Big Data Analytics

Learning theory

Souhaib Ben Taieb

March 14, 2021

University of Mons

Table of contents

Table of contents

The VC dimension

The VC dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{\text{VC}}(\mathcal{H})$ or d_{VC}, is the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$, i.e. **the most points $\mathcal{H}$ can shatter**. If $m_{\mathcal{H}}(N) = 2^N$ for all N, then $d_{\text{VC}} = \infty$.

- $d_{\text{VC}}(\mathcal{H}) \leq N' \implies$??
- $d_{\text{VC}}(\mathcal{H}) > N' \implies$??

The VC dimension

The VC dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{\mathrm{VC}}(\mathcal{H})$ or d_{VC}, is the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$, i.e. **the most points** $\mathcal{H}$ **can shatter**. If $m_{\mathcal{H}}(N) = 2^N$ for all N, then $d_{\mathrm{VC}} = \infty$.

- $d_{\mathrm{VC}}(\mathcal{H}) < N' \implies N'$ is a break point for $\mathcal{H}$
- $d_{\mathrm{VC}}(\mathcal{H}) \geq N' \implies \mathcal{H}$ can shatter at least N' points
  ($d_{\mathrm{VC}}(\mathcal{H}) \geq k \implies k$ is not a break point for $\mathcal{H}$)

In other words, $d_{\mathrm{VC}}(\mathcal{H}) = k^* - 1$ where k^* is the smallest break point for $\mathcal{H}$.

The growth function

The growth function

In terms of a break point k:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

In terms of the VC dimension d_{VC}:

$$m_{\mathcal{H}}(N) \leq \underbrace{\sum_{i=0}^{d_{\mathrm{VC}}} \binom{N}{i}}_{\text{maximum power is } N^{d_{\mathrm{VC}}}}$$

We can prove by induction that

$$m_{\mathcal{H}}(N) \leq N^{d_{\mathrm{VC}}} + 1,$$

i.e d_{VC} is the order of the polynomial bound on $m_{\mathcal{H}}(N)$.

Examples

What is $d_{\mathsf{VC}}(\mathcal{H})$ for the following examples?

- $\mathcal{H}$ is positive rays
- $\mathcal{H}$ is 2D perceptrons
- $\mathcal{H}$ is convex sets

Examples

Examples

- $\mathcal{H}$ is positive rays:

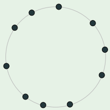
$$d_{\mathrm{VC}} = 1$$

- $\mathcal{H}$ is 2D perceptrons:

$$d_{\mathrm{VC}} = 3$$

- $\mathcal{H}$ is convex sets:

$$d_{\mathrm{VC}} = \infty$$



VC dimension and learning

The VC inequality states that

$$\mathbb{P}[|E_{\text{in}} - E_{\text{out}}| > \epsilon] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}.$$

Equivalently, with probability at least $1 - \delta$, we have

$$|E_{\text{out}} - E_{\text{in}}| \leq \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}} = \Omega(N, \mathcal{H}, \delta).$$

The VC *generalization bound* is given by:

$$E_{\text{out}} \leq E_{\text{in}} + \Omega(N, \mathcal{H}, \delta).$$

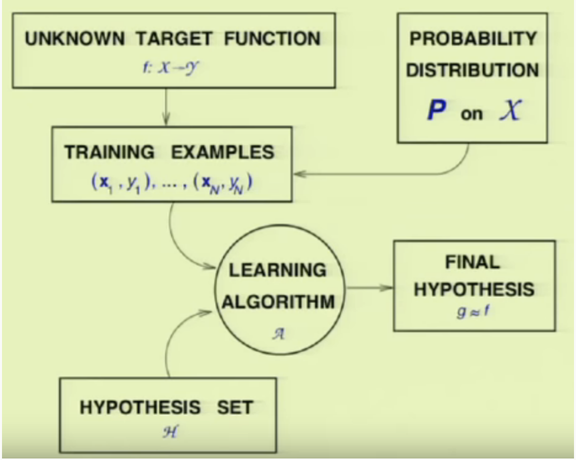
VC dimension and learning

- $d_{\mathrm{VC}}(\mathcal{H}) < \infty$ (finite)
  - $\rightarrow$ $m_{\mathcal{H}}(N)$ is bounded by a polynomial in N
  - $\rightarrow$ $\ln m_{\mathcal{H}}(2N)$ grows logarithmically in N regarless of the order of the polynomial, and so, it will be crushed by the $\frac{1}{N}$ factor
- $d_{\mathrm{VC}}(\mathcal{H}) = \infty$ (infinite)
  - $\rightarrow$ $m_{\mathcal{H}}(N)$ is exponential in N

VC dimension and learning

$$d_{\text{VC}}(\mathcal{H}) \text{ finite } \implies g \in \mathcal{H} \text{ will generalize}$$

How is this statement related to the learning diagram?



VC dimension and learning

VC dimension and learning

$d_{\mathrm{VC}}(\mathcal{H})$ is finite $\implies$ $g \in \mathcal{H}$ will generalize

- Independent of the **learning algorithm**
- Independent of the **input distribution**
- Independent of the **target function**

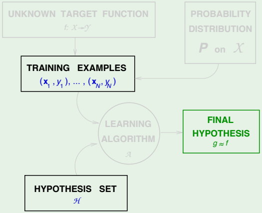
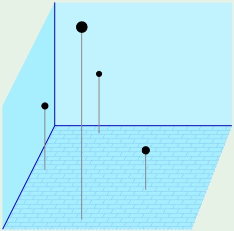


Table of contents

VC dimension of perceptrons

For $d = 2$, $d_{\mathrm{VC}} = 3$

In general, $\boxed{d_{\mathrm{VC}} = d + 1}$



We will prove two directions:

$$d_{\mathrm{VC}} \leq d + 1$$

$$d_{\mathrm{VC}} \geq d + 1$$

VC dimension of perceptrons - first direction

Let us construct a set of $N = d + 1$ points in $\mathbb{R}^d$ shattered by the perceptron.

$$\mathrm{X} = \begin{bmatrix} — \mathbf{x}_1^\intercal — \\ — \mathbf{x}_2^\intercal — \\ — \mathbf{x}_3^\intercal — \\ \vdots \\ — \mathbf{x}_{d+1}^\intercal — \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & & 0 \\ & \vdots & & \ddots & 0 \\ 1 & 0 & \dots & 0 & 1 \end{bmatrix}$$

X is invertible

VC dimension of perceptrons - first direction

Can we shatter this data set?

For any $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$, can we find a vector $\mathbf{w}$ satisfying

$$\text{sign}(\mathrm{X}\mathbf{w}) = \mathbf{y}$$

Easy! Just make $\mathrm{X}\mathbf{w} = \mathbf{y}$

which means $\mathbf{w} = \mathrm{X}^{-1}\mathbf{y}$

VC dimension of perceptrons - first direction

We can shatter these $d+1$ points

This implies what?

[a] $d_{\text{VC}} = d+1$

[b] $d_{\text{VC}} \geq d+1$

[c] $d_{\text{VC}} \leq d+1$

[d] No conclusion

VC dimension of perceptrons - first direction

We can shatter these $d+1$ points

This implies what?

[a] $d_{\mathrm{VC}} = d+1$

[b] $d_{\mathrm{VC}} \geq d+1$ ✓

[c] $d_{\mathrm{VC}} \leq d+1$

[d] No conclusion

VC dimension of perceptrons - second direction

Now, to show that $d_{\text{vc}} \leq d + 1$

We need to show that:

[a] There are $d + 1$ points we cannot shatter

[b] There are $d + 2$ points we cannot shatter

[c] We cannot shatter *any* set of $d + 1$ points

[d] We cannot shatter *any* set of $d + 2$ points

VC dimension of perceptrons - second direction

Now, to show that $d_{\text{vc}} \le d + 1$

We need to show that:

[a] There are $d+1$ points we cannot shatter

[b] There are $d+2$ points we cannot shatter

[c] We cannot shatter *any* set of $d+1$ points

[d] We cannot shatter *any* set of $d+2$ points ✓

VC dimension of perceptrons - second direction

Take any $d+2$ points. For any $d+2$ points, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}, \boldsymbol{x}_{d+2}$, we have more points than dimensions (since each $\boldsymbol{x}_j \in \mathbb{R}^{d+1}$, $j = 1, \ldots, d+2$).

When we have more vectors than dimensions, these vectors must be linearly dependent. In other words, we must have

$$\boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{x}_i,$$

where $a_i \in \mathbb{R}$.

Furthermore, since the first coordinate of $\boldsymbol{x}_j$ is always one, we must have

$$\boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{x}_i,$$

where <u>not all the a_i's are zeros</u>.

VC dimension of perceptrons - second direction

Recall that we have

$$\boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{x}_i,$$

where not all the a_i's are zeros, and $j = 1, \ldots, d+2$.

Let us construct a dichotomy that the perceptron cannot implement:

- $y_i = \text{sign}(a_i)$ if $\boldsymbol{x}_i$ has non-zero a_i
  (for zero a_i, you can choose $+1$ or -1 for y_i)
- $y_j = -1$ for $\boldsymbol{x}_j$

. Why?

VC dimension of perceptrons - second direction

$$\boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{x}_i, \implies \boldsymbol{w}^T \boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{w}^T \boldsymbol{x}_i$$

In other words, the signal for $\boldsymbol{x}_j$ is a linear combination of the signals for the $\boldsymbol{x}_i$ with coefficients a_i.

- We know that $y_i = \text{sign}(\boldsymbol{w}^T \boldsymbol{x}_i)$. However, for $a_i \neq 0$, we forced $y_i = \text{sign}(a_i)$. In other words, we have $y_i = \text{sign}(\boldsymbol{w}^T \boldsymbol{x}_i) = \text{sign}(a_i)$, or equivalently $a_i \boldsymbol{w}^T \boldsymbol{x}_i > 0$
- This forces $\sum_{i \neq j} a_i \boldsymbol{w}^T \boldsymbol{x}_i > 0$ since $a_i \boldsymbol{w}^T \boldsymbol{x}_i > 0$ for all $a_i \neq 0$, and all $a_i = 0$ do not contribute to the sum.
- However, $\sum_{i \neq j} a_i \boldsymbol{w}^T \boldsymbol{x}_i = \boldsymbol{w}^T \boldsymbol{x}_j$. Therefore, we must have $y_j = \text{sign}(\boldsymbol{w}^T \boldsymbol{x}_j) = +1$.

$\rightarrow$ It is impossible to obtain $y_j = -1$. We found a data set that cannot be shattered by the perceptron.

VC dimension of perceptrons - summary

Putting it together

We proved $d_{\text{VC}} \leq d + 1$ and $d_{\text{VC}} \geq d + 1$

$$\boxed{d_{\text{VC}} = d + 1}$$

What is $d + 1$ in the perceptron?

It is the number of parameters $w_0, w_1, \cdots, w_d$

Table of contents

1. Degrees of freedom

Parameters create degrees of freedom

# of parameters: **analog** degrees of freedom

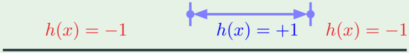
d_{VC}: equivalent **'binary'** degrees of freedom

Examples

The usual suspects

Positive rays ($d_{\mathrm{VC}} = 1$):

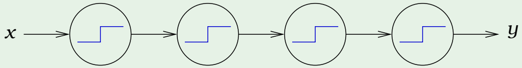
Positive intervals ($d_{\mathrm{VC}} = 2$):



Not just parameters

Not just parameters

Parameters may not contribute degrees of freedom:



d_{VC} measures the **effective** number of parameters

- $d = 1$
- w_0, w_1 (four times) $\rightarrow$ eight parameters
- Not eight degrees of freedom

Sample complexity - Number of data points needed

The VC inequality gives

$$\mathbb{P}[|E_{\text{out}} - E_{\text{in}}| > \epsilon] \leq 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\epsilon^2 N}.$$

This can be rephrased as follows. Pick a **tolerance level** δ, for example $\delta = 0.01$, and assert with probability at least $1 - \delta$ that

$$|E_{\text{out}} - E_{\text{in}}| \leq \sqrt{\frac{8}{N}\ ln\left(\frac{4 m_{\mathcal{H}}(2N)}{\delta}\right)}.$$

If we want the generalization to be **at most** ε, it suffices to make

$$\sqrt{\frac{8}{N}\ ln\left(\frac{4 m_{\mathcal{H}}(2N)}{\delta}\right)} \leq \varepsilon$$

Sample complexity - Number of data points needed

It folllows that

$$N \geq \frac{8}{\varepsilon^2} ln \left(\frac{4m_{\mathcal{H}}(2N)}{\delta} \right)$$

suffices to obtain generalization error at most ε (with probability at least $1 - \delta$).

Using the VC dimension, we can write

$$N \geq \frac{8}{\varepsilon^2} ln \left(\frac{4((2N)^{d_{\mathrm{VC}}} + 1)}{\delta} \right).$$

Sample complexity - Number of data points needed

- The **sample complexity** denotes how many training examples N are needed to achieve a certain generalization performance, specified by two parameters ϵ and δ.
- The *error tolerance* ε determines the allowed generalization error.
- The *confidence parameter* δ determines how often the error tolerance ε is violated.
- How fast N grows as ε and δ become smaller indicates how much data is needed to get good generalization.

Sample complexity - Number of data points needed

Suppose we have $d_{\text{VC}} = 3$ and want the generalization error to be at most 0.1 with confidence 90%. How big a dataset do we need?

$$N \geq \frac{8}{0.1^2} ln \left(\frac{4((2N)^3 + 1)}{0.1} \right)$$

Using an iterative process, we converge to $N \approx 30,000$.

- $d_{\text{VC}} = 3 \implies N \approx 30,000$
- $d_{\text{VC}} = 4 \implies N \approx 40,000$
- $d_{\text{VC}} = 5 \implies N \approx 50,000$
- ...

The inegality suggests that N is approximately proportional to d_{VC}, as has been observed in practice. The constant of proportionality it suggests is $10,000$, which is a gross overestimate; a more practical constant of proportionality is closer to 10.

Sample complexity - Number of data points needed

2. Number of data points needed

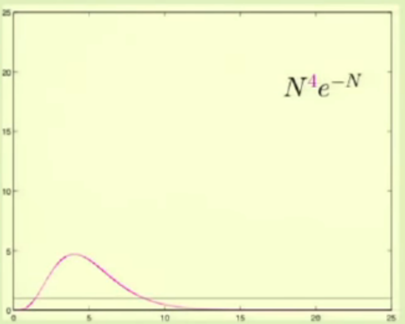
Two small quantities in the VC inequality:

$$\mathbb{P}\left[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\right] \leq \underbrace{4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}}_{\delta}$$

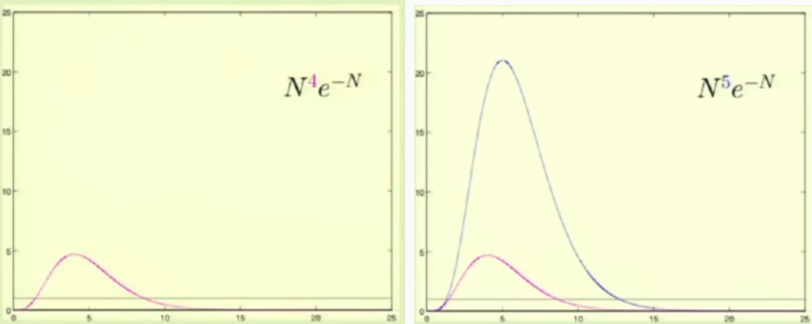
If we want certain ϵ and δ, how does N depend on d_{VC}?

Let us look at $N^{d}e^{-N}$

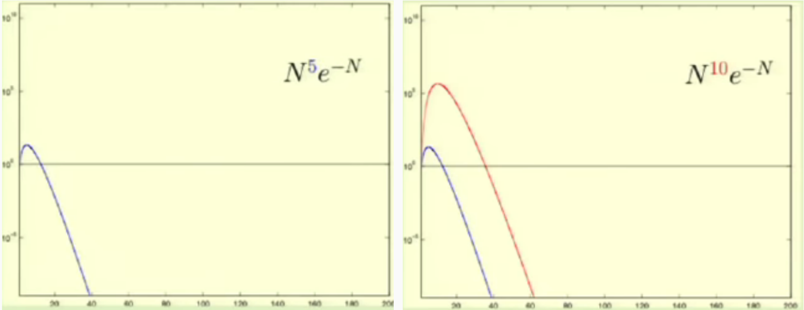
Sample complexity - Number of data points needed



Sample complexity - Number of data points needed

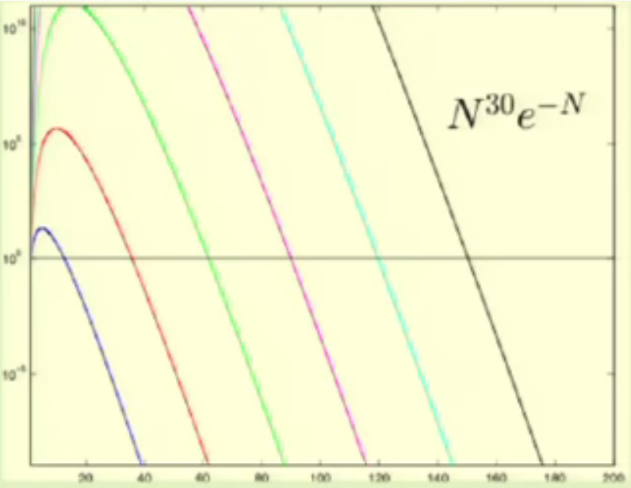


Sample complexity - Number of data points needed



(log scale)

Sample complexity - Number of data points needed



(log scale)

Fix $N^{d}e^{-N}$ to a small value. How does N change with d?

Sample complexity - Number of data points needed

- The previous observation is in terms of the **bound** which is based on theoretical derivations.
- The problem is that we can have $P_1 \leq A$ and $P_2 \leq B$ with $A \leq B$, while $P_1 \geq P_2$.
- We would like to make a statement about the actual quantity (not the bounds).
- How many observations N do I need to arrive in the comfort zone (bound less than one)?
  - It depends on many parameters (ϵ, δ, etc).
  - Practical observation: the actual quantity we are trying to bound follows the same monotonicity as the bound.
  - Rule of thumb: $N \geq 10\ d_{\text{VC}}$