Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology

Technical University of Munich

Vapnik-Chervonenkis (VC) dimension

Outline

- Previously: Uniform convergence bound for infinite $\mathcal{H}$ using growth function $\tau_{\mathcal{H}}(\cdot)$
  (worst case $\tau_{\mathcal{H}}(m) \leq 2^m$, bound is useless)

- This lecture: VC dimension of $\mathcal{H}$
  - Quantifies complexity of hypothesis class

- Sauer's lemma
  - Bound $\tau_{\mathcal{H}}(\cdot)$ in terms of VC dimension
  - If $\mathrm{VCdim}(\mathcal{H}) = d < \infty$, then $\tau_{\mathcal{H}}(m) = O(m^d) \ll 2^m$ (for large m)

- Examples: VC dimension and generalisation error bound for linear classifier, 1-NN, neural networks

Shattering of a set

Shattering of a set

Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ and $C = \{x_1, \ldots, x_m\} \in \mathcal{X}^m$. We say C is shattered by $\mathcal{H}$ if $|\mathcal{H}_{|C}| = 2^m$.

Equivalently,
for every possible labelling $s \in \{\pm 1\}^m$ of instances in C, there is a $h_s \in \mathcal{H}$ such that that $h_s(x_i) = s_i$ for $i = 1, \ldots, m$.

Vapnik Chervonenkis (VC) dimension

VC dimension

VC dimension of a non-empty $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ is the cardinality of the largest possible subset of $\mathcal{X}$ that can be shattered by $\mathcal{H}$, that is,

$$\mathrm{VCdim}(\mathcal{H}) = \max\{m \in \mathbb{N} \;:\; \tau_{\mathcal{H}}(m) = 2^m\}.$$

If $\mathcal{H}$ can shatter arbitrarily large sets, then $\mathrm{VCdim}(\mathcal{H}) = \infty$.

- Alternative view: $\mathrm{VCdim}(\mathcal{H}) = d \leq \infty$ if
  - there exists some set $C \in \mathcal{X}^d$ that can be shattered by $\mathcal{H}$
  - no set of cardinality $d + 1$ can be shattered by $\mathcal{H}$

VC dimension for finite $\mathcal{H}$

- State an upper bound on $\text{VCdim}(\mathcal{H})$ in terms of $|\mathcal{H}|$
  - Answer: $\text{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$
  - Recall $\tau_{\mathcal{H}}(m) \leq |\mathcal{H}|$
  - From definition, $\text{VCdim}(\mathcal{H}) = d$ satisfies $2^d = \tau_{\mathcal{H}}(d) \leq |\mathcal{H}|$
- Is above bound tight? Is it equality in some case?
  - Yes
  - Let $\mathcal{H} = \{h_1(x) = \text{sign}(x), h_2(x) = -\text{sign}(x)\} \subset \{\pm 1\}^{\mathbb{R}}$
  - Verify that $\mathcal{H}$ can shatter only one point $\implies \text{VCdim}(\mathcal{H}) = 1 = \log_2(|\mathcal{H}|)$

VC dimension for decision stump

- $\mathcal{H}_{ds-1} = \left\{ h(x) = b \cdot \text{sign}(x - t) \ : \ b \in \{\pm 1\}, t \in \mathbb{R} \right\}$
- Compute $\text{VCdim}(\mathcal{H}_{ds-1})$
- Approach 1 (using definition):
  - Recall from previous lecture: $\tau_{\mathcal{H}_{ds-1}}(m) = 2m$
  - $\tau_{\mathcal{H}_{ds-1}}(2) = 4 = 2^2$, but $\tau_{\mathcal{H}_{ds-1}}(3) = 6 < 2^3$
  - So $\text{VCdim}(\mathcal{H}_{ds-1}) = 2$

VC dimension for decision stump

- Approach 2 (using alternative view):
  - Take any $x_1 < x_2$. There are 4 possible labellings

$$\underbrace{- \quad -}_{\text{use } b=1, t>x_2} \qquad \underbrace{- \quad +}_{\text{use } b=1, x_1<t<x_2} \qquad \underbrace{+ \quad -}_{\text{use } b=-1, x_1<t<x_2} \qquad \underbrace{+ \quad +}_{\text{use } b=1, t<x_1}$$

  - So $\mathcal{H}_{ds-1}$ can shatter $\{x_1, x_2\}$
  - Take any $x_1 < x_2 < x_3$ (if they are not distinct we cannot shatter them)
  - 8 possible labelling, but we cannot correctly label following configurations:

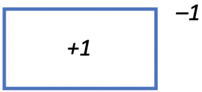
$$- \quad + \quad - \qquad\qquad + \quad - \quad +$$

  - Any set of size 3 cannot be shattered by $\mathcal{H}_{ds-1}$. So $\text{VCdim}(\mathcal{H}_{ds-1}) = 2$

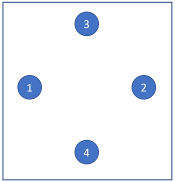
VC dimension of axis parallel rectangles in $\mathbb{R}^2$

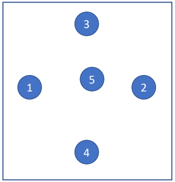
- $\mathcal{X} = \mathbb{R}^2$, and $\mathcal{H}$ class of all axis parallel rectangles of form

$$h_{a,b,c,d}\left(x^{(1)}, x^{(2)}\right) = \begin{cases} +1 & \text{if } a \leq x^{(1)} \leq b \text{ and } c \leq x^{(2)} \leq d, \\ -1 & \text{otherwise.} \end{cases}$$



- Show that $\mathrm{VCdim}(\mathcal{H}) = 4$
  - Can shatter 4 points shown on right (need only one such set to exist)
  - There can be 4 points that are not shattered
  - For any 5 points, there are 4 points that define an axis-parallel rectangle containing all points
  - Cannot label the 4 points as $+1$, and 5^{th} as -1

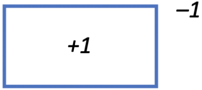




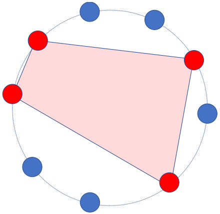
Convex polygons in $\mathbb{R}^2$

- For any convex polygon C in $\mathbb{R}^2$, define
$$h_C = \begin{cases} +1 & \text{if } x \in C \\ -1 & \text{otherwise.} \end{cases}$$

$C =$ square



- $\mathcal{H} = \{h_C \ : \ C \text{ is a convex polygon}\}$
  Show that $\text{VCdim}(\mathcal{H}) = \infty$
  - Take m distinct points on a circle
  - Can be shattered for any m



- What happens if we restrict the number of sides of polygon?

Further examples

- Try out other examples by yourself:
  - Signed axis parallel rectangles (allow -1 inside)
  - Convex polygons with at most k edges
  - In some cases, you may only find upper bounds
- Later in this section
  - VC dimension of linear classifiers
  - VC dimension of 2-layer neural networks (simplified)
  - Hypothesis class for nearest neighbour, and its VC dimension

Bound on growth function in terms of VC dimension

Theorem VC.1 (Sauer's lemma)

Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ *be non-empty with* $\mathrm{VCdim}(\mathcal{H}) = d < \infty$. *For all* $m \in \mathbb{N}$,

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

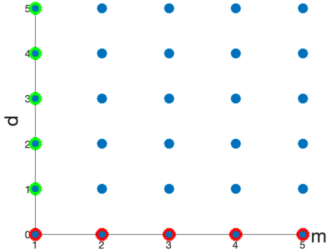
A simpler bound which holds for all $m \geq d \geq 1$

$$\tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^{d}$$

Use above inequality to derive generalisation error bound for ERM

Proof of Sauer's lemma

- Proof is by induction on m and d.
- **Two base cases:** $d = 0, m \geq 1$ and $m = 1, d \geq 1$
- $d = 0, m \geq 1$:
  - $d = 0 \implies |\mathcal{H}| = 1$ and $\tau_{\mathcal{H}}(m) = 1 = \binom{m}{0}$
- $d \geq 1, m = 1$:
  - $d \geq 1 \implies |\mathcal{H}| \geq 2 \implies$ there is $x \in \mathcal{X}$ such that $|\mathcal{H}_{|\{x\}}| = 2$
  - $\tau_{\mathcal{H}}(m) = 2 = \binom{m}{0} + \binom{m}{1}$ for $m = 1$



Proof of Sauer's lemma

- **Induction** for $m > 1$ and $d > 0$
- From inductive hypothesis:
    - $\tau_{\mathcal{H}}(m') \leq \sum_{i=0}^{d'} \binom{m'}{i}$ holds for

$$(m', d') = (m-1, d-1) \qquad \text{and} \qquad (m', d') = (m-1, d)$$

Proof of Sauer's lemma

- Let $C = (x_1, x_2, \ldots, x_m)$ and denote $C' = (x_2, \ldots, x_m)$
- For every $(y_2, \ldots, y_m) \in \mathcal{H}_{|C'}$ there can be only two possibilities:
  - both $(-1, y_2, \ldots, y_m)$ and $(+1, y_2, \ldots, y_m)$ are in $\mathcal{H}_{|C}$
  - either $(-1, y_2, \ldots, y_m) \in \mathcal{H}_{|C}$ or $(+1, y_2, \ldots, y_m) \in \mathcal{H}_{|C}$
- Let $Y = \big\{(y_2, \ldots, y_m) \in \mathcal{H}_{|C'} \ : \ (-1, y_2, \ldots, y_m), (+1, y_2, \ldots, y_m) \in \mathcal{H}_{|C}\big\}$.

$$|\mathcal{H}_{|C}| = |\mathcal{H}_{|C'}| + |Y|$$

  - Will bound the size of each of the two sets

Proof of Sauer's lemma

- Bounding $|\mathcal{H}_{|C'}|$:
  - $\text{VCdim}(\mathcal{H}) = d$ and $|C'| = m - 1$
  - $|\mathcal{H}_{|C'}| \leq \tau_{\mathcal{H}}(m-1) \leq \sum_{i=0}^{d} \binom{m-1}{i}$ (by induction hypothesis)
- Bounding $|Y|$:
  - View $Y \subset \{\pm 1\}^{C'}$ as a hypothesis class, and show $\text{VCdim}(Y) \leq d - 1$
  - Proof by contradiction. If $\text{VCdim}(Y) = d$, then Y shatters a set $C'' \subset C$ of size d
  - So $C'' \cup \{x_1\}$ is shattered by $\mathcal{H} \implies \text{VCdim}(\mathcal{H}) \geq d + 1$ (contradiction)
  - $\text{VCdim}(Y) \leq d - 1 \implies |Y| \leq \sum_{i=0}^{d-1} \binom{m-1}{i}$

Proof of Sauer's lemma

- Bounding $|\mathcal{H}_{|C}| = |\mathcal{H}_{|C'}| + |Y|$:

$$|\mathcal{H}_{|C}| \leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i}$$

$$= \binom{m-1}{0} + \sum_{i=1}^{d} \left(\binom{m-1}{i} + \binom{m-1}{i-1} \right) = \sum_{i=0}^{d} \binom{m}{i} \quad \text{as } \binom{m}{i} = \binom{m-1}{i} + \binom{m-1}{i-1}$$

- Above is true for every $C \in \mathcal{X}^m \implies$ bound holds for $\tau_{\mathcal{H}}(m)$

Proof of Sauer's lemma

Derivation of simpler bound:

$$
\begin{aligned}
\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} &\leq \sum_{i=0}^{d} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} && \text{we assume } m \geq d \\
&= \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^{i} 1^{d-i} \\
&\leq \left(\frac{m}{d}\right)^{d} \left(1 + \frac{d}{m}\right)^{m} \\
&\leq \left(\frac{em}{d}\right)^{d} && \text{since } \left(1 + \frac{x}{n}\right)^{n} \leq e^{x}
\end{aligned}
$$

VC dimension of linear classifiers in $\mathbb{R}^p$

- Class of linear classifiers over $\mathcal{X} = \mathbb{R}^p$

$$\mathcal{H}_{lin} = \left\{ \text{sign}(\langle w, x \rangle + b) \;:\; w \in \mathbb{R}^p, b \in \mathbb{R} \right\} \qquad \ldots \langle w, x \rangle = w^T x$$

- ERM over $\mathcal{H}_{lin}$ related to SVMs, perceptron

- $\text{VCdim}(\mathcal{H}_{lin}) = p + 1$

- Generalisation error bound for ERM over $\mathcal{H}_{lin}$: w.p. $1 - \delta$

$$L_{\mathcal{D}}(\widehat{h}) \;\leq\; L_{\mathcal{D}}(\mathcal{H}_{lin}) + 2\sqrt{\frac{8}{m}\left(\ln\left(\left(\frac{2em}{p+1}\right)^{p+1}\right) + \ln\left(\frac{4}{\delta}\right)\right)} \;\leq\; L_{\mathcal{D}}(\mathcal{H}_{lin}) + O\left(\sqrt{\frac{p \ln m}{m}}\right)$$

$\mathcal{H}_{lin}$ shatters $p + 1$ points

- Verify that $\mathcal{H}$ some set of shatters $p + 1$ points:

- Take the set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p, \mathbf{0}\}$ $\ldots \mathbf{e}_i = i^{th}$ standard basis vector

Linear classifiers cannot shatter $p+2$ points

- Proof by contradiction. Assume $x_1, x_2, \ldots, x_{p+2} \in \mathbb{R}^p$ can be shattered.
- Consider the set of $p+1$ linear equations

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_{p+2} \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{p+2} \end{pmatrix} = \mathbf{0}$$

  - $p+2$ variables and $p+1$ equations $\implies$ there is a solution $(a_1, \ldots, a_{p+2}) \neq \mathbf{0}$
  - Let $I_+ = \{i : a_i > 0\}$ and $I_- = \{i : a_i < 0\}$. Verify that

$$\sum_{i \in I_+} a_i = \sum_{i \in I_-} |a_i| \qquad \text{and} \qquad \sum_{i \in I_+} a_i x_i = \sum_{i \in I_-} |a_i| x_i$$

Linear classifiers cannot shatter $p + 2$ points

- Assuming points can be shattered, there is $w, b \in \mathcal{H}$ such that

$$\langle w, x_i \rangle + b \begin{cases} > 0 & \text{for } i \in I_+ \\ < 0 & \text{for } i \in I_- \end{cases}$$

- Hence

$$\begin{aligned} 0 < \sum_{i \in I_+} a_i(\langle w, x_i \rangle + b) &= \left\langle w, \sum_{i \in I_+} a_i x_i \right\rangle + b \sum_{i \in I_+} a_i \\ &= \left\langle w, \sum_{i \in I_-} |a_i| x_i \right\rangle + b \sum_{i \in I_-} |a_i| = \sum_{i \in I_-} |a_i|(\langle w, x_i \rangle + b) < 0 \end{aligned}$$

- Contradiction $(0 < 0) \implies \mathcal{H}$ cannot shatter $p + 2$ points

VC dimension of 1-nearest neighbour

- Recall 1-NN predictor $\widehat{h}_S(x) = y_{\pi_1(x)}$ $\qquad$ $\pi_1(x) =$ NN of x in S

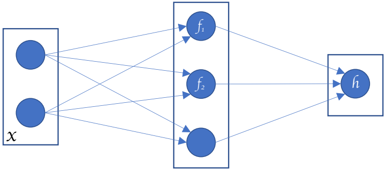
- Define hypothesis class of 1-NN as

$$\mathcal{H}_{1-NN} = \left\{ h_S(x) = y_{\pi_1(x)} \mid S \in (\mathcal{X} \times \mathcal{Y})^m, m \in \mathbb{N} \right\}$$

- Claim: $\text{VCdim}(\mathcal{H}_{1-NN}) = \infty$
  - To shatter any set C of size m, use predictors $h_{S_1}, \ldots, h_{S_{2^m}}$ where $S_1, \ldots, S_{2^m}$ has features same as C and all the 2^m labellings

- No uniform convergence for 1-NN for finite m

VC dimension of 2-layers neural network



- Consider simplified 2-layer network
  - Input: $x \in \mathbb{R}^p$
  - N units in the hidden layer, each corresponding to function

$$f_i(x) = \text{sign}(\langle w_i, x\rangle + b_i), \qquad i = 1, \ldots, N.$$

    Define $f(x) = (f_1(x), \ldots, f_N(x)) \in \{\pm 1\}^N$

  - Output: $h(x) = \text{sign}(\langle w, f(x)\rangle + b)$

- $\mathcal{H} = \{h(x) \;:\; \text{parameterised by } w \in \mathbb{R}^N,\; w_1, \ldots, w_N \in \mathbb{R}^p,\; b, b_1, \ldots, b_N \in \mathbb{R}\}$

VC dimension of 2-layers neural network

- $\mathrm{VCdim}(\mathcal{H}) = O(pN \log_2(pN))$
- Generalisation error bound for ERM over $\mathcal{H}$: w.p. $1-\delta$

$$L_{\mathcal{D}}(\widehat{h}) \;\leq\; L_{\mathcal{D}}(\mathcal{H}) + O\left(\sqrt{\frac{pN\ln(pN)\ln m + \ln\frac{1}{\delta}}{m}}\right)$$

- Key idea for computing $\mathrm{VCdim}(\mathcal{H})$:
  - Neural network is combination of several linear classifiers
  - Need ways to compute growth function of combinations

Growth function of combined classes

Lemma VC.2 (Concatenating classifiers)

$\mathcal{G}' \subseteq \mathcal{Y}'^{\mathcal{X}}$ *and* $\mathcal{G}'' \subseteq \mathcal{Y}''^{\mathcal{X}}$ *be two classes. Define* $\mathcal{G} = \mathcal{G}' \times \mathcal{G}'' \subseteq (\mathcal{Y}' \times \mathcal{Y}'')^{\mathcal{X}}$ *as*

$$\mathcal{G} = \{(g'(\cdot), g''(\cdot)) \ : \ g' \in \mathcal{G}', g'' \in \mathcal{G}''\}$$

Growth functions satisfy $\tau_{\mathcal{G}}(m) \leq \tau_{\mathcal{G}'}(m)\tau_{\mathcal{G}''}(m)$

Lemma VC.3 (Composition of classifiers)

$\mathcal{G}' \subseteq \mathcal{Y}^{\mathcal{X}}$ *and* $\mathcal{G}'' \subseteq \mathcal{Z}^{\mathcal{Y}}$ *be two classes. Define* $\mathcal{G} = \mathcal{G}'' \circ \mathcal{G}' \subseteq \mathcal{Z}^{\mathcal{X}}$ *as*

$$\mathcal{G} = \{g''(g'(\cdot)) \ : \ g' \in \mathcal{G}', g'' \in \mathcal{G}''\}$$

Growth functions satisfy $\tau_{\mathcal{G}}(m) \leq \tau_{\mathcal{G}'}(m)\tau_{\mathcal{G}''}(m)$

Computing VC dimension of neural network

- Hypothesis class: $\mathcal{H} = \mathcal{H}' \circ (\mathcal{H}_1 \times \ldots \mathcal{H}_N)$
  - $\mathcal{H}_i \subseteq \{\pm 1\}^{\mathbb{R}^p}$ hypothesis class corresponding to i-th hidden unit
  - $\mathrm{VCdim}(\mathcal{H}_1) = \ldots = \mathrm{VCdim}(\mathcal{H}_N) = p + 1$

$$\tau_{\mathcal{H}_i}(m) \leq \left(\frac{em}{p+1}\right)^{p+1} < (me)^{p+1}$$

  - $\mathcal{H}' \subseteq \{\pm 1\}^{\mathbb{R}^N}$ hypothesis class corresponding to output unit
  - $\mathrm{VCdim}(\mathcal{H}') = N + 1$

$$\tau_{\mathcal{H}'}(m) \leq \left(\frac{em}{N+1}\right)^{N+1} < (me)^{N+1}$$

Computing VC dimension of neural network

- Using growth function bound for compositions

$$\begin{aligned}\tau_{\mathcal{H}}(m) &\leq \tau_{\mathcal{H}'}(m) \cdot \tau_{\mathcal{H}_1}(m) \cdot \ldots \cdot \tau_{\mathcal{H}_N}(m) \\ &< (me)^{N(p+1)+N+1} \\ &< m^{8pN} \qquad \text{for } m > e, p \geq 1\end{aligned}$$

- Recall: $\mathrm{VCdim}(\mathcal{H}) = d \implies 2^d = \tau_{\mathcal{H}}(d)$ and $\tau_{\mathcal{H}}(m) < 2^m$ for all $m > d$
  - Find m such that $\tau_{\mathcal{H}}(m) < 2^m \implies d < m$
- For $c > 0$, $x > \max\{2, 3c\}$ and $m = 3cx \log_2 x \implies 2^m > m^{cx}$ (try to verify this)
  - Assume $x = Np > 24$ (here, $c = 8$):
    $\tau_{\mathcal{H}}(m) < m^{8pN} < 2^m$ for $m = 24Np \log_2(Np) \implies \mathrm{VCdim}(\mathcal{H}) = O(pN \log_2(pN))$
  - For $Np \leq 24$ (means $Np = O(1)$):
    $\tau_{\mathcal{H}}(m) < m^{8 \cdot 24} < 2^m$ for $m >$ large enough constant $\implies \mathrm{VCdim}(\mathcal{H}) = O(1)$