CS group meeting 03/31/2017

The Nature of Statistical Learning Theory V. Vapnik

Lecture 3

$Q(z,\alpha)$ $\alpha \in \Lambda$ is a set of indicator functions

$H^\Lambda(\ell)$ VC entropy

$H^\Lambda_{ann}(\ell)$ annealed entropy

$G^\Lambda(\ell)$ Growth function.

Distribution-dependent bounds

Theorem 1

$$\mathbb{P}\left(\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z) - \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i,\alpha)\right| > \varepsilon\right)$$

$$\leq 4\exp\left(\left(\frac{H^\Lambda_{ann}(2\ell)}{\ell} - \varepsilon^2\right)\ell\right)$$

Theorem 2

$$\mathbb{P}\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z) - \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)\,dF(z)}} > \varepsilon\right\}$$

$$\leq 4\exp\left\{\left(\frac{H^\Lambda_{ann}(2\ell)}{\ell} - \frac{\varepsilon^2}{4}\right)\ell\right\}$$

These two bounds are non-trivial if

$$\lim_{\ell \to \infty} \frac{H^{\Lambda}_{ann}(\ell)}{\ell} = 0$$

<u>Distribution - Independent Bounds</u>

For <u>any</u> distribution function $F(z)$

$$\mathbb{P}\left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z,\alpha)\,dF(z) - \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i,\alpha) \right| > \varepsilon \right\}$$

$$\leq 4 \exp\left\{ \left(\frac{G^{\Lambda}(2\ell)}{\ell} - \varepsilon^2 \right) \ell \right\}$$

$$\mathbb{P}\left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z,\alpha)\,dF(z) - \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)\,dF(z)}} > \varepsilon \right\}$$

$$\leq 4 \exp\left\{ \left(\frac{G^{\Lambda}(2\ell)}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\}$$

These inequalities are non-trivial if

$$\lim_{\ell \to \infty} \frac{G^{\Lambda}(\ell)}{\ell} = 0$$

$Q(z,\alpha)$ $\alpha \in \Lambda$ is a set of real functions

where $A = \inf_{\alpha, z} Q(z,\alpha) \leq Q(z,\alpha) \leq \sup_{\alpha, z} Q(z,\alpha) = B$

indicators $I(z,\alpha,\beta) = \mathbb{1}\left(Q(z,\alpha) \geq \beta\right)$

$(A, B) = \mathcal{B}$ $\beta \in \mathcal{B}$, $\alpha \in \Lambda$

In the case where $Q(z,\alpha)$, $\alpha \in \Lambda$ are indicator functions, the set of indicators $I(z,\alpha,\beta)$, $\alpha \in \Lambda$, $\beta \in (0,1)$ coincides with the set $Q(z,\alpha)$, $\alpha \in \Lambda$.

$H^{\Lambda,\mathcal{B}}(l)$ VC entropy of $\left\{ I(z,\alpha,\beta), \begin{matrix}\alpha \in \Lambda \\ \beta \in \mathcal{B}\end{matrix} \right\}$

$H^{\Lambda,\mathcal{B}}_{ann}(l)$ annealed VC entropy of the same set above

$G^{\Lambda,\mathcal{B}}(l)$ Growth function of the same set above

<u>Theorem 3</u> ① If $Q(z,\alpha)$ $\alpha \in \Lambda$ is a set of totally bounded function, then

$$\mathbb{P}\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z)-\frac{1}{\ell}\sum_{i=1}^{\ell}Q(z_i,\alpha)\right|>\varepsilon\right\}$$

$$\leq 4\exp\left\{\left(\frac{H_{ann}^{\Lambda,B}(2\ell)}{\ell}-\frac{\varepsilon^2}{(B-A)^2}\right)\ell\right\}$$

② $0\leq Q(z,\alpha)\leq B$ a set of totally bounded non-negative functions

$$\mathbb{P}\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z)-\frac{1}{\ell}\sum_{i=1}^{\ell}Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)\,dF(z)}}>\varepsilon\right\}$$

$$\leq 4\exp\left\{\left(\frac{H_{ann}^{\Lambda,B}(2\ell)}{\ell}-\frac{\varepsilon^2}{4B}\right)\ell\right\}$$

③ $0\leq Q(z,\alpha)$ $\alpha\in\Lambda$ is a set of functions such that for some $p>2$ the p-th normalized moments

$$m_p(\alpha)=\sqrt[p]{\int Q^p(z,\alpha)\,dF(z)}$$

Then

$$\mathbb{P}\left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z,\alpha)\,dF(z) - \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i,\alpha)}{\sqrt[p]{\int Q^p(z,\alpha)\,dF(z)}} > a(p)\varepsilon \right\}$$

$$\leq 4 \exp\left\{ \left(\frac{H_{ann}^{\Lambda,\mathcal{B}}(2\ell)}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\}$$

where $$a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}}$$

These bounds become non-trivial of $\lim_{\ell \to \infty} \frac{H_{ann}^{\Lambda,B}(\ell)}{\ell} = 0$.

I will skip the corresponding inequalities for distribution - independent bounds.

Bounds on the generalization ability of learning machines.

Two major questions concerning generalization ability

(A) What actural risk $R(\alpha_\ell)$ is provided by the function $Q(z, \alpha_\ell)$ that achieves minimal empirical risk $R_{emp}(\alpha_\ell)$?

(B) How close is this risk to the minimum possible $\inf_\alpha R(\alpha)$, $\alpha \in \Lambda$ for the given set of functions?

Let
$$\mathcal{E} = 4 \frac{G^{\Lambda, B}(2\ell) - \ln\left(\frac{\eta}{4}\right)}{\ell}$$

We work with distribution-independent bounds.

These bounds are nontrivial if $\mathcal{E} < 1$

<u>Case 1</u> The set of totally bounded functions

$A \le Q(z, \alpha) \le B, \quad \alpha \in \Lambda$

(A) With probability at least $1-\eta$, simultaneously for all $Q(z,\alpha)$, $\alpha \in \Lambda$ (including the function that minimizes the empirical risk)

$$R(\alpha) \le R_{emp}(\alpha) + \frac{(B-A)}{2}\sqrt{\mathcal{E}}$$

$$R_{emp}(\alpha) - \frac{(B-A)}{2}\sqrt{\mathcal{E}} \le R(\alpha)$$

(B) With probability of at least $1-2\eta$ for the function $Q(z, \alpha_\ell)$ that minimizes the empirical risk

$$R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha) \le (B-A)\sqrt{\frac{-\ln \eta}{2\ell}} + \frac{B-A}{2}\sqrt{\mathcal{E}}$$

The proof of this fact needs a Barrey-Essen inequality (see Vapnik, V. N. Statistical Learning Theory) and we omit it.

<u>Case 2</u> The set of totally bounded non-negative functions, $0 \le Q(z,\alpha) \le B$, $\alpha \in \Lambda$

(A) With probability of at least $1-\eta$ simultaneously for all functions $Q(z,\alpha) \le B$, $\alpha \in \Lambda$ (including the function that minimizes the empirical risk)

$$R(\alpha) \le R_{emp}(\alpha) + \frac{B\mathcal{E}}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}(\alpha)}{B\mathcal{E}}}\right)$$

(B) With probability of at least $1-2\eta$ for the function $Q(z,\alpha_\ell)$ that minimizes the empirical risk

$$R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha) \le B\sqrt{\frac{-\ln \eta}{2\ell}} + \frac{B\mathcal{E}}{2}\left(1 + \sqrt{1 + \frac{4}{\mathcal{E}}}\right)$$

Case 3 The set of unbounded non-negative bounds

$0 \le Q(z,\alpha) \quad \alpha \in \Lambda$

We are given a pair (p, τ) such that the inequality holds:

$$\sup_{\alpha \in \Lambda} \frac{\left(\int Q^p(z,\alpha)\, dF(z)\right)^{1/p}}{\int Q(z,\alpha)\, dF(z)} \le \tau < \infty, \quad p > 1 \qquad (*)$$

$p > 2$ case

(A) With probability of at least $1-\eta$ we have

$$R(\alpha) \le \frac{R_{emp}(\alpha)}{\left(1 - a(p)\tau\sqrt{\mathcal{E}}\right)_+} \quad \text{where}$$

$$a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}}$$

holds simultaneously ~~the~~ for all $0 \le Q(z,\alpha)$ satisfying $(*)$

(B) With probability of at least $1-2\eta$ we've

$$\frac{R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha)}{\inf_{\alpha \in \Lambda} R(\alpha)} \le \frac{\tau a(p)\sqrt{\mathcal{E}}}{\left(1 - \tau a(p)\sqrt{\mathcal{E}}\right)_+} + O\left(\frac{1}{\ell}\right)$$

holds for the function $Q(z, \alpha_\ell)$ that minimizes the empirical risk.

These estimates evaluate how close the risk obtained by using the ERM principle is to the smallest possible risk.

To make the above bounds about generalization ability of learning machines to be constructive rather than conceptual, we introduce the new notion of <u>VC-dimension</u> (Vapnik-Chervonenkis dimension)

▷ The VC dimension of a set of indicator functions (Vapnik-Chervonenkis, 1968, 1971)

Let $Q(z,\alpha)$ $\alpha\in\Lambda$ be a set of indicator functions, then the VC-dimension of the set $\{Q(z,\alpha), \alpha\in\Lambda\}$ is the maximum number h of vectors $z_1 \ldots z_h$ that can be separated into 2^h possible ways using the functions of this set.

(= the maximum number of vectors that can be shattered by the set of functions)

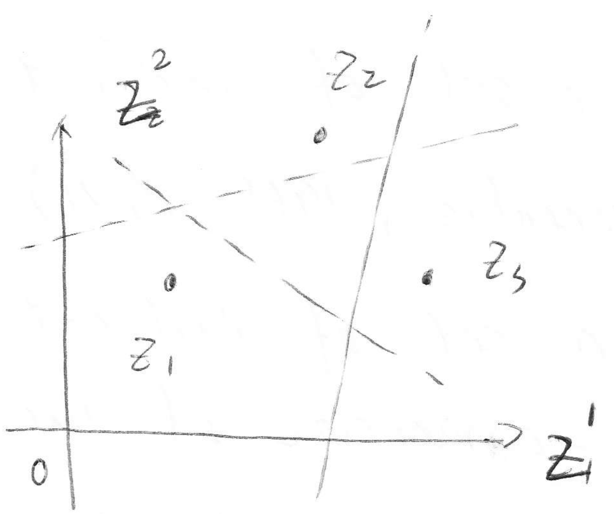
If for any n there exists a set of n vectors ~~to~~ which can be shattered by the set $Q(z,\alpha)$, $\alpha \in \Lambda$ then the VC dimension is equal to infinity.

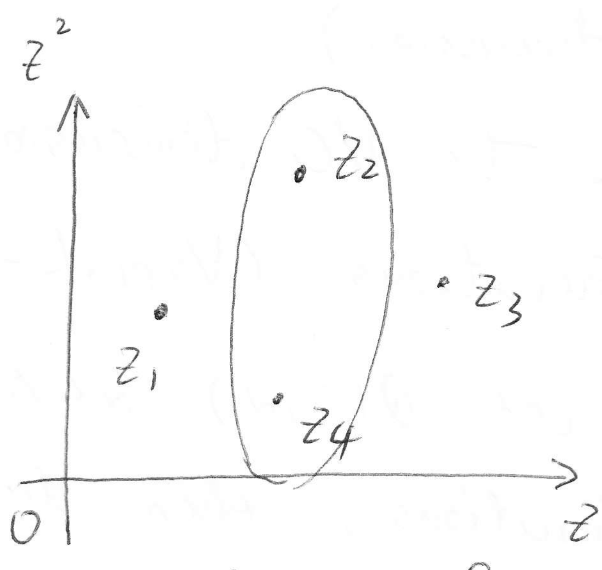
<u>Example</u> $\quad Q(z,\alpha) = \mathbb{1}\left\{ \sum_{p=1}^{n} \alpha_p z_p + \alpha_0 > 0 \right\}$

is a set of indicator functions, $\alpha = (\alpha_1 \ldots \alpha_n)$

$z = (z_1 \ldots z_n)$ then VC-dimension $= h =$

$n+1$.





▷ The VC-dimension of a set of real functions (Vapnik 1979)

Let $A \leq Q(z,\alpha) \leq B$, $\alpha \in \Lambda$ be a set of real functions bounded by constants A and B (A can be $-\infty$ and B can be ∞)

Introduce a set of indicators associated with

$Q(z,\alpha)$, $\alpha \in \Lambda$ as

$$I(z,\alpha,\beta) = \mathbb{1}\{Q(z,\alpha) - \beta > 0\}$$

$\alpha \in \Lambda$, $\beta \in (A, B)$

The VC-dimension of a set of real functions $\{Q(z,\alpha), \alpha \in \Lambda\}$ is the VC-dimension of a set of corresponding indicators $\{I(z,\alpha,\beta), \alpha \in \Lambda, \beta \in (A,B)\}$

<u>Example</u> $$Q(z,\alpha) = \sum_{p=1}^{n} \alpha_p z_p + \alpha_0$$

$\alpha_0 \ldots \alpha_n \in (-\infty, +\infty)$

$z = (z_1 \ldots z_n)$ VC-dimension $= n+1$.

Since $$I(z,\alpha,\beta) = \mathbb{1}\left\{\sum_{p=1}^{n} \alpha_p z_p + (\alpha_0 - \beta) > 0\right\}$$

Here VC-dimension = # of free parameters

In general THIS IS NOT TRUE.

<u>Example</u>. The VC-dimension of the following set of functions $f(z,\alpha) = \mathbb{1}\{\sin \alpha z > 0\}$ $\alpha \in R^1$ is infinite.

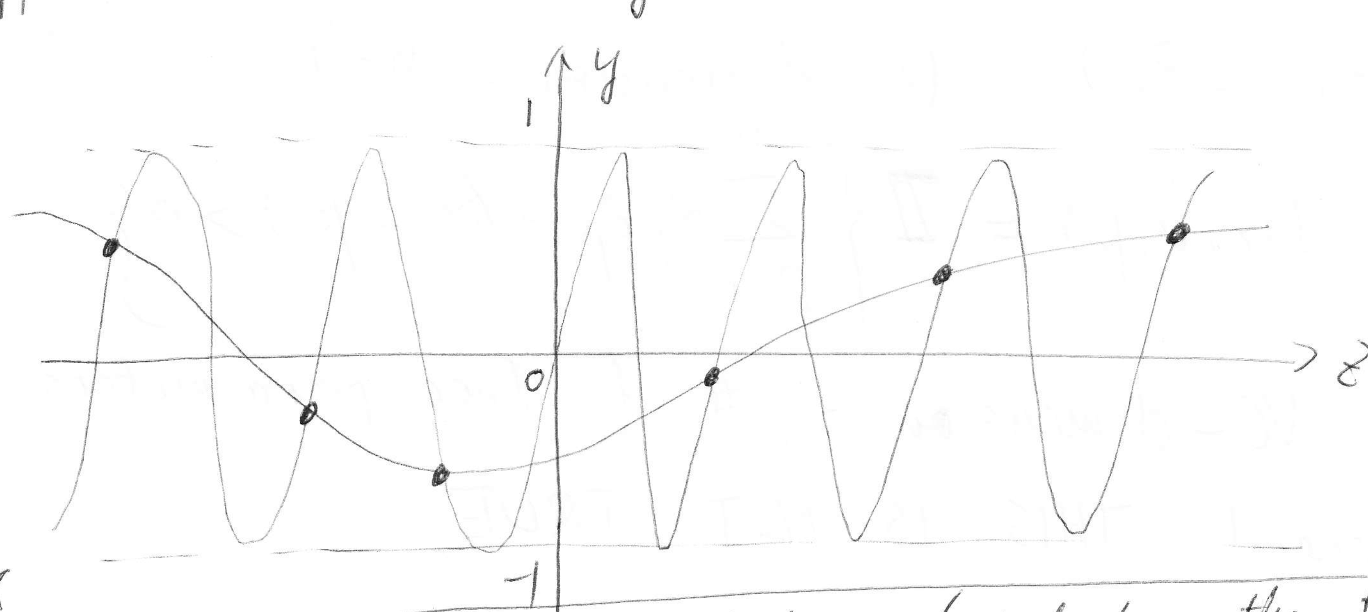
Consider $z_1 = 10^{-1}, \ldots, z_\ell = 10^{-\ell}$

they can be separated (shattered) by considering

$$\alpha = \pi \left(\sum_{i=1}^{\ell} (1-\delta_i) 10^i + 1 \right), \quad i = 1, 2, \ldots, \ell$$
$$\delta_i = 0 \text{ or } 1$$

Then $f(z_i, \alpha) = \delta_i$.

<u>Fact</u>. Choosing an appropriate coefficient α one can for any number of appropriate chosen points approximate values of any function in $(-1, +1)$



☆ $\boxed{\text{VC dimension can be both much larger/much less than \# parameters}}$

How does the notion of VC-dimension help us to make the bounds about generalization ability of learning machines into constructive bounds?

Theorem about the Structure of the growth function

Any growth function either satisfies the equality

$$G^\Lambda(\ell) = \ell \ln 2$$

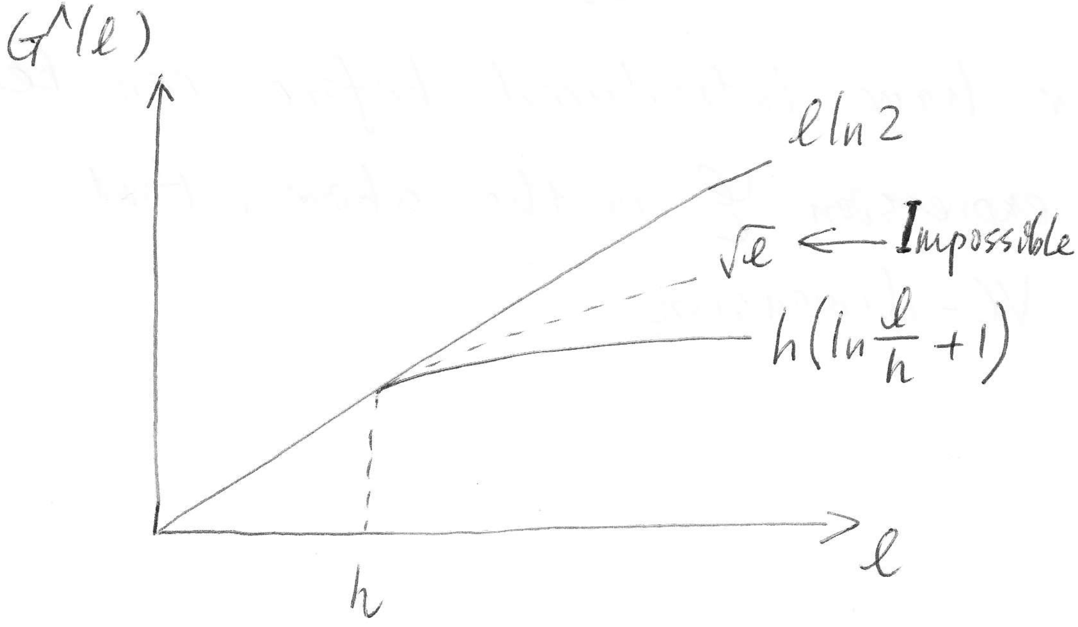
or is bounded by the inequality

$$G^\Lambda(\ell) \leq h\left(\ln \frac{\ell}{h} + 1\right)$$

where h is an integer such that when $\ell = h$

$$G^\Lambda(h) = h \ln 2$$

$$G^\Lambda(h+1) < (h+1) \ln 2$$



"Growth function is either linear or is bounded by a logarithmic function"

One can show that the VC-dimension of the set of indicator functions $Q(z,\alpha)$ $\alpha \in \Lambda$ is infinite of the Growth function for this set of functions is linear; One can also show that the VC dimension of the set of indicator functions $Q(z,\alpha)$ $\alpha \in \Lambda$ is finite and equals h if the corresponding Growth function is bounded by a logarithmic function with coefficient h

Let us consider sets of functions which possess a finite VC dimension h. In this case

$$G^{\Lambda}(\ell) \leq h\left(\ln \frac{\ell}{h} + 1\right), \quad \ell > h$$

Recall then

$$\mathcal{E} = 4 \, \frac{h\left(\ln \frac{2\ell}{h} + 1\right) - \ln\left(\frac{\eta}{4}\right)}{\ell}$$

All estimates we have introduced before can be replaced by the expression $\mathcal{E}$ in the above, that is calculated from VC-dimension