PAC-Bayes

Peng Lingwei

September 10, 2019

Contents

31 PAC-Bayes

We assign a prior distribution P on $\mathcal{H}$. The $PAC-Bayes$ returns a posterior probability Q over $\mathcal{H}$.

1. $l(Q,z) = \mathbb{E}_{h\sim Q}\left[l(h,z)\right]$;
2. $L_S(Q) = \mathbb{E}_{h\sim Q}\left[L_S(h)\right]$
3. $L_\mathcal{D}(Q) = \mathbb{E}_{h\sim Q}\left[L_\mathcal{D}(h)\right]$

Theorem 1.

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{\forall Q, L_\mathcal{D}(Q) \le L_S(Q) + \sqrt{\frac{D_{KL}(Q\|P)+\ln m/\delta}{2(m-1)}}\right\} \ge 1-\delta$$

Proof. Let $\Delta(h) = L_\mathcal{D}(h) - L_S(h)$. And construct function

$$\begin{aligned}
f(S) =& \sup_Q \left(2(m-1)\mathbb{E}_{h\sim Q}(\Delta(h))^2 - D_{KL}(Q\|P)\right) \\
=& \sup_Q \left(\mathbb{E}_{h\sim Q}\left[\ln\left(e^{2(m-1)\Delta^2(h)}P(h)/Q(h)\right)\right]\right) \\
\le& \sup_Q \left(\ln \mathbb{E}_{h\sim Q}\left[e^{2(m-1)\Delta^2(h)}P(h)/Q(h)\right]\right) \\
=& \ln \mathbb{E}_{h\sim P}\left[e^{2(m-1)\Delta^2(h)}\right] \\
\mathbb{E}_{S\sim\mathcal{D}^m}\left[e^{f(S)}\right] \le& \mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{h\sim P}\left[e^{2(m-1)\Delta^2(h)}\right] \\
=& \mathbb{E}_{h\sim P}\mathbb{E}_{S\sim\mathcal{D}^m}\left[e^{2(m-1)\Delta^2(h)}\right]
\end{aligned}$$

If $\mathbb{E}_{S\sim\mathcal{D}^m}\left[e^{2(m-1)\Delta^2(h)}\right] \le m$,

$$E_{S\sim\mathcal{D}^m}\left[e^{f(S)}\right] \le m \Rightarrow \mathbb{P}_{S\sim\mathcal{D}^m}\left[f(S)\ge\epsilon\right] \le \frac{m}{e^\epsilon}$$

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{2(m-1)\mathbb{E}_{h\sim Q}(\Delta(h))^2 - D_{KL}(Q\|P) \le \ln(m/\delta)\right\} \ge 1-\delta$$

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{(\mathbb{E}_{h\sim Q}\Delta(h))^2 \le \mathbb{E}_{h\sim Q}(\Delta(h))^2 \le \frac{\ln(m/\delta)+D_{KL}(Q\|P)}{2(m-1)}\right\} \ge 1-\delta$$

□

(PAC-Bayes rules)

$$\min_Q \left(L_S(Q) + \sqrt{\frac{D_{KL}(Q\|P)+\ln(m/\delta)}{2(m-1)}}\right)$$

Lemma 1.

$$\mathbb{P}\left[X\ge\epsilon\right] \le e^{-2m\epsilon^2} \Rightarrow \mathbb{E}\left[e^{2(m-1)X^2}\right] \le m$$

Proof. I have no idea at all. □

I have doubt on this theorem.

31.1 General PAC-Bayesian Theorem

Theorem 2. *Let* $\Delta : [0,1] \times [0,1] \to \mathbb{R}$,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{\forall Q \in \mathcal{H}, \Delta(L_S(Q), L_{\mathcal{D}}(Q)) \leq \frac{1}{m}\left[KL(Q\|P) + \ln\frac{\mathcal{I}_\Delta(m)}{\delta}\right]\right\} \leq \delta$$

where

$$\mathcal{I}_\Delta(m) = \sup_{r\in[0,1]}\left[\sum_{k=0}^{m} \mathbb{C}_m^k r^k (1-r)^{m-k} e^{m\Delta(\frac{k}{m}, r)}\right]$$

Proof.

- $\forall \phi : \mathcal{H} \to \mathbb{R}, \mathbb{E}_{h\sim Q}\phi(h) \leq KL(Q\|P) + \ln\left(\mathbb{E}_{h\sim P}e^{\phi(h)}\right)$
- $\mathbb{P}_{S\sim\mathcal{D}^m}\left\{L_S(h) = \frac{k}{m}\right\} = \mathbb{C}_m^k (L_{\mathcal{D}}(h))^k (1 - L_{\mathcal{D}}(h))^{m-k} = Bin(k; m, L_{\mathcal{D}}(h))$

$$\begin{aligned}
&m\Delta(\mathbb{E}_S(Q), \mathbb{E}_{\mathcal{D}}(Q)) \leq m\mathbb{E}_{h\sim Q}\Delta(L_S(h), L_{\mathcal{D}}(h)) \leq KL(Q\|P) + \ln\mathbb{E}_{h\sim P}e^{m\Delta(L_S(h), L_{\mathcal{D}}(h))} \\
\leq_{1-\delta}&KL(Q\|P) + \ln\frac{1}{\delta}\mathbb{E}_{S'\sim\mathcal{D}^m}\mathbb{E}_{h\sim P}e^{m\cdot\Delta(L_S(h), L_{\mathcal{D}}(h))} \\
\leq&KL(Q\|P) + \ln\frac{1}{\delta}\mathbb{E}_{h\sim P}\mathbb{E}_{S'\sim\mathcal{D}^m}e^{m\cdot\Delta(L_S(h), L_{\mathcal{D}}(h))} \\
=&KL(Q\|P) + \ln\frac{1}{\delta}\mathbb{E}_{h\sim P}\sum_{k=0}^{m} Bin(k; m, L_{\mathcal{D}}(h))e^{m\cdot\Delta(\frac{k}{m}, L_{\mathcal{D}}(h))} \\
\leq&KL(Q\|P) + \ln\frac{1}{\delta}\sup_{r\in[0,1]}\sum_{k=0}^{m} Bin(k; m, r)e^{m\cdot\Delta(\frac{k}{m}, r)} \\
=&KL(Q\|P) + \ln\frac{1}{\delta}\mathcal{I}_\Delta(m).
\end{aligned}$$

□

Corollary 1. ***(Langford and Seeger).*** $\Delta(L_S(Q), L_{\mathcal{D}}(Q)) = kl(L_S(Q), L_{\mathcal{D}}(Q))$, *where* $kl(q,p) = q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}$.

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{kl(L_S(Q), L_{\mathcal{D}}(Q)) \leq \frac{1}{m}\left[KL(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right]\right\} \geq 1 - \delta$$

Proof.

$$\begin{aligned}
&\mathbb{E}_{S'\sim\mathcal{D}^m}\mathbb{E}_{h\sim P}e^{m\cdot kl(L_S(h), L_{\mathcal{D}}(h))} \\
=&\mathbb{E}_{h\sim P}\mathbb{E}_{S'\sim\mathcal{D}^m}\left(\frac{L_S(h)}{L_{\mathcal{D}}(h)}\right)^{mL_S(h)}\left(\frac{1-L_S(h)}{1-L_{\mathcal{D}}(h)}\right)^{m(1-L_S(h))} \\
=&\mathbb{E}_{h\sim P}\sum_{k=0}^{m}\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_S(h) = \frac{k}{m}\right)\left(\frac{\frac{k}{m}}{L_{\mathcal{D}}(h)}\right)^k\left(\frac{1-\frac{k}{m}}{1-L_{\mathcal{D}}(h)}\right)^{m-k} \\
=&\sum_{k=0}^{m}\mathbb{C}_m^k\left(\frac{k}{m}\right)^k\left(1-\frac{k}{m}\right)^{m-k} \leq 2\sqrt{m}
\end{aligned}$$

□

Corollary 2. ***(Cartoni).*** $\Delta(L_S(Q), L_{\mathcal{D}}(Q)) = \mathcal{F}(L_{\mathcal{D}}(Q)) - CL_S(Q)$.

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{\forall Q \text{ on } \mathcal{H}, L_{\mathcal{D}}(Q) \leq \frac{1}{1-e^{-C}}\left\{1-\exp\left[-\left(CL_S(Q)+\frac{1}{m}\left[KL(Q\|P)+\ln\frac{1}{\delta}\right]\right)\right]\right\}\right\} \geq 1-\delta$$

Proof.

$$\begin{aligned}
&\mathbb{E}_{h\sim P}\mathbb{E}_{S'\sim\mathcal{D}^m} e^{m(\mathcal{F}(L_{\mathcal{D}}(h))-CL_{S'}(h))} \\
=&\mathbb{E}_{h\sim P}\sum_{k=0}^{m}\mathbb{P}_{S'\sim\mathcal{D}^m}\left(L_{S'}(h)=\frac{k}{m}\right)e^{m(\mathcal{F}(L_{\mathcal{D}}(h))-CL_{S'}(h))} \\
=&\mathbb{E}_{h\sim P}e^{m(\mathcal{F}(L_{\mathcal{D}}(h)))}\sum_{k=0}^{m}\mathbb{C}_m^k(L_{\mathcal{D}})^k(1-L_{\mathcal{D}}(h))^{m-k}e^{-Ck} \\
=&\mathbb{E}_{h\sim P}e^{m\mathcal{F}(L_{\mathcal{D}}(h))}\left(L_{\mathcal{D}}(h)e^{-C}+1-L_{\mathcal{D}}(h)\right)^m
\end{aligned}$$

We choose $\mathcal{F}(R)$ satisfies

$$e^{\mathcal{F}(R)}(Re^{-C}+1-R)=1,$$

then, with probability larger than $1-\delta$, we have

$$\mathcal{F}(L_{\mathcal{D}}(Q)) - CL_S(Q) \leq \frac{1}{m}\left(KL(Q\|P)+\ln\frac{1}{\delta}\right),$$

□

The preceeding corollary gives us a new cost function:

$$mCL_S(Q) + KL(Q\|P)$$

31.2 Applications

Definition 1. ***(Majority vote and majority vote risk).***

$$B_Q(x) = sign\left(\mathbb{E}_{h\sim Q}h(x)\right)$$

$$L_{\mathcal{D}}(B_Q) = \mathbb{P}_{(x,y)\sim\mathcal{D}}\left(B_Q(x)\neq y\right) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_{h\sim Q}\left[yh(x)\leq 0\right]$$

Corollary 3.

$$L_{\mathcal{D}}(B_Q) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(1-yh(x)\geq 1) \leq \mathbb{E}_{(x,y)\sim D}(1-yB_Q(x)) = 2L_{\mathcal{D}}(Q)$$

The following is an example of linear classifiers.

1. $\phi(x) = (\phi_1(x),\ldots,\phi_N(x))$, or implicitly given by $k(x,x') = \phi(x)\cdot\phi(x')$;
2. $h_v(x) = sign(\langle v, \phi(x)\rangle) \in \mathcal{H}$;
3. $Q_w(v) = \left(\frac{1}{\sqrt{2\pi}}\right)^N \exp\left(-\frac{1}{2}\|v-w\|^2\right)$
4. $B_{Q_w}(x) = sign(\mathbb{E}_{v\sim Q_w} sign(\langle v,\phi(x)\rangle)) = sign(\langle w,\phi(x)\rangle) = h_w(x)$
5. The prior P_{w_p} is also an isotorpic Gaussian centered on w_p. Consequently:

$$KL(Q_w\|P_{w_p}) = \frac{1}{2}\|w-w_p\|^2$$

6. Gibbs's risk (0–1 risk):

$$L_{(x,y)}(Q_w) = \int Q_w(v) 1_{[yv^T\phi(x)<0]} dv = \Phi\left(\frac{yw^T\phi(x)}{\|\phi(x)\|}\right)$$

where

$$\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^\infty \exp\left(-\frac{1}{2}x^2\right) dx.$$

7. The cost function is

$$CmL_S(Q_w) + KL(Q_w \| P_{w_p}) = C\sum_{i=1}^m \Phi\left(\frac{y_i w^T \phi(x_i)}{\|\phi(x_i)\|}\right) + \frac{1}{2}\|w - w_p\|^2$$

8. If $w_p = 0$ (absence of prior knowledge), we get the cost function alike

$$C\sum_{i=1}^m \max\big(0, 1 - y_i w^T \phi(x_i)\big) + \frac{1}{2}\|w\|^2,$$

which is SVM minimizes.