Generalisation Bounds (4): PAC Bayesian Bounds

Qinfeng (Javen) Shi

The Australian Centre for Visual Technologies,
The University of Adelaide, Australia

31 Aug. 2012

Course Outline

Generalisation Bounds:

1. Basics
2. VC dimensions and bounds
3. Rademacher complexity and bounds
4. PAC Bayesian Bounds (Today)
5. . . .

Recap: Risk

Given $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ sampled from a unknown but fixed distribution $P(x, y)$, the goal is to learn a hypothesis function $g : \mathcal{X} \to \mathcal{Y}$, for now assume $\mathcal{Y} = \{-1, 1\}$.

A typical $g(x) = \text{sign}(\langle \phi(x), w \rangle)$, where $\text{sign}(z) = 1$ if $z > 0$, $\text{sign}(z) = -1$ otherwise.
Generalisation error

$$R(g) = \mathbb{E}_{(x,y)\sim P}[\mathbf{1}_{g(x)\neq y}]$$

Empirical risk for zero-one loss (*i.e.* training error)

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{g(x_i)\neq y_i}.$$

Recap: VC bound

For VC dimension h, we know that $\forall n \geq h$, the growth function (*i.e.* #outputs) $S_{\mathcal{G}}(n) \leq (\frac{en}{h})^h$. Thus

Theorem (VC bound)

For any $\delta \in (0,1)$, with probability at least $1-\delta$, $\forall g \in \mathcal{G}$

$$R(g) \leq R_n(g) + 2\sqrt{2\frac{h \log \frac{2en}{h} + \log(\frac{2}{\delta})}{n}}.$$

Problems:

- data dependency come through training error
- very loose

Recap: Rademacher bound

Theorem (Rademacher)

Fix $\delta \in (0,1)$ and let $\mathcal{G}$ be a set of functions mapping from Z to $[a, a+1]$. Let $S = \{z_i\}_{i=1}^n$ be drawn i.i.d. from P. Then with probability at least $1-\delta$, $\forall g \in \mathcal{G}$,

$$\mathbb{E}_P[g(z)] \leq \hat{\mathbb{E}}[g(z)] + \mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{\ln(2/\delta)}{2n}}$$

$$\leq \hat{\mathbb{E}}[g(z)] + \hat{\mathcal{R}}_n(\mathcal{G}, S) + 3\sqrt{\frac{\ln(2/\delta)}{2n}},$$

where $\hat{\mathbb{E}}[g(z)] = \frac{1}{n}\sum_{i=1}^n g(z_i)$.

Recap: Rademacher Margin bound

Theorem (Margin)

Fix $\gamma > 0, \delta \in (0,1), \forall g \in \mathcal{G}$, let $\{(x_i, y_i)\}_{i=1}^n$ be drawn i.i.d. from $P(X, Y)$ and let $\xi_i = (\gamma - y_i g(x_i))_+$. Then with probability at least $1 - \delta$ over sample of size n, we have

$$P(y \neq g(x)) \leq \frac{1}{n\gamma}\sum_{i=1}^{n}\xi_i + \frac{4}{n\gamma}\sqrt{tr(\mathbf{K})} + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Problems:

- data dependency only come through training error and margin
- tighter than VC bound, but still loose

PAC-bayes bounds

Assume $\hat{Q}$ is the prior distribution over classifier $g \in \mathcal{G}$ and Q is any (could be the posterior) distribution over the classifier.

PAC-bayes bounds on:

- Gibbs classifier: $G_Q(x) = g(x), g \sim Q$
  risk: $R(G_Q) = \mathbb{E}_{(x,y)\sim P, g\sim Q}[\mathbf{1}_{g(x)\neq y}]$
  (McAllester98,99,01,Germain *et al.* 09)
- Average classifier: $B_Q(x) = \mathrm{sgn}[\mathbb{E}_{g\sim Q}\, g(x)]$
  risk: $R(B_Q) = \mathbb{E}_{(x,y)\sim P}[\mathbf{1}_{\mathbb{E}_Q[g(x)]\neq y}]$
  (Langford01, Zhu&Xing09)
- Single classifier: $g \in \mathcal{G}$.
  risk: $R(g) = \mathbb{E}_{(x,y)\sim P}[\mathbf{1}_{g(x)\neq y}]$
  (Langford01,McAllester07)

Relation between Gibbs, Average and Single classifier Risks

$R(G_Q)$ (original PAC-Bayes bounds)

$\Downarrow \because R(B_Q)/2 \leq R(G_Q)$

$R(B_Q)$ (PAC-Bayes margin bound for boostings)

$\Downarrow$ via picking a "good" prior $\hat{Q}$ and posterior Q over g

$R(g)$ (PAC-Bayes margin bound for SVMs)

PAC-Bayesian bound on Gibbs Classifier (1)

Theorem (Gibbs (McAllester99,03))

For any distribution P, for any set $\mathcal{G}$ of the classifiers, any prior distribution $\hat{Q}$ of $\mathcal{G}$, any $\delta \in (0, 1]$, we have

$$\Pr_{S \sim P^n} \left\{ \forall Q \text{ on } \mathcal{G} : R(G_Q) \leq R_S(G_Q) + \sqrt{\frac{1}{2n-1} \left[KL(Q || \hat{Q}) + \ln \frac{1}{\delta} + \ln n + 2 \right]} \right\} \geq 1 - \delta.$$

where $KL(Q||\hat{Q}) = \mathbb{E}_{g \sim Q} \ln \frac{Q(g)}{\hat{Q}(g)}$ is the KL divergence.

PAC-Bayesian bound on Gibbs Classifier (2)

Theorem (Gibbs (Seeger02 and Langford05))

For any distribution P, for any set $\mathcal{G}$ of the classifiers, any prior distribution $\hat{Q}$ of $\mathcal{G}$, any $\delta \in (0, 1]$, we have

$$\Pr_{S \sim P^n} \left\{ \forall Q \text{ on } \mathcal{G} : kl(R_S(G_Q), R(G_Q)) \leq \frac{1}{n}\left[KL(Q \| \hat{Q}) + \ln \frac{n+1}{\delta} \right] \right\} \geq 1 - \delta.$$

where

$$kl(q, p) = q \ln \frac{q}{p} + (1 - q) \ln \frac{1 - q}{1 - p}.$$

PAC-Bayesian bound on Gibbs Classifier (3)

Since

$$kl(q,p) \geq (q-p)^2,$$

The theorem Gibbs (Seeger02 and Langford05) yields

$$\Pr_{S\sim P^n}\left\{\forall Q \text{ on } \mathcal{G} : R(G_Q)) \leq R_S(G_Q) + \sqrt{\frac{1}{n}\left[\mathrm{KL}(Q||\hat{Q}) + \ln\frac{n+1}{\delta}\right]}\right\} \geq 1-\delta.$$

PAC-Bayesian bound on Average Classifier

Theorem (Average (Langford *et al.* 01))

For any distribution P, for any set $\mathcal{G}$ of the classifiers, any prior distribution $\hat{Q}$ of $\mathcal{G}$, any $\delta \in (0, 1]$, and any $\gamma > 0$, we have

$$\Pr_{S \sim P^n} \left\{ \forall Q \text{ on } \mathcal{G} : R(B_Q) \leq \Pr_{(\mathbf{x},y) \sim S} \left(y \, \mathbb{E}_{g \sim Q}[g(x)] \leq \gamma \right) + O\left(\sqrt{\frac{\gamma^{-2} KL(Q || \hat{Q}) \ln n + \ln n + \ln \frac{1}{\delta}}{n}} \right) \right\} \geq 1 - \delta.$$

Zhu& Xing09 extended to structured output case.

PAC-Bayesian bound on Single Classifier

Assume $g(x) = \langle w, \phi(x) \rangle$ and rewrite $R(g)$ as $R(w)$.

Theorem (Single (McAllester07))

For any distribution P, for any set $\mathcal{G}$ of the classifiers, any prior distribution $\hat{Q}$ over w, any $\delta \in (0, 1]$, and any $\gamma > 0$, we have

$$\Pr_{S \sim P^n} \Bigg\{ \forall w \sim \mathcal{W} : R(w) \leq \Pr_{(\mathbf{x},y) \sim S} \Big(y \langle w, \phi(x) \rangle \leq \gamma \Big) + O\left(\sqrt{\frac{\gamma^{-2} \frac{\|w\|^2}{2} \ln(n|\mathcal{Y}|) + \ln n + \ln \frac{1}{\delta}}{n}} \right) \Bigg\} \geq 1 - \delta.$$

Proofs

Germain *et al.* icml09 (Thm 2.1) significantly simplified the proof of PAC-Bayes bounds. Here $R_S(g) = \frac{1}{n}\sum_{(x,y)\in S} \mathbf{1}_{g(x)\neq y}$.

Theorem (Simplified PAC-Bayes (Germain09))

For any distribution P, for any set $\mathcal{G}$ of the classifiers, any prior distribution $\hat{Q}$ of $\mathcal{G}$, any $\delta \in (0,1]$, and any convex function $\mu : [0,1] \times [0,1] \to \mathbb{R}$, we have

$$\Pr_{S\sim P^n}\left\{\forall Q \text{ on } \mathcal{G} : \mu(R_S(G_Q), R(G_Q)) \leq \frac{1}{n}\left[KL(Q||\hat{Q}) + \ln\left(\frac{1}{\delta}\mathbb{E}_{S\sim P^n}\mathbb{E}_{g\sim\hat{Q}}\, e^{n\mu(R_S(g),R(g))}\right)\right]\right\} \geq 1-\delta,$$

where $KL(Q||\hat{Q}) = \mathbb{E}_{g\sim Q} \ln\frac{Q(g)}{\hat{Q}(g)}$ *is the KL divergence.*

Proof of Gibbs (Seeger02 and Langford05)

Let $\mu(q, p) = \mathrm{kl}(q, p)$, where

$$kl(q, p) = q \ln \frac{q}{p} + (1 - q) \ln \frac{1 - q}{1 - p}.$$

The fact that

$$\mathbb{E}_{S \sim P^n} \mathbb{E}_{g \sim \hat{Q}} e^{n\mathrm{kl}(R_S(g), R(g))} \leq n + 1.$$

The Simplified PAC-Bayes theorem yields PAC-bayes bound on Gibbs Classifier (Seeger02 and Langford05).

Proof of Gibbs (McAllester99,03)

Let $\mu(q, p) = 2(q - p)^2$, the theorem will yield the PAC-Bayes bound of McAllester99,03.

Proof of Single (McAllester07)

It's essentially how to get a bound on Single Classifier, from a existing bound on Average Classifier.

By choosing the weight prior $\hat{Q}(\mathbf{w}) = \frac{1}{Z}\exp(-\frac{\|\mathbf{w}\|^2}{2})$ and the posterior $Q(\mathbf{w}') = \frac{1}{Z}\exp(-\frac{\|\mathbf{w}'-\mathbf{w}\|^2}{2})$, one can show $\mathbb{E}_{(x,y)\sim P}[\mathbf{1}_{y\langle w,\phi(x)\rangle\leq 0}] = \mathbb{E}_{(x,y)\sim P}[\mathbf{1}_{\mathbb{E}_Q\, y\, \mathbb{E}_{g\sim Q}[g(x)]\leq 0}]$ by symmetry argument proposed in Langford *et al.* 01 and McAllester07. The fact that $\mathrm{KL}(Q||\hat{Q}) = \frac{\|\mathbf{w}\|^2}{2}$ yields the theorem of Single (McAllester07).

Proof of the Simplified PAC-Bayes thm (1)

To prove

$$\Pr_{S\sim P^n}\left\{\forall Q : \mu(R_S(G_Q), R(G_Q)) \leq \frac{1}{n}\left[\mathrm{KL}(Q||\hat{Q}) + \ln\left(\frac{1}{\delta}\,\mathbb{E}_{S\sim P^n}\,\mathbb{E}_{g\sim\hat{Q}}\, e^{n\mu(R_S(g),R(g))}\right)\right]\right\} \geq 1-\delta,$$

Realise that $\mathbb{E}_{g\sim\hat{Q}}\, e^{n\mu(R_S(g),R(g))}$ is a random variable (due to randomness of S). Let $z = \mathbb{E}_{g\sim\hat{Q}}\, e^{n\mu(R_S(g),R(g))}$. Obviously z can take only non-negative values.

Proof of the Simplified PAC-Bayes thm (2)

Markov inequality states that for any $a > 0$, $\Pr(z \geq a) \leq \frac{\mathbb{E}[z]}{a}$. This is because

$$\mathbb{E}[z] = \int_0^{\infty} zp(z)dz = \int_0^{a} zp(z)dz + \int_a^{\infty} zp(z)dz$$
$$\geq 0 + \int_a^{\infty} zp(z)dz \geq a \int_a^{\infty} p(z)dz = a\Pr(z \geq a)$$

Let $a = \frac{\mathbb{E}[z]}{\delta}$, we have $\Pr(z \geq \frac{\mathbb{E}[z]}{\delta}) \leq \delta$. Thus

$$\Pr(z < \frac{\mathbb{E}[z]}{\delta}) \leq 1 - \delta \tag{1}$$

Proof of the Simplified PAC-Bayes thm (3)

Recall $z = \mathbb{E}_{g \sim \hat{Q}} e^{n\mu(R_S(g), R(g))}$, eq(1) is

$$\Pr_{S \sim P^n}\left(\mathbb{E}_{g \sim \hat{Q}} e^{n\mu(R_S(g), R(g))} \leq \frac{\mathbb{E}_{S \sim P^n}[\mathbb{E}_{g \sim \hat{Q}} e^{n\mu(R_S(g), R(g))}]}{\delta} \right) \geq 1 - \delta$$

Taking log on both sides in Pr yields

$$\Pr_{S \sim P^n}\left(\ln\left[\mathbb{E}_{g \sim \hat{Q}} e^{n\mu(R_S(g), R(g))} \right] \leq \ln\left[\frac{\mathbb{E}_{S \sim P^n}[\mathbb{E}_{g \sim \hat{Q}} e^{n\mu(R_S(g), R(g))}]}{\delta} \right] \right) \geq 1 - \delta$$

Proof of the Simplified PAC-Bayes thm (4)

Since $\mathbb{E}_{g \sim \hat{Q}} f(g) = \mathbb{E}_{g \sim Q} \frac{\hat{Q}(g)}{Q(g)} f(g)$ (same trick in importance sampling)

$$\Pr_{S \sim P^n} \left(\forall Q : \ln \left[\mathbb{E}_{g \sim Q} \frac{\hat{Q}(g)}{Q(g)} e^{n\mu(R_S(g), R(g))} \right] \leq \ln \left[\frac{\mathbb{E}_{S \sim P^n} [\mathbb{E}_{g \sim \hat{Q}} e^{n\mu(R_S(g), R(g))}]}{\delta} \right] \right)$$

$$\geq 1 - \delta$$

$$\begin{aligned}
&\ln \left[\mathbb{E}_{g \sim Q} \frac{\hat{Q}(g)}{Q(g)} e^{n\mu(R_S(g), R(g))} \right] \\
&\geq \mathbb{E}_{g \sim Q} \ln \left[\frac{\hat{Q}(g)}{Q(g)} e^{n\mu(R_S(g), R(g))} \right] \quad \text{(concavity of log)} \\
&\geq -\mathbb{E}_{g \sim Q} \ln\left(\frac{Q(g)}{\hat{Q}(g)}\right) + \mathbb{E}_{g \sim Q}[n\mu(R_S(g), R(g))] \\
&\geq -\mathrm{KL}(Q || \hat{Q}) + n\mu(\mathbb{E}_{g \sim Q} R_S(g), \mathbb{E}_{g \sim Q} R(g)) \quad \text{(convexity of } \mu\text{)} \\
&\geq -\mathrm{KL}(Q || \hat{Q}) + n\mu(R_S(G_Q), R(G_Q))
\end{aligned}$$

Proof of the Simplified PAC-Bayes thm (5)

$\forall Q$, with probability at least $1-\delta$, below holds

$$-\mathrm{KL}(Q||\hat{Q}) + n\mu(R_S(G_Q), R(G_Q)) \leq \ln(\frac{1}{\delta}\mathbb{E}_{S\sim P^n}\mathbb{E}_{g\sim\hat{Q}}\, e^{n\mu(R_S(g),R(g))})$$

Thus

$$\mu(R_S(G_Q), R(G_Q)) \leq \frac{1}{n}\Big[\mathrm{KL}(Q||\hat{Q}) + \ln(\frac{1}{\delta}\mathbb{E}_{S\sim P^n}\mathbb{E}_{g\sim\hat{Q}}\, e^{n\mu(R_S(g),R(g))})\Big],$$

which is

$$\Pr_{S\sim P^n}\Big\{\forall Q : \mu(R_S(G_Q), R(G_Q)) \leq \frac{1}{n}\Big[\mathrm{KL}(Q||\hat{Q}) + \ln(\frac{1}{\delta}\mathbb{E}_{S\sim P^n}\mathbb{E}_{g\sim\hat{Q}}\, e^{n\mu(R_S(g),R(g))})\Big]\Big\} \geq 1-\delta,$$

Related concepts

- Regret bounds for online learning (will be covered in the next talk)