Exercise 1

Prove that for any finite hypothesis class $\mathcal{H}$ and any description language $d : \mathcal{H} \to \{0,1\}^*$, the VC-dimension of $\mathcal{H}$ satisfies

$$\mathrm{VCdim}(\mathcal{H}) \;\leq\; 2 \sup\{\, |d(h)| : h \in \mathcal{H} \,\},$$

where $|d(h)|$ denotes the description length of h.

Furthermore, if d is a prefix-free description language, then

$$\mathrm{VCdim}(\mathcal{H}) \;\leq\; \sup\{\, |d(h)| : h \in \mathcal{H} \,\}.$$

Exercise 1 Solution

Problem overview

This exercise shows that the VC-dimension of a finite hypothesis class is bounded by the maximum number of bits required to describe its hypotheses. In particular, it connects:

- description length (compressibility),
- hypothesis complexity,
- VC dimension,
- the theoretical foundations of Occam's Razor, MDL, and SRM.

Setup and notation

Let $\mathcal{H}$ be a finite hypothesis class. Each hypothesis $h \in \mathcal{H}$ is a binary classifier.

Let

$$d : \mathcal{H} \to \{0,1\}^*$$

be a description language that uniquely identifies each hypothesis. That is, implicitly,

$$h_1 \neq h_2 \;\Rightarrow\; d(h_1) \neq d(h_2).$$

Denote the description length by

$$|d(h)|,$$

and define

$$n := \sup\{\, |d(h)| : h \in \mathcal{H} \,\}.$$

Since $\mathcal{H}$ is finite, this supremum is in fact a maximum.

Part (a): General description language

We first consider the case where no additional structure is assumed on d.

Counting argument Each hypothesis is represented by a distinct binary string of length at most n. The total number of binary strings of length at most n is

$$\sum_{i=0}^{n} 2^i = 2^{n+1} - 1.$$

Hence,

$$|\mathcal{H}| \leq 2^{n+1} - 1.$$

VC-dimension bound For any finite hypothesis class,

$$\mathrm{VCdim}(\mathcal{H}) \leq \lfloor \log_2 |\mathcal{H}| \rfloor.$$

Combining the two inequalities gives

$$\mathrm{VCdim}(\mathcal{H}) \leq \log_2(2^{n+1} - 1) < n + 1 \leq 2n,$$

which proves

$$\boxed{\mathrm{VCdim}(\mathcal{H}) \leq 2 \sup\{\, |d(h)| : h \in \mathcal{H} \,\}.}$$

Part (b): Prefix-free description language

We now assume that the description language d is prefix-free.

Key structural property A prefix-free code corresponds to a set of non-overlapping leaves in a binary tree. This property allows us to apply the Kraft inequality.

Kraft inequality Since d is prefix-free,

$$\sum_{h \in \mathcal{H}} 2^{-|d(h)|} \leq 1.$$

Because $|d(h)| \leq n$ for all $h \in \mathcal{H}$, we have

$$\sum_{h \in \mathcal{H}} 2^{-|d(h)|} \geq |\mathcal{H}| \cdot 2^{-n}.$$

Combining the two inequalities yields

$$|\mathcal{H}| \cdot 2^{-n} \leq 1 \quad \Rightarrow \quad |\mathcal{H}| \leq 2^n.$$

VC-dimension bound Applying the standard VC bound for finite classes,

$$\mathrm{VCdim}(\mathcal{H}) \leq \log_2 |\mathcal{H}| \leq n.$$

Thus,

$$\boxed{\mathrm{VCdim}(\mathcal{H}) \leq \sup\{\, |d(h)| : h \in \mathcal{H}\,\}.}$$

Interpretation

This result formalizes the principle that

> hypothesis classes that can be described succinctly cannot shatter large sets.

In particular:

- shorter descriptions imply fewer hypotheses,
- fewer hypotheses imply smaller VC dimension,
- prefix-free descriptions yield a strictly stronger bound.

This provides a precise mathematical bridge between description length, VC dimension, and the theoretical justification of Occam's Razor, MDL, and Structural Risk Minimization.

Edge Case: $n = 0$

Suppose

$$n = \max_{h \in \mathcal{H}} |d(h)| = 0.$$

Interpretation

- Every hypothesis has description length 0 bits.
- The only binary string of length 0 is the empty string ε.

Constraint imposed by the description language In this problem, the description language d must uniquely identify hypotheses, that is, it must be injective:

$$h_1 \neq h_2 \;\Rightarrow\; d(h_1) \neq d(h_2).$$

However:

- there exists only one binary string of length 0,
- distinct hypotheses cannot be assigned distinct codes.

Therefore, the only possible case is

$$|\mathcal{H}| \leq 1.$$

VC dimension If the hypothesis class contains zero or one hypothesis, it cannot shatter any nonempty set. Hence,

$$\mathrm{VCdim}(\mathcal{H}) = 0.$$

Verification of the bounds

- For part (a), the bound states
$$\mathrm{VCdim}(\mathcal{H}) \leq 2n.$$
Substituting $n = 0$ yields
$$\mathrm{VCdim}(\mathcal{H}) \leq 0,$$
which holds with equality.
- For part (b), the prefix-free bound gives
$$\mathrm{VCdim}(\mathcal{H}) \leq n = 0,$$
which also holds exactly.

Conclusion The extreme case $n = 0$ is consistent with both bounds and confirms that the injectivity requirement of the description language is essential even in degenerate settings.

Exercise 6

In this question we wish to show that the algorithm Memorize is a consistent learner for every class of binary-valued functions over any countable domain.

Let $\mathcal{X}$ be a countable domain and let D be a probability distribution over $\mathcal{X}$.

1. Let $\{x_i : i \in \mathbb{N}\}$ be an enumeration of the elements of $\mathcal{X}$ such that for all $i \leq j$,

   $$D(\{x_i\}) \geq D(\{x_j\}).$$

   Prove that

   $$\lim_{n\to\infty} \sum_{i\geq n} D(\{x_i\}) = 0.$$

2. Given any $\epsilon > 0$, prove that there exists $\epsilon_D > 0$ such that

   $$D\big(\{x \in \mathcal{X} : D(\{x\}) < \epsilon_D\}\big) < \epsilon.$$

3. Prove that for every $\eta > 0$, if n is such that

   $$D(\{x_i\}) < \eta \quad \text{for all } i > n,$$

   then for every $m \in \mathbb{N}$,

   $$\Pr_{S\sim D^m}\Big[\exists x_i : \big(D(\{x_i\}) > \eta \text{ and } x_i \notin S\big)\Big] \leq ne^{-\eta m}.$$

4. Conclude that if $\mathcal{X}$ is countable, then for every probability distribution D over $\mathcal{X}$ there exists a function

   $$m_D : (0,1) \times (0,1) \to \mathbb{N}$$

   such that for every $\epsilon, \delta > 0$, if $m > m_D(\epsilon, \delta)$ then

   $$\Pr_{S\sim D^m}\Big[D(\{x : x \notin S\}) > \epsilon\Big] < \delta.$$

5. Prove that Memorize is a consistent learner for every class of binary-valued functions over any countable domain.

Exercise 6 Solution

(1) Tail mass goes to zero, and why the uncountable case breaks

Part 1: Showing the tail mass tends to 0

Common Setup

Let $\mathcal{X}$ be a countable input space. Hence we may enumerate it as

$$\mathcal{X} = \{x_1, x_2, x_3, \dots\}.$$

Let D be a probability distribution over $\mathcal{X}$, satisfying

$$D(x_i) \geq 0, \qquad \sum_{i=1}^{\infty} D(x_i) = 1.$$

A sample $S \sim D^m$ consists of m independent draws from D (with repetitions allowed). The notation "$x \in S$" means that x appears at least once in the sample.

0. Definition of a Consistent Learner

A learning algorithm A is said to be *consistent* if

$$\forall D,\ \forall f,\ \forall \epsilon > 0, \qquad \lim_{m \to \infty} \mathbb{P}_{S \sim D^m}[\, L_D(A(S)) > \epsilon\,] = 0.$$

Here:

- A is a learning algorithm,
- $h = A(S)$ is the hypothesis learned from sample S,
- $L_D(h)$ is the true risk:
$$L_D(h) = \mathbb{P}_{x \sim D}[h(x) \neq f(x)],$$
- the probability is taken over the randomness of the sample S.

This definition does not require a uniform PAC-style bound for all sample sizes, but only convergence to zero as $m \to \infty$.

1. Tail Mass of a Countable Distribution

Assume without loss of generality that the elements are ordered so that

$$i \leq j \ \Rightarrow\ D(x_i) \geq D(x_j).$$

We claim that

$$\lim_{n\to\infty} \sum_{i\geq n} D(x_i) = 0.$$

Define the partial sums

$$s_n = \sum_{i=1}^{n} D(x_i).$$

Then (s_n) is nondecreasing and bounded above by 1. By the monotone convergence theorem,

$$\lim_{n\to\infty} s_n = 1.$$

Therefore,

$$\sum_{i\geq n} D(x_i) = 1 - \sum_{i=1}^{n-1} D(x_i) \xrightarrow[n\to\infty]{} 0.$$

2. Small Point Masses Have Small Total Probability

For every $\epsilon > 0$, there exists $\epsilon_D > 0$ such that

$$D(\{x : D(x) < \epsilon_D\}) < \epsilon.$$

Using the ordering from Section 1, choose N such that

$$\sum_{i>N} D(x_i) < \epsilon.$$

Set $\epsilon_D = D(x_N)$. Then for all $i > N$, $D(x_i) \leq \epsilon_D$, hence

$$\{x : D(x) < \epsilon_D\} \subseteq \{x_{N+1}, x_{N+2}, \ldots\},$$

which implies

$$D(\{x : D(x) < \epsilon_D\}) \leq \sum_{i>N} D(x_i) < \epsilon.$$

3. Probability That a High-Probability Point Is Missing

Assume that

$$D(x_i) < \eta \quad \forall i > n.$$

Equivalently, there are at most n points with probability larger than η.

Consider the event

$$E = \{\exists i \leq n : \; x_i \notin S\}.$$

For each $i \le n$,

$$\mathbb{P}(x_i \notin S) = (1 - D(x_i))^m \le (1 - \eta)^m.$$

By the union bound,

$$\mathbb{P}(E) \le \sum_{i=1}^{n} (1 - \eta)^m = n(1 - \eta)^m.$$

Using the inequality $1 - t \le e^{-t}$,

$$\mathbb{P}(E) \le ne^{-\eta m}.$$

4. Bounding the Probability of Missing Large Probability Mass

Define the event

$$A = \{D(\{x : x \notin S\}) > \epsilon\}.$$

By Section 2, choose $\eta > 0$ and N such that

$$\sum_{i>N} D(x_i) < \epsilon/2.$$

Decompose the mass of unseen points:

$$D(\{x : x \notin S\}) = \sum_{\substack{i \le N \\ x_i \notin S}} D(x_i) + \sum_{\substack{i > N \\ x_i \notin S}} D(x_i).$$

The second term is at most $\epsilon/2$. Hence, if A occurs, at least one of the first N points must be missing:

$$A \subseteq \{\exists i \le N : \ x_i \notin S\}.$$

Therefore,

$$\mathbb{P}(A) \le \mathbb{P}(\exists i \le N : \ x_i \notin S) \le Ne^{-\eta m}.$$

Choosing

$$m \ge \frac{\log(N/\delta)}{\eta}$$

ensures

$$\mathbb{P}(A) < \delta.$$

5. Consistency of the Memorization Algorithm

Given a labeled sample

$$S = \{(x^{(1)}, f(x^{(1)})), \ldots, (x^{(m)}, f(x^{(m)}))\},$$

define the memorization hypothesis h_S by

$$h_S(x) = \begin{cases} f(x), & x \in S, \\ 0, & x \notin S. \end{cases}$$

Errors can occur only on points not seen in the sample. Thus,

$$\{x : h_S(x) \neq f(x)\} \subseteq \{x : x \notin S\},$$

and

$$L_D(h_S) \leq D(\{x : x \notin S\}).$$

From Section 4, for any $\epsilon > 0$ and $\delta > 0$, there exists m such that

$$\mathbb{P}[D(\{x : x \notin S\}) > \epsilon] < \delta.$$

Hence,

$$\mathbb{P}[L_D(h_S) > \epsilon] < \delta,$$

which implies

$$\lim_{m\to\infty} \mathbb{P}[L_D(h_S) > \epsilon] = 0.$$

Therefore, the memorization algorithm is a consistent learner.

Part 2: Why this strategy breaks on uncountable domains (continuous distributions)

Context The countable-domain proof strategy for Memorize relies on the existence of *atoms* (points with positive probability):

$$D(\{x\}) > 0 \quad \text{for some } x.$$

Then one can select finitely many high-probability points so that almost all probability mass is covered, and argue these points appear in the sample with high probability.

Key structural fact for continuous distributions If $\mathcal{X}$ is uncountable and D is continuous (e.g. $U[0, 1]$), then

$$\boxed{D(\{x\}) = 0 \quad \text{for all } x \in \mathcal{X}.}$$

So there are *no* high-probability points to “capture.”

Finite samples have zero probability mass For a sample of size m,

$$S = \{X_1, \dots, X_m\}$$

is a finite set. Under a continuous distribution,

$$D(S) = D\Big(\bigcup_{k=1}^{m} \{X_k\}\Big) \leq \sum_{k=1}^{m} D(\{X_k\}) = 0.$$

Hence,

$$D(\mathcal{X} \setminus S) = 1 - D(S) = 1.$$

That is, almost every test point is unseen:

$$\Pr_{X \sim D}[X \in S] = D(S) = 0 \quad \Rightarrow \quad \Pr_{X \sim D}[X \notin S] = 1.$$

Why MEMORIZE fails to be consistent The MEMORIZE algorithm predicts the memorized label on points in S, but for $x \notin S$ it outputs a fixed default label (say 0). Since $X \notin S$ almost surely, the prediction is almost always the default label.

Consider a simple realizable counterexample:

$$\mathcal{X} = [0,1], \quad D = U[0,1], \quad f(x) = \mathbf{1}\{x > 1/2\}.$$

If MEMORIZE predicts 0 on all unseen points, then for a fresh test point $X \sim U[0,1]$, we have $X \notin S$ with probability 1, so the prediction is 0. Therefore,

$$L_D(h_S) = \Pr[h_S(X) \neq f(X)] = \Pr[f(X) = 1] = \Pr[X > 1/2] = \frac{1}{2}.$$

This lower bound does not depend on m, so the error does not vanish and MEMORIZE cannot be consistent under such continuous distributions.

Problem 6 summary

- **Countable:** atoms with $D(\{x\}) > 0$ exist ⇒ finitely many "important" points can capture most mass.
- **Uncountable + continuous:** $D(\{x\}) = 0$ for all points ⇒ almost every test point is unseen ⇒ MEMORIZE fails.