Tutorial

Characterizing the Generalization Error of Machine Learning Algorithms via Information Measures

Gholamali Aminian, Yuheng Bu, Iñaki Esnaola, and Samir M. Perlaza

2024 IEEE Information Theory Workshop

The 20th of November, 2024

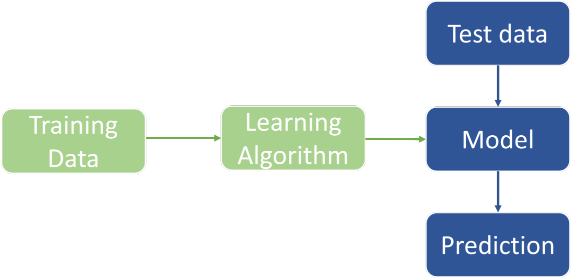
Shenzhen, China

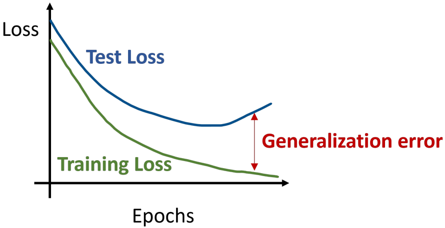
Slides for Part I

Table of Contents (Part I)

Supervised Learning

Generalization Error





Generalization error = Population risk (Test Loss) - Empirical risk (Training Loss)

Supervised Learning

Problem Formulation

- Training data set $S = \{Z_1, \cdots, Z_n\}$, $Z_i = \{X_i, Y_i\} \in \mathcal{Z}$ generated from P_S
- Parameters (weights) of learning model $w \in \mathcal{W}$, e.g., $\hat{Y} = f(X; w)$
- Nonnegative loss function $\ell : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$, e.g., $\ell(w, z) = (y - f(x; w))^2$

Empirical risk (training loss):

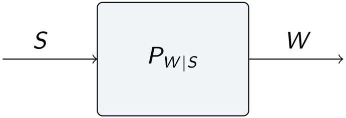
$$L_E(w, s) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i), \qquad \forall w \in \mathcal{W}$$

Population risk (test loss):

$$L_P(w, P_S) \triangleq \mathbb{E}_{P_S}[L_E(w, S)], \qquad \forall w \in \mathcal{W}$$

Learning algorithm can be modeled as randomized mapping: $P_{W|S}$.

- Randomness in initialization
- Stochastic gradient descent (SGD)
- Empirical Risk Minimization (ERM) is a special case



Generalization error:

$$\mathrm{gen}(P_{W|S}, P_S) \triangleq L_P(W, P_S) - L_E(W, S),$$

with W generated from $P_{W|S}$

Generalization Error in Supervised Learning

Different Types of Bounds

- **Single-draw Generalization Error Upper Bound:** Under joint distribution of $P_{W,S}$, following upper bound holds with probability at least $(1-\delta)$,

$$\mathrm{gen}(P_{W|S}, P_S) \leq g(\delta, n),$$

  for a given real function g and $\delta \in (0,1)$,

- **PAC-Bayesian Generalization Error Upper Bound:** Under distribution P_S, following upper bound holds with probability at least $(1-\delta)$,

$$\mathbb{E}_{P_{W|S}}[\mathrm{gen}(P_{W|S}, P_S)] \leq f(\delta, n),$$

  for a given real function f and $\delta \in (0,1)$,

- **Expected Generalization error Upper Bound:** The expectation of generalization error with respect to joint distribution $P_{W,S}$

$$\overline{\mathrm{gen}}(P_{W|S}, P_S) \triangleq \mathbb{E}_{P_{W,S}}[L_P(W, P_S) - L_E(W, S)] \leq h(n),$$

  for a given real function h.

Table of Contents (Part I)

If the **induced function** class $\mathcal{F}_{\ell,W} := \{\ell(w,\cdot) : w \in W\}$ is not 'too rich,' then

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}} |L_P(w, P_S) - L_E(w, S)|\right] \leq \frac{\mathrm{Comp}(\mathcal{F}_{\ell,W})}{\sqrt{n}},$$

where $\mathrm{Comp}(\mathcal{F}_{\ell,W})$ measures complexity of $\mathcal{F}_{\ell,W}$ and does not depend on μ (distribution-free)

Some examples:

- Cardinality of $\mathcal{F}_{\ell,W}$
- VC-dimension [Vapnik, 1999]
- Natarajan-dimension [Holden and Niranjan, 1995]
- Empirical Rademacher complexity [Bartlett and Mendelson, 2002]

We can **always** bound the generalization error as

$$\overline{\mathrm{gen}}(P_{W|S}, P_S) \leq \mathbb{E}\left[\sup_{w \in \mathcal{W}} |L_P(w, P_S) - L_E(w, S)|\right]$$

... but this bound:

- relies on restricting the complexity of the hypothesis space
- ignores the learning algorithm, $P_{W|S}$
- may be too conservative if algorithm does not explore the entire $\mathcal{W}$ due to computational budget.

Learning does not require uniform convergence

One can construct examples of (ℓ, W), where uniform convergence does not hold (the upper bound does not converge to 0 as $n \to \infty$), yet learning still takes place [Shalev-Shwartz and Ben-David, 2014].

Algorithm-dependent Bounds

Uniform Stability

Stability quantifies the sensitivity of algorithm $P_{W|S}$ to local modifications

- replace Z_i with Z_i' in the training data S

$$(Z_1, \cdots, Z_{i-1}, Z_i, Z_{i+1}, \cdots, Z_n) \xrightarrow{P_{W|S}} W$$

$$(Z_1, \cdots, Z_{i-1}, Z_i', Z_{i+1}, \cdots, Z_n) \xrightarrow{P_{W|S}} W^{(i)}$$

- For any learning algorithm

$$\overline{\mathrm{gen}}(P_{W|S}, P_S) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(W, Z_i') - \ell(W^{(i)}, Z_i')]$$

Definition ([Bousquet and Elisseeff, 2002] Uniform Stability)

$P_{W|S}$ is ε-uniformly stable if $\sup_z \mathbb{E}[\ell(W, z) - \ell(W^{(i)}, z)] \leq \varepsilon$.

The stability of learning algorithm $P_{W|S}$ leads to generalization.

- Population risk is the expectation of $\ell(w, s)$ under product of the marginal distributions $P_W P_S$
- Empirical risk is the expectation of $\ell(w, s)$ under joint distribution $P_{W|S} P_S$

Lemma ([Xu and Raginsky, 2017])

Suppose $\ell(w, Z)$ is σ-sub-Gaussian under $Z \sim \mu$ for all $w \in \mathcal{W}$, then

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \sqrt{\frac{2\sigma^2}{n}\mathrm{I}(S; W)},$$

where σ-sub-Gaussian means

$$\log\left(\mathbb{E}\left[e^{\lambda(X-\mathbb{E}(X))}\right]\right) \leq \frac{\sigma^2}{2}\lambda^2$$

- Depends on every ingredient in the supervised learning problem
- Reducing dependence between W and S leads to better generalization bound

Information-theoretic Bounds

The proof is based Donsker-Varadhan variational representation of KL divergence:

$$\mathrm{KL}(P\|Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp f(X)],$$

where $\mathcal{F}$ denotes the set of functions $f : \mathcal{X} \to \mathbb{R}$.

Proof.

- $L_E(w, S)$ is $\frac{\sigma}{\sqrt{n}}$-sub Gaussian for any fixed w.
- Set $f(w, s) = \lambda L_E(w, s) - \lambda \mathbb{E}_S[L_E(w, S)]$ in Donsker-Varadhan

Thus,

$$\begin{aligned}
\mathrm{I}(S; W) &= \mathrm{KL}(P_{W,S} \| P_W P_S) \\
&\geq \mathbb{E}_{P_{W,S}}[\lambda f(W, S)] - \log(\mathbb{E}_{P_{\bar{W}} P_{\bar{S}}} e^{\lambda f(\bar{w}, \bar{s})}) \\
&\geq \lambda \mathbb{E}_{P_{W,S}}[L_E(W, S)] - \lambda \mathbb{E}[L_E(\bar{W}, \bar{S})] - \frac{\lambda^2 \sigma^2}{2n}
\end{aligned}$$

This inequality holds for all $\lambda \in \mathbb{R}$, optimizing over the λ gives the final bound. $\square$

Summary of Existing Generalization Bounds

Traditional ways of bounding generalization errors are not satisfying:

- Do not fully characterize all aspects of learning algorithm
  - only measuring complexity of **functional space** $\mathcal{W}$, e.g., VC dimension
  - only exploring properties of **learning algorithm**, e.g., uniform stability
- Information-theoretical bounds
  - depending on input distribution P_S
  - depending on learning algorithm $P_{W|S}$

  can still be loose.

Our method differs from previous generalization bounds

- instead of a loose bound for general learning algorithms
- **exact** characterization of a specific learning algorithm that has *better structure*

References I

Bartlett, P. L. and Mendelson, S. (2002).

Rademacher and gaussian complexities: Risk bounds and structural results.

Journal of Machine Learning Research, 3(Nov):463–482.

Bousquet, O. and Elisseeff, A. (2002).

Stability and generalization.

Journal of machine learning research, 2(Mar):499–526.

Holden, S. B. and Niranjan, M. (1995).

On the practical applicability of vc dimension bounds.

Neural Computation, 7(6):1265–1288.

References II

Shalev-Shwartz, S. and Ben-David, S. (2014).

Understanding machine learning: From theory to algorithms.

Cambridge university press.

Vapnik, V. N. (1999).

An overview of statistical learning theory.

IEEE transactions on neural networks, 10(5):988–999.

Xu, A. and Raginsky, M. (2017).

Information-theoretic analysis of generalization capability of learning algorithms.

In *Advances in Neural Information Processing Systems*, pages 2524–2533.