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/s̃hais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf
[2] http://users.isr.ist.utl.pt/w̃urmd/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): $f : \mathcal{X} \to \mathcal{Y}$.

  The i.i.d. assumption: the training samples are independently and identically distributed.

  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.

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

2 From ERM to PAC

2.1 ERM (Empirical Risk Minimization) may lead to overfitting

Since the generalization error is intractable, turn to minimize the **empirical risk**:

$$L_S(h) \overset{def}{=} \frac{|\{(x_i, y_i) \in S : h(x_i) \neq y_i\}|}{m} \tag{2}$$

Consider a 'lazy' learner h, which predict $y = y_i$ iff. $x = x_i$, and 0 otherwise. It has 1/2 probability to fail for unseen instances, i.e., $L_{\mathcal{D},f}(h) = 1/2$, while $L_S(h) = 0$. Hence, it is an excellent learner on the training set, but a poor learner in the universe case. This phenomenon is called 'overfitting'. The lesson behind this learner is: without restriction on the hypothesis set, ERM can lead to overfitting.

2.2 ERM with restricted hypothesis set (inductive bias)

Instead of $h_S \in \arg\min L_S(h)$, ERM with restricted hypothesis set return the following hypothesis:

$$h_S \in \arg\min_{h \in \mathcal{H}} L_S(h) \tag{3}$$

Start from an ideal case, in which the **realizability assumption** holds, i.e., there exists $h^* \in \mathcal{H}$, such that $L_{\mathcal{D},f}(h^*) = 0$.

It implies that $L_S(h^*) = 0$, $L_S(h_S) = 0$. However, we are interested in $L_{\mathcal{D},f}(h_S)$.

2.3 PAC (Probably Approximately Correct) learnability

Definition: Training on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ samples, there exists an algorithm to be able to achieve **accuracy** at least $1 - \epsilon$ with **confidence** at least $1 - \delta$.

Theorem 1 *Finite hypothesis classes are PAC learnable, and the sample complexity is:*

Proof

3 Summary

Now that, we have come to some important conclusions under the PAC learning framework:

1. No universal learner;
2. Inductive bias is neccessary to avoid overfitting;
3. Sample complexity is function about hypothesis set, confidence level and error, interestingly, it is nothing to do with the dimension of feature space;
4. Inductive bias controls the balance of approximation error and estimation error.

We have reached the fundamental question in learning theory: **Over which hypothesis classes, ERM learning will not result in overfitting (or, PAC learnable)?** Currently, we just confirm the PAC learnability for finite classes. In the next chapter, the most important part in learning theory, VC-dimension, will gives a more precise answer.