

Generalization bounds for neural networks

Nong Minh Hieu
NTU School of Physical & Mathematical Sciences

Statistical Learning Theory - An overview

Notation : We will refer to \mathcal{X} as the **space of features** and \mathcal{Y} as the **space of labels/ground truths**.

Motivation : In supervised machine learning, we are given a sample $S = \left\{ (x_i, y_i) \right\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ and tasked with finding a map $h : \mathcal{X} \rightarrow \mathcal{Y}$.

Problem : It is impossible to find all possible functions h . Hence, we have to hypothesize that h belongs to some class \mathcal{H} . We call \mathcal{H} a **Hypothesis class**.

Statistical Learning Theory - An overview

Examples of hypothesis classes : We encounter some common hypothesis classes in several statistical learning problems. For example:

- *Regression* : We hypothesize that the maps $h : \mathcal{X} \rightarrow \mathcal{Y}$ is linear.

$$\mathcal{H} = \left\{ X \mapsto \beta X : \beta^T, X \in \mathbb{R}^d \right\}$$

- *Classification* : One possible hypothesis class is the set of logit functions applied to linear functions.

$$\mathcal{H} = \left\{ X \mapsto \sigma(\beta X) : \beta^T, X \in \mathbb{R}^d, \sigma(z) = \frac{1}{1 + e^{-z}} \right\}$$

Empirical Risk Minimization

Risk : Once we have settled on a hypothesis class. We need a criterion to choose a specific function from the class. We call this criterion **risk**. There are multiple notions of risk but for the sake of simplicity, we consider the risk for binary classification ($\mathcal{Y} = \{-1, 1\}$).

$$R(h) = P(h(X) \neq Y) = \mathbb{E}_{XY}[\mathbf{1}\{h(X) \neq Y\}], \quad h \in \mathcal{H}$$

One problem with the above criterion is that:

- We are assuming a distribution P_{XY} over $\mathcal{X} \times \mathcal{Y}$.
- We have no way of knowing P_{XY} .

Empirical Risk Minimization

Empirical Risk : Since we cannot find P_{XY} easily, we use the notation of **empirical risk** instead:

$$\hat{R}(h) = \frac{1}{|S|} \sum_{(x_i, y_i) \in S} \mathbf{1}\{h(x_i) \neq y_i\}, \quad h \in \mathcal{H}$$

The process of finding the best $h \in \mathcal{H}$ using the above criterion is called **empirical risk minimization (ERM)**.

Motivation : the solution for ERM is only considered the solution of the learning problem if and only if we can guarantee that the **generalization gap** is small enough. The gap is formalized using the following probability:

$$P\left(R(h) - \hat{R}(h) \leq \epsilon\right) \geq 1 - \delta$$

Where $\delta \in (0, 1)$ is an arbitrary granularity and $\epsilon > 0$ is the gap that depends on the granularity, sample training size and the complexity of the hypothesis class.

Theorem (One-sided Rademacher bound)

Let $\mathcal{H} \subseteq [a, b]^{\mathcal{X}}$ be a class of function with output in the interval $[a, b]$. Then, with probability of at least $1 - \delta$, $\delta \in (0, 1)$, we have:

$$R(h) \leq \hat{R}(h) + 2\mathfrak{R}_n(\mathcal{H}) + (b - a)\sqrt{\frac{\log 1/\delta}{2n}}$$

Where \mathfrak{R}_n denotes the **Rademacher complexity** of the hypothesis class.