Chapter TWO VC-dimension

Siheng Zhang
zhangsiheng@cvte.com

September 7, 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 $\mathcal{H} = \{h_a(x) = \mathbb{1}_{[x \leq a]}, a \in \mathbb{R}\}$. Let a^* be the threshold such that $L_{\mathcal{D}}(h^*) = 0$. Let $a_0 < a^* < a_1$ such that:

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

If $\mathcal{D}_x(-\infty, a^*) \leq \epsilon$, 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

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

Each term on the right-side is bounded by $(1 - \epsilon)^m \leq e^{-\epsilon m}$. 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.

1.2 The VC-dimension

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