CSIE5002 Prediction, learning, and games

Lecture 2: Introduction to statistical learning

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

25.02.2019

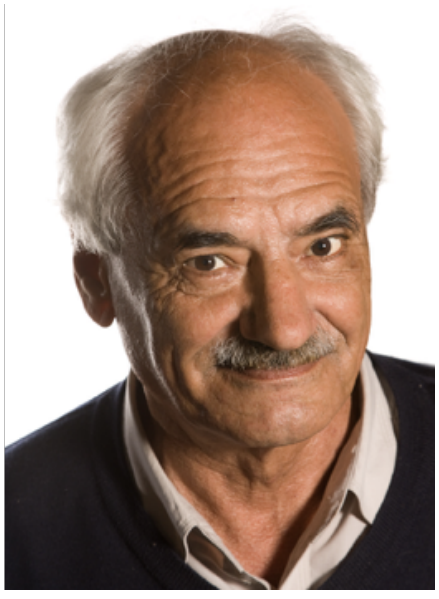
Department of Computer Science and Information Engineering
National Taiwan University

Abstract

Statistical learning and online learning are closely related, though the relation can be subtle. This lecture provides an introduction to standard topics in statistical learning theory.

Table of contents

Pioneers of statistical learning theory



V. Vapnik and A. Chervenenkis

- Vapnik-Chervenenkis theory.
- Support vector machine.

L. Valiant

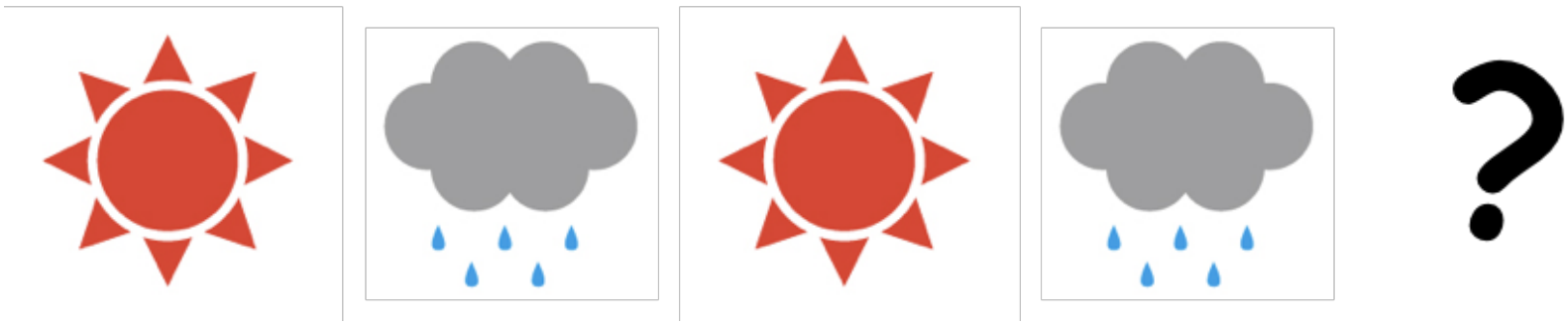
- PAC learnability

V. Vapnik. *The Nature of Statistical Learning Theory*. 2000.

L. Valiant. *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. 2013.

Standard model of machine learning

Machine learning problems (1/3)

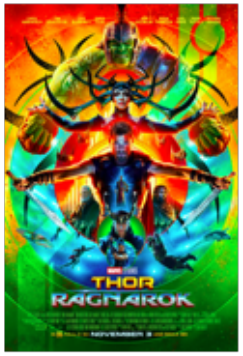


Time series prediction.

Machine learning problems (2/3)

Classification and regression.

Machine learning problems (3/3)

				
Alice	5	?	3	5
Bob	4	2	5	3
Charles	3	4	4	2

Matrix completion.

What is machine learning?

Goal. To predict the unknown given the data.

Requirements of a good model.

- The unknown can be different from the data.
- The unknown should be related to the data.
- There should be a measure of the prediction accuracy.
- The assumptions should be as weak as possible.

Standard model of machine learning

- Data: I.i.d. random variables $z_1, \ldots, z_n \in \mathcal{Z}$ following an unknown probability distribution P.
- Hypothesis class: A set $\mathcal{H}$.
- Loss: A function $\rho : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$.
- Risk: The expected loss $R : h \mapsto \mathsf{E}_z\, \rho(h, z)$, where z is a random variable following P.
- Goal: Given the data, find the optimal hypothesis

$$h^\star \in \arg\min_h \{\, R(h) \mid h \in \mathcal{H} \,\}\,.$$

S. Shalev-Shwartz *et al.* Learnability, stability, and uniform convergence. 2010.

Applications of the standard model

Binary classification. $z = (x, y) \in \mathcal{X} \times \{0, 1\}$,
$\mathcal{H} \coloneqq \{h : \mathcal{X} \to \{0, 1\}\}$ (or a subset), $\rho(h, z) \coloneqq \mathbb{1}_{\{h(x) \neq y\}}$.

Least squares regression. $z = (x, y) \in \mathcal{X} \times \mathbb{R}$,
$\mathcal{H} \coloneqq \{h : \mathcal{X} \to \mathbb{R}\}$ (or a subset), $\rho(h, z) \coloneqq (y - h(x))^2$.

K-means clustering/vector quantization. $z = \mathbb{R}^p$,
$\mathcal{H} \coloneqq \{\{c_1, \ldots, c_K\} \mid c_i \in \mathbb{R}^p\}$, $\rho(h, z) \coloneqq \min_{c \in h} \|z - c\|_2^2$.

Density estimation. $\mathcal{H}$ a subset of probability density functions on $\mathcal{Z}$, $\rho(h, z) \coloneqq -\log h(z)$.

S. Shalev-Shwartz *et al.* Learnability, stability, and uniform convergence. 2010.

Probably approximately correct (PAC) learnability

Definition. (PAC learnability) A hypothesis class is called *agnostic PAC learnable*, if and only if there exists a function

$$n_{\mathcal{H}} :]0,1[^2 \to \mathbb{N}$$

and an algorithm, such that for every probability distribution P and every $(\varepsilon, \delta) \in]0,1[^2$, the algorithm returns a hypothesis $\hat{h}$ satisfying

$$R(\hat{h}) \leq \min_h \{\, R(h) \mid h \in \mathcal{H} \,\} + \varepsilon$$

with probability at least $1 - \delta$, given data of size $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$.

L. G. Valiant. A theory of the learnable. 1984.

D. Hausler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. 1992.

Equivalent formulation

PAC learnability briefly restated. There exists an algorithm such that

$$\forall P : n \geq n_{\mathcal{H}}(\varepsilon, \delta) \Rightarrow \mathsf{P}\left(R(\hat{h}) \leq \min_{h} \{\, R(h) \mid h \in \mathcal{H} \,\} + \varepsilon \right) \geq 1 - \delta.$$

Equivalent formulation. There exists an algorithm such that

$$\forall P : \mathsf{P}\left(R(\hat{h}) \leq \min_{h} \{\, R(h) \mid h \in \mathcal{H} \,\} + \varepsilon_n \right) \geq 1 - \delta,$$

for some $\varepsilon_n \to 0$ as $n \to \infty$.

Characteristics of PAC learnability

PAC learnability briefly restated. There exists an algorithm such that

$$\forall P : n \geq n_{\mathcal{H}(\varepsilon,\delta)} \Rightarrow \mathsf{P}\left(R(\hat{h}) \leq \min_{h} \{ R(h) \mid h \in \mathcal{H} \} + \varepsilon \right) \geq 1-\delta.$$

- *Agnostic*: Comparison only with the optimal hypothesis in $\mathcal{H}$, which may yield a high unsatisfactory risk.
- *Probably*: Risk unbounded with non-zero probability.
- *Approximately correct*: Smallest possible risk approximated, but not necessarily achieved.

Estimation & approximation errors

The gap to the *optimal risk* can be decomposed as

$$R(\hat{h}) - \min_h R(h) = \underbrace{\left[R(\hat{h}) - \min_{h\in\mathcal{H}} R(h)\right]}_{\textit{Estimation error}} + \underbrace{\left[\min_{h\in\mathcal{H}} R(h) - \min_h R(h)\right]}_{\textit{Approximation error}}.$$

Remark. If the hypothesis class $\mathcal{H}$ is agnostic PAC learnable, then with high probability, the estimation error vanishes as the data size goes to infinity.

Remark. The approximation error is determined by $\mathcal{H}$ and P and not affected by the data size.

Another perspective: Generalization error

We expect that a good learning algorithm, after seeing the data, can *generalize* and predict accurately the unknown.

Definition. The *empirical risk* yielded by a hypothesis h is the empirical average of the loss on the data:

$$R_n(h) := \frac{1}{n}\sum_{i=1}^{n} \rho(h, z_i), \quad \forall h \in \mathcal{H}.$$

Definition. The *generalization error* of a hypothesis h is given by

$$G(h) := R(h) - R_n(h).$$

Remark. A small generalization error implies that if the empirical risk is small, then the risk is also small; then, *overfitting* will not happen.

Empirical risk minimization

Learning problem

Machine learning problem briefly restated. Let z be a random variable whose probability distribution is unknown. Solve the optimization problem

$$h^\star \in \arg\min_h \{\, R(h) := \mathsf{E}\, \rho(h, z) \mid h \in \mathcal{H} \,\},$$

given independent random variables $z_1, \ldots, z_n$ following the distribution of z.

Issue. The risk function $R(h)$ cannot be exactly evaluated.

Empirical risk minimization

Empirical risk minimization (ERM). Compute

$$\hat{h}_n := \arg\min_h \left\{ R_n(h) := \frac{1}{n} \sum_{i=1}^{n} \rho(h, z_i) \;\middle|\; h \in \mathcal{H} \right\}$$

as an approximate of $h^\star$.

Remark. For any given $h \in \mathcal{H}$, the weak law of large numbers says that $R_n(h)$ converges to $R(h)$ in probability as n goes to infinity, i.e.,

$$\lim_{n \to \infty} \mathsf{P}\left(|R_n(h) - R(h)| > \varepsilon \right) = 0, \quad \forall \varepsilon > 0.$$

V. Vapnik. Principles of risk minimization for learning theory. 1991.

Uniform convergence

To ensure that the ERM problem asymptotically approximates the original risk minimization problem, a sufficient condition is the *uniform law of large numbers*:

$$\lim_{n\to\infty} \mathsf{P}\left(\sup_h \{\, |R_n(h) - R(h)| \mid h \in \mathcal{H} \,\} > \varepsilon\right) = 0, \quad \forall \varepsilon > 0.$$

Remark. To check learnability by the ERM, we need to characterize the rate of uniform convergence.

V. N. Vapnik and A. Ya. Chervenenkis. On the uniform convergence of the frequencies of occurrence of events to their probabilities. 1968.

Finite hypothesis class case

Proposition 1. Suppose that the hypothesis class $\mathcal{H}$ contains m elements. Suppose that the loss function ρ takes values in $[0, 1]$. Then, for every δ in $]0, 1[$, the empirical risk minimizer $\hat{h}_n$ satisfies

$$R(\hat{h}_n) \leq \min_{h \in \mathcal{H}} R(h) + \sqrt{\frac{2 \log(2m/\delta)}{n}},$$

with probability at least $1 - \delta$.

Remark. Keep in mind that the estimation error is of $O(\sqrt{n^{-1} \log m})$.

Preliminaries: Union bound

Lemma. (Union bound) Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two events. Then, it holds that

$$\mathsf{P}\left(\mathcal{E}_1 \cup \mathcal{E}_2\right) \leq \mathsf{P}\left(\mathcal{E}_1\right) + \mathsf{P}\left(\mathcal{E}_2\right).$$

Proof. Notice that

$$\mathcal{E}_1 \cup \mathcal{E}_2 = \mathcal{E}_1 \cup \left(\mathcal{E}_2 \setminus \mathcal{E}_1\right)$$

and that

$$\mathcal{E}_1 \cap \left(\mathcal{E}_2 \setminus \mathcal{E}_1\right) = \emptyset.$$

Therefore, we have

$$\mathsf{P}\left(\mathcal{E}_1 \cup \mathcal{E}_2\right) = \mathsf{P}\left(\mathcal{E}_1\right) + \mathsf{P}\left(\mathcal{E}_2 \setminus \mathcal{E}_1\right).$$

The lemma follows as $\left(\mathcal{E}_2 \setminus \mathcal{E}_1\right) \subseteq \mathcal{E}_2$.

Preliminaries: Hoeffding's inequality (1/2)

Lemma. (Hoeffding's lemma) Let ξ be a random variable taking values in $[a, b]$. Then, it holds that

$$\log \mathsf{E}\, \mathrm{e}^{\lambda(\xi - \mathsf{E}\,\xi)} \leq \frac{\lambda^2 (b-a)^2}{8}, \quad \forall \lambda \in \mathbb{R}.$$

Theorem. (Hoeffding's inequality) Let $\xi_1, \ldots, \xi_n$ be independent random variables taking values in $[a, b]$. Then, it holds that

$$\mathsf{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mathsf{E}\,\xi_i) \right| \geq t \right) \leq 2\mathrm{e}^{-\frac{2nt^2}{(b-a)^2}}, \quad \forall t > 0.$$

W. Hoeffding. Probability inequalities for sums of bounded random variables. 1963.

Preliminaries: Hoeffding's inequality (2/2)

Proof. (Hoeffding's inequality) We write

$$
\begin{aligned}
&\mathsf{P}\left(\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mathsf{E}\,\xi_i) \geq t\right) \\
&= \mathsf{P}\left(\mathrm{e}^{\lambda\sum_{i=1}^{n}(\xi_i - \mathsf{E}\,\xi_i)} \geq \mathrm{e}^{\lambda n t}\right) \\
&\leq \frac{\mathsf{E}\,\mathrm{e}^{\lambda\sum_{i=1}^{n}(\xi_i - \mathsf{E}\,\xi_i)}}{\mathrm{e}^{\lambda n t}} && \text{(HW \#0)} \\
&= \frac{\prod_{i=1}^{n}\mathsf{E}\,\mathrm{e}^{\lambda(\xi_i - \mathsf{E}\,\xi_i)}}{\mathrm{e}^{\lambda n t}} && \text{(statistical independence)} \\
&\leq \mathrm{e}^{-\lambda n t}\prod_{i=1}^{n}\mathrm{e}^{\frac{\lambda^2(b-a)^2}{8}} && \text{(Hoeffding's lemma)} \\
&= \mathrm{e}^{-\lambda n t + \frac{n\lambda^2(b-a)^2}{8}}
\end{aligned}
$$

The theorem follows by optimizing over λ and the union bound.

Proof of Proposition 1 (1/2)

Proof. (Proposition 1) The following decomposition is key:

$$R(\hat{h}_n) - R(h^\star) = \Big[R(\hat{h}_n) - R_n(\hat{h}_n)\Big] + \Big[R_n(\hat{h}_n) - R_n(h^\star)\Big] + \left[R_n(h^\star) - R(h^\star)\right].$$

Notice that $R_n(\hat{h}_n) - R_n(h^\star) \leq 0$. Then, we have

$$\begin{aligned} R(\hat{h}_n) - R(h^\star) &\leq \sup_{h\in\mathcal{H}} \left(R(h) - R_n(h)\right) + \sup_{h\in\mathcal{H}} \left(R_n(h) - R(h)\right) \\ &\leq 2 \sup_{h\in\mathcal{H}} |R(h) - R_n(h)|. \end{aligned}$$

Proof of Proposition 1 (2/2)

Proof continued. (Proposition 1) By the union bound and Hoeffding's inequality, we write

$$\begin{aligned}
\mathsf{P}\left(R(\hat{h}_n) - R(h^\star) \geq t\right) &\leq \mathsf{P}\left(\sup_{h\in\mathcal{H}} |R(h) - R_n(h)| \geq \frac{t}{2}\right) \\
&\leq \sum_{j=1}^{m} \mathsf{P}\left(|R(h_j) - R_n(h_j)| \geq \frac{t}{2}\right) \\
&\leq m\left(2\mathrm{e}^{-\frac{nt^2}{2}}\right), \quad \forall t > 0.
\end{aligned}$$

Then, we obtain

$$\mathsf{P}\left(R(\hat{h}_n) \leq R(h^\star) + \sqrt{\frac{2\log(2m/\delta)}{n}}\right) \geq 1 - \delta.$$

Generalization error

Similarly, we can bound the generalization error. Notice that we only need a one-sided upper bound of $R(\hat{h}_n) - R_n(\hat{h}_n)$.

Proposition 2. For every $\delta \in]0, 1[$, the empirical risk minimizer $\hat{h}_n$ satisfies

$$R(\hat{h}_n) \leq R_n(\hat{h}_n) + \sqrt{\frac{\log(m/\delta)}{2n}}.$$

with probability at least $(1 - \delta)$.

Proof. Notice that

$$R(\hat{h}_n) - R_n(\hat{h}_n) \leq \sup_{h \in \mathcal{H}} \left(R(h) - R_n(h) \right).$$

Rademacher complexity

Observation

To bound the generalization error, it suffices to bound

$$\begin{aligned}\sup_{h\in\mathcal{H}} R(h) - R_n(h) &= \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \left[\mathsf{E}\,\rho(h,z) - \rho(h,z_i)\right] \\ &= \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \left[\mathsf{E}\,f(z) - f(z_i)\right],\end{aligned}$$

where

$$\mathcal{F} := \{\, f : z \mapsto \rho(h,z) \mid h \in \mathcal{H} \,\}.$$

Empirical and Rademacher process

Definition. An *empirical process* is the stochastic process $\{ G_f \mid f \in \mathcal{F} \}$ indexed by a function class $\mathcal{F}$ defined as

$$G_f := \frac{1}{n} \sum_{i=1}^{n} (\mathsf{E}\, f(z) - f(z_i)) \,,$$

where $z, z_1, \ldots, z_n$ are i.i.d. random variables.

Discussion. Suppose that f are bounded. It is possible that $G_f = \Theta(1)$. However, we expect that $G_f = o(1)$.

Symmetrization

Definition. A *Rademacher random variable* σ takes values in $\{ \pm 1 \}$ with equal probability.

Definition. The associated *Rademacher process* is the stochastic process $\{ S_f \mid f \in \mathcal{F} \}$ given by

$$S_f := \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i),$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. *Rademacher random variables* independent of $z_1, \ldots, z_n$.

Lemma. (Symmetrization) It holds that

$$\mathrm{E} \sup_{f \in \mathcal{F}} G_f \leq 2 \mathrm{E} \sup_{f \in \mathcal{F}} S_f.$$

E. Giné and J. Zinn. 1984. Some limit theorems for empirical processes.

Proof of the symmetrization lemma (1/2)

Proof. (Symmetrization) Let $y_1, \ldots, y_n$ be i.i.d. copies of z, independent of $z_1, \ldots, z_n$. We write

$$\begin{aligned} \mathsf{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \left[\mathsf{E}\, f(z) - f(z_i) \right] &= \mathsf{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \left[\mathsf{E}\, f(y_i) - f(z_i) \right] \\ &\leq \mathsf{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \left[f(y_i) - f(z_i) \right]. \end{aligned}$$

Notice that the probability distributions of $f(y_i) - f(z_i)$ and $\sigma_i \left[f(y_i) - f(z_i) \right]$ are the same. Therefore, we obtain

$$\mathsf{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \left[\mathsf{E}\, f(z) - f(z_i) \right] = \mathsf{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i \left[f(y_i) - f(z_i) \right].$$

Proof of the symmetrization lemma (2/2)

Proof continued. (Symmetrization) Notice that the probability distributions of $\sigma_i f(y_i)$ and $-\sigma_i f(z_i)$ are the same. Then, we write

$$\begin{aligned}\mathsf{E} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} [\mathsf{E}\, f(z) - f(z_i)] &\leq \mathsf{E} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i f(y_i) + \mathsf{E} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} -\sigma_i f(z_i) \\ &\leq 2\mathsf{E} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i f(z_i).\end{aligned}$$

This proves the lemma.

Rademacher complexity

Definition. (Rademacher complexity) The associated *Rademacher complexity* is given by

$$C_n(\mathcal{F}) := \mathsf{E} \sup_{f \in \mathcal{F}} S_f = \mathsf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i).$$

Theorem 1. Suppose that $\mathcal{F}$ consists of functions taking values in $[0, 1]$. Then, with probability at least $1 - \delta$, it holds that

$$\mathsf{E}\, f(z) \leq \frac{1}{n} \sum_{i=1}^{n} f(z_i) + 2C_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

V. Koltchinskii. Rademacher penalties and structural risk minimization. 2001.

Proof o Theorem 1: Prelude

A random variable that depends in a "smooth" way on the influence of many independent variables (but not too much on any of them) is essentially constant. — Michel Talagrand

---

M. Talagrand. A new look at independence. 1996.

Proof of Theorem 1 (1/3)

Theorem. (McDiarmid's inequality) Let $\xi_1, \ldots, \xi_n$ be independent random variables taking values in $\mathcal{X}$. Let $\varphi : \mathcal{X}^n \to \mathbb{R}$. Suppose that

$$|\varphi(\xi_1, \ldots, \xi_i, \ldots, \xi_n) - \varphi(\xi_1, \ldots, \xi_i', \ldots, \xi_n)| \leq c_i, \quad \forall \xi_i' \in \mathcal{X},$$

$c_1, \ldots, c_n > 0$. Then, for every $\varepsilon > 0$, it holds that

$$\varphi(\xi_1, \ldots, \xi_n) \leq \mathsf{E}\, \varphi(\xi_1, \ldots, \xi_n) + \varepsilon$$

with probability at least $1 - \mathrm{e}^{-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}}$.

C. McDiarmid. On the method of bounded differences. 1989.

Proof of Theorem 1 (2/3)

Proof. (Theorem 1) Define

$$\varphi(z_1, \ldots, z_n) \coloneqq \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left[\mathsf{E}\, f(z) - f(z_i) \right].$$

Then, we write

$$\begin{aligned}
&|\varphi(z_1, \ldots, z_j, \ldots, z_n) - \varphi(z_1, \ldots, z_j', \ldots, z_n)| \\
&\quad \leq \sup_{f \in \mathcal{F}} \frac{1}{n} \left[\sum_{i \neq j} \left[f(z_i) - f(z_i) \right] + f(z_j) - f(z_j') \right] \\
&\quad \leq \frac{1}{n}, \quad \forall j.
\end{aligned}$$

Proof of Theorem 1 (3/3)

Proof continued. (Theorem 1) Applying McDiarmid's inequality, we obtain

$$\varphi(z_1, \ldots, z_n) \leq \mathsf{E}\,\varphi(z_1, \ldots, z_n) + \sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1 - \delta$. It remains to apply the symmetrization lemma, which says that

$$\mathsf{E}\,\varphi(z_1, \ldots, z_n) \leq 2C_n(\mathcal{F}).$$

Application: Binary classification

Corollary 1. Consider the binary classification problem, where $z = (x, y) \in \mathcal{X} \times \{\pm 1\}$, $\mathcal{H}$ is a set of $\{\pm 1\}$-valued functions, and the loss is given by

$$\rho(h, z) := \mathbb{1}_{\{h(x) \neq y\}}.$$

Then, it holds that

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

with probability at least $1 - \delta$, where

$$C_n(\mathcal{H}) := \mathsf{E} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i).$$

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

Proof of Corollary 1 (1/2)

Lemma. Consider the function class

$$\mathcal{F} := \{\, f : z \mapsto \rho(h, z) \mid h \in \mathcal{H} \,\}.$$

It holds that

$$C_n(\mathcal{F}) = \frac{1}{2} C_n(\mathcal{H}).$$

Proof. Notice that

$$f(z) = \mathbb{1}_{\{\, h(x) \neq y \,\}} = \frac{1 - yh(x)}{2}.$$

Then, we write

$$C_n(\mathcal{F}) = \mathsf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left[\sigma_i - \frac{\sigma_i y_i h(x_i)}{2} \right] = \frac{1}{2} C_n(\mathcal{H}).$$

Proof of Corollary 1 (2/2)

Proof. (Corollary 1) Consider the function class

$$\mathcal{F} \coloneqq \{\, f : z \mapsto \rho(h, z) \mid h \in \mathcal{H} \,\}.$$

Then, we have

$$\sup_{h \in \mathcal{H}} \left[R(h) - R_n(h) \right] = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left[\mathsf{E}\, f(z) - f(z_i) \right].$$

It remains to apply Theorem 1, and the lemma in the previous page.

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

Conclusions

Summary

- We have introduced the standard model of statistical learning.
- We have introduced the ERM approach.
- The success of the ERM approach can be guaranteed if uniform convergence can hold.
- Uniform convergence can be established by the union bound with Hoeffding's inequality, or symmetrization with the McDiarmid's inequality.

Next lecture

- More complexity measures.