Chapter TWO VC-dimension

Siheng Zhang
zhangsiheng@cvte.com

September 8, 2020

The notes is mainly based on the following books:

- Understanding Machine Learning: From Theory to Algorithms, Shai Shalev-Shwartz and Shai Ben-David, 2014 [1]
- pattern recognition and machine learning, Christopher M. Bishop, 2006 [2]
- Probabilistic Graphical Models: Principles and Techniques, Daphne Koller and Nir Friedman, 2009 [3]
- Graphical Models, Exponential Families, and Variational Inference, Martin J. Wainwright and Michael I. Jordan, 2008 [4]

This part corresponds to **Chapter 2-5 in UML**, and mainly answers the following questions:

- What can we know about the generalization error?
- How does the hypothesis set (in application, the choice of classifier/regressor or so on) reflect our prior knowledge, or, inductive bias?

[1] https://www.cs.huji.ac.il/s̃hais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf
[2] http://users.isr.ist.utl.pt/w̃urmd/Livros/school/Bishop - Pattern Recognition And Machine Learning - Springer 2006.pdf
[3] https://mitpress.mit.edu/books/probabilistic-graphical-models
[4] https://people.eecs.berkeley.edu/w̃ainwrig/Papers/WaiJor08_FTML.pdf

Contents

1 The VC-dimension

1.1 Shattering

Consider the set of threshold functions over the real line $\mathcal{H} = \{h_a(x) = \mathbb{1}_{[x \le a]}, a \in \mathbb{R}\}$. Let a^* be the threshold such that $L_{\mathcal{D}}(h^*) = 0$. Let $a_0 < a^* < a_1$ such that:

$$\mathbb{P}_{x \sim \mathcal{D}_x}[x \in (a_0, a^*)] = \mathbb{P}_{x \sim \mathcal{D}_x}[x \in (a^*, a_1)] = \epsilon$$

If $\mathcal{D}_x(-\infty, a^*) \le \epsilon$, we set $a_0 = -\infty$, and similarly for a_1.

Given a training set S, let $b_0 = \max\{x : (x, 1) \in S\}$ (if no example is positive then $b_0 = -\infty >$, and $b_1 = \min\{x : (x, 0) \in S\}$ (if no example is negative then $b_1 = \infty$). Let b_S be the threshold of an ERM hypothesis h_S, which implies $b_S \in (b_0, b_1)$, then we have

$$\mathbb{P}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(h_S) < \epsilon] \le \mathbb{P}_{S \sim \mathcal{D}^m}[b_0 < a_0] + \mathbb{P}_{S \sim \mathcal{D}^m}[b_1 > a_1]$$

Each term on the right-side is bounded by $(1-\epsilon)^m \le e^{-\epsilon m}$. Let $m > \log(2/\delta)/\epsilon$, then the left-side is bounded by δ. As a result, the hypothesis class is PAC-learnable.

The example above shows that: **finiteness is not a necessary condition for learnability**, and hence we turn to the definition of **shattering**, which describes the ability of a hypothesis set to cover the training set.

The definition of VC-dimension is motivated from the No-Free-Lunch theorem: without restricting the hypothesis class, for any learning algorithm, an **adversary** can construct a distribution for which the learning algorithm will perform poorly, while there is another learning algorithm that will succeed on the same distribution. To make any algorithm fail, the **adversary** used the power of choosing a target function from the set of all possible labelling functions.

When considering PAC learnability of a hypothesis class $\mathcal{H}$, the **adversary** is restricted to constructing distributions for which some hypothesis $h \in \mathcal{H}$ achieves a zero risk. Since we are considering distributions that are concentrated on elements of C, we should study how $h \in \mathcal{H}$ behaves on C.

Definition (Restriction of $\mathcal{H}$ to C): The restriction of $\mathcal{H}$ to C is the set of functions from C to $\{0, 1\}$ that can be derived from $\mathcal{H}$. That is,

$$\mathcal{H}_C = \{(h(c_1), \cdots, h(c_m)) : h \in \mathcal{H}\} \tag{1}$$

where we represent each function from C to $\{0, 1\}$ as a vector in $\{0, 1\}^{|C|}$.

Definition (Shattering): A hypothesis class $\mathcal{H}$ shatters a finite set $C \in \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0, 1\}$. That is, $|\mathcal{H}_C| = 2^{|C|}$.

1.2 The VC-dimension

Definition (VC-dimension): The VC-dimension of a hypothesis class $\mathcal{H}$, denoted VCdim$(\mathcal{H})$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite VC-dimension.

1.2.1 Examples

To calculate the VC-dimension for a hypothesis set, we should show that:

- There **exists** a subset of size d that can be shattered;
- **Every** subset of size $d + 1$ can not be shattered.

1 Threshold functions

2 Fundermental theorem of PAC learning

3 Effective size of a hypothesis class

4 Non-uniform learnability

"non-uniform learnability" allows the sample size to be non-uniform with respect to the different hypotheses with which the learner is competing.

A hypothesis is (ϵ, δ)-competitive with another if

5 Summary

6 Exercises and solutions

To be continue...
Chapter 3. Bayesian-PAC
Chapter 4. Generalization in Deep Learning