CSE522, Winter 2011, Learning Theory	Lecture 14 - 02/15/2011

PAC-Bayes Analysis

Lecturer: Ofer Dekel *Scribe: Krishnamurthy Dvijotham*

1 Recap of PAC-Bayes Theory

PAC-Bayes theory [McA03] was developed by McAllester initially as an attempt to explain Bayesian learning from a learning theory perspective, but the tools developed later proved to be useful in a much more general context. PAC-Bayes theory gives the tightest known generalization bounds for SVMs, with fairly simple proofs. PAC-Bayesian analysis applies directly to algorithms that output *distributions* on the hypothesis class, rather than a single best hypothesis. However, it is possible to de-randomize the PAC-Bayes bound to get bounds for algorithms that output deterministic hypothesis.

2 PAC-Bayes Generalization Bound

We will consider the binary classification task with an input space $\mathcal{X}$ and label set $\mathcal{Y} = \{+1, -1\}$. Let $\mathcal{D}$ be the (unknown) true on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H}$ be a hypothesis class of functions $f : \mathcal{X} \mapsto \mathcal{Y}$. Let $\mathcal{P}$ be the space of probability distributions on $\mathcal{H}$. We consider $0, 1$-valued loss functions $l : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \mapsto \{0, 1\}$.

Definition 1. *Let $Q \in \mathcal{P}$. Define:*

$$\text{Risk of } Q \; l(Q; \mathcal{D}) = E_{(x,y)\sim\mathcal{D}} E_{h\sim Q} \left[l(h; (x, y)) \right]$$

$$\text{Emperical Risk of } Q \; l(Q; D) = \frac{1}{|D|} \sum_{(x,y)\in D} E_{h\sim Q} \left[l(h; (x, y)) \right]$$

For $0, 1$-valued loss functions, $l(Q; D), l(Q; \mathcal{D}) \in [0, 1]$. Thus, they can be interpreted as the parameter of a Bernoulli random variable. Given, $P, Q \in \mathcal{P}$, we measure the distance between them using the KL-divergence:

$$\mathrm{KL}\left(l(Q; \mathcal{D}) \parallel l(P; \mathcal{D})\right) = l(Q; \mathcal{D}) \log \left(\frac{l(Q; \mathcal{D})}{l(P; \mathcal{D})} \right) + (1 - l(Q; \mathcal{D})) \log \left(\frac{1 - l(Q; \mathcal{D})}{1 - l(P; \mathcal{D})} \right)$$

Note that the KL-divergence is jointly convex in both its arguments (this follows from the convexity of the function $x \log(x/y)$ over $0 \le x, y \le 1$). We'll use this fact in the proofs later. We analyze algorithms with the following structure:

1: Choose a **prior distribution** $P \in \mathcal{P}$ *before* seeing any data.
2: Observe data D and choose posterior $Q \in \mathcal{P}$. Q can depend on D, P.
3: Output Q

Note: The distribution Q need not be a Bayesian posterior, it can be **any** distribution. It is allowed to depend on P, D but need not. We will later talk about constructing distribution-dependent priors P where the algorithm is **not allowed** to use P.
Note: We use probability distributions with two different semantics: P encodes our **subjective a-priori belief** about what hypotheses are true and $\mathcal{D}$ describes the randomness in the real-world.

Theorem 2. *(McAllester) $\forall \mathcal{D}, \forall \mathcal{H} \forall P \in \mathcal{P} \forall \delta > 0$, we have with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$: $\forall Q \in \mathcal{P}$ (posterior distribution on $\mathcal{H}$ that depends on S),*

$$\mathrm{KL}\left(l(Q; S) \parallel l(Q : \mathcal{D})\right) \le \frac{\mathrm{KL}\left(Q \parallel P\right) + \log\left(\frac{m+1}{\delta}\right)}{m}$$

Proof. Define

$$Z = \mathop{\mathrm{E}}_{h\sim P}\left[\exp\left(m\mathrm{KL}\left(l(h;S) \parallel l(h;\mathcal{D})\right)\right)\right]$$

We shall prove this theorem in 2 parts:

1 With probability at least $1-\delta$, $\mathrm{KL}\left(l(Q;S) \parallel l(Q;\mathcal{D})\right) \leq \frac{\mathrm{KL}(Q\|P)+\log\left(\frac{\mathrm{E}_S[Z]}{\delta}\right)}{m}$

2 $\mathrm{E}_S\left[Z\right] \leq m+1$

Proof of Part 1

Using Markov's inequality, we have: $\forall a \Pr\left[Z > a\right] \leq \frac{\mathrm{E}_S[Z]}{a}$. Plugging in $a = \frac{\mathrm{E}_S[Z]}{\delta}$, we get

$$\Pr\left[Z > \frac{\mathrm{E}_S\left[Z\right]}{a}\right] \leq \delta$$

Note that the probability is only over sampling of $h \sim P$. Rewriting this, we have $w.p \geq 1-\delta \quad Z \leq \frac{\mathrm{E}_S[Z]}{a}$ which is equivalent to

$$w.p \geq 1-\delta \quad \log(Z) \leq \log\left(\frac{\mathrm{E}_S\left[Z\right]}{a}\right)$$

Thus, $w.p \geq 1-\delta$, we have:

$$\begin{aligned}
\log(Z) &= \log\left(\mathop{\mathrm{E}}_{h\sim P}\left[\exp\left(m\mathrm{KL}\left(l(h;S) \parallel l(h;\mathcal{D})\right)\right)\right]\right) \\
&= \log\left(\mathop{\mathrm{E}}_{h\sim Q}\left[\frac{P(h)}{Q(h)}\exp\left(m\mathrm{KL}\left(l(h;S) \parallel l(h;\mathcal{D})\right)\right)\right]\right) \quad \text{(Change of Measure)} \\
&\geq \mathop{\mathrm{E}}_{h\sim Q}\left[\log\left(\frac{P(h)}{Q(h)}\right) + m\mathrm{KL}\left(l(h;S) \parallel l(h;\mathcal{D})\right)\right] \quad \text{(Concavity of log)} \\
&= -\mathrm{KL}\left(Q \parallel P\right) + m\mathop{\mathrm{E}}_{h\sim Q}\left[\mathrm{KL}\left(l(h;S) \parallel l(h;\mathcal{D})\right)\right] \quad \text{(Definition of KL)} \\
&\geq -\mathrm{KL}\left(Q \parallel P\right) + m\mathrm{KL}\left(l(Q;S) \parallel l(Q;\mathcal{D})\right) \quad \text{(Convexity of KL)}
\end{aligned}$$

Rearranging terms, we get $w.p \geq 1-\delta$,

$$\mathrm{KL}\left(l(Q;S) \parallel l(Q;\mathcal{D})\right) \leq \frac{\mathrm{KL}\left(Q \parallel P\right) + \log(Z)}{m}$$

Proof of Part 2

Let $l(h;S) = a_h, l(h;\mathcal{D}) = b_h$.

$$\begin{aligned}
\mathop{\mathrm{E}}_{S}\left[Z\right] &= \mathop{\mathrm{E}}_{S}\left[\mathop{\mathrm{E}}_{h\sim P}\left[\exp\left(m(a_h\log(a_h/b_h) + (1-a_h)\log((1-a_h)/(1-b_h))\right)\right)\right] \\
&= \mathop{\mathrm{E}}_{S}\left[\mathop{\mathrm{E}}_{h\sim P}\left[\left(\frac{a_h}{b_h}\right)^{ma_h}\left(\frac{1-a_h}{1-b_h}\right)^{m(1-a_h)}\right]\right]
\end{aligned}$$

a_h can take $m+1$ values: $\left\{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\right\}$ and has a binomial distibution with parameter b_h. Thus,

$$\begin{aligned}
&\mathop{\mathrm{E}}_{S}\left[\left(\frac{a_h}{b_h}\right)^{ma_h}\left(\frac{1-a_h}{1-b_h}\right)^{m(1-a_h)}\right] \\
&= \sum_{k=0}^{m}\binom{m}{k}b_h^k(1-b_h)^{m-k}\left(\frac{k/m}{b_h}\right)^k\left(\frac{1-k/m}{(1-b_h)}\right)^{m-k} \\
&= \sum_{k=0}^{m}\binom{m}{k}\left(\frac{k}{m}\right)^k\left(1-\frac{k}{m}\right)^{m-k}
\end{aligned}$$

We know that $\binom{m}{k}\left(\frac{k}{m}\right)^k\left(1-\frac{k}{m}\right)^{m-k}$ is the probability that a binomial random variable with parameter $\frac{k}{m}, k, m$ is equal to k, and hence is smaller than 1. Thus, the sum over k is smaller than $m+1$. Thus $\mathrm{E}_S[Z] \leq m+1$. One can actually show a tighter bound: $\mathrm{E}_S[Z] \in [\sqrt{m}, \sqrt{2m}]$ using a more careful analysis. ☐

We now prove some corollaries to relate the KL-divergence bound to the kinds of additive bounds we have seen before.

Lemma 3. *If* $a, b \in [0,1]$ *and* $\mathrm{KL}(a \parallel b) \leq x$, *then*

$$b \leq a + \sqrt{\frac{x}{2}}, b \leq a + 2x + \sqrt{2ax}$$

Proof. **Proof of First Inequality**
Consider the function $f(a) = \mathrm{KL}(a \parallel b) - 2(a-b)^2$.

$$f'(a) = \log\left(\frac{a}{1-a}\right) - \log\left(\frac{b}{1-b}\right) - 4(a-b)$$

$$f''(a) = \frac{1}{a(1-a)} - 4$$

$a(1-a)$ achieves its maximum of $1/4$ at $a = 1/2$ and hence $f''(a) \geq 0 \, \forall a \in [0,1]$. $f'(a) = 0$ at $a = b$ and $f'' \geq 0$, therefor, b is the minimum of $f(a)$ and $f(b) = 0$. Hence $f(a) \geq 0 \forall a \in [0,1]$. Hence $x \geq \mathrm{KL}(a \parallel b) \geq 2(a-b)^2$. $G(b) = 2b^2 - 4ab + 2a^2 - x \leq 0$. G is a convex quadratic in b and hence if $G(b) \leq 0$, then b must lie between the roots of G and hence be smaller than the larger root of G. Thus,

$$b \leq a + \sqrt{a^2 - \frac{2a^2 - x}{2}} = a + \sqrt{\frac{x}{2}}$$

Proof of Second Inequality
If $a \geq b$ then the inequality is obviously true. Suppose that $b > a$. Then consider the function $f(a) = \mathrm{KL}(a \parallel b) - \frac{(a-b)^2}{2b}$.

$$f'(a) = \log\left(\frac{a}{1-a}\right) - \log\left(\frac{b}{1-b}\right) - \frac{a-b}{b}$$

$$f''(a) = \frac{1}{a} + \frac{1}{1-a} - \frac{1}{b}$$

Since $b > a, 1/a > 1/b$ and hence $f''(a) > 0$. $f'(b) = 0, f(b) = 0$ and hence $f(a) > 0 \forall a \in (a, b)$. Thus, if $b > a$, $x \geq \mathrm{KL}(a \parallel b) \geq \frac{(a-b)^2}{2b}$. Thus, we get

$$G(b) = b^2 - (2a + 2x)b + a^2 \leq 0$$

Thus, as before, b is smaller than the larger root of G, ie,

$$a + x + \sqrt{(a+x)^2 - a^2} = a + x + \sqrt{x^2 + 2ax} \leq a + x + x + \sqrt{2ax} = a + 2x + \sqrt{2ax}$$

where we used the sub-additivity of the square root function.

☐

Corollary 4. $\forall \mathcal{D}, \forall \mathcal{H} \forall P \in \mathcal{P} \forall \delta > 0$, *we have the following bounds with probability at least* $1-\delta$ *over* $S \sim \mathcal{D}^m$:

$$\forall Q \in \mathcal{P} \quad l(Q; \mathcal{D}) \leq l(Q; S) + \sqrt{\frac{\mathrm{KL}(Q \parallel P) + \log\left(\frac{m+1}{\delta}\right)}{m}}$$

$$\forall Q \in \mathcal{P} \quad l(Q;\mathcal{D}) \leq l(Q;S) + 2\left(\frac{\mathrm{KL}\left(Q \parallel P\right) + \log\left(\frac{m+1}{\delta}\right)}{m}\right) + \sqrt{2l(Q;S)\left(\frac{\mathrm{KL}\left(Q \parallel P\right) + \log\left(\frac{m+1}{\delta}\right)}{m}\right)}$$

Proof. These follow directly by plugging the KL bounds from lemma 3 into the PAC Bayes bound from theorem 2. □

References

[McA03] D. McAllester. Simplified PAC-Bayesian Margin Bounds. In *Learning theory and Kernel machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003: proceedings*, page 203. Springer Verlag, 2003.