Generalisation Bounds (1): Basics

Qinfeng (Javen) Shi

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

13 April 2012

Course Outline

Generalisation Bounds:

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

History

- Pionneered by Vapnik and Chervonenkis (1968, 1971), Sauer (1972), Shelah (1972) as Vapnik-Chevonenkis-Sauer Lemma
- Introduced in the west by Valiant (1984) under the name of "probably approximately correct" (PAC) Typical results state that with probability at least $1-\delta$ (probably), any classifier from hypothesis class which has low training error will have low generalisation error (approximately correct).
- Learnability and the VC dimension by Blumer *et al.* (1989), forms the basis of statistical learning theory
- Generalisation bounds, (1) SRM, Shawe-Taylor, Bartlett, Williamson, Anthony, (1998), (2) Neural Networks, Bartlett (1998).
- Soft margin bounds, Cristianini, Shawe-Taylor (2000), Shawe-Taylor, Cristianini (2002)

History

- Apply Concentration inequalities, Boucheron *et al.* (2000), Bousquet, Elisseff (2001)
- Rademacher complexity, Koltchinskii, Panchenko (2000), Kondor, Lafferty (2002), Bartlett, Boucheron, Lugosi (2002), Bartlett, Mendelson (2002)
- PAC-Bayesian Bound proposed by McAllester (1999), improved by Seeger (2002) in Gaussian processes, applied to SVMs by Langford, Shawe-Taylor (2002), Tutorial by Langford (2005), greatly simplified proof by Germain *et al.* (2009).

Good books/tutorials

- J Shawe-Taylor, N Cristianini's book "Kernel Methods for Pattern Analysis", 2004
- V Vapnik's books "The nature of statistical learning theory", 1995 and "Statistical learning theory", 1998
- Bousquet *et al.*'s ML summer school tutorial "Introduction to Statistical Learning Theory", 2004
- . . .

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. Given a loss $\ell(x, y, f)$,
(True) Risk

$$R(w, \ell) = \mathbb{E}_{(x,y)\sim P}\, \ell(x, y, w)$$

Empirical Risk

$$R_n(w, \ell) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i, w)$$

The hinge loss $\ell(x, y, w) = [1 - y \langle \phi(x), w \rangle]_+$.
The zero-one loss $\ell(x, y, w) = \mathbf{1}_{g(x) \neq y}$.

Generalisation error

Generalisation error is the error rate over all possible testing data from the distribution P, that is the risk w.r.t. zero loss,

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

(Zero-one) Empirical risk

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

which is in fact the training error.

Regularised ERM

Regularised empirical risk minimisation

$$g_n = \operatorname*{argmin}_{g \in \mathcal{G}} R_n(g) + \lambda\Omega(g),$$

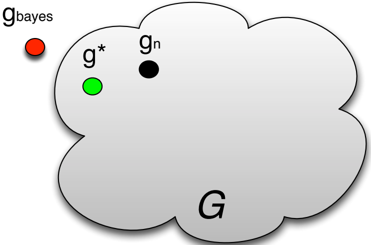
where $\Omega(g)$ is the regulariser, *e.g.* $\Omega(g) = \|g\|^2$. $\mathcal{G}$ is the hypothesis set. Unfortunately, above is not convex. It turns out that one can optimise

$$w_n = \operatorname*{argmin}_{w \in \mathcal{W}} R_n(w, \ell) + \lambda\Omega(w),$$

as long as ℓ is a surrogate loss of the zero-one loss.

Approximation error and estimation error

$$g_{bayes} = \operatorname*{argmin}_{g} R(g)$$
$$g^* = \operatorname*{argmin}_{g \in \mathcal{G}} R(g)$$
$$g_n = \operatorname*{argmin}_{g \in \mathcal{G}} R_n(g)$$



$$R(g_n) - R(g_{bayes}) = \underbrace{[R(g^*) - R(g_{bayes})]}_{\textit{approximation error}} + \underbrace{[R(g_n) - R(g^*)]}_{\textit{estimation error}}$$

Typical error bounds:

$$R(g_n) \leq R_n(g_n) + B_1(n, \mathcal{G}) \tag{1}$$
$$R(g_n) \leq R(g^*) + B_2(n, \mathcal{G}) \tag{2}$$
$$R(g_n) \leq R(g_{bayes}) + B_3(n, \mathcal{G}), \tag{3}$$

where $B(n, \mathcal{G}) \geq 0$ (and usually $B(n, \mathcal{G}) \to 0$ as $n \to +\infty$).

From Hoeffding's ineqaulity to a bound (1)

How to get $R(g) \leq R_n(g) + B(n, \mathcal{G})$?

From Hoeffding's ineqaulity to a bound (2)

Theorem (Hoeffding)

Let $Z_1, \cdots, Z_n$ be n i.i.d. random variables with $f(Z) \in [a, b]$. Then for all $\epsilon > 0$, we have

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}[f(Z)]\right| > \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

Let $Z = (X, Y)$ and $f(Z) = \mathbf{1}_{g(X)\neq Y}$, we have

$$R(g) = \mathbb{E}(f(Z)) = \mathbb{E}_{(X,Y)\sim P}[\mathbf{1}_{g(X)\neq Y}]$$

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n} f(Z_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{g(X_i)\neq Y_i}$$

$$\Rightarrow \Pr(|R(g) - R_n(g)| > \epsilon) \leq 2\exp\left(-2n\epsilon^2\right)$$

From Hoeffding's ineqaulity to a bound (3)

$$\Pr(|R(g) - R_n(g)| > \epsilon) \leq 2\exp\left(-2n\epsilon^2\right)$$

Let $\delta = 2\exp\left(-2n\epsilon^2\right) \Rightarrow \epsilon = \sqrt{\log(2/\delta)/2n}$.

$\Rightarrow$ For training examples $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, and for a hypothesis g, for any $\delta \in (0, 1)$ with probability at least $1 - \delta$,

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

Union bound over finite many hypotheses

Let consider a finite hypothesis set $\mathcal{G} = \{g_1, \cdots, g_N\}$.
Union bound

$$\Pr(\bigcup_{i=1}^{N} A_i) \leq \sum_{i=1}^{N} \Pr(A_i)$$

$$\Pr(R(g) - R_n(g) > \epsilon) \leq 2\exp\left(-2n\epsilon^2\right) \Rightarrow$$

$$\begin{aligned}\Pr(\exists g \in \mathcal{G} : R(g) - R_n(g) > \epsilon) &\leq \sum_{i=1}^{N} \Pr(R(g_i) - R_n(g_i) > \epsilon) \\ &\leq 2N\exp\left(-2n\epsilon^2\right)\end{aligned}$$

Let $\delta = 2N\exp\left(-2n\epsilon^2\right)$, we have, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\forall g \in \mathcal{G}, R(g) \leq R_n(g) + \sqrt{\frac{\log N + \log(\frac{1}{\delta})}{2n}}$$

Estimation Error bound (1)

Now we have

$$\forall g \in \mathcal{G}, R(g) \leq R_n(g) + B_1(n, \mathcal{G}).$$

How do we get bound like

$$R(g_n) \leq R(g^*) + B_2(n, \mathcal{G})?$$

The latter is interesting because $R(g_n) - R(g^*)$ is the estimation error. In fact, if the former is obtained, we can get the latter using the former.

Estimation Error bound (2)

Derivation: By definition, we know $R_n(g^*) \geq R_n(g_n)$, so

$$
\begin{aligned}
R(g_n) &= R(g_n) - R(g^*) + R(g^*) \\
&\leq R_n(g^*) - R_n(g_n) + R(g_n) - R(g^*) + R(g^*) \\
&= (R_n(g^*) - R(g^*)) + R(g_n) - R_n(g_n) + R(g^*) \\
&\leq |R(g^*) - R_n(g^*)| + |R(g_n) - R_n(g_n)| + R(g^*) \\
&\leq 2 \sup_{g \in \mathcal{G}} |R(g) - R_n(g)| + R(g^*) \\
&\leq R(g^*) + 2B_1(n, \mathcal{G})
\end{aligned}
$$

For any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,

$$
\forall g \in \mathcal{G}, R(g) \leq R(g^*) + 2\sqrt{\frac{\log N + \log(\frac{1}{\delta})}{2n}}
$$

Problem

The bound we have shown only works for a finite hypothesis set $\mathcal{G} = \{g_1, \cdots, g_N\}$. Obviously $\sqrt{\frac{\log N + \log(\frac{1}{\delta})}{2n}}$ does not exist if $N = +\infty$. This is because we were counting the number of hypothesis when applying the union bound technique

$$\begin{aligned}\Pr(\exists g \in \mathcal{G} : R(g) - R_n(g) > \epsilon) &\leq \sum_{i=1}^{N} \Pr(R(g_i) - R_n(g_i) > \epsilon) \\ &\leq 2N \exp\left(-2n\epsilon^2\right)\end{aligned}$$

A simple fix (for countably infinite)

For any g, $\delta_g := \Pr\left(R(g) - R_n(g) > \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right)$

$$\Rightarrow \ \epsilon \leq \sqrt{\frac{\log(\frac{2}{\delta_g})}{2n}} \Rightarrow \Pr(\exists g \in \mathcal{G} : R(g) - R_n(g) > \epsilon) \leq \sum_{g \in \mathcal{G}} \delta_g$$

If $\sum_{g \in \mathcal{G}} \delta_g < +\infty$, let $\delta = \sum_{g \in \mathcal{G}} \delta_g$.

$$\Pr\left(\exists g \in \mathcal{G} : R(g) - R_n(g) > \sqrt{\frac{\log(\frac{2}{\delta_g})}{2n}}\right) \leq \delta$$

Let $P(g) := \delta_g / \delta$, we have $\log(\frac{2}{\delta_g}) = \log(\frac{1}{P(g)}) + \log(\frac{2}{\delta})$.
Thus for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\forall g \in \mathcal{G}, R(g) \leq R(g^*) + 2\sqrt{\frac{\log(\frac{1}{P(g)}) + \log(\frac{2}{\delta})}{2n}}$$

Hint for better remedy

Problem: a $g \in \mathcal{G}$, such that $P(g) \approx 0$, increases the bound tremendously (thus useless).

Another way: Though there are infinite many g in $\mathcal{G}$, there are only two possible outputs for a x, because $g(x) \in \{-1, +1\}$. What matters is the number of different prediction outputs ($\leq 2^n$), not the cardinality of $\mathcal{G}$.

Next talk: This gives a hint for bounds and techniques for infinite hypothesis set, some of which (including VC dimension, VC bound) will be covered in the next talk.