Statistical Learning

Tianjiao Nie

September 16, 2025

Contents

1 Introduction

Lemma 1.1 (No free lunch)**.** For every learner A and training set size m, there exists $(\mathcal{D}, f)$ such that

$$\mathrm{Pr}_{S\sim\mathcal{D}^m}[L_{\mathcal{D},f}(h_S) \geq 1/8] \geq 1/7.$$

Remark 1.2. This is a no-go result for learning that is extremely general in the sense that there is no prior knowledge on the target to learn.

2 Probably Approximately Correct Learning

Definition 2.1. We say that $(\mathcal{D}, f, \mathcal{H})$ is realizable if there exists $h \in \mathcal{H}$ such that $L_{\mathcal{D},f}(h) = 0$.

Definition 2.2. A hypothesis class $\mathcal{H}$ is called *PAC learnable* (in the realizable setting) if there exists a function $m_{\mathcal{H}} : (0,1) \times (0,1) \to \mathbb{N}$ and a learning algorithm $S \mapsto h_S \in \mathcal{H}$ with the following property. For every $\epsilon, \delta \in (0,1)$, every distribution $\mathcal{D}$ on $\mathcal{X}$, and every labeling function $f : \mathcal{X} \to \mathcal{Y}$, if the realizability condition holds for $(\mathcal{D}, f, \mathcal{H})$, then running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ many i.i.d. samples from $\mathcal{D}$ labelled by f gives a hypothesis $h \in \mathcal{H}$ such that with probability at least $1 - \delta$ over the choice of the m samples, we have $L_{\mathcal{D},f}(h) \leq \epsilon$. The function $m_{\mathcal{H}}$ is called the *sample complexity* of learning $\mathcal{H}$. In other words, we need estimations of the form

$$\mathrm{Pr}_{S\sim\mathcal{D}^m}[L_{\mathcal{D},f}(h_S) \leq \epsilon] \geq 1 - \delta.$$

Now we define our first learning algorithm.

Definition 2.3. Let $\mathcal{H}$ be a finite hypothesis class. Let $S = \big((x_i, y_i)\big)_{1\leq i\leq m}$ be a given training set. The *empirical loss* of $h \in \mathcal{H}$ is defined as

$$L_S(h) = \frac{1}{m}\big|\{i \mid h(x_i) \neq y_i\}\big|.$$

The *empirical risk minimization* (ERM) learner outputs an $\mathrm{ERM}_{\mathcal{H}}(S) \in \mathcal{H}$ that minimizes the empirical loss.

Lemma 2.4. Let $\mathcal{H}$ be a finite hypothesis class. Let $\epsilon, \delta \in (0,1)$. If $m \geq \ln\bigl(|\mathcal{H}|/\delta\bigr)/\epsilon$, then for every $(\mathcal{D}, f)$ such that $(\mathcal{D}, f, \mathcal{H})$ is realizable, we have

$$\mathrm{Pr}_{S \sim \mathcal{D}^m}\bigl[L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) \leq \epsilon\bigr] \geq 1 - \delta.$$

In particular, the finite hypothesis class $\mathcal{H}$ is PAC learnable with sample complexity

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

Proof. Unwinding the definitions, we need to show that

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\}) \leq \delta.$$

Let $\mathcal{H}_B = \{h \in \mathcal{H} \mid L_{D,f}(h) > \epsilon\}$ be the set of "bad" hypothesis. Let $M = \{S \mid \exists h \in \mathcal{H}_B, L_S(h) = 0\}$ be the set of "misleading" samples. From the realizability condition we see that

$$\{S \mid L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\} \subset M = \bigcup_{h \in \mathcal{H}_B} \{S \mid L_S(h) = 0\}.$$

Hence

$$\begin{aligned}\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\}) &\leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S \mid L_S(h) = 0\}) \\ &\leq |\mathcal{H}_B| \max_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S \mid L_S(h) = 0\}).\end{aligned}$$

Since the elements of S are sampled independently, we have

$$\mathcal{D}^m(\{S \mid L_S(h) = 0\}) = \mathcal{D}(\{x \mid h(x) = f(x)\})^m = (1 - L_{\mathcal{D},f}(h))^m \leq (1 - \epsilon)^m \leq \exp(-\epsilon m)$$

for $h \in \mathcal{H}_B$. Therefore $\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\}) \leq |\mathcal{H}| \exp(-\epsilon m) \leq \delta$. □

3 The Vapnik–Chervonenkis Dimension

Definition 3.1. Assume that $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{H}$ be a hypothesis class. Let $C = \{x_1, \ldots, x_m\} \subset \mathcal{X}$. Let $\mathcal{H}_C$ be the restriction of $\mathcal{H}$ to C. Note that every restriction $h_C \in \mathcal{H}_C$ can be represented as a vector $(h(x_1), \ldots, h(x_m)) \in \{0,1\}^m$. Hence $|\mathcal{H}_C| \leq 2^m$. We say that C is *shattered* by $\mathcal{H}$ if $|\mathcal{H}_C| = 2^m$. The *Vapnik–Chervonenkis (VC) dimension* of $\mathcal{H}$ is defined as

$$\mathrm{VCDim}(\mathcal{H}) = \sup\{|C| \mid C \subset \mathcal{X} \text{ shattered by } \mathcal{H}\}.$$

Example 3.2. Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{x \mapsto \mathrm{sign}(x - \theta) \mid \theta \in \mathbb{R}\}$. It's clear that $C = \{0\}$ is shattered by $\mathcal{H}$, and that every set of size 2 is not shattered by $\mathcal{H}$. Hence $\mathrm{VCDim}(\mathcal{H}) = 1$.

Example 3.3. Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{1_{[a,b]} \mid a < b\}$. It's clear that $C = \{0, 1\}$ is shattered by $\mathcal{H}$, and that every set of size 3 is not shattered by $\mathcal{H}$. Hence $\mathrm{VCDim}(\mathcal{H}) = 2$.

Example 3.4. The VC dimension of axis-aligned rectangles in $\mathbb{R}^d$ is $2d$.

Example 3.5. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{H} = \{x \mapsto \mathrm{sign}(w^\top x) \mid w \in \mathbb{R}^d\}$. The set $\{e_1, \ldots, e_d\}$ is shattered by $\mathcal{H}$. Let $C = \{x_1, \ldots, x_{d+1}\} \subset \mathbb{R}^d$ be a set of size $d + 1$. Then $\alpha_1 x_1 + \cdots + \alpha_{d+1} x_{d+1} = 0$ for some $\alpha_1, \ldots, \alpha_{d+1}$ not all zero. Let $I = \{i \mid \alpha_i > 0\}$ and $J = \{i \mid \alpha_i < 0\}$. Then

$$\sum_{i \in I} \alpha_i x_i = -\sum_{j \in J} \alpha_j x_j.$$

Assume that C is shattered by $\mathcal{H}$. Then there exists $h \in \mathcal{H}$ such that $h(x_i) = 1$ if and only if $i \in I$. In other words, there exists $w \in \mathbb{R}^d$ such that $w^\top x_i > 0$ if and only if $i \in I$. This leads to a contradiction

$$0 < \sum_{i \in I} \alpha_i w^\top x_i = \sum_{j \in J} -\alpha_j w^\top x_j < 0.$$

Therefore $\mathrm{VCDim}(\mathcal{H}) = d$.

Lemma 3.6. Let $\mathcal{H}$ be a finite hypothesis class.

1. $\mathrm{VCDim}(\mathcal{H}) \leq \log_2|\mathcal{H}|$.
2. The gap between $\mathrm{VCDim}(\mathcal{H})$ and $\log_2|\mathcal{H}|$ can be arbitrarily large.

Proof. Here we only give a construction for (2). Let $k \geq 1$. Take $\mathcal{X} = \{1, 2, \ldots, k\}$ and

$$\mathcal{H} = \{x \mapsto \mathrm{sign}(x - \theta + 0.5) \mid \theta \in \{1, \ldots, k\}\}.$$

We have $\mathrm{VCDim}(\mathcal{H}) = 1$ and $\log_2|\mathcal{H}| = \log_2 k$. □

Lemma 3.7. Let $\mathcal{H}$ be a hypothesis class of binary classifiers with VC dimension $d < \infty$. Then $\mathcal{H}$ is PAC learnable with sample complexity bounded by

$$C_1 \frac{d + \ln(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \ln(1/\epsilon) + \ln(1/\delta)}{\epsilon}$$

where $C_1, C_2 > 0$ are absolute constants. Moreover, this sample complexity can be achieved by the ERM learning algorithm.