CSIE5002 Prediction, learning, and games

Lecture 3: Introduction to statistical learning II

Yen-Huan Li (yenhuan.li@csie.ntu.edu.tw)

04.03.2019

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

This lecture is a continuation of Lecture 2. In particular, this lecture introduces several standard complexity measures in statistical learning theory.

Related advanced topics (1/2)

- Generic chaining
  - M. Talagrand. *Upper and Lower Bounds for Stochastic Processes*. 2014.
  - W. Bednorz and R. Latala. On the boundedness of Bernoulli processes. 2014.

- Local Rademacher complexity
  - V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. 2006.
  - P. Bartlett *et al.* Local Rademacher complexities. 2005.

Related advanced topics (2/2)

- Learning with sparsity
  - V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. 2011.
  - M. Wainwright. *High-Dimensional Statistics*. 2019.

- Generalization error analysis of deep neural networks
  - P. Bartlett *et al.* Spectrally-normalized margin bounds for neural networks. 2017.
  - N. Golowich *et al.* Size-independent sample complexity of neural networks. 2018.

Table of contents

Empirical Rademacher complexity

Recap: Result for binary classification

Theorem 1. Consider the binary classification problem with the 0-1 loss, where $\mathcal{H}$ is a class of $\{\pm 1\}$-valued functions. Then, for every $\delta \in]0,1[$, it holds with probability at least $(1-\delta)$ that

$$R(h) \leq R_n(h) + C_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}, \quad \forall h \in \mathcal{H}.$$

Question. How do we compute the Rademacher complexity?

M. Mohri *et al.* *Foundations of Machine Learning*. 2012.

Empirical Rademacher complexities

Let $z_1, \ldots, z_n$ be i.i.d. random variables taking values in $\mathcal{Z}$. Let $\mathcal{F}$ be class of functions mapping from $\mathcal{Z}$ to $\mathbb{R}$.

Definition. (Empirical Rademacher complexity) The associated *empirical Rademacher complexity (ERC)* of a function class $\mathcal{F}$ is given by

$$\hat{C}_n(\mathcal{H}) \coloneqq \mathsf{E}_{\sigma_1,\ldots,\sigma_n} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i),$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher r.v.'s independent of $z_1, \ldots, z_n$.

Remark. Then, we have $C_n(\mathcal{H}) = \mathsf{E}_{z_1,\ldots,z_n} \hat{C}_n(\mathcal{H})$.

V. Koltchinskii. Rademacher penalties and structural risk minimization. 2001.
P. Bartlett *et al.* Rademacher and Gaussian complexities: Risk bounds and structural results. 2002.

Concentration of the ERC

Observation. As $\hat{C}_n(\mathcal{H}) = \mathsf{E}\, C_n(\mathcal{H})$, we expect that $\hat{C}_n(\mathcal{H})$ is close to $C_n(\mathcal{H})$ when n is large enough.

Proposition 1. (Concentration of the ERC) Suppose that $\mathcal{F}$ is a class of functions from $\mathcal{Z}$ to $[0,1]$. Then, it holds with probability at least $(1-\delta)$ that

$$C_n(\mathcal{F}) \leq \hat{C}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Question. What is the probability space in the proposition?

Proof of Proposition 1

Proof. (Proposition 1) Define the function

$$\varphi(\sigma_1, \dots, \sigma_n) \coloneqq \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i).$$

Then, by McDiarmid's inequality, it holds with probability at least $(1 - \delta)$ that

$$\varphi(\sigma_1, \dots, \sigma_n) \leq \mathsf{E}\, \varphi(\sigma_1, \dots, \sigma_n) + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

The proposition follows.

Generalization error in terms of the ERC

Corollary 1. Consider the binary classification problem with the 0-1 loss, where $\mathcal{H}$ is a class of $\{\pm 1\}$-valued functions. Then, for every $\delta \in]0,1[$, it holds with probability at least $(1-\delta)$ that

$$R(h) \leq R_n(h) + \hat{C}_n(\mathcal{H}) + 3\sqrt{\frac{\log(1/\delta)}{2n}}, \quad \forall h \in \mathcal{H}.$$

Proof. Recall the proof of Theorem 1. With probability at least $(1 - 1/(2\delta))$, it holds that

$$R(h) \leq R_n(h) + 2C_n(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{2n}}, \quad \forall h \in \mathcal{H}.$$

It remains to apply Proposition 1 and the fact that

$$C_n(\mathcal{F}) = C_n\mathcal{H}.$$

VC-dimension

Growth function

Definition. (Growth function) The *growth function* of a class $\mathcal{F}$ of functions defined on $\mathcal{Z}$ is given by

$$G_n(\mathcal{F}) := \max_{z_1,\ldots,z_n \in \mathcal{Z}} |\{ (f(z_1), \ldots, f(z_n)) \mid f \in \mathcal{F} \}|,$$

where $|\cdot|$ denotes the cardinality function.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. 1971.

Massart's lemma

Lemma 1. (Massart's lemma) Let $\mathcal{X} \subset \mathbb{R}^n$ be such that $|\mathcal{X}| < +\infty$. Let $\sigma \in \mathbb{R}^n$ be a vector of i.i.d. Rademacher r.v.'s. Then, it holds that

$$\mathsf{E}_\sigma \sup_{x \in \mathcal{X}} \frac{1}{n} \langle \sigma, x \rangle \leq \frac{r\sqrt{2 \log |\mathcal{X}|}}{n},$$

where

$$r := \max_{x \in \mathcal{X}} \|x\|_2.$$

P. Massart. Some applications of concentration inequalities to statistics. 2000.
M. Mohri *et al. Foundations of Machine Learning*. 2012.

Proof of Massart's lemma (1/2)

Proof. (Massart's lemma) For every $\lambda > 0$, we write

$$
\begin{aligned}
\mathsf{E}_\sigma \sup_{x \in \mathcal{X}} \lambda \langle \sigma, x \rangle &= \log \exp \left(\mathsf{E}_\sigma \sup_{x \in \mathcal{X}} \lambda \langle \sigma, x \rangle \right) \\
&\leq \log \mathsf{E} \exp \left(\sup_{x \in \mathcal{X}} \lambda \langle \sigma, x \rangle \right) \\
&\leq \log \sum_{x \in \mathcal{X}} \mathsf{E} \exp \left(\lambda \langle \sigma, x \rangle \right) \\
&\leq \log \sum_{x \in \mathcal{X}} \mathsf{E} \prod_{i=1}^{n} \mathrm{e}^{\lambda \sigma^{(i)} x^{(i)}} \\
&= \log \sum_{x \in \mathcal{X}} \prod_{i=1}^{n} \mathsf{E} \, \mathrm{e}^{\lambda \sigma^{(i)} x^{(i)}}.
\end{aligned}
$$

Proof of Massart's lemma (2/2)

Proof continued. (Massart's lemma) Notice that $\sigma^{(i)}x^{(i)} \in [-|x^{(i)}|, |x^{(i)}|]$. Then, by Hoeffding's lemma, we have

$$\mathsf{E}\, \mathrm{e}^{\lambda \sigma^{(i)} x^{(i)}} \leq \exp\left[\frac{\lambda^2 (2|x_i|)^2}{8}\right] = \exp\left(\frac{\lambda^2 x_i^2}{2}\right), \quad \forall 1 \leq i \leq n,$$

and hence

$$\begin{aligned} \mathsf{E}_\sigma \sup_{x \in \mathcal{X}} \lambda \langle \sigma, x \rangle &\leq \log \sum_{x \in \mathcal{X}} \exp\left(\frac{\lambda^2 \sum_{i=1}^n x_i^2}{2}\right) \\ &\leq \log \sum_{x \in \mathcal{X}} \exp\left(\frac{\lambda^2 r^2}{2}\right) \\ &= \log |\mathcal{X}| + \frac{\lambda^2 r^2}{2}. \end{aligned}$$

Optimizing over λ, the lemma follows.

Applications of the growth function

Proposition 2. Let $\mathcal{F}$ be a class of functions taking values in $[-1, 1]$. Then, it holds that

$$C_n(\mathcal{F}) \leq \sqrt{\frac{2G_n(\mathcal{F})}{n}}.$$

Proof. Exercise.

Corollary 2. Consider the binary classification problem with the 0-1 loss, with hypotheses taking values in $\{\pm 1\}$. Then, with probability at least $(1-\delta)$, it holds that

$$R(h) \leq R_n(h) + \sqrt{\frac{2\log G_n(\mathcal{H})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}, \quad \forall h \in \mathcal{H}.$$

Vapnik-Chervonenkis dimension

Definition. The *Vapnik-Chervonenkis dimension (VC-dimension)* of a hypothesis class $\mathcal{H}$ of $\{\pm 1\}$-valued functions is given by

$$\mathrm{VC}(\mathcal{H}) := \max\{\, n \mid G_n(\mathcal{H}) = 2^n \,\}\,.$$

Example. The VC-dimension of the class of linear classifiers on $\mathbb{R}^p$ equals $(p+1)$.

Example. The VC-dimension of axis-aligned rectangles in $\mathbb{R}^2$ equals 4.

M. Mohri *et al. Foundations of Machine Learning*. 2012.

Vapnik-Chervonenkis-Sauer lemma

Lemma 2. (Vapnik-Chervonenkis-Sauer lemma) Let $\mathcal{H}$ be a hypothesis class of VC-dimension d. Then, it holds that

$$G_n(\mathcal{H}) \leq \sum_{i=0}^{d} \binom{n}{i}.$$

Proof. Check the textbook by Mohri *et al.*

M. Mohri *et al. Foundations of Machine Learning*. 2012.

L. Bottou. On the Vapnik-Chervonenkis-Sauer lemma.

Applications of the VC-dimension (1/2)

Corollary 3. Let $\mathcal{H}$ be a hypothesis class of VC-dimension d. Then, it holds that

$$G_n(\mathcal{H}) \leq \left(\frac{\mathrm{e}n}{d}\right)^d .$$

Corollary 4. Consider the binary classification problem with the 0-1 loss, with hypotheses taking values in $\{\pm 1\}$. Then, with probability at least $(1-\delta)$, it holds that

$$R(h) \leq R_n(h) + \sqrt{\frac{2d\log(\mathrm{e}n/d)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}, \quad \forall h \in \mathcal{H}.$$

Applications of the VC-dimension (2/2)

The following is called *the fundamental theorem of PAC learning* in the textbook by Shalev-Shwartz and Ben-David.

Theorem 2. Let $\mathcal{H}$ be a hypothesis class of $\{\pm 1\}$-valued functions. Then, the hypothesis class $\mathcal{H}$ is agnostic PAC learnable, if and only if its VC-dimension is finite. Moreover, learnability can be achieved by empirical risk minimization.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning*. 2014.

Covering number

Preliminary: Metric space

Definition. A *metric space* (E, d) is a set E with a function $d : E \times E \to \mathbb{R}$, such that the following hold for every $x, y, z \in E$.

- (non-negativity) $d(x, y) \geq 0$; $d(x, y) = 0$ if and only if $x = y$.
- (symmetry) $d(x, y) = d(y, x)$.
- (triangle inequality) $d(x, y) + d(y, z) \leq d(x, z)$.

Example. A *normed space* $(E, \|\cdot\|)$ is a metric space (E, d) with

$$d(x, y) \coloneqq \|x - y\|, \quad \forall x, y \in E.$$

Covering number

Let (E, d) be a metric space. Let $\mathcal{U} \subseteq E$.

Definition. An *ε-cover* (aka an ε-net) of the set $\mathcal{U}$ is another set $\mathcal{V} \subseteq E$, such that

$$\sup_{u \in \mathcal{U}} \inf_{v \in \mathcal{V}} d(u, v) \leq \varepsilon.$$

Definition. The *ε-covering number* of the set $\mathcal{U}$ is given as

$$N(\varepsilon, \mathcal{U}, d) := \inf \left\{ |\mathcal{V}| \mid \mathcal{V} \text{ is an } \varepsilon\text{-cover of } \mathcal{U} \right\}.$$

The quantity $\log N(\varepsilon, \mathcal{U}, d)$ is sometimes called the *metric entropy*.

A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. 1959.

Bounding the ERC in terms of the entropy integral

Theorem 3. (Entropy integral bound) Let $\mathcal{F}$ be a class of functions with the norm

$$\|f\|_{L_2(P_n)} \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^{n}[f(z_i)]^2}, \quad \forall f \in \mathcal{F}.$$

It holds that

$$\hat{C}_n(\mathcal{F}) \leq \inf_{\varepsilon \geq 0}\left\{ 4\varepsilon + 12\int_{\varepsilon}^{+\infty}\sqrt{\frac{\log N(\varepsilon, \mathcal{F}, \|\cdot\|_{L_2(P_n)})}{n}}\,\mathrm{d}\varepsilon \right\}.$$

R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. 1967.

K. Sridharan. Note of refined Dudley integral covering number bound. 2010.

P. L. Bartlett *et al.* Spectrally-normalized margin bounds for neural networks. 2017.

Well-known formulation

Corollary 5. Let $\mathcal{F}$ be a class of functions with the norm

$$\|f\|_{L_2(P_n)} \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[f(z_i)\right]^2}, \quad \forall f \in \mathcal{F}.$$

It holds that

$$\hat{C}_n(\mathcal{F}) \leq 12\int_0^{+\infty}\sqrt{\frac{\log N(\varepsilon, \mathcal{F}, \|\cdot\|_{L_2(P_n)})}{n}}\,\mathrm{d}\varepsilon.$$

Key idea: Chaining (1/2)

Let (E, d) be a metric space and $\mathcal{T} \subseteq E$. Let $\{\, \xi_t \mid t \in \mathcal{T} \,\}$ be a stochastic process. Let $\mathcal{N}$ be an ε-cover of $\mathcal{T}$, and

$$\pi(t) := \arg\min_{s \in \mathcal{N}} d(s, t).$$

Then, we have

$$\begin{aligned} \mathsf{E} \sup_{t \in \mathcal{T}} \xi_t &= \mathsf{E} \sup_{t \in \mathcal{T}} \left(\xi_{\pi(t)} + \xi_t - \xi_{\pi(t)} \right) \\ &\leq \mathsf{E} \sup_{t \in \mathcal{T}} \xi_{\pi(t)} + \mathsf{E} \sup_{t \in \mathcal{T}} \left(\xi_t - \xi_{\pi(t)} \right). \end{aligned}$$

This is sometimes called an *ε-net argument*.

R. van Handel. *Probability in High Dimension*. 2016.

Key idea: Chaining (2/2)

Similarly, let $\mathcal{N}_k$, $k \in \mathbb{N} \cup \{ 0 \}$, be an ε_k-cover of $\mathcal{T}$ such that $(\varepsilon_k)_{k \in \mathbb{N}}$ is a decreasing sequence. Define

$$\pi_k(t) := \arg\min_{s \in \mathcal{N}_k} d(s,t), \quad \forall t \in \mathcal{T}.$$

The *chaining argument* considers the decomposition

$$\begin{aligned} \mathsf{E} \sup_{t \in \mathcal{T}} \xi_t &= \mathsf{E} \sup_{t \in \mathcal{T}} \left[\xi_{\pi_0(t)} + \sum_{k=1}^{K} \left(\xi_{\pi_k(t)} - \xi_{\pi_{k-1}(t)} \right) + \left(\xi_t - \xi_{\pi_n(t)} \right) \right] \\ &\leq \mathsf{E} \sup_{t \in \mathcal{T}} \xi_{\pi_0(t)} + \\ &\quad \sum_{k=1}^{K} \mathsf{E} \sup_{t \in \mathcal{T}} \left(\xi_{\pi_k(t)} - \xi_{\pi_{k-1}(t)} \right) + \\ &\quad \mathsf{E} \sup_{t \in \mathcal{T}} \left(\xi_t - \xi_{\pi_n(t)} \right). \end{aligned}$$

Proof of the entropy integral bound (1/5)

Proof. (Theorem 3) Define

$$\varepsilon_0 \coloneqq \sup_{f \in \mathcal{F}} \|f\|_{L_2(P_n)},$$

and

$$\varepsilon_k \coloneqq 2^{-k} \varepsilon_0, \quad \forall k \in \mathbb{N}.$$

Let $\mathcal{N}_k$ be an ε_k-cover of $\mathcal{F}$ that achieves the ε_k-covering number. Choose $\mathcal{N}_0 = \{ 0 \}$ for convenience. Define

$$\hat{f}_k \coloneqq \arg\min_{\varphi \in \mathcal{N}_k} \|\varphi - f\|_{L_2(P_n)}.$$

Proof of the entropy integral bound (2/5)

Proof continued. (Theorem 3) Then, we have

$$
\begin{aligned}
\hat{C}_n(\mathcal{F}) &= \mathsf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \\
&= \mathsf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left\{ f(z_i) - \hat{f}_N(z_i) + \sum_{k=1}^{K} \left[\hat{f}_k(z_i) - \hat{f}_{k-1}(z_i) \right] \right\} \\
&\leq \mathsf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left[f(z_i) - \hat{f}_N(z_i) \right] \\
&\quad \sum_{k=1}^{K} \mathsf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left[\hat{f}_k(z_i) - \hat{f}_{k-1}(z_i) \right]
\end{aligned}
$$

Proof of the entropy integral bound (3/5)

Proof continued. (Theorem 3) The first term can be bounded by the Cauchy-Schwartz inequality as

$$\begin{aligned} \mathsf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left[f(z_i) - \hat{f}_N(z_i) \right] &\leq 1 \times \sup_{f \in \mathcal{F}} \|f - \hat{f}_N\|_{L_2(P_n)} \\ &\leq \varepsilon_N. \end{aligned}$$

For each $k \in \mathbb{N}$, by Massart's lemma, we have

$$\mathsf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left[\hat{f}_k(z_i) - \hat{f}_{k-1}(z_i) \right] \leq r_k \sqrt{\frac{2 \log \left(|\mathcal{N}_k| |\mathcal{N}_{k-1}| \right)}{n}},$$

where

$$r_k := \sup_{f \in \mathcal{F}} \|\hat{f}_k - \hat{f}_{k-1}\|_{L_2(P_n)}.$$

Proof of the entropy integral bound (4/5)

Proof continued. (Theorem 3) The rest is tedious. By the triangle inequality, we have

$$r_k \le \sup_{f \in \mathcal{F}} \|\hat{f}_k - f + f - \hat{f}_{k-1}\|_{L_2(P_n)} \le \varepsilon_k + \varepsilon_{k-1} \le 3\varepsilon_k.$$

Then, we write

$$\begin{aligned}
&\mathsf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left[\hat{f}_k(z_i) - \hat{f}_{k-1}(z_i)\right] \\
&\quad \le 3\varepsilon_k \sqrt{\frac{2\log\left(|\mathcal{N}_k||\mathcal{N}_k|\right)}{n}} \\
&\quad \le 3 \times 2(\varepsilon_k - \varepsilon_{k+1}) \sqrt{\frac{4 \log |\mathcal{N}_k|}{n}} \\
&\quad = 12(\varepsilon_k - \varepsilon_{k+1}) \sqrt{\frac{\log N(\varepsilon_k, \mathcal{F}, \|\cdot\|_{L_2(P_n)})}{n}}.
\end{aligned}$$

Proof of the entropy integral bound (5/5)

Proof continued. (Theorem 3) Therefore, we obtain

$$\sum_{k=1}^{K} \mathsf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left[\hat{f}_k(z_i) - \hat{f}_{k-1}(z_i) \right]$$
$$\leq 12 \int_{\varepsilon_{K+1}}^{\varepsilon_0} \sqrt{\frac{\log N(\varepsilon, \mathcal{F}, \|\cdot\|_{L_2(P_n)})}{n}} \, \mathrm{d}\varepsilon.$$

For every $\alpha > 0$, choose K such that $\alpha \leq \varepsilon_{K+1} \leq 2\alpha$. Then, we have $\varepsilon_N \leq 4\alpha$, and

$$\hat{C}_n(\mathcal{F}) \leq 4\alpha + 12 \int_{\alpha}^{\varepsilon_0} \sqrt{\frac{\log N(\varepsilon, \mathcal{F}, \|\cdot\|_{L_2(P_n)})}{n}} \, \mathrm{d}\varepsilon.$$

It remains to optimize over α.

Conclusions

Comparison of complexity measures

- The empirical Rademacher complexity is data dependent.

- The Rademacher complexity is distribution dependent.

- The VC-dimension and covering number are worst-case bounds.

Summary

- The generalization error can be bounded via the Rademacher and empirical Rademacher complexities.
- The Rademacher complexity can be approximated by the empirical Rademacher complexity (ERC).
- The Rademacher complexity can be bounded from above via the VC-dimension (for $\{\pm 1\}$-valued hypotheses).
- The ERC (and hence the Rademacher complexity) can be bounded from above via the covering number.

Next lecture

- Model selection
- PAC Bayes.
- Multiplicative weight update.