CUHK STAT6050: Statistical Learning Theory Spring 2022

Lecture 1: Introduction to Statistical Learning Theory

Lecturer: Ben Dai

"*There is Nothing More Practical Than A Good Theory.*" — Kurt Lewin

1 Overview

In [Von Luxburg and Schölkopf, 2011]: *"Statistical learning theory is regarded as one of the most beautifully developed branches of artificial intelligence. It provides the theoretical basis for many of today's machine learning algorithms. The theory helps to explore what permits to draw valid conclusions from empirical data."*

This course mainly focuses on the subset of statistical learning theory which is highly related to supervised statistical methodologies. Following are some specific purposes:

- (Justification). Theoretical analysis of machine learning methods with a large-scale dataset. The methods can be arbitrary, ranging from parametric models to deep neural networks. For example, to asymptotically show that Method A is better than Method B; to find conditions under which Method A is better; or to determine whether a method is the best one. (**Asymptotics**; **excess risk bound**; **Consistency**; **Convergence rate**; **Minimax rate**.)
- (Explore new methods). Most machine learning methods are motivated by { **intuition** | **numerical studies** | **theory** }. Statistical learning theory is one of the most important ways to motivate a useful method. For example, SVM (**VC-dimension**), new surrogate losses in classification (**Fisher/excess risk consistency**), random forest (**bias-variance trade-off**), local smoothing (**nonparametric statistics**), ...

2 Framework

The content of this section is:

- Define a **risk** function to measure the performance of a decision function.
- Define the **Bayes rule** and an **excess risk** to measure "efficiency" of a decision function.

A **risk** function is introduced to measure predictive performance. Given a decision function f, its predictive performance is computed as

$$R(f) = \mathbb{E}\Big(l\big(\mathbf{Y}, f(\mathbf{X})\big)\Big).$$

Table 1: Notations in supervised learning

		Dataset
$\mathscr{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1,\cdots,n}$	$\triangleq$	Training set with n samples, where $(\mathbf{X}_i, \mathbf{Y}_i) \overset{d}{=} (\boldsymbol{X}, \boldsymbol{Y})(i = 1, \cdots, n)$ are i.i.d. random samples on a probability space with the probability measure $\mathbb{P}$.
$\mathbf{X}$	$\triangleq$	Features or inputs of a sample. $\mathbf{X} \in \mathscr{X} \subset \mathbb{R}^d$ is a d-length (random) vector.
$\mathbf{Y}$	$\triangleq$	Response or outcome of a sample. $\mathbf{Y} \in \mathscr{Y} \subset \mathbb{R}^K$ is a K-length vector.
		Learning paradigm
$f(\cdot)$	$\triangleq$	A decision function. $f : \mathscr{X} \to \mathbb{R}^K; \mathbf{x} \to f(\mathbf{x})$ maps the input (feature) space to the outcome space, say the decision function is $f(\mathbf{x})$ given a sample $\mathbf{X} = \mathbf{x}$.
$l(\cdot, \cdot)$	$\triangleq$	A loss function. $l : \mathscr{Y} \times \mathbb{R}^K \to \mathbb{R}; \big(\mathbf{y}, f(\mathbf{x})\big) \to l\big(\mathbf{y}, f(\mathbf{x})\big)$ measure the discrepancy between the true outcome and the decision function.
$R(f)$	$\triangleq$	The risk of the decision function. $$R(f) \triangleq \mathbb{E}\Big(l\big(\mathbf{Y}, f(\boldsymbol{X})\big)\Big)$$

Note that the expectation is taken w.r.t. both $\mathbf{X}$ and $\mathbf{Y}$. Given a testing dataset $\mathscr{T}_m = (\mathbf{X}_j^{\text{te}}, \mathbf{Y}_j^{\text{te}})_{j=1,\cdots,m}$, the risk function is empirically evaluated as an averaged loss:

$$\widehat{R}_m(f) = \frac{1}{m} \sum_{j=1}^{m} l\big(\mathbf{y}_j^{\text{te}}, f(\mathbf{x}_j^{\text{te}})\big).$$

The risk function can be used to check the performance of a decision function, yet we want to further investigate its "efficiency". To this end, we first introduce the best decision function, namely Bayes decision function (rule), then compute the discrepancy to measure "efficiency".

Definition 2.1 (Bayes decision rule)**.** A Bayes (decision) rule is defined as the smallest risk achievable by any measurable decision function, that is,

$$f^* = \arg\min_f R(f),$$

where the minimum is taken over all possible measurable functions.

We illustrate the risk function and its Bayes rule by following two examples.

Lemma 2.2 (Mis-classification error)**.** *The misclassification error (MCE) in binary classification ($Y \in \{-1,+1\}$) is defined as:*

$$R(f) = \mathbb{P}\Big(Y \neq \operatorname{sgn}(f(\mathbf{X}))\Big) = \mathbb{E}\Big(\mathbf{1}\big(Y \neq \operatorname{sgn}(f(\mathbf{X}))\big)\Big) = \mathbb{E}\Big(\mathbf{1}\big(Yf(\mathbf{X}) \leq 0\big)\Big),$$

and f^* is a Bayes rule iff

$$\operatorname{sgn}\big(f^*(\mathbf{x})\big) = \operatorname{sgn}\big(\mathbb{P}_{Y|\mathbf{X}}(Y=1|\mathbf{X}=\mathbf{x}) - 1/2\big).$$

Remark 2.3. f^* in binary classification is *non-identifiable*.

Lemma 2.4 (Mean squared error)**.** *The mean squared error (MSE) in (multi-outcome) regression ($\mathbf{Y} \in \mathbb{R}^K$) is defined as:*

$$R(f) = \mathbb{E}\Big(\big(\mathbf{Y} - f(\mathbf{X})\big)^2\Big),$$

and the Bayes rule is defined as:

$$f^*(\mathbf{x}) = \mathbb{E}\big(\mathbf{Y}|\mathbf{X}=\mathbf{x}\big).$$

Once the Bayes rule is obtained, we can define the best risk as $R^* = R(f^*)$, which is the best performance you can achieve. To measure "efficiency", the **excess risk** is introduced:

$$\mathscr{E}(f) = R(f) - R^*.$$

Note that $\mathscr{E}(f) \geq 0$, since $R(f) \geq R^*$. Now, we want to check the performance and efficiency of our finite-sample estimator via ERM.

Before that, we would like to point out a probabilistic perspective of ERM. Note that our final goal is to find a minimizer of the risk function at the population level

$$\min_f R(f) = \min_f \mathbb{E}\big(l(\mathbf{Y}, f(\mathbf{X}))\big).$$

Two issues are likely to stand out. (i) We have no idea about calculating the expectation, since we don't want to make any assumption on data distribution. (ii) The minimum is taken over all measurable functions, which is infeasible to optimize.

To address (i), the strategy of ERM is to replace the population mean by the empirical average on a training dataset. This is the key to "*learning from data*": good performance in training set yields good performance in testing set or in population. The assumption of this framework is that the training set and testing set are i.i.d. samples[1]. To address (ii), we introduce a candidate class $\mathscr{F}$, usually a function space index by some parameters, yet it can be a general functional space as in nonparametric methods.

Now, the formulation of ERM is given as:

$$\widehat{f}_n = \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} l\big(\mathbf{y}_i, f(\mathbf{x}_i)\big), \tag{1}$$

[1] One may check "transfer learning" when training set and testing set have different distributions.

where $\widehat{f}_n$ is the final estimator we obtained from the training set. Then, we aim to quantify the performance of $\widehat{f}_n$:

$$\mathscr{E}(\widehat{f}_n) = R(\widehat{f}_n) - R^*.$$

Remark 2.5. $\mathscr{E}(\widehat{f}_n)$ is random, and its randomness is caused by $\widehat{f}_n$, which is estimated from random samples in the training set $\mathscr{D}_n$.

To measure the performance based on the random criteria, we introduce three concepts:

- **Consistency.** $\widehat{f}_n$ is consistent w.r.t. the risk $R(\cdot)$ if

$$R(\widehat{f}_n) \xrightarrow{\mathbb{P}} R^*, \quad \text{as} \quad n \to \infty.$$

- **Convergence rate.** Suppose that $\delta_n \to 0$, and $\widehat{f}_n$ satisfies that

$$\mathscr{E}(\widehat{f}_n) = R(\widehat{f}_n) - R^* = O_P(\delta_n),$$

  then δ_n is the convergence rate of $\mathscr{E}(\widehat{f}_n)$.

- **Probabilistic bounds.** For any $\varepsilon > 0$, there exists $N_0(\varepsilon)$, for $n > N_0(\varepsilon)$

$$\mathbb{P}\big(\mathscr{E}(\widehat{f}_n) \geq \delta_n'(\varepsilon)\big) \leq \varepsilon,$$

  provided that some $\delta_n'(\varepsilon) \to 0$, as $n \to \infty$.

Remark 2.6. Probabilistic bound $\Longrightarrow$ Convergence rate $\Longrightarrow$ Consistency,
where each step provides progressively coarser information.

Example 2.7 (Toy example)**.** ***Data.*** *Suppose* $(Y_1, \cdots, Y_n)$ *is a sequence of i.i.d. random samples with* $\mathbb{E}(Y_i) = \mu = 0$ *and* $\mathbf{Var}(Y_i) = \sigma = 1$. ***Risk.*** $R(\theta) = \mathbb{E}l(Y, \theta) = \mathbb{E}\big((Y-\theta)^2\big)$.

Bayes decision function: $\theta^* = \mathbb{E}(Y) = \mu$.
Empirical estimator: $\widehat{\theta} = \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ *is a function of* $(Y_1, \cdots, Y_n)$.
Then, the excess risk is

$$\mathscr{E}(\widehat{\theta}) = R(\widehat{\theta}) - R^* = \mathbb{E}\big((Y-\widehat{\theta})^2\big) - \mathbb{E}\big((Y-\mu)^2\big) = \mathbb{E}(\widehat{\theta}^2) = \widehat{\theta}^2.$$

Note that the expectation is taken w.r.t. Y, which is independent of $(Y_1, \cdots, Y_n)$.

- *Probabilistic bound. For any* $\delta > 0$,

$$\mathbb{P}\big(\mathscr{E}(\widehat{\theta}) \geq \delta^2\big) = \mathbb{P}\big(\widehat{\theta}^2 \geq \delta^2\big) = \mathbb{P}\big(|\widehat{\theta}| \geq \delta\big) \leq \frac{1}{n\delta^2},$$

  where the last inequality follows from Chebyshev's inequality. Alternatively, we can say, for any $\varepsilon > 0$,

$$\mathbb{P}\big(\mathscr{E}(\widehat{\theta}) \geq \frac{1}{\varepsilon^2 n}\big) \leq \varepsilon.$$

- *Convergence rate and excess risk consistency.*

$$\mathscr{E}(\widehat{\theta}) = O_P(1/n).$$

A O_P and o_P notations

Definition A.1 (O_P notation)**.** For a set of random variables $\{X_n\}_{n\in\mathbb{N}}$ and a set of constants $\{a_n\}_{n\in\mathbb{N}}$, we say $X_n = O_P(a_n)$ as $n \to \infty$, if for any $\varepsilon > 0$, there exists a finite constant $\delta(\varepsilon) > 0$ and $N_0(\varepsilon) \in \mathbb{N}$ such that $\mathbb{P}\big(|X_n/a_n| \geq \delta(\varepsilon)\big) < \varepsilon$, for any $n > N_0(\varepsilon)$.

Definition A.2 (o_P notation)**.** For a set of random variables $\{X_n\}_{n\in\mathbb{N}}$ and a set of constant $\{a_n\}_{n\in\mathbb{N}}$, we say $X_n = o_P(a_n)$ as $n \to \infty$, if for any $\varepsilon, \delta > 0$, there exists $N_0(\varepsilon, \delta) \in \mathbb{N}$ such that $\mathbb{P}\big(|X_n/a_n| \geq \delta\big) < \varepsilon$, for any $n > N_0(\varepsilon, \delta)$.

Equivalently, we can say $X_n = o_P(a_n)$ as $n \to \infty$ if $X_n/a_n \to 0$ in probability, that is,

$$\lim_{n\to\infty} \mathbb{P}\big(|X_n/a_n| \geq \delta\big) = 0,$$

for any $\delta > 0$.

Intuition: The sequence $\{X_n/a_n\}$ is stochastically bounded (or a reasonable random variable)—it does not diverge to infinity in probability.

Example A.3. *1. If $X_n = X \sim N(0,1)$ for all n, then $X_n = O_P(1)$.*

2. If $X_n = \bar{X}_n$ is the sample mean of $\{X_i\}_{i=1}^n$, under the conditions of CLT, then $\bar{X}_n = O_P(1/\sqrt{n})$ and $\bar{X}_n = o_P(1)$.

References

[Von Luxburg and Schölkopf, 2011] Von Luxburg, U. and Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic*, volume 10, pages 651–706. Elsevier.