VC-Dimension Summary

Pingbang Hu

University of Illinois Urbana-Champaign

May 30, 2024



VC-Dimension: Definition

Intuitively, VC-dimension $VC(\mathcal{H})$ of a hypothesis class \mathcal{H} is defined as

▶ largest size of a subset C of X that can be binary labeled **arbitrarily** (*shattered*) by H.

Another useful way to think about this is by its original (combinatorial) formulation:

- ▶ Think about *sets*: binary-labeling $C \Leftrightarrow \text{partitioning } C$ into two subsets.
- ▶ Relate back to our formulation: the vector $(h(c_1), ..., h(c_m))$ is an *indication vector*.
- ▶ Shattering then means \mathcal{H} can partition C arbitrarily.

Intuition

The complexity of $\mathcal H$ on $\mathcal X$ is determined by $VC(\mathcal H)$.

Pingbang Hu (UIUC)

¹In general, for \mathbb{R} -valued functions, one may consider the *fat-shattering dimension*.

VC-Dimension: From Finite Class to Infinite Class

$VC(\mathcal{H})$ characterizes *PAC learnability* of \mathcal{H} since:

- ▶ If $VC(\mathcal{H}) = \infty$: observing any finitely many samples and their labeling doesn't help.
- ▶ If VC(\mathcal{H}) < ∞ : effectively there's only $O(|S|^{VC(\mathcal{H})})$ possibilities for $S \subseteq \mathcal{X}$:
 - ⇒ Uniform convergence is guaranteed: polynomial blow-up of complexity is slow enough.
 - $\Rightarrow \mathcal{H}$ is PAC learnable.

In view of the uniform convergence proof, $VC(\mathcal{H})$ also characterizes the *sample complexity*:

- ▶ By bounding the *Rademacher complexity*; more on that later.
- \triangleright Effectively measures the same thing as VC(·), but in a **probabilistic way**.

Remark

Even $|\mathcal{H}| = \infty$, as long as $VC(\mathcal{H}) < \infty$, it's PAC learnable.

Sauer's Lemma

Lemma (Pajor's lemma)

For any $C \subseteq \mathcal{X}$ with |C| = m, $|\mathcal{H}_C| \le |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|$.

Proof idea.

Induction works since we can divide \mathcal{H} into two class based on their values on, a particular one element. This reduces the problem to a smaller size, where the induction hypothesis applies. \Box

This leads to the following.

Lemma (Sauer-Shelah-Perles lemma)

Let \mathcal{H} be a hypothesis class with $VC(\mathcal{H}) = d < \infty$. Then, for all $m, \tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} {m \choose i}$.

Proof idea.

With Pajor's lemma, by counting the results follow.

Uniform Convergence

Lemma (Uniform convergence for finite VC-Dimension)

For every \mathcal{D} and $\delta \in (0,1)$, with probability at least $1-\delta$,

$$|L_{\mathcal{D}}(h) - L_{\mathcal{S}}(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}.$$

Proof idea.

Standard inequality techniques like Jensen's inequality and the *symmetrization trick*, where:

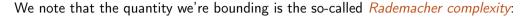
- \blacktriangleright we realize that we can <u>swap</u> S and S' (introduced by $L_{\mathcal{D}}$)
- result is *symmetric*, so we can write everything into one (i.i.d.) quantity with σ_i 's.

Then Hoeffding's inequality bounds the probability, hence the expectation as we want.

This with Sauer's lemma leads to the fundamental theorem of statistical learning theory.

 Pingbang Hu (UIUC)
 VC-Dimension [SB14]
 May 30, 2024
 7 / 10

And More



$$\mathbb{E}_{S,S'\sim\mathcal{D},oldsymbol{\sigma}\sim U^m_\pm}\left[\sup_{h\in\mathcal{H}}rac{1}{m}\left|\sum_{i=1}^m\sigma_i(\ell(h,z_i')-\ell(h,z_i))
ight|
ight].$$

This is the central object in statistical learning theory. More generally:

▶ It's in the form of the *expectation of supermum* of an *empirical process*.

Remark (General theory of empirical process)

More general treatments can be given under the empirical process framework.

Involving ϵ -packing/covering, bracketing bound, Dudley's integral bound, etc.



References



https://www.cambridge.org/core/product/identifier/9781107298019/type/book (visited on 09/03/2023).

Pingbang Hu (UIUC) VC-Dimension [SB14] May 30, 2024 10 / 10