Chapter ONE Probably Approximately Correct (PAC)

Siheng Zhang
zhangsiheng@cvte.com

September 2, 2020

The notes is mainly based on the following book

- 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]

Corresponding to Chapter 2-5 in UML.
This part mainly answers the question:

- 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/ shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf
[2] http://users.isr.ist.utl.pt/ wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf
[3] https://mitpress.mit.edu/books/probabilistic-graphical-models
[4] https://people.eecs.berkeley.edu/w̃ainwrig/Papers/WaiJor08_FTML.pdf

Contents

1 Formulation

1.1 The learner's input, output, and evaluation

- **input**:
  - Domain Set: instance $x \in \mathcal{X}$.
  - Label Set: label $y \in \mathcal{Y}$. Currently, just consider the binary classification task.
  - Training data: $S = ((x_1, y_1), \cdots, (x_m, y_m))$ is a finite sequence.
- **output**: hypothesis (or classifier, regressor) $h : \mathcal{X} \to \mathcal{Y}$.
- **data generation model**: Assume that the instances are generated by some probability distribution $\mathcal{D}$, and there is some 'correct' labeling function (currently): $f : \mathcal{X} \to \mathcal{Y}$.

  remark1: The learner is blind to the data generation model.

  remark2: usually called 'training set', but must be 'training sequence', because the same sample may repeat, and some training algorithms is order-sensitive.
- **Generalization error**: *a.k.a*, true error/risk.

$$L_{\mathcal{D},f}(h) \stackrel{def}{=} \mathop{\mathbb{P}}_{x\sim\mathcal{D}}[h(x) \neq f(x)] \stackrel{def}{=} \mathcal{D}(x : h(x) \neq f(x)) \tag{1}$$

1.2 From Empirical Risk Minimization (ERM) to Probably Approximately Correct (PAC)

1.2.1 ERM may lead to overfitting