# 泛化理论

## 第四章 PAC-Bayesian

## §4.1.1 PAC-Bayesian Bound

@ 滕佳烨

[ref] McAllester, D. A. (1999, July). PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory* (pp. 164-170).

**Recall:**

Uniform Convergence: $L(\hat{f}) - \hat{L}(\hat{f}) \leq \sup_{f\in\mathcal{F}} \left| L(f) - \hat{L}(f) \right|.$

Stability-based bound: similar dataset returns similar models

**Today's topic:**

**PAC-Bayesian bound**: another approach to generalization (on stochastic parameter)

## Stochastic Parameter (Bayesian)

***One sentence: the parameter is not a single value but drawn from a distribution***

**Standard training:** starting from an initialization, by an algorithm, returns a model, evaluate on the model.

$$\theta_0 \to \mathcal{A} \to \hat{\theta} \to \ell(\hat{\theta}).$$

**Bayesian training:** starting from an initial distribution (prior), by an algorithm, returns a trained distribution (posterior), evaluate on the model by expectation.

$$\theta \sim P \to \mathcal{A} \to \theta \sim Q \to \mathbb{E}_Q \ell(\theta).$$

**Goal:** to bound the generalization gap $L_D(Q) - L_S(Q)$,
where $L_D(Q) = \mathbb{E}_Q \mathbb{E}_z \ell(\theta; z)$ denotes the test error,
and $L_S(Q) = \mathbb{E}_Q \frac{1}{n} \sum_i \ell(\theta; z_i)$ denotes the training error,

**Remark:** PAC-Bayesian has something different from Bayesian, the prior is not the initial distribution. (but we still call it PAC-Bayesian.)

## PAC-Bayesian bound

***One sentence: if prior P (any given prior) is close to Q, the generalization is good.***

**Theorem (PAC-Bayesian).** Given prior distribution $P$, for bounded loss $\ell \in [0,1]$, with probability at least $1-\delta$ (prob over training), for all posterior distribution $Q$,

$$L_D(Q) - L_S(Q) \leq \sqrt{\frac{KL(Q||P) + \log\left(\frac{n}{\delta}\right)}{2(n-1)}}.$$

Remark: $P$ does not to be the initial distribution. All we need is that $P$ is independent of the training process.

The bound is approximately $\sqrt{\frac{KL(Q||P)}{n}}$. Therefore, if the trained distribution is close to our prior on it, the generalization bound is small.

**Take-away messages**

(a) Bayesian training: prior, posterior, parameter drawn from a distribution.
(b) PAC-Bayesian bound: “prior” P, posterior Q.
If P and Q is close (KL), the bound is small.
For any posterior $Q$, we have

$$L_D(Q) - L_S(Q) \leq \sqrt{\frac{KL(Q||P) + \log\left(\frac{n}{\delta}\right)}{2(n-1)}}.$$

All the slides will be available at [www.tengjiaye.com/generalization](www.tengjiaye.com/generalization) soon.

@ 滕佳烨

Thanks!