

31 PAC-Bayes

The Minimum Description Length (MDL) and Occam's razor principles allow a potentially very large hypothesis class but define a hierarchy over hypotheses and prefer to choose hypotheses that appear higher in the hierarchy. In this chapter we describe the PAC-Bayesian approach that further generalizes this idea. In the PAC-Bayesian approach, one expresses the prior knowledge by defining prior distribution over the hypothesis class.

31.1 PAC-Bayes Bounds

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} . That is, we assign a probability (or density if \mathcal{H} is continuous) $P(h) \geq 0$ for each $h \in \mathcal{H}$ and refer to $P(h)$ as the prior score of h . 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 Q . In the context of a supervised learning problem, where \mathcal{H} contains functions from \mathcal{X} to \mathcal{Y} , one can think of Q as defining a randomized prediction rule as follows. Whenever we get a new instance \mathbf{x} , we randomly pick a hypothesis $h \in \mathcal{H}$ according to Q and predict $h(\mathbf{x})$. We define the loss of Q on an example z to be

$$\ell(Q, z) \stackrel{\text{def}}{=} \mathbb{E}_{h \sim Q} [\ell(h, z)].$$

By the linearity of expectation, the generalization loss and training loss of Q can be written as

$$L_{\mathcal{D}}(Q) \stackrel{\text{def}}{=} \mathbb{E}_{h \sim Q} [L_{\mathcal{D}}(h)] \quad \text{and} \quad L_S(Q) \stackrel{\text{def}}{=} \mathbb{E}_{h \sim Q} [L_S(h)].$$

The following theorem tells us that the difference between the generalization loss and the empirical loss of a posterior Q is bounded by an expression that depends on the Kullback-Leibler divergence between Q and the prior distribution P . The Kullback-Leibler is a natural measure of the distance between two distributions. The theorem suggests that if we would like to minimize the generalization loss of Q , we should jointly minimize both the empirical loss of Q and the Kullback-Leibler distance between Q and the prior distribution. We will

later show how in some cases this idea leads to the regularized risk minimization principle.

THEOREM 31.1 *Let \mathcal{D} be an arbitrary distribution over an example domain Z . Let \mathcal{H} be a hypothesis class and let $\ell : \mathcal{H} \times Z \rightarrow [0, 1]$ be a loss function. Let 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 $S = \{z_1, \dots, z_m\}$ sampled according to \mathcal{D} , for all distributions Q over \mathcal{H} (even such that depend on S), we have*

$$L_{\mathcal{D}}(Q) \leq L_S(Q) + \sqrt{\frac{D(Q||P) + \ln m/\delta}{2(m-1)}},$$

where

$$D(Q||P) \stackrel{\text{def}}{=} \mathbb{E}_{h \sim Q} [\ln(Q(h)/P(h))]$$

is the Kullback-Leibler divergence.

Proof For any function $f(S)$, using Markov's inequality:

$$\mathbb{P}_S[f(S) \geq \epsilon] = \mathbb{P}_S[e^{f(S)} \geq e^\epsilon] \leq \frac{\mathbb{E}_S[e^{f(S)}]}{e^\epsilon}. \quad (31.1)$$

Let $\Delta(h) = L_{\mathcal{D}}(h) - L_S(h)$. We will apply Equation (31.1) with the function

$$f(S) = \sup_Q \left(2(m-1) \mathbb{E}_{h \sim Q} (\Delta(h))^2 - D(Q||P) \right).$$

We now turn to bound $\mathbb{E}_S[e^{f(S)}]$. The main trick is to upper bound $f(S)$ by using an expression that does not depend on Q but rather depends on the prior probability P . To do so, fix some S and note that from the definition of $D(Q||P)$ we get that for all Q ,

$$\begin{aligned} 2(m-1) \mathbb{E}_{h \sim Q} (\Delta(h))^2 - D(Q||P) &= \mathbb{E}_{h \sim Q} [\ln(e^{2(m-1)\Delta(h)^2} P(h)/Q(h))] \\ &\leq \ln \mathbb{E}_{h \sim Q} [e^{2(m-1)\Delta(h)^2} P(h)/Q(h)] \\ &= \ln \mathbb{E}_{h \sim P} [e^{2(m-1)\Delta(h)^2}], \end{aligned} \quad (31.2)$$

where the inequality follows from Jensen's inequality and the concavity of the log function. Therefore,

$$\mathbb{E}_S[e^{f(S)}] \leq \mathbb{E}_S \mathbb{E}_{h \sim P} [e^{2(m-1)\Delta(h)^2}]. \quad (31.3)$$

The advantage of the expression on the right-hand side stems from the fact that we can switch the order of expectations (because P is a prior that does not depend on S), which yields

$$\mathbb{E}_S[e^{f(S)}] \leq \mathbb{E}_{h \sim P} \mathbb{E}_S [e^{2(m-1)\Delta(h)^2}]. \quad (31.4)$$

Next, we claim that for all h we have $\mathbb{E}_S[e^{2(m-1)\Delta(h)^2}] \leq m$. To do so, recall that Hoeffding's inequality tells us that

$$\mathbb{P}_S[\Delta(h) \geq \epsilon] \leq e^{-2m\epsilon^2}.$$

This implies that $\mathbb{E}_S[e^{2(m-1)\Delta(h)^2}] \leq m$ (see Exercise 1). Combining this with Equation (31.4) and plugging into Equation (31.1) we get

$$\mathbb{P}_S[f(S) \geq \epsilon] \leq \frac{m}{e^\epsilon}. \quad (31.5)$$

Denote the right-hand side of the above δ , thus $\epsilon = \ln(m/\delta)$, and we therefore obtain that with probability of at least $1 - \delta$ we have that for all Q

$$2(m-1) \mathbb{E}_{h \sim Q}(\Delta(h))^2 - D(Q||P) \leq \epsilon = \ln(m/\delta).$$

Rearranging the inequality and using Jensen's inequality again (the function x^2 is convex) we conclude that

$$\left(\mathbb{E}_{h \sim Q} \Delta(h) \right)^2 \leq \mathbb{E}_{h \sim Q} (\Delta(h))^2 \leq \frac{\ln(m/\delta) + D(Q||P)}{2(m-1)}. \quad (31.6)$$

□

Remark 31.1 (Regularization) The PAC-Bayes bound leads to the following learning rule:

Given a prior P , return a posterior Q that minimizes the function

$$L_S(Q) + \sqrt{\frac{D(Q||P) + \ln m/\delta}{2(m-1)}}. \quad (31.7)$$

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

31.2 Bibliographic Remarks

PAC-Bayes bounds were first introduced by McAllester (1998). See also (McAllester 1999, McAllester 2003, Seeger 2003, Langford & Shawe-Taylor 2003, Langford 2006).

31.3 Exercises

1. Let X be a random variable that satisfies $\mathbb{P}[X \geq \epsilon] \leq e^{-2m\epsilon^2}$. Prove that $\mathbb{E}[e^{2(m-1)X^2}] \leq m$.

2. • Suppose that \mathcal{H} is a finite hypothesis class, set the prior to be uniform over \mathcal{H} , and set the posterior to be $Q(h_S) = 1$ for some h_S and $Q(h) = 0$ for all other $h \in \mathcal{H}$. Show that

$$L_{\mathcal{D}}(h_S) \leq L_S(h) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(m/\delta)}{2(m-1)}}.$$

Compare to the bounds we derived using uniform convergence.

- Derive a bound similar to the Occam bound given in Chapter 7 using the PAC-Bayes bound