ensure that the empirical loss approximates the expected loss with high probability. For a bounded loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [a,b]$, the variance of the loss is influenced by the range (b-a).

Step 2: Application of Hoeffding's Inequality

Hoeffding's inequality states that if X1, X2,....Xm are independent random variables bounded by [a,b], their empirical mean approximates their expected mean with probability:

$$P\left(|\hat{L} - \mathbb{E}[L]| \ge \epsilon\right) \le 2\exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right)$$

To ensure that the probability of deviation is at most , we set:

$$2\exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right) \le \delta.$$

Rearranging the inequality:

$$m \geq \frac{2\log(2/\delta)(b-a)^2}{\epsilon^2}.$$

For a finite hypothesis class of size |H| , applying the union bound over all hypotheses gives:

$$m \geq \frac{2\log(2|\mathcal{H}|/\delta)(b-a)^2}{\epsilon^2}$$

Since sample complexity must be an integer, we take the ceiling function:

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

More details on the solution:

Sample complexity measures how many samples are needed to guarantee an error at most ε with probability 1 - δ.

2. Rescaling the Loss Function

- If the loss function was originally in [0,1], but now in [a, b], the new loss function can be rescaled by defining:

$$\ell' = \frac{\ell - a}{b - a}$$

where ℓ' is the normalized loss in the range [0,1].

- The difference b - a scales the loss values. The variance of the loss function changes by a factor of (b-a)^2.

3. Applying Standard Sample Complexity Bound

From standard learning theory, the upper bound on sample complexity for a bounded loss function in [0,1] is:

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

Since we have a loss function in [a, b], the effective error tolerance ε needs to be rescaled by the factor (b-a):

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

Conclusion: This result shows that the sample complexity scales quadratically with the range (b - a) of the loss function, meaning that a wider range increases the number of required samples. The dependence on ε-2 and log(|H|)is consistent with uniform convergence bounds in statistical learning theory.

The proof follows from substituting the scaled loss function into the standard bound for sample complexity.

This result shows that if the loss function is scaled, the sample complexity increases proportionally to the squared range of the loss function.

Example: Binary Classification with 0-1 Loss

Consider a binary classification setting where the loss function is the 0-1 loss:

$$\ell(h,(x,y)) = \begin{cases} 0, & \text{if } h(x) = y, \\ 1, & \text{if } h(x) \neq y. \end{cases}$$

Here, the loss function is bounded in the range [0,1]. According to our result, the sample complexity satisfies:

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

This bound tells us that the number of samples required to ensure a small generalization error ε with high probability (1 – delta) depends on the size of the hypothesis space (|H|)

Now, let's generalize this idea.

Example: Loss Function in the Range [a, b]

Now, suppose we use a different loss function, say a scaled version of the hinge loss:

$$\ell_{\text{scaled}}(h,(x,y)) = a + (b-a)\max(0, 1 - y\langle w, x\rangle)$$

Here, the loss values are now in [a, b] rather than [0,1].

Applying our proven result:

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

This means:

- If we increase the range (b-a) of the loss function, the sample complexity increases. A larger range introduces higher variance, requiring more samples for reliable learning.

- If we reduce the range (b-a), fewer samples are required. This is why many algorithms normalize loss functions to [0,1] to keep sample complexity small.

Key Points

1. Why Does $(b\text{-}a)^2$ Appear?

- The loss function's range affects how much variance exists in the estimates. A larger range requires more samples to achieve the same accuracy.

2. Why Log(|H|/δ)?

- This accounts for the number of hypotheses we need to consider in our learning process. If we have more hypotheses, we need more samples to ensure that the worst-case deviation remains small.

3. Why $1/\varepsilon^2$?

- This follows from concentration inequalities (Hoeffding's bound) that describe how many samples are needed to achieve a small error ε.

Q2. Read Section 6.1 through 6.4

Key Concept:

VC-Dimension

The **Vapnik-Chervonenkis (VC) dimension** is a measure of the capacity or complexity of a hypothesis class in machine learning. It defines the largest set of points that can be **shattered** by the hypothesis class, meaning that for any possible binary labeling of those points, there exists a hypothesis in the class that correctly classifies them all.

The **VC-dimension** of a hypothesis class **H** is the largest number of points that can be **shattered** by **H**.

- A set of points is **shattered** if, for **every** possible labeling of those points, there exists a hypothesis in **H** that correctly classifies all of them.
- If no number of points can be shattered, the VC-dimension is **0**.

$$VCdim(H) = d \quad \text{if H can shatter any set of size } d \text{ but not any set of size } d+1.$$

Mathematical Definition:

Intuition Behind VC-Dimension

- A hypothesis class with higher VC-dimension has a greater ability to fit complex patterns but risks **overfitting** to data.
- A lower VC-dimension indicates a more constrained model, which may lead to **underfitting** but better **generalization**.
- A hypothesis class is **probably approximately correct (PAC) learnable** if and only if its VC-dimension is **finite**. This means that with enough training samples, the empirical error will converge to the true error.

Examples of VC-Dimension

1. **Threshold Functions on the Real Line:**
    - A set with one point can always be classified as either 0 or 1, so it is **shattered**.
    - However, with two points, the classification (0,1) cannot be achieved because a single threshold must separate the points.
    - Hence, the **VC-dimension of threshold functions is 1**

2. **Intervals on the Real Line:**
    - A class of interval functions (e.g., indicator functions that return 1 if a point is within an interval and 0 otherwise) can **shatter** any set of two points, but not three.
    - Therefore, the **VC-dimension of interval classifiers is 2**

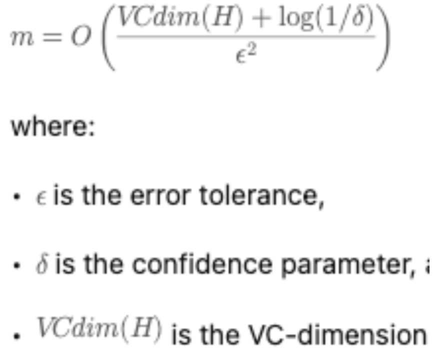
3. **Axis-Aligned Rectangles in 2D:**

- Four points in a 2D space can be arranged so that an axis-aligned rectangle can classify them in all **16 possible ways (2^4 = 16)**.
- However, five points cannot be shattered because one of them will always force an unwanted classification.
- Hence, the **VC-dimension of axis-aligned rectangles in 2D is 4**

4. **Finite Hypothesis Classes:**

- The VC-dimension of a finite class **H** is at most **$\log_2$(|H|)**, but in some cases, it can be significantly lower.

The **fundamental theorem of learning theory** states that a hypothesis class **H** is PAC-learnable if and only if its VC-dimension is finite. It also provides a **quantitative bound** on the number of training samples required for generalization:

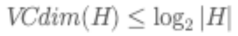


$$m = O\left(\frac{VCdim(H) + \log(1/\delta)}{\epsilon^2}\right)$$

where:

- ϵ is the error tolerance,
- δ is the confidence parameter, a
- $VCdim(H)$ is the VC-dimension

- If **H** is a **finite hypothesis class** with **|H| hypotheses**, then its VC-dimension is at most:



$$VCdim(H) \leq \log_2 |H|$$

- However, in many cases, the VC-dimension is much lower.

6. Infinite-Size Classes Can Be Learnable

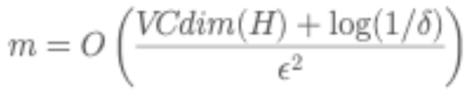
Can infinite hypothesis classes be learnable?

- The answer is **yes**, but only under certain conditions. The key insight is that the **size of a hypothesis class alone does not determine learnability**—instead, its **capacity to fit data (expressed by VC-dimension) does**.

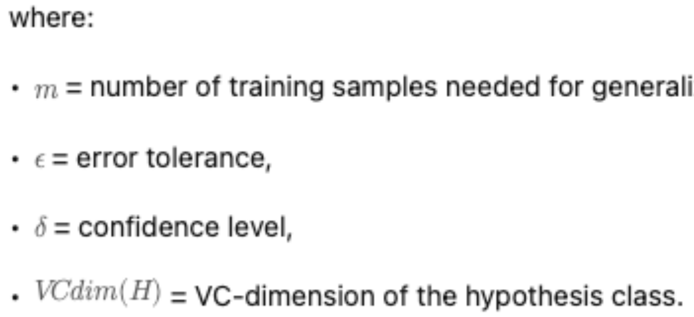
7. The Fundamental Theorem of PAC Learning

A key result in statistical learning theory is that **a hypothesis class is Probably Approximately Correct (PAC) learnable if and only if it has finite VC-dimension**.

Sample Complexity Bound:



$$m = O\left(\frac{VCdim(H) + \log(1/\delta)}{\epsilon^2}\right)$$



where:

- m = number of training samples needed for generalization,
- ϵ = error tolerance,
- δ = confidence level,
- $VCdim(H)$ = VC-dimension of the hypothesis class.

This theorem guarantees that if a class has **finite VC-dimension**, the empirical risk (error on training data) will converge to the true risk (error on unseen data) with high probability.

8. Sauer's Lemma and the Growth Function

• Sauer's Lemma states that **if a hypothesis class has a finite VC-dimension, the number of distinct labelings it can produce grows polynomially rather than exponentially**.

• The **growth function** measures how many different classifications can be realized on **m** points. If the growth function is polynomial, the class is PAC learnable.

9. Uniform Convergence and Learnability

• The **VC-dimension controls how well empirical error (on training data) approximates true error (on unseen data).**

• If a class has finite VC-dimension, **uniform convergence** occurs, meaning training error becomes a good estimate of generalization error.

• If a class has infinite VC-dimension, it cannot be PAC learned.

Key Understanding:

• The VC-dimension quantifies the expressiveness of a hypothesis class.

• PAC Learnability is characterized by finite VC-dimension.

• If a hypothesis class has infinite VC-dimension, it cannot be PAC-learned.

• The sample complexity required for learning is directly related to the VC-dimension.

• A finite VC-dimension ensures learnability in the PAC setting.

• The number of training samples needed for learning depends on the VC-dimension.

• If the VC-dimension is infinite, the hypothesis class is not PAC learnable.