

Lecture 6: PAC Bayesian Theory

Tianpei Xie

Feb. 10th., 2023

Contents

1	Bayesian Learning	2
1.1	Bayesian Predictor	2
1.2	Generalized Bayesian Learning	3
1.3	Gibbs Posterior	3
2	PAC Bayesian Theory	3
2.1	PAC Bayesian Inequalities	3
2.2	PAC Bayesian Inequalities for Other Divergences	7

1 Bayesian Learning

1.1 Bayesian Predictor

- **Remark (*Data*)**

Define an **observation** as a d -dimensional vector x . The *unknown* nature of the observation is called a **class**, denoted as y . The domain of observation is called an **input space** or **feature space**, denoted as $\mathcal{X} \subset \mathbb{R}^d$, whereas the domain of class is called the **target space**, denoted as \mathcal{Y} . For **classification task**, $\mathcal{Y} = \{1, \dots, M\}$; and for **regression task**, $\mathcal{Y} = \mathbb{R}$. Denote a collection of n **samples** as

$$\mathcal{D} \equiv \mathcal{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n)).$$

Note that \mathcal{D}_n is a finite **sub-sequence** in $(\mathcal{X} \times \mathcal{Y})^n$.

- **Definition (*Concept Class as a Function Class*)**

A **concept** $c : \mathcal{X} \rightarrow \mathcal{Y}$ is the *input-output association* from the nature and is *to be learned* by *a learning algorithm*. Denote \mathcal{C} as the *set of all concepts* we wish to learn as the **concept class**. That is, $\mathcal{C} \subseteq \{c : \mathcal{X} \rightarrow \mathcal{Y}\} = \mathcal{Y}^{\mathcal{X}}$. Concept class \mathcal{C} is a *function class*.

- **Definition (*Hypothesis and Hypothesis Class*)**

The learner is requested to output a *prediction rule*, $h : \mathcal{X} \rightarrow \mathcal{Y}$. This function is also called a **predictor**, a **hypothesis**, or a **classifier**. The predictor can be used to predict the label of new domain points.

Note that \mathcal{H} and \mathcal{C} may not overlap, since the concept class is unknown to learner.

- **Definition (*Bayesian Hypothesis*)**

Assume instead that the hypothesis h is *random*. That is, let $(\mathcal{H}, \mathcal{H}, \mathbb{P})$ be a probability space with probability measure \mathbb{P} . We refer \mathbb{P} as the **prior distribution of hypothesis** $h \in \mathcal{H}$. The corresponding **randomized hypothesis** h is called **Bayesian hypothesis**.

- **Definition (*Bayesian Learning and Generalization Error*)**

Following the Bayesian reasoning approach, *the output of the learning algorithm is not necessarily a single hypothesis*. Instead, the learning process defines a **posterior probability** over \mathcal{H} , which we denote by \mathbb{Q} . Note that the posterior distribution is absolutely continuous with respect to prior \mathbb{P} , i.e. $\mathbb{Q} \ll \mathbb{P}$.

In the context of a *supervised learning problem*, where \mathcal{H} contains functions from \mathcal{X} to \mathcal{Y} , one can think of \mathbb{Q} as defining a randomized prediction rule as follows. Whenever we get a new instance x , we **randomly** pick a hypothesis $h \in \mathcal{H}$ according to \mathbb{Q} and predict $h(x)$. We define *the loss of* \mathbb{Q} on an example z to be

$$L(\mathbb{Q}, z) := \mathbb{E}_{h \sim \mathbb{Q}} [\ell(h, z)] \quad (1)$$

where $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}_+$ is a loss function. **The generalization loss** and **training loss of** \mathbb{Q} can be written as

$$L_{\mathcal{P}}(\mathbb{Q}) := \mathbb{E}_{h \sim \mathbb{Q}} [L_{\mathcal{P}}(h)] = \mathbb{E}_{h \sim \mathbb{Q}} [\mathbb{E}_{Z \sim \mathcal{P}} [\ell(h, Z)]] = \mathbb{E}_{Z \sim \mathcal{P}} [L(\mathbb{Q}, Z)] \quad (2)$$

$$L_{\mathcal{D}}(\mathbb{Q}) := \mathbb{E}_{h \sim \mathbb{Q}} [L_{\mathcal{D}}(h)] = \mathbb{E}_{h \sim \mathbb{Q}} \left[\frac{1}{m} \sum_{i=1}^m \ell(h, Z_i) \right] = \frac{1}{m} \sum_{i=1}^m L(\mathbb{Q}, Z_i) \quad (3)$$

1.2 Generalized Bayesian Learning

1.3 Gibbs Posterior

2 PAC Bayesian Theory

2.1 PAC Bayesian Inequalities

- **Theorem 2.1** (*Catoni's PAC Bayesian Inequality*)[Catoni, 2003, Alquier, 2021]

Let \mathcal{P} be an arbitrary distribution over an example domain \mathcal{Z} . Let \mathcal{H} be a hypothesis class and let $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow [0, 1]$ be a loss function. Let \mathbb{P} be a prior distribution over \mathcal{H} and let $\delta \in (0, 1)$. Then, with probability of at least $1 - \delta$ over the choice of an i.i.d. training set $\mathcal{D} = \{z_1, \dots, z_m\}$ sampled according to \mathcal{P} , for all distributions \mathbb{Q} over \mathcal{H} (even such that depend on \mathcal{D}) and for all $\lambda > 0$, we have

$$L_{\mathcal{P}}(\mathbb{Q}) \leq L_{\mathcal{D}}(\mathbb{Q}) + \frac{\text{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(1/\delta)}{\lambda} + \frac{\lambda}{8m} \quad (4)$$

where $\text{KL}(\mathbb{Q} \parallel \mathbb{P}) = \mathbb{E}_{\mathbb{Q}}[\log(\mathbb{Q}/\mathbb{P})]$ is the Kullback-Leibler divergence.

Proof: 1. Recall *the duality formulation* of logarithmic moment generating function for random variable M :

$$\log \mathbb{E}_{\mathbb{P}}[e^{\lambda M}] = \sup_{\mathbb{Q} \ll \mathbb{P}} \{\lambda \mathbb{E}_{\mathbb{Q}}[M] - \text{KL}(\mathbb{Q} \parallel \mathbb{P})\}$$

Let $M := \Delta(h)$ where $\Delta(h) := (L_{\mathcal{P}}(h) - L_{\mathcal{D}}(h))$. For all $\mathbb{Q} \ll \mathbb{P}$, we have

$$\log \mathbb{E}_{\mathbb{P}}[e^{\lambda \Delta(h)}] \geq \{\lambda \mathbb{E}_{\mathbb{Q}}[\Delta(h)] - \text{KL}(\mathbb{Q} \parallel \mathbb{P})\}. \quad (5)$$

It follows that $\Delta(h) := (L_{\mathcal{P}}(h) - L_{\mathcal{D}}(h)) \equiv \Delta(h, \mathcal{D})$. Taking exponential and expectation with respect to sample \mathcal{D} on both sides of inequality yields

$$\mathbb{E}_{\mathcal{D}}[e^{\sup_{\mathbb{Q} \ll \mathbb{P}} \{\lambda \mathbb{E}_{\mathbb{Q}}[\Delta(h)] - \text{KL}(\mathbb{Q} \parallel \mathbb{P})\}}] \leq \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\mathbb{P}}[e^{\lambda \Delta(h)}]] \quad (6)$$

The advantage of the expression on the right-hand side stems from the fact that we can switch the order of expectations (because \mathbb{P} is a prior that **does not depend on sample \mathcal{D}**), which yields

$$\mathbb{E}_{\mathcal{D}}[e^{\sup_{\mathbb{Q} \ll \mathbb{P}} \{\lambda \mathbb{E}_{\mathbb{Q}}[\Delta(h)] - \text{KL}(\mathbb{Q} \parallel \mathbb{P})\}}] \leq \mathbb{E}_{\mathbb{P}}[\mathbb{E}_{\mathcal{D}}[e^{\lambda \Delta(h)}]] \quad (7)$$

2. Next, **for any hypothesis** $h \in \mathcal{H}$, we bound the expectation term $\mathbb{E}_{\mathcal{D}}[e^{\lambda \Delta(h)}]$. Since $L_{\mathcal{D}}(h) = \frac{1}{m} \sum_{i=1}^m \ell(h, z_i) \in [0, 1]$, a.s., from *Hoeffding's lemma*

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}[e^{\lambda m \Delta(h)}] &\leq \exp\left(\frac{m\lambda^2}{8}\right) \\ \Rightarrow \mathbb{E}_{\mathcal{D}}[e^{\lambda \Delta(h)}] &\leq \exp\left(\frac{\lambda^2}{8m}\right) \end{aligned} \quad (8)$$

Combining (8) with Equation (7), we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} \left[e^{\sup_{\mathbb{Q} \ll \mathbb{P}} \{\lambda \mathbb{E}_{\mathbb{Q}}[\Delta(h)] - \text{KL}(\mathbb{Q} \parallel \mathbb{P})\}} \right] \leq \exp \left(\frac{\lambda^2}{8m} \right) \\ \Rightarrow & \mathbb{E}_{\mathcal{D}} \left[\exp \left(\sup_{\mathbb{Q} \ll \mathbb{P}} \left\{ \lambda \mathbb{E}_{\mathbb{Q}}[\Delta(h)] - \text{KL}(\mathbb{Q} \parallel \mathbb{P}) - \frac{\lambda^2}{8m} \right\} \right) \right] \leq 1 \end{aligned} \quad (9)$$

3. Finally, we obtain the result by applying *Chernoff's method*. Specifically, by *Markov's inequality*,

$$\begin{aligned} & \mathcal{P}_{\mathcal{D}} \left\{ \sup_{\mathbb{Q} \ll \mathbb{P}} \left\{ \lambda \mathbb{E}_{\mathbb{Q}}[\Delta(h)] - \text{KL}(\mathbb{Q} \parallel \mathbb{P}) - \frac{\lambda^2}{8m} \right\} \geq \epsilon \right\} \\ & \leq e^{-\epsilon} \mathbb{E}_{\mathcal{D}} \left[\exp \left(\sup_{\mathbb{Q} \ll \mathbb{P}} \left\{ \lambda \mathbb{E}_{\mathbb{Q}}[\Delta(h)] - \text{KL}(\mathbb{Q} \parallel \mathbb{P}) - \frac{\lambda^2}{8m} \right\} \right) \right] \\ & \leq e^{-\epsilon} \end{aligned} \quad (10)$$

Denote the right-hand side of the above δ , thus $\epsilon = \log(1/\delta)$. After rearranging the term, we therefore obtain that with probability of at least $1 - \delta$ we have that for all $\mathbb{Q} \ll \mathbb{P}$, and for all λ

$$\mathbb{E}_{\mathbb{Q}}[\Delta(h)] \leq \frac{\text{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(1/\delta)}{\lambda} + \frac{\lambda}{8m}. \quad \blacksquare$$

- The first PAC-Bayesian Inequality is from McAllester [McAllester, 2003].

Theorem 2.2 (McAllester's PAC Bayesian Inequality) [McAllester, 2003, Shalev-Shwartz and Ben-David, 2014, Rasmussen and Williams, 2005, Alquier, 2021]

Under the same condition as in (4), then, with probability of at least $1 - \delta$, for all distributions $\mathbb{Q} \ll \mathbb{P}$ over \mathcal{H} , we have

$$\mathbb{E}_{h \sim \mathbb{Q}}[L_{\mathcal{P}}(h)] \leq \mathbb{E}_{h \sim \mathbb{Q}}[L_{\mathcal{D}}(h)] + \sqrt{\frac{\text{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(1/\delta) + \log(m) + 2}{2m - 1}} \quad (11)$$

Proof: The proof is similar to above. In this time, we want to show that

$$\mathbb{E}_{\mathcal{D}} \left[e^{(2m-1)\Delta(h)^2} \right] \leq 4m, \quad (12)$$

where $\Delta(h) := |L_{\mathcal{P}}(h) - L_{\mathcal{D}}(h)|$. Since the loss function is bounded within $[0, 1]$ almost surely, by *Hoeffding's inequality*

$$\mathcal{P}_{\mathcal{D}} \{ \Delta(h) \geq x \} \leq 2 \exp(-2mx^2).$$

Note that $\mathcal{P}_{\mathcal{D}} \{ \Delta \geq x \} = \int_x^\infty f(\Delta) d\Delta$ where $f(\Delta) \equiv \frac{d\mathcal{P}_{\mathcal{D}}(\Delta)}{d\Delta}$ is the density function. Since the tail is dominated by Gaussian tail, the density function is also dominated by Gaussian density

$$\begin{aligned} & \int_x^\infty f(\Delta) d\Delta \leq 2e^{-2mx^2} \\ \Rightarrow & f(\Delta) \leq 8m\Delta e^{-2m\Delta^2}. \end{aligned}$$

Therefore, the expectation

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}} \left[e^{(2m-1)\Delta(h)^2} \right] &= \int_0^\infty e^{(2m-1)\Delta^2} f(\Delta) d\Delta \\
&\leq \int_0^\infty e^{(2m-1)\Delta^2} 8m\Delta e^{-2m\Delta^2} d\Delta \\
&= 8m \int_0^\infty e^{-\Delta^2} \Delta d\Delta \\
&= 4m.
\end{aligned}$$

With inequality (12), we use the dual formulation of log-MGF,

$$\log \mathbb{E}_{\mathbb{P}} \left[e^{\lambda M} \right] = \sup_{\mathbb{Q} \ll \mathbb{P}} \{ \mathbb{E}_{\mathbb{Q}} [\lambda M] - \text{KL}(\mathbb{Q} \parallel \mathbb{P}) \}$$

and let $M := \Delta^2$ and $\lambda := (2m-1)$, so that we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}} \left[e^{\sup_{\mathbb{Q} \ll \mathbb{P}} \{ \mathbb{E}_{\mathbb{Q}} [(2m-1)\Delta^2] - \text{KL}(\mathbb{Q} \parallel \mathbb{P}) \}} \right] &\leq \mathbb{E}_{\mathcal{D}} \left[\mathbb{E}_{\mathbb{P}} \left[e^{(2m-1)\Delta(h)^2} \right] \right] \\
&\leq \mathbb{E}_{\mathbb{P}} \left[\mathbb{E}_{\mathcal{D}} \left[e^{(2m-1)\Delta(h)^2} \right] \right] \\
&\leq 4m \quad (\text{by bound (12)}). \tag{13}
\end{aligned}$$

By Markov's inequality

$$\begin{aligned}
&\mathcal{P} \left\{ \sup_{\mathbb{Q} \ll \mathbb{P}} \{ \mathbb{E}_{\mathbb{Q}} [(2m-1)\Delta^2] - \text{KL}(\mathbb{Q} \parallel \mathbb{P}) \} \geq \epsilon \right\} \\
&\leq e^{-\epsilon} \mathbb{E}_{\mathcal{D}} \left[e^{\sup_{\mathbb{Q} \ll \mathbb{P}} \{ \mathbb{E}_{\mathbb{Q}} [(2m-1)\Delta^2] - \text{KL}(\mathbb{Q} \parallel \mathbb{P}) \}} \right] \\
&\leq 4me^{-\epsilon}.
\end{aligned}$$

Denote the RHS as δ , so $\epsilon = \log(4m/\delta)$. We have with probability as least $1 - \delta$, for all $\mathbb{Q} \ll \mathbb{P}$,

$$\begin{aligned}
(2m-1)\mathbb{E}_{\mathbb{Q}} [\Delta^2] - \text{KL}(\mathbb{Q} \parallel \mathbb{P}) &\leq \log \frac{4m}{\delta} \\
\Rightarrow (\mathbb{E}_{\mathbb{Q}} [\Delta])^2 &\leq \mathbb{E}_{\mathbb{Q}} [\Delta^2] \leq \frac{\text{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(1/\delta) + \log(m) + 2}{2m-1}
\end{aligned}$$

The leftmost inequality is due to Jenson's inequality on $\phi(x) := x^2$. We have proved the result. \blacksquare

- **Remark** Note that this bound (11) cannot be obtained from (4) by minimizing λ since the optimal $\lambda^* = \sqrt{(\text{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(1/\delta))8m}$ depends on \mathbb{Q} , which is not allowed.

A natural idea is to propose a *finite grid* $\Lambda \subset (0, +\infty)$ and to minimize over this grid, which can be justified by a union bound argument. This way we pay the rise for an additional term $\log(m)$ in the bound, i.e. $\sqrt{\frac{\text{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(1/\delta)}{2m}} \rightarrow \sqrt{\frac{\text{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(1/\delta) + \log(m)}{2m-1}}$.

- **Remark (*Generalization Error Bound of Posterior by KL Divergence*)**
The McAllester's PAC Bayesian theorem tells us that the difference between the generalization loss and the empirical loss of a **posterior** \mathbb{Q} is bounded by an expression that depends on **the Kullback-Leibler divergence** between \mathbb{Q} and the prior distribution \mathbb{P} .

- **Remark (*Agnostic PAC Bound vs. PAC Bayesian Bound*)**

We can compare the PAC bound and PAC-Bayesian bound. With probability at least $1 - \delta$,

$$\begin{aligned} \text{(Agnostic PAC Bound)} \quad L_{\mathcal{P}}(h) &\leq L_{\mathcal{D}}(h) + \sqrt{\frac{\log |\mathcal{H}| + \log(1/\delta)}{2m}} \\ \text{(PAC-Bayesian Bound)} \quad \mathbb{E}_{h \sim \mathbb{Q}} [L_{\mathcal{P}}(h)] &\leq \mathbb{E}_{h \sim \mathbb{Q}} [L_{\mathcal{D}}(h)] + \sqrt{\frac{\text{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(m/\delta)}{2m}} \end{aligned}$$

- **Remark (*Bayesian Learning as Minimum Description Length*)**

As in the MDL paradigm, we define a *hierarchy* over hypotheses in our class \mathcal{H} . Now, the hierarchy takes the form of a prior distribution over \mathcal{H} so that the preferred hypothesis has higher chance being selected.

The *McAllester's PAC Bayesian bound* is like the MDL paradigm with the complexity of hypothesis encoded by the KL-divergence.

- **Remark (*Regularization*).**

The **PAC-Bayes bound** leads to the following learning rule:

Given a prior \mathbb{P} , return a posterior \mathbb{Q} that minimizes the function

$$\mathbb{E}_{h \sim \mathbb{Q}} [L_{\mathcal{D}}(h)] + \sqrt{\frac{\text{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(m/\delta)}{2m - 1}} \quad (14)$$

This rule is similar to the regularized risk minimization principle. That is, we *jointly minimize the empirical loss* of \mathbb{Q} on the sample *and the Kullback-Leibler "distance"* between \mathbb{Q} and \mathbb{P} .

- For the special case of 0-1 loss, we can the following improved bound:

Theorem 2.3 (*Seeger's PAC Bayesian Inequality*)[Seeger, 2002, Maurer, 2004, Rasmussen and Williams, 2005, Alquier, 2021]

Let \mathcal{P} be an arbitrary distribution over an example domain \mathcal{Z} . Let \mathcal{H} be a hypothesis class and let $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \{0, 1\}$ be a loss function. Let \mathbb{P} be a prior distribution over \mathcal{H} and let $\delta \in (0, 1)$. Then, with probability of at least $1 - \delta$ over \mathcal{D} , for all distributions $\mathbb{Q} \ll \mathbb{P}$ over \mathcal{H} , we have

$$\text{KL}_{\text{Ber}}(L_{\mathcal{D}}(\mathbb{Q}) \parallel L_{\mathcal{P}}(\mathbb{Q})) \leq \frac{\text{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(1/\delta) + \log(2\sqrt{m})}{m} \quad (15)$$

where $\text{KL}_{\text{Ber}}(p \parallel q)$ is the Kullback-Leibler divergence for Bernoulli random variable

$$\text{KL}_{\text{Ber}}(p \parallel q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

- **Remark** This bound is based on the following inequality (see [Maurer, 2004]):

$$\mathbb{E}_{\mathcal{D}} \left[e^{m \text{KL}_{\text{Ber}}(\hat{\mu}_m \parallel \mu)} \right] \leq 2\sqrt{m}, \quad (16)$$

where $\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^m X_i$ where $X_i \in [0, 1]$ almost surely and X_1, \dots, X_m are i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$. The inequality is sharp since the equality is attained by Bernoulli random variable. The original inequality in [Seeger, 2002] is

$$\mathbb{E}_{\mathcal{D}} \left[e^{m \text{KL}_{\text{Ber}}(\hat{\mu}_m \parallel \mu)} \right] \leq m + 1.$$

- **Remark** By *Pinsker's inequality*,

$$(L_{\mathcal{P}}(\mathbb{Q}) - L_{\mathcal{D}}(\mathbb{Q}))^2 \leq \mathbb{KL}_{Ber}(L_{\mathcal{D}}(\mathbb{Q}) \parallel L_{\mathcal{P}}(\mathbb{Q}))$$

which recovers the inequality (11).

- **Remark** We can rewrite (15) explicitly as

$$\mathcal{P}_{\mathcal{D}} \left\{ L_{\mathcal{P}}(\mathbb{Q}) \leq \mathbb{KL}_{Ber}^{-1} \left(L_{\mathcal{D}}(\mathbb{Q}) \parallel \frac{\mathbb{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(2\sqrt{m}/\delta)}{m} \right) \right\} \geq 1 - \delta \quad (17)$$

where

$$\mathbb{KL}_{Ber}^{-1}(q \parallel b) = \sup \{p \in [0, 1] : \mathbb{KL}_{Ber}(p \parallel q) \leq b\}.$$

- **Corollary 2.4** [Alquier, 2021]

For any $\delta > 0$, any $\lambda \in (0, 2)$, with probability at least $1 - \delta$,

$$L_{\mathcal{P}}(\mathbb{Q}) \leq \left(1 - \frac{\lambda}{2}\right)^{-1} L_{\mathcal{D}}(\mathbb{Q}) + \left[\lambda \left(1 - \frac{\lambda}{2}\right)\right]^{-1} \frac{\mathbb{KL}(\mathbb{Q} \parallel \mathbb{P}) + \log(2\sqrt{m}/\delta)}{m} \quad (18)$$

2.2 PAC Bayesian Inequalities for Other Divergences

References

- Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.
- Olivier Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*, 840, 2003.
- Andreas Maurer. A note on the PAC Bayesian theorem. *arXiv preprint cs/0411099*, 2004.
- David A McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.
- Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.
- Matthias Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.