Lecture 12: Learning Theory

Qinfeng (Javen) Shi

29 Oct 2015

Intro. to Stats. Machine Learning
COMP SCI 4401/7401

Course info

- Honours/Master/PhD projects
- eSELT (19 October to 6 November)
- Exam

Table of Contents I

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)
  — with probability at least $1 - \delta$ (probably), any classifier from hypothesis class/set, if the classifier has low training error, it 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
- Online course "Learning from the Data", by Yaser Abu-Mostafa in Caltech.
- Bousquet *et al.*'s ML summer school tutorial "Introduction to Statistical Learning Theory", 2004
- ...

Generalisation error

$\{(x_1, y_1), \cdots, (x_n, y_n) \sim P(X, Y)\}$[1], hypothesis function $g : \mathcal{X} \to \mathcal{Y}$, $\mathcal{Y} = \{-1, 1\}$.

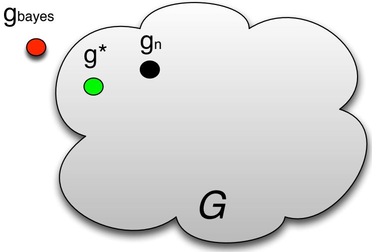
Generalisation error: error over all possible testing data from P, i.e. risk w.r.t. zero one loss $R(g) = \mathbb{E}_{(x,y)\sim P}[\mathbf{1}_{g(x)\neq y}]$.

Training error, i.e. empirical risk w.r.t. zero one loss $R_n(g) = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}_{g(x_i)\neq y_i}]$.

[1] To simplify notation and make the results more general, we don't use boldface to distinguish vectors and scalers *i.e.* x, y, w can be vectors too.

Approximation error and estimation error

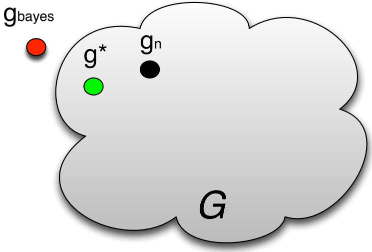
$$g_{bayes} = \operatorname*{argmin}_{g} R(g)$$
$$g^* = \operatorname*{argmin}_{g \in \mathcal{G}} R(g)$$
$$g_n = \operatorname*{argmin}_{g \in \mathcal{G}} R_n(g)$$



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

Generalisation bounds

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



Generalisation bounds:

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

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

Capacity measure

Capacity/complexity of $\mathcal{G} \downarrow \Rightarrow B(n, \mathcal{G}) \downarrow$

How to measure the capacity/complexity of $\mathcal{G}$?

- Counting the hypotheses in $\mathcal{G}$, *i.e.* $|\mathcal{G}|$.

Capacity measure

Capacity/complexity of $\mathcal{G} \downarrow \Rightarrow B(n, \mathcal{G}) \downarrow$

How to measure the capacity/complexity of $\mathcal{G}$?

- Counting the hypotheses in $\mathcal{G}$, *i.e.* $|\mathcal{G}|$.
- Counting all possible outputs of the hypotheses

Capacity measure

Capacity/complexity of $\mathcal{G} \downarrow \Rightarrow B(n, \mathcal{G}) \downarrow$

How to measure the capacity/complexity of $\mathcal{G}$?

- Counting the hypotheses in $\mathcal{G}$, *i.e.* $|\mathcal{G}|$.
- Counting all possible outputs of the hypotheses
- Ability to fit noise

Capacity measure

Capacity/complexity of $\mathcal{G} \downarrow \Rightarrow B(n, \mathcal{G}) \downarrow$

How to measure the capacity/complexity of $\mathcal{G}$?

- Counting the hypotheses in $\mathcal{G}$, *i.e.* $|\mathcal{G}|$.
- Counting all possible outputs of the hypotheses
- Ability to fit noise
- Divergence of the prior and posterior distributions over classifiers (omitted)
- $\cdots$

Counting the hypotheses

(a.k.a Hoeffding's inequality bound) For training examples $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, for a finite hypothesis set $\mathcal{G} = \{g_1, \cdots, g_N\}$, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

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

Proof (1)– Hoeffding's inequality

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

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

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

$$b = 1, a = 0$$

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

Proof (2) – for a hypothesis

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

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 $1 - \delta$,

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

Proof (3) – 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 $1-\delta$,

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

Counting outputs

What if there are infinite many hypotheses $N = \infty$?

Counting outputs

What if there are infinite many hypotheses $N = \infty$?

$$\sqrt{\frac{\log N + \log(\frac{1}{\delta})}{2n}} = \infty$$

Counting outputs

What if there are infinite many hypotheses $N = \infty$?

Observation:

1. For any x, only two possible outputs ($g(x) \in \{-1, +1\}$);
2. For any n training data at most 2^n different outputs of $g(x)$.

What matters is the "expressive power" (Blumer *et al.* 1986,1989)(*e.g.* the number of different prediction outputs), not the cardinality of $\mathcal{G}$.

Growth function

Definition (Growth function)

The growth function (a.k.a Shatter coefficient) of $\mathcal{F}$ with n points is

$$S_{\mathcal{F}}(n) = \sup_{(z_1,\cdots,z_n)} \left| \left\{ \left(f(z_1), \cdots, f(z_n) \right) \right\}_{f \in \mathcal{F}} \right|.$$

i.e. maximum number of ways that n points can be classified by the hypothesis set $\mathcal{F}$.

Note: g can be a f, and $\mathcal{G}$ can be a $\mathcal{F}$.

Growth function

If no restriction on g, we know

$$\sup_{(z_1, z_2, z_3)} \left| \left\{ \Big(g(z_1), g(z_2), g(z_3) \Big) \right\} \right| = 2^3$$

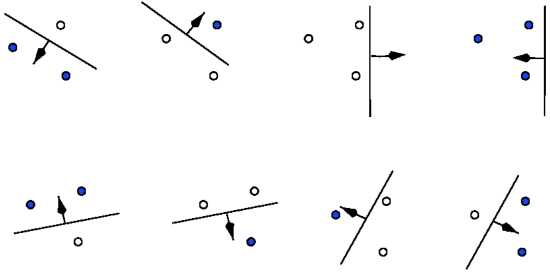
When we restrict $g \in \mathcal{G}$,

$$S_{\mathcal{G}}(3) = \sup_{(z_1, z_2, z_3)} \left| \left\{ \Big(g(z_1), g(z_2), g(z_3) \Big) \right\}_{g \in \mathcal{G}} \right|.$$

i.e. counting all possible outputs that $\mathcal{G}$ can express.

Growth function

The growth function $S_{\mathcal{G}}(3) = 8$, if $\mathcal{G}$ is the set of linear decision functions shown in the image below[2].



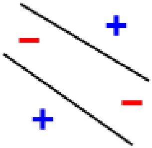
[2] The image is from http://www.svms.org/vc-dimension/

Growth function

How about $S_{\mathcal{G}}(4)$?

Growth function

How about $S_{\mathcal{G}}(4)$?



One g can not classify 4 points above correctly (two gs or a curve needed), which means $S_{\mathcal{G}}(4) < 2^4$.

Picture courtesy of wikipedia

VC dimension (1)

Definition (VC dimension)

The VC dimension (often denoted as h) of a hypothesis set $\mathcal{G}$, is the largest n such that

$$S_{\mathcal{G}}(n) = 2^n.$$

$h = 3$ for $\mathcal{G}$ being the set of linear decision functions in 2-D.

VC dimension (2)

Lemma

Let $\mathcal{G}$ be a set of functions with finite VC dimension h. Then for all $n \in \mathbb{N}$,

$$S_{\mathcal{G}}(n) \leq \sum_{i=0}^{h} \binom{n}{i},$$

and for all $n \geq h$,

$$S_{\mathcal{G}}(n) \leq \left(\frac{en}{h}\right)^{h}.$$

VC dimension (3)

Theorem (Growth function bound)

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

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

Thus for all $n \geq h$, since $S_{\mathcal{G}}(n) \leq (\frac{en}{h})^h$, we have

Theorem (VC bound)

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

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

VC dimension (4)

Assume $x \in \mathbb{R}^d, \Phi(x) \in \mathbb{R}^D$ (Note D can be $+\infty$).

- linear $\langle x, w \rangle$, $h = d + 1$
- polynomial $(\langle x, w \rangle + 1)^p$, $h = \binom{d+p-1}{p} + 1$
- Gaussian RBF $\exp\left(-\frac{\|x-x'\|^2}{\sigma^2}\right)$, $h = +\infty$.
- Margin γ, $h \leq \min\{D, \lceil \frac{4R^2}{\gamma^2} \rceil\}$, where the radius $R^2 = \max_{i=1}^n \langle \Phi(x_i), \Phi(x_i) \rangle$ (assuming data are already centered)

Ability to fit noise

Definition (Rademacher complexity)

Given $S = \{z_1, \cdots, z_n\}$ from a distribution P and a set of real-valued functions $\mathcal{G}$, the empirical Rademacher complexity of $\mathcal{G}$ is the random variable

$$\hat{\mathcal{R}}_n(\mathcal{G}, S) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right| \right],$$

where $\boldsymbol{\sigma} = \{\sigma_1, \cdots, \sigma_n\}$ are independent uniform $\{\pm 1\}$-valued (Rademacher) random variables. The Rademacher complexity of $\mathcal{G}$ is

$$\mathcal{R}_n(\mathcal{G}) = \mathbb{E}_S[\hat{\mathcal{R}}_n(\mathcal{G}, S)] = \mathbb{E}_{S\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right| \right]$$

First sight

$\sup_{g \in \mathcal{G}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right|$

- measures the best correlation between $g \in \mathcal{G}$ and random label (*i.e.* noise) $\sigma_i \sim U(\{-1, +1\})$.
- ability of $\mathcal{G}$ to fit noise.
- the smaller, the less chance of detected pattern being spurious
- if $|\,\mathcal{G}\,| = 1$, $\mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right| \right] = 0.$

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* $1-\delta$, $\forall g \in \mathcal{G}$,

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

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

Note: $\hat{\mathcal{R}}_n(\mathcal{G}, S)$ is computable whereas $\mathcal{R}_n(\mathcal{G})$ is not.

Example

Let $S = \{(x_i, y_i)\}_{i=1}^n \sim P^n$. $y_i \in \{-1, +1\}$
One form of soft margin binary SVMs is

$$\min_{w,\gamma,\xi} -\gamma + C \sum_{i=1}^{n} \xi_i \tag{4}$$

$$\text{s.t. } y_i \langle \phi(x_i), w \rangle \geq \gamma - \xi_i, \xi_i \geq 0, \|w\|^2 = 1$$

- The Rademacher Margin bound (next slide) applies.
- $\hat{\mathcal{R}}_n(\mathcal{G}, S)$ is essential, where
  $\mathcal{G} = \{-yf(x; w), f(x; w) = \langle \phi(x_i), w \rangle, \|w\|^2 = 1\}$.

Rademacher Margin bound

Theorem (Margin)

Fix $\gamma > 0, \delta \in (0,1)$, let $\mathcal{G}$ be the class of functions mapping from $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ given by $g(x,y) = -yf(x)$, where f is a linear function in a kernel-defined feature space with norm at most 1. Let $S = \{(x_i, y_i)\}_{i=1}^n$ be drawn i.i.d. from $P(X,Y)$ and let $\xi_i = (\gamma - y_i f(x_i))_+$. Then with probability at least $1-\delta$ over sample of size n, we have

$$\mathbb{E}_P[\mathbf{1}_{y \neq \mathrm{sgn}(f(x))}] \leq \frac{1}{n\gamma}\sum_{i=1}^n \xi_i + \frac{4}{n\gamma}\sqrt{\mathrm{tr}(\mathbf{K})} + 3\sqrt{\frac{\ln(2/\delta)}{2n}},$$

- data dependency come through training error and margin
- tighter than VC bound ($\frac{4}{n\gamma}\sqrt{tr(\mathbf{K})} \leq \frac{4}{n\gamma}\sqrt{nR^2} \leq 4\sqrt{\frac{R^2}{n\gamma^2}}$)

Recap today's content

Capacity/complexity of $\mathcal{G} \downarrow \Rightarrow B(n, \mathcal{G}) \downarrow$

How to measure the capacity/complexity of $\mathcal{G}$?

- Counting the hypotheses in $\mathcal{G}$, *i.e.* $|\mathcal{G}|$.
- Counting all possible outputs of the hypotheses
- Ability to fit noise

Next

Course Revision/Review for Exam