CSIE5002 Prediction, learning, and games

Lecture 4: Introduction to statistical learning III

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

11.03.2019

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

This lecture is a continuation of Lecture 1 and Lecture 2. In particular, this lecture introduces the notions of PAC-Bayes analyses and model selection.

Related advanced topics

- PAC-Bayes with data-dependent priors
  - O. Catoni. 2007. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*.
  - G. Lever *et al.* 2013. Tighter PAC-Bayes bounds through distribution-dependent priors.
  - G. K. Dziugaite and D. M. Roy. 2018. Data-dependent PAC-Bayes priors via differential privacy.

- Deep learning
  - B Neyshabur *et al.* 2017. Exploring generalization in deep learning.
  - G. K. Dziugaite and D. M. Roy. 2018. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of entropy-SGD and data-dependent priors.

Table of contents

PAC-Bayes analyses

Starting point

Proposition 1. (Proposition 2 in Lecture 2) Let $\mathcal{H}$ be a finite hypothesis class. Suppose that the loss functions take values in $[0,1]$. Then for every $\delta \in]0,1[$, it holds with probability at least $(1-\delta)$ that

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

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

$$\begin{aligned} \mathsf{P}\left(\exists h \in \mathcal{H} : R(h) - R_n(h) > t\right) &\leq \sum_{h \in \mathcal{H}} \mathsf{P}\left(R(h) - R_n(h) > t\right) \\ &\leq |\mathcal{H}| \mathrm{e}^{-2nt^2}. \end{aligned}$$

Countable hypothesis class

The previous proof cannot be directly applied when the hypothesis class is countable.

Proposition 2. Let $\mathcal{H}$ be a *countable* hypothesis class. Suppose that the loss functions take values in $[0, 1]$. Then for any *probability distribution π on $\mathcal{H}$* and $\delta \in]0, 1[$, it holds with probability at least $(1 - \delta)$ that

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

Remark. The probability distribution π is called the *prior* in PAC-Bayes literature.

Proof of Proposition 2

Proof. (Proposition 2) Let $t : \mathcal{H} \to]0, +\infty[$. By the union bound and Hoeffding's inequality, we write

$$\begin{aligned}
&\mathsf{P}\left(\exists h \in \mathcal{H} : R(h) - R_n(h) > t(h)\right) \\
&\quad\leq \sum_{h \in \mathcal{H}} \mathsf{P}\left(R(h) - R_n(h) > t(h)\right) \\
&\quad\leq \sum_{h \in \mathcal{H}} \mathrm{e}^{-2nt(h)^2}.
\end{aligned}$$

Fix some $t > 0$. Set

$$t(h) \coloneqq \sqrt{t^2 + \frac{\log(1/\pi(h))}{2n}}.$$

Then, we have

$$\mathsf{P}\left(\exists h \in \mathcal{H} : R(h) - R_n(h) > t(h)\right) \leq \sum_{h \in \mathcal{H}} \pi(h)\mathrm{e}^{-2nt^2} = \mathrm{e}^{-2nt^2}.$$

Looseness of the union bound

The union bound can be loose, because...

Gibbs Predictor

Definition. (Gibbs predictor) A *Gibbs predictor* is a random hypothesis following some probability distribution $\hat{\pi}$ on $\mathcal{H}$.

Remark. The probability distribution $\hat{\pi}$ is called the *posterior distribution* in PAC-Bayes literature.

Remark. Standard proof techniques in PAC-Bayes analyses consider *data-independent priors* and *possibly data-dependent posterior distributions*.

PAC-Bayes generalization bound

Definition. (Relative entropy) Let π and $\hat{\pi}$ be two probability density functions on $\mathcal{H}$. The relative entropy is given by

$$D(\hat{\pi}||\pi) := \mathsf{E}_{\hat{\pi}} \frac{\hat{\pi}(h)}{\pi(h)}.$$

Theorem 1. Let π be a data-independent probability distribution on a hypothesis class $\mathcal{H}$; let $\hat{\pi}$ be a possibly data-dependent probability distribution on $\mathcal{H}$. Let $\hat{h}$ be the Gibbs predictor following $\hat{\pi}$. Suppose that the loss functions take values in $[0, 1]$. Then for every $\eta > 0$ and $\delta \in]0, 1[$, it holds with probability at least $(1 - \delta)$ that

$$\mathsf{E}_{\hat{\pi}}\, R(\hat{h}) \leq \mathsf{E}_{\hat{\pi}}\, R_n(\hat{h}) + \frac{\eta}{8} + \frac{1}{\eta n}\left[D(\hat{\pi}\|\pi) + \log\frac{1}{\delta}\right].$$

T. van Erven. 2014. PAC-Bayes mini-tutorial: A continuous union bound.

Choosing the posterior

$$\mathsf{P}\left(\mathsf{E}_{\hat{\pi}}\, R(\hat{h}) \leq \mathsf{E}_{\hat{\pi}}\, R_n(\hat{h}) + \frac{\eta}{8} + \frac{1}{\eta n}\left[D(\hat{\pi}\|\pi) + \log\frac{1}{\delta}\right]\right) \geq 1 - \delta.$$

Question. The prior π is arbitrary. How do we choose the posterior $\hat{\pi}$?

Theorem 2. (Gibbs variational principle) The generalization error bound is minimized by the *Gibbs distribution*:

$$\hat{\pi}(h) \propto \exp\left[-\eta n R_n(h) \pi(h)\right], \quad \forall h \in \mathcal{H}.$$

Question. What happens when $\eta \to +\infty$?

Preliminary: Mix risk

Definition. (Mix risk) Denote the loss function by $\rho : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$. Let z be a random variable taking values on $\mathcal{Z}$. The associated *mix risk* with parameter $\eta > 0$ is given by

$$M_\eta(h) := -\frac{1}{\eta} \log \mathsf{E}\, \mathrm{e}^{-\eta \rho(z,h)}, \quad \forall h \in \mathcal{H}.$$

Remark. Suppose that ρ takes values in $[a, b]$ for some $a, b \in \mathbb{R}$, $a < b$. Then, by Hoeffding's inequality, we have

$$R(h) \leq M_\eta(h) + \eta \frac{(b-a)^2}{8}, \quad \forall \eta > 0.$$

Therefore, we may bound the generalization error by *relating the mix risk with the empirical risk.*

Proof of Theorem 1 (1/2)

The proof follows that in (van Erven 2014).

Proof. (Theorem 1) It suffices to prove that for any $\eta > 0$ and $\delta \in]0, 1[$, we have

$$\mathsf{E}_{\hat{\pi}} M_\eta(\hat{h}) \leq \mathsf{E}_{\hat{\pi}} R_n(\hat{h}) + \frac{1}{\eta n} \left[D(\hat{\pi} \| \pi) + \log \frac{1}{\delta} \right],$$

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

Notice that

$$\mathrm{e}^{-\eta n M_\eta(h)} = \mathsf{E}\, \mathrm{e}^{-\eta n R_n(h)}, \quad \forall h \in \mathcal{H}.$$

T. van Erven. 2014. PAC-Bayes mini-tutorial: A continuous union bound.

Proof of Theorem 1 (2/2)

Proof continued. (Theorem 1) Denote by $P^{\otimes n}$ the joint probability distribution of the data. By Jensen's inequality, we write

$$\begin{aligned} 1 &= \mathsf{E}_\pi \, \mathsf{E}_{P^{\otimes n}} \, \exp\left[\eta n M_\eta(\hat{h}) - \eta n R_n(\hat{h})\right] \\ &= \mathsf{E}_{P^{\otimes n}} \, \mathsf{E}_\pi \, \exp\left[\eta n M_\eta(\hat{h}) - \eta n R_n(\hat{h})\right] \\ &= \mathsf{E}_{P^{\otimes n}} \, \mathsf{E}_{\hat{\pi}} \, \exp\left[\eta n M_\eta(\hat{h}) - \eta n R_n(\hat{h}) - \log \frac{\hat{\pi}(\hat{h})}{\pi(\hat{h})}\right] \\ &\geq \mathsf{E}_{P^{\otimes n}} \, \exp\left[\eta n \mathsf{E}_{\hat{\pi}} M_n(\hat{h}) - \eta n \mathsf{E}_{\hat{\pi}} \, R_n(\hat{h}) - D(\hat{\pi} \| \pi)\right]. \end{aligned}$$

By Markov's inequality, we obtain

$$\mathsf{P}\left(\eta n \mathsf{E}_{\hat{\pi}} M_n(\hat{h}) - \eta n \mathsf{E}_{\hat{\pi}} \, R_n(\hat{h}) - D(\hat{\pi} \| \pi) > t\right) \leq \frac{1}{\mathrm{e}^t}.$$

Selecting the "learning rate" (1/2)

For every $\eta > 0$ and $\delta \in]0,1[$,

$$\mathsf{P}\left(\mathsf{E}_{\hat{\pi}}\, R(\hat{h}) \le \mathsf{E}_{\hat{\pi}}\, R_n(\hat{h}) + \frac{\eta}{8} + \frac{1}{\eta n}\left[D(\hat{\pi}\|\pi) + \log\frac{1}{\delta}\right]\right) \ge 1 - \delta.$$

Question. How do we choose the learning rate η?

Corollary 1. Follow the setup in Theorem 1. Let $\alpha > 1$, and $v > u > 0$. For any $\delta \in]0, 1[$, it holds with probability at least $(1 - \delta)$ that for any $\eta \in [u, v]$,

$$\mathsf{E}_{\hat{\pi}} \, R(\hat{h}) \leq \mathsf{E}_{\hat{\pi}} \, R_n(\hat{h}) + \frac{\eta}{8} + \frac{\alpha}{\eta n} \left[D(\hat{\pi} \| \pi) + \log \frac{1}{\delta} + \log \left\lceil \log_\alpha \frac{v}{u} \right\rceil \right].$$

Remark. Finding u and v such that the interval $[u, v]$ contains the optimal η is ad hoc.

T. van Erven. 2014. PAC-Bayes mini-tutorial: A continuous union bound.

Proof of Corollary 1

Proof. (Corollary 1) It suffices to prove that

$$\mathsf{E}_{\hat{\pi}}\, M_\eta(\hat{h}) \le \mathsf{E}_{\hat{\pi}}\, R_n(\hat{h}) + \frac{1}{\eta n}\left[D(\hat{\pi}\|\pi) + \log\frac{1}{\delta} + \log\left\lceil \log_\alpha \frac{v}{u}\right\rceil\right].$$

Define $\mathcal{U} \coloneqq \{\, u, u\alpha, u\alpha^2, \dots, u\alpha^{\lceil \log_\alpha(v/u)\rceil - 1}\,\}$. By the union bound, it holds with probability at least $(1-\delta)$ that

$$\mathsf{E}_{\hat{\pi}}\, M_\eta(\hat{h}) \le \mathsf{E}_{\hat{\pi}}\, R_n(\hat{h}) + \frac{1}{\eta n}\left[D(\hat{\pi}\|\pi) + \log\frac{1}{\delta} + \log\left\lceil \log_\alpha \frac{v}{u}\right\rceil\right],$$

for all $\eta \in \mathcal{U}$. It remains to notice that for any $\zeta \in [u, v]$, there must exist some $\eta \in \mathcal{U}$ such that $\eta \le \zeta \le \alpha\eta$.

Model selection

Trade-off between estimation and approximation errors

Consider the standard setting of statistical learning. Recall that

$$R(\hat{h}) - \min_{h} R(h) = \left[R(\hat{h}) - \min_{h \in \mathcal{H}} R(h)\right] + \left[\min_{h \in \mathcal{H}} R(h) - \min_{h} R(h)\right].$$

The first term on the RHS is called the *estimation error*; the second term is called the *approximation error*.

Question. How do we choose the hypothesis class $\mathcal{H}$?

Structural risk minimization

Let $\mathcal{H}$ be the hypothesis class.

Structural risk minimization. Let $\mathcal{H}_1, \mathcal{H}_2 \ldots$ be subsets of $\mathcal{H}$, such that their union equals $\mathcal{H}$. Let $\mathrm{pen}_n : \mathbb{N} \to \mathbb{R}$ be the *penalty function*. Denote by $\hat{h}_{n,k}$ the empirical risk minimizer of $\mathcal{H}_k$. The *structural risk minimization* approach is the following.

1. Solve the optimization problem

$$\hat{k} \in \arg\min_{k \in \mathbb{N}} R_n(\hat{h}_{n,k}) + \mathrm{pen}_n(k).$$

2. Output $\hat{h}_n = \hat{h}_{n,\hat{k}}$.

V. N. Vpanik. 1998. *Statistical Learning Theory*.

Oracle inequality

We expect that the output $\hat{h}_n$ satisfies

$$R(\hat{h}_n) \leq C \inf_{k \in \mathbb{N}} \left\{ \inf_{h \in \mathcal{F}_k} R(h) + \beta_n(k) \right\},$$

with high probability, for some $\beta_n : \mathbb{N} \to \mathbb{R}$. Such an inequality is called an *oracle inequality*. We expect that

$$\lim_{n \to \infty} \beta_n(k) \to 0, \quad \forall k \in \mathbb{N}.$$

If $C = 1$, then we call the oracle inequality *sharp*.

Question. How do we interpret such an inequality?

Penalization by error estimates (1/2): Rough idea

- We would like to find the $\hat{h}_{n,\hat{k}}$ which yields the smallest risk among $\hat{h}_{n,k}$'s.
- However, for every n and k, we do not have access to the exact value of $R(\hat{h}_{n,k})$.

Suppose that there are good risk estimates $\gamma_{n,k}$ for each $\mathcal{H}_k$:

$$\forall k \in \mathbb{N} : \gamma_{n,k} \approx R(\hat{h}_{n,k}), \quad \text{w.h.p.}$$

Then we may set

$$\mathrm{pen}_n(k) \approx \gamma_{n,k} - R_n(\hat{h}_{n,k}).$$

Penalization by error estimates (2/2)

Theorem 3. Let $\hat{h}_n$ be the output of a learning method. Suppose there exist random variables $\gamma_{n,k}$, such that

$$\mathsf{P}\left(R(\hat{h}_n) \geq \gamma_{n,k} + \varepsilon\right) \leq c\mathrm{e}^{-2m\varepsilon^2}, \quad \forall \varepsilon > 0,$$

where the numbers c and m may depend on n. Set

$$\mathrm{pen}_n(k) \coloneqq \gamma_{n,k} - R_n(\hat{h}_{n,k}) + \sqrt{\frac{\log k}{m}}, \quad \forall k \in \mathbb{N}.$$

Then, it holds that

$$\mathsf{E}\, R(\hat{h}_n) \leq \inf_{k \in \mathbb{N}} \left[\inf_{h \in \mathcal{H}_k} R(h) + \mathsf{E}\, \mathrm{pen}_n(k)\right] + \sqrt{\frac{\log(c\mathrm{e})}{2m}}.$$

P. L. Bartlett *et al.* 2002. Model selection and error estimation.

Proof of Theorem 3 (1/4): Preliminary

Lemma 1. Let ξ be a non-negative random variable. Then, we have

$$\mathsf{E}\,\xi = \int_0^\infty \mathsf{P}\left(\xi > t\right) \,\mathrm{d}t.$$

Corollary 2. Let $c > 0$ and $n \in \mathbb{N}$. Let ζ be a non-negative random variable satisfying $\mathsf{P}\left(\zeta > t\right) \leq c\mathrm{e}^{-2nt^2}$. Then, it holds that

$$\mathsf{E}\,\zeta \leq \sqrt{\frac{\log(c\mathrm{e})}{2n}}.$$

Proof. Notice that $\mathsf{E}\,\zeta \leq \sqrt{\mathsf{E}\,\zeta^2}$. Apply Lemma 1 to bound $\mathbb{E}\,\zeta^2$.

L. Devroye *et al.* 1996. *A Probabilistic Theory of Pattern Recognition.*

Proof of Theorem 3 (2/4)

Lemma 2. Define

$$\tilde{R}_n(\hat{h}_{n,k}) \coloneqq R_n(\hat{h}_{n,k}) + \mathrm{pen}_n(k) = \gamma_{n,k} + \sqrt{\frac{\log k}{m}}, \quad \forall k \in \mathbb{N}.$$

For any $\varepsilon > 0$, it holds that

$$\mathsf{P}\left(R(\hat{h}_n) - \tilde{R}_n(\hat{h}_n) > \varepsilon\right) \leq 2c\mathrm{e}^{-2m\varepsilon^2}.$$

Proof. (Lemma 2) By the union bound, we write

$$\begin{aligned}\mathsf{P}\left(R(\hat{h}_n) - \tilde{R}_n(\hat{h}_n) > \varepsilon\right) &\leq \sum_{k\in\mathbb{N}} \mathsf{P}\left(R(\hat{h}_{n,k}) - \tilde{R}_n(\hat{h}_{n,k}) > \varepsilon\right) \\ &= \sum_{k\in\mathbb{N}} \mathsf{P}\left(R(\hat{h}_{n,k}) - \gamma_{n,k} > \varepsilon + \sqrt{\frac{\log k}{m}}\right).\end{aligned}$$

Proof of Theorem 3 (3/4)

Proof. (Theorem 3) Define

$$R_k^\star := \min_{h \in \mathcal{H}_k} R(h).$$

We decompose the expected *excess risk* as

$$\mathsf{E}\left[R(\hat{h}_n) - R_k^\star\right] = \mathsf{E}\left[R(\hat{h}_n) - \tilde{R}_n(\hat{h}_n)\right] + \mathsf{E}\left[\tilde{R}_n(\hat{h}_n) - R_k^\star\right].$$

By Corollary 2 and Lemma 2, we have

$$\mathsf{E}\left[R(\hat{h}_n) - \tilde{R}_n(\hat{h}_n)\right] \leq \sqrt{\frac{\log(ce)}{2m}}.$$

It remains the bound the second term $\mathsf{E}\left[\tilde{R}_n(\hat{h}_n) - R_k^\star\right]$.

Proof of Theorem 3 (4/4)

Proof continued. (Theorem 3) We write

$$\begin{aligned}
\mathsf{E}\left[\tilde{R}_n(\hat{h}_n) - R_k^\star\right] &\le \mathsf{E}\,\tilde{R}_n(\hat{h}_{n,k}) - R_k^\star \\
&= \mathsf{E}\,R_n(\hat{h}_n) - R_k^\star + \mathsf{E}\,\mathrm{pen}_n(k) \\
&\le \mathsf{E}\,R_n(h_k^\star) - R_k^\star + \mathsf{E}\,\mathrm{pen}_n(k) \\
&= \mathsf{E}\,\mathrm{pen}_n(k).
\end{aligned}$$

Then, we have

$$\mathsf{E}\,R(\hat{h}_n) \le R_k^\star + \mathsf{E}\,\mathrm{pen}_n(k) + \sqrt{\frac{\log(c\mathrm{e})}{2m}}, \quad \forall k \in \mathbb{N}.$$

High-probability guarantee

Theorem 4. Consider the assumptions in Theorem 3. Then, for any $\varepsilon > 0$, it holds that

$$R(\hat{h}_n) \leq \inf_{k \in \mathbb{N}} \left[\inf_{h \in \mathcal{H}_k} R(h) + \mathrm{pen}_n(k) + \sqrt{\frac{\log k}{n}} \right] + \varepsilon,$$

with probability at least

$$1 - 2c\mathrm{e}^{-m\varepsilon^2/2} - 2c\mathrm{e}^{-n\varepsilon^2/2}.$$

Proof. Check the reference.

P. L. Bartlett *et al.* 2002. Model selection and error estimation.

Conclusions

Conclusions

- The PAC-Bayes approach is *non-Bayesian*. The prior does not represent a belief, and posterior is not updated following the Baye's rule.

- The PAC-Bayes approach considers the expected risk of a Gibbs predictor.

- Structural risk minimization is an approach to *balancing estimation and approximation errors*.

- Generalization error bounds can lead to structural risk minimization methods via *penalization*.

Important techniques

- Markov's inequality, the union bound, and expectation as integral of tail probabilities.

- Mix risk as a surrogate of the risk.

Next lecture

- Multiplicative weight update (voting, gambling, etc.).