Machine Learning Theory (CSC 482A/581A) - Lectures 23 & 24

Nishant Mehta

1 From "deterministic" predictors to randomized predictors

The risk guarantees we have developed so far all apply to learning algorithms that output a single hypothesis $\hat{f}$. What if we were to generalize our notion of learning algorithms to include those which, given a training sample, output a distribution $\hat{\Pi}$ over a set of hypotheses $\mathcal{F}$; then, each time a test input x is received and a prediction needs to be made, a hypothesis f is sampled from $\hat{\Pi}$ and the prediction made is $f(x)$. Since the predictions are random, even conditional on the training sample and the test input, we refer to such a predictor as a *randomized predictor*.

PAC-Bayesian bounds are a type of risk guarantee for such randomized predictors; this style of bounds is rooted in radically different techniques than those we have seen thus far in this course. The name "PAC-Bayesian" (often abbreviated to PAC-Bayes) derives from PAC learning — we still will seek PAC-style (ε, δ)-bounds — and from Bayesian estimators — statistical estimators which make predictions according to a posterior that is based on a prior distribution Π over $\mathcal{F}$ (which does not depend on the training sample) and an analogue of a likelihood function (which does depend on the training sample).

The name PAC-Bayes can be misleading:

- Regarding "PAC": we make no assumption that we are in the PAC learning setting (i.e. the realizable setting).
- Regarding "Bayes": Bayesian inference is based on a subjective belief, in the form of a prior distribution Π over $\mathcal{F}$, over the likelihood of each hypotheses being the correct rule for labeling the data; in PAC-Bayesian analysis, the bounds we derive will be correct regardless of whether or not we truly believe that, before seeing the training sample, the correct hypothesis was drawn according to our prior distribution; moreover, it can well be that none of the hypotheses is the correct one.

We will restrict our discussion to the setting of classification for simplicity, but PAC-Bayesian bounds can in fact be derived for general loss functions (thus extending to regression as well) and a PAC-Bayesian-style analysis even can be extended to online learning.

Before introducing a first PAC-Bayesian bound, we introduce the notion of a Gibbs classifier.

Definition 1. The *Gibbs classifier* with respect to probability distribution Q over $\mathcal{F}$ is the classifier which, for each an unlabeled example x, predicts by first drawing $f \sim Q$ and then outputting label $f(x)$. The risk of the Gibbs classifier is defined as its expected risk with respect to Q:

$$R(Q) = \mathsf{E}_{f \sim Q}[R(f)].$$

2 PAC-Bayesian bounds

We frequently need to refer to the KL divergence between two Bernoulli distributions, and so let's introduce notation for this special case. Denote by $\mathrm{d_{KL}}(p, q)$ the KL-divergence of Bernoulli(p) from Bernoulli(q), so that

$$\mathrm{d_{KL}}(p, q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

In the theorem below (and throughout this lecture), Π will be a prior distribution over $\mathcal{F}$ which does not depend on the training sample, $\hat{\Pi}$ will be a posterior distribution over $\mathcal{F}$ which can (and should) depend on the training sample.

Theorem 1. *With probability at least $1-\delta$ over the training sample, for all $\hat{\Pi}$,*

$$\mathrm{d_{KL}}\left(\mathsf{E}_{f\sim\hat{\Pi}}\left[\hat{R}_n(f)\right], \mathsf{E}_{f\sim\hat{\Pi}}[R(f)]\right) \leq \frac{1}{n}\left(\mathrm{D_{KL}}(\hat{\Pi} \,\|\, \Pi) + \log(n+1) + \log\frac{1}{\delta}\right).$$

The above theorem is not easy to interpret, but it does provide a quantifiably useful bound. Treated as a function of its second parameter (with the first parameter fixed), the function $\mathrm{d_{KL}}$ is invertible, and hence one can re-express the above bound as an upper bound on $\mathsf{E}_{f\sim\hat{\Pi}}[R(f)]$, the risk of the Gibbs classifier with respect to posterior $\hat{\Pi}$. Still, with a little more work, we can obtain a more easily interpretable (but looser) bound. For this, we make use of an inequality called *Pinsker's inequality*:

Lemma 1. *For probability distributions P and Q, it holds that*

$$\int_{x\in\mathcal{X}} |P(x) - Q(x)| \leq \sqrt{2\,\mathrm{D_{KL}}(P \,\|\, Q)}.$$

It is simple to verify that for P and Q Bernoulli distributions with success probabilities p and q respectively, we have

$$\int_{x\in\mathcal{X}} |P(x) - Q(x)| = 2|p-q|,$$

and so in this speical case,

$$2|p-q| \leq \sqrt{2\,\mathrm{d_{KL}}(p \,\|\, q)} \quad \Rightarrow \quad 2|p-q|^2 \leq \mathrm{d_{KL}}(P \,\|\, Q).$$

We thus have the following corollary of Theorem 1:

Corollary 1. *With probability at least $1-\delta$ over the training sample, for all $\hat{\Pi}$,*

$$\left|\mathsf{E}_{f\sim\hat{\Pi}}\left[\hat{R}_n(f)\right] - \mathsf{E}_{f\sim\hat{\Pi}}[R(f)]\right| \leq \sqrt{\frac{1}{2n}\left(\mathrm{D_{KL}}(\hat{\Pi} \,\|\, \Pi) + \log(n+1) + \log\frac{1}{\delta}\right)}. \tag{1}$$

Before proving Theorem 1, let's see why the above corollary might be useful (aside from just being yet another risk bound). The bound (1) motivates a learning algorithm that tries to minimize

$$\mathsf{E}_{f\sim\hat{\Pi}}\left[\hat{R}_n(f)\right] + \sqrt{\frac{1}{2n}\left(\mathrm{D_{KL}}(\hat{\Pi} \,\|\, \Pi)\right)}.$$

This is a form of regularized empirical risk minimization, where the regularizer (the second term) depends on our choice of prior Π. Moreover, in the special case where $\mathcal{F}$ is finite, we select the uniform prior over $\mathcal{F}$ (so that $\Pi(f) = \frac{1}{|\mathcal{F}|}$ for all $f \in \mathcal{F}$), and we take $\hat{\Pi}$ to be a Dirac distribution supported on a single hypothesis $\hat{f}$ (so that $\hat{\Pi}(\hat{f}) = 1$), we have (using $0 \cdot \log 0 = 0$)

$$\mathrm{D}_{\mathrm{KL}}(\hat{\Pi} \,\|\, \Pi) = \sum_{f \in \mathcal{F}} \hat{\Pi}(f) \log \frac{\hat{\Pi}(f)}{\Pi(f)} = \hat{\Pi}(\hat{f}) \log \frac{\hat{\Pi}(\hat{f})}{\Pi(\hat{f})} = \log \frac{1}{1/|\mathcal{F}|} = \log |\mathcal{F}|.$$

The KL divergence complexity term thus recovers the standard complexity term of $\log |\mathcal{F}|$ for finite classes. But the KL divergence complexity term can be even smaller if $\hat{\Pi}$ is more diffuse. As an exercise, consider what happens if $\hat{\Pi}$ is the uniform distribution over k hypotheses $\hat{f}_1, \ldots, \hat{f}_k$ (where, as before, each hypothesis depends on the training sample).

In order to prove Theorem 1, we will use a powerful inequality colloquially referred to as the *change of measure inequality*[1].

Lemma 2. *Consider a function $\psi \colon \mathcal{F} \to \mathbb{R}$ and let P and Q be probability distributions over $\mathcal{F}$ with corresponding probability density functions p and q respectively. Then*

$$\mathsf{E}_{f \sim Q} [\psi(f)] - \mathrm{D}_{\mathrm{KL}}(Q \,\|\, P) \le \log \mathsf{E}_{f \sim P} [\exp(\psi(f))] .$$

$$\mathsf{E}_{f \sim Q} [\psi(f)] - \mathrm{D}_{\mathrm{KL}}(Q \,\|\, P) \le \log \mathsf{E}_{f \sim P} [\exp(\psi(f))] .$$

Proof.

$$\begin{aligned} \log \mathsf{E}_{f \sim P} [\exp(\psi(f))] &= \log \mathsf{E}_{f \sim Q} \left[\exp\left(\psi(f) + \log \frac{p(f)}{q(f)} \right) \right] \\ &\ge \log \exp \left(\mathsf{E}_{f \sim Q} \left[\psi(f) + \log \frac{p(f)}{q(f)} \right] \right) \qquad \text{(Jensen's inequality)} \\ &= \mathsf{E}_{f \sim Q} [\psi(f)] - \mathrm{D}_{\mathrm{KL}}(Q \,\|\, P). \end{aligned}$$

□

Proof of Theorem 1. First, we make use of the fact that the KL divergence is jointly convex in its arguments. We will not prove this here, but a proof is not difficult. Then, from Jensen's inequality,

$$\lambda \, \mathrm{d}_{\mathrm{KL}} \left(\mathsf{E}_{f \sim \hat{\Pi}} \left[\hat{R}_n(f) \right], \mathsf{E}_{f \sim \hat{\Pi}} [R(f)] \right) \le \mathsf{E}_{f \sim \hat{\Pi}} \left[\lambda \, \mathrm{d}_{\mathrm{KL}} \left(\hat{R}_n(f), R(f) \right) \right] . \tag{2}$$

Applying Lemma 2 with $P = \Pi$, $Q = \hat{\Pi}$, and $\psi(f) = \lambda \, \mathrm{d}_{\mathrm{KL}}(\hat{R}_n(f), R(f))$ yields

$$\mathsf{E}_{f \sim \hat{\Pi}} \left[\lambda \, \mathrm{d}_{\mathrm{KL}} \left(\hat{R}_n(f), R(f) \right) \right] \le \log \mathsf{E}_{f \sim \Pi} \left[\exp \left(\lambda \, \mathrm{d}_{\mathrm{KL}} \left(\hat{R}_n(f), R(f) \right) \right) \right] + \mathrm{D}_{\mathrm{KL}}(\hat{\Pi} \,\|\, \Pi). \tag{3}$$

Next, let's obtain a high probability upper bound on the first term on the right-hand side. To do this, we will use Markov's inequality:

[1] Because this inequality holds for all probability distributions Q, we can also take the supremum (with respect to Q) of the left hand side; taking $q(f) = \frac{e^{\psi(f)} p(f)}{\mathsf{E}_{g \sim P}\, e^{\psi(g)}}$, the inequality becomes an equality. This is known as the dual formula to the Donsker-Varadhan variational formula; the Donsker-Varadhan variational formula itself is a re-expression of KL-divergence as $\mathrm{D}_{\mathrm{KL}}(P \,\|\, Q) = \sup_{\psi} \left\{ \mathsf{E}_{f \sim P}[\psi(f)] - \log \mathsf{E}_{f \sim Q} \left[e^{\psi(f)} \right] \right\}$.

Let X be a nonnegative random variable. Then for any $t > 0$,

$$\Pr(X \geq t) \leq \frac{\mathsf{E}[X]}{t}.$$

To see why Markov's inequality holds, observe that for $t > 0$ it holds that $\mathbf{1}\left[X \geq t\right] \leq \frac{X}{t}$. Hence,

$$\Pr(X \geq t) = \mathsf{E}[\mathbf{1}\left[X \geq t\right]] \leq \frac{\mathsf{E}[X]}{t}.$$

A useful special case of Markov's inequality is

$$\Pr\left(X \geq \frac{\mathsf{E}[X]}{\delta}\right) \leq \delta.$$

Applying the above special case for $X = \mathsf{E}_{f\sim\Pi}\left[\exp\left(\lambda\, \mathrm{d}_{\mathrm{KL}}\left(\hat{R}_n(f), R(f)\right)\right)\right]$, we have, with probability at least $1 - \delta$ over the training sample,[2]

$$\begin{aligned}\mathsf{E}_{f\sim\Pi}\left[\exp\left(\lambda\, \mathrm{d}_{\mathrm{KL}}\left(\hat{R}_n(f), R(f)\right)\right)\right] &\leq \frac{1}{\delta}\cdot \mathsf{E}_{Z^n}\left[\mathsf{E}_{f\sim\Pi}\left[\exp\left(\lambda\, \mathrm{d}_{\mathrm{KL}}\left(\hat{R}_n(f), R(f)\right)\right)\right]\right] \\ &= \frac{1}{\delta}\cdot \mathsf{E}_{f\sim\Pi}\left[\mathsf{E}_{Z^n}\left[\exp\left(\lambda\, \mathrm{d}_{\mathrm{KL}}\left(\hat{R}_n(f), R(f)\right)\right)\right]\right].\end{aligned}$$

Therefore, from (3), we see that with probability at least $1 - \delta$, *for all* $\hat{\Pi}$ *simultaneously (!)*,

$$\begin{aligned}&\mathsf{E}_{f\sim\hat{\Pi}}\left[\lambda\, \mathrm{d}_{\mathrm{KL}}\left(\hat{R}_n(f), R(f)\right)\right] \\ &\qquad\leq \log \mathsf{E}_{f\sim\Pi}\left[\mathsf{E}_{Z^n}\left[\exp\left(\lambda\, \mathrm{d}_{\mathrm{KL}}\left(\hat{R}_n(f), R(f)\right)\right)\right]\right] + \log\frac{1}{\delta} + \mathrm{D}_{\mathrm{KL}}(\hat{\Pi}\,\|\,\Pi). \qquad (4)\end{aligned}$$

Next, we bound the quantity inside the logarithm on the right-hand side above. It suffices to obtain a bound which holds for any fixed $f \in \mathcal{F}$, and this is what we will do. Since we are using zero-one loss, observe that

$$\mathsf{E}_{Z^n}\left[e^{\lambda\, \mathrm{d}_{\mathrm{KL}}\left(\hat{R}_n(f), R(f)\right)}\right] = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}e^{\lambda\, \mathrm{d}_{\mathrm{KL}}(\frac{k}{n}, p)},$$

where $p = R(f)$. That is, we need only bound the expectation of $\mathrm{d}_{\mathrm{KL}}(\frac{k}{n}, p)$ where k is distributed according to a Binomial distribution with success probability p and n trials.

If we take $\lambda = n$, the above is equal to

$$\begin{aligned}&\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}e^{n\, \mathrm{d}_{\mathrm{KL}}(\frac{k}{n}, p)} \\ &= \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\exp\left(k\log\frac{k/n}{p} + (n-k)\log\frac{(n-k)/n}{1-p}\right),\end{aligned}$$

which simplifies to

$$\sum_{k=0}^{n}\binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k}.$$

[2] Exchanging the expectations to yield the equality is permissible because of Tonelli's theorem, which allows for interchanging integrals in the case of nonnegative functions.

This quantity can readily be upper bounded as follows[3]:

$$\begin{aligned}\sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^{k} \left(\frac{n-k}{n}\right)^{n-k} &\leq \sum_{j=0}^{n} \left(\sum_{k=0}^{n} \binom{n}{k} \left(\frac{j}{n}\right)^{k} \left(\frac{j-k}{n}\right)^{n-k} \right) \\ &= n+1.\end{aligned}$$

Thus, we have that

$$\log \mathsf{E}_{Z^n} \left[e^{\lambda \, \mathrm{d}_{\mathrm{KL}}\left(\hat{R}_n(f), R(f)\right)} \right] \leq \log(n+1).$$

Applying this bound in (4) with $\lambda = n$ and further applying (2) yields the result:

With probability at least $1 - \delta$ over the training sample,

$$\mathrm{d}_{\mathrm{KL}} \left(\mathsf{E}_{f \sim \hat{\Pi}} \left[\hat{R}_n(f) \right], \mathsf{E}_{f \sim \hat{\Pi}}[R(f)] \right) \leq \frac{1}{n} \left(\mathrm{D}_{\mathrm{KL}}(\hat{\Pi} \,\|\, \Pi) + \log(n+1) + \log \frac{1}{\delta} \right).$$

□

Proceeding differently via a route proposed by Olivier Catoni, one also can show the following PAC-Bayesian bound:

For any $\lambda > \frac{1}{2}$, with probability at least $1 - \delta$, for all $\hat{\Pi}$,

$$\mathsf{E}_{f \sim \hat{\Pi}}[R(f)] \leq \frac{1}{1 - \frac{1}{2\lambda}} \left(\mathsf{E}_{f \sim \hat{\Pi}} \left[\hat{R}_n(f) \right] + \frac{\lambda}{n} \left(\mathrm{D}_{\mathrm{KL}}(\hat{\Pi} \,\|\, \Pi) + \log \frac{1}{\delta} \right) \right).$$

This bound is interesting for at least three different reasons.

First, this bound holds for loss functions taking values in $[0, 1]$, not just zero-one loss. A PAC-Bayesian risk guarantee can also be obtained for arbitrary bounded nonnegative loss functions by appropriately scaling the loss down to the scale $[0, 1]$, applying the above bound on the rescaled loss, and then undoing the rescaling; for a loss taking values in $[0, L_{\max}]$, the effect of this "rescaling trick" is to add a multiplier of $L_{\max}$ in front of the term $\frac{\lambda}{n}$.

Second, because this bound holds for all posteriors, similar to before, a sensible choice is to select to posterior $\hat{\Pi}$ that minimizes the right-hand side. It is known that

$$\mathsf{E}_{f \sim \hat{\Pi}}[\hat{R}_n(f)] + \frac{\lambda}{n} \mathrm{D}_{\mathrm{KL}}(\hat{\Pi} \,\|\, \Pi)$$

is minimized by taking

$$\hat{\pi}(f) = \frac{e^{-\frac{n}{\lambda} \hat{R}_n(f)} \pi(f)}{\mathsf{E}_{g \sim \Pi} \left[e^{-\frac{n}{\lambda} \hat{R}_n(g)} \right]}.$$

In online learning, this predictor is called the exponentially weighted average forecaster, or "exponential weights" for short.

Third, in the special case of PAC learning (i.e. classification in the realizable case), it is possible to select to select a posterior distribution $\hat{\Pi}$ for which $\mathsf{E}_{f \sim \hat{\Pi}}[\hat{R}_n(f)] = 0$; indeed, ERM itself, which corresponds to placing a Dirac distribution on $\hat{f}_{\mathrm{ERM}}$, obtains zero empirical risk in this case. In this setting, we can thus set λ to some moderate value like $\lambda = 1$ to obtain the following risk bound with a fast rate of convergence:

[3] This argument was taken from Cover and Thomas (2006). In fact, a considerably better bound of $2\sqrt{n}$ holds for $n \geq 8$ (and the same bound holds with slightly larger constants even for $n \geq 2$).

For any $\lambda > \frac{1}{2}$, with probability at least $1 - \delta$, for all $\hat{\Pi}$ for which $\mathsf{E}_{f \sim \hat{\Pi}}[\hat{R}_n(f)] = 0$,

$$\mathsf{E}_{f \sim \hat{\Pi}}[R(f)] \leq \frac{2}{n} \left(\mathrm{D}_{\mathrm{KL}}(\hat{\Pi} \,\|\, \Pi) + \log \frac{1}{\delta} \right).$$

3 Obtaining a bound for a non-randomized predictor

When using the Gibbs classifier, for every test example we need to draw a hypothesis f from our posterior distribution $\hat{\Pi}$. What if we instead would like the hypothesis we use for prediction to be deterministic, conditional on the training sample? For this, we can use the majority vote predictor, defined as:

$$\textsc{mv}_{\hat{\Pi}}(x) = \operatorname*{arg\,max}_{y \in \{-1,+1\}} \hat{\Pi}\left(\{f \in \mathcal{F} : f(x) = y\}\right),$$

with ties broken arbitrarily.

It is easy to see that the risk of the majority vote classifier with respect to $\hat{\Pi}$ cannot be more than twice the risk of the Gibbs classifier: on any example where the majority vote classifier $\textsc{mv}_{\hat{\Pi}}$ makes a mistake, at least half of the probability mass with respect to the $\hat{\Pi}$ must have also made a mistake. Therefore, we have

$$R(\textsc{mv}_{\hat{\Pi}}) \leq 2\,\mathsf{E}_{f \sim \hat{\Pi}}\left[R(f)\right].$$

This bound can be quite loose in practice. In many situations, the majority vote classifier can actually *outperform* the Gibbs classifier. Precisely this situation occurred in our analysis of AdaBoost.

References

Thomas M Cover and Joy A Thomas. *Elements of information theory.* John Wiley & Sons, second edition, 2006.