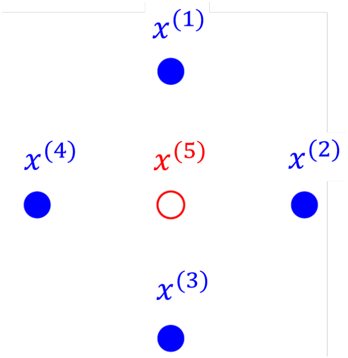
Introduction to Machine Learning

PAC Learning and VC Dimension



Learning goals

- Know PAC learning
- Know that there is no "universal" learner which works on every task (no free lunch)
- Know that complexity of a hypothesis space can be measured by VC dimension
- Know that a hypothesis space is PAC learnable iff it has finite VC dimension

PAC LEARNING

A hypothesis space $\mathcal{H}$ over a data space $\mathcal{X} \times \mathcal{Y}$ is agnostic PAC learnable, if there exists a function $n_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property:
For every ϵ, δ and for **every data distribution** $\mathbb{P}_{xy}$, when running the algorithm on $n \geq n_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples from $\mathbb{P}_{xy}$, the learner returns a model $\hat{f}$ such that, which probability at least $1 - \delta$ (over the choice of the n training examples), it holds

$$\mathcal{R}(\hat{f}) \leq \min_{f \in \mathcal{H}} \mathcal{R}(f) + \epsilon$$

- PAC = Probably (δ) Approximately (ϵ) Correct learning.
- It implies that our learner, given enough samples, always return an "approximately" correct function.
- $n_{\mathcal{H}}(\epsilon, \delta)$ is the sample complexity of our learner, how many samples do we need to obtain a PAC solution.
- PAC gives us finite-sample bounds on **arbitrary** data distributions!

FINITE SPACES ARE AGNOSTIC PAC LEARNABLE

Every finite hypothesis space is agnostic PAC learnable, using empirical risk minimization, with sample complexity

$$n_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{2 \log(2|\mathcal{H}|/\delta)}{\epsilon^2} \rceil$$

- Proof: See "Understanding Machine Learning", chapter 4.
- While many spaces $\mathcal{H}$ are infinite, we can discretize them to get a certain impression of their sample complexity.
  - Assume that parameters are floats on a 32bit machine, then there are at most 2^{32} different values for each parameter.
  - If we have d parameters, that's 2^{32d} functions in $\mathcal{H}$.
  - That gives us $n_{\mathcal{H}} \leq \frac{64d + \log 2/\delta}{\epsilon^2}$
  - For 10 parameters and $\epsilon = \delta = 0.05$ that is ca. $n = 250.000$.

NO FREE LUNCH

Let $\mathcal{I}$ be any learning algorithm for binary classification, with respect to 0-1 loss over domain $\mathcal{X}$. Let the training set size $n \leq |\mathcal{X}|/2$. Then a data distribution $\mathbb{P}_{xy}$ exists, such that

1. There exists a function f with $\mathcal{R}_{\mathbb{P}}(f) = 0$
2. With probability at least $1/7$ (over the choice of $\mathcal{D}_n$) we have that $\mathcal{R}_{\mathbb{P}}(I(\mathcal{D}_n)) \geq 1/8$.

- Proof: See "Understanding Machine Learning", chapter 5.
- This implies that for every learner, there is a task on which it fails, even though it could be learned by another learner. So there is no "universal" learner.
- This implies that if $\mathcal{X}$ is infinite, the space $\mathcal{H}$ of all functions is not PAC learnable. Learning without any assumptions does not work.

VC DIMENSION

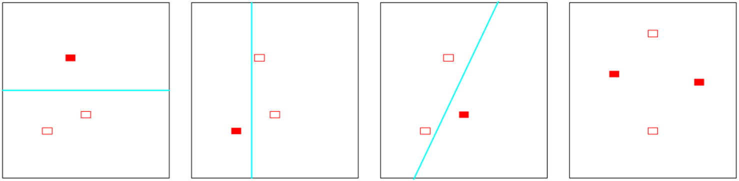
A general measure of the complexity of a function space is the **Vapnik-Chervonenkis (VC)** dimension.

The **VC dimension** of a class of binary-valued functions $\mathcal{H} = \{h : \mathcal{X} \rightarrow \{0, 1\}\}$ is defined to be the largest number of points in $\mathcal{X}$ (in some configuration) that can be "shattered" by members of $\mathcal{H}$. We write $VC_p(\mathcal{H})$, where p denotes the dimension of the input space.

A set of points is said to be **shattered** by a class of functions if a member of this class can perfectly separate them no matter how we assign binary labels to the points.

Note: If the VC dimension of a hypothesis class is d, it does not mean that **all** sets of size d can be shattered. Rather, it simply means that there is at least **one** such set which can be shattered and **no** set of size $d + 1$ which can be shattered.

VC DIMENSION OF HYPERPLANES



For $\mathbf{x} \in \mathbb{R}^2$, the class of linear indicator functions
$\mathcal{H} = \{h : \mathbb{R}^2 \to \{0, 1\} \mid h(\mathbf{x} \mid \theta_0, \boldsymbol{\theta}) = \mathbb{I}[\mathbf{x}^T\boldsymbol{\theta} - \theta_0 > 0]\}$

- can shatter 3 points: No matter how we assign labels to the configuration of three points shown above, we can find a linear line separating them perfectly;
- cannot shatter a configuration of 4 points.

Hence, $VC_2(\mathcal{H}) = 3$.

VC DIMENSION OF HYPERPLANES

Theorem: The VC dimension of the class of homogeneous halfspaces, $\mathcal{H} = \{h : \mathbb{R}^p \to \{-1, 1\} \mid h(\mathbf{x}) = \text{sign}(\mathbf{x}^T\boldsymbol{\theta})\}$, in $\mathbb{R}^p$ is p.

Proof: p as a lower bound:
Consider the set of standard basis vectors $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \dots, \mathbf{e}^{(p)}$ in $\mathbb{R}^p$. For every possible labeling $y^{(1)}, y^{(2)}, \dots, y^{(p)} \in \{-1, +1\}$, if we set $\boldsymbol{\theta} = (y^{(1)}, y^{(2)}, \dots, y^{(p)})^\top$, then $h(\mathbf{e}^{(i)}) = \text{sgn}\left(\boldsymbol{\theta}^\top \mathbf{e}^{(i)}\right) = \text{sgn}\left(y^{(i)}\right) = y^{(i)}$ for all i. Therefore, the p points are shattered.

p as an upper bound:

- Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(p+1)}$ be a set of $p + 1$ vectors in $\mathbb{R}^p$.
- Because any set of $p + 1$ vectors in $\mathbb{R}^p$ is linearly dependent, there must exist real numbers $a_1, a_2, \dots, a_{p+1} \in \mathbb{R}$, not all of them zero, such that $\sum_{i=1}^{p+1} a_i\mathbf{x}^{(i)} = 0$.

VC DIMENSION OF HYPERPLANES

Let $I = \{i : a_i > 0\}$ and $J = \{j : a_j < 0\}$. Either I or J is nonempty. If we assume both I and J are nonempty, then:

- $\sum_{i \in I} a_i \mathbf{x}^{(i)} = \sum_{j \in J} |a_j| \mathbf{x}^{(j)}$
- Let us assume $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p+1)}$ are shattered by $\mathcal{H}$.
- There must exist a vector $\boldsymbol{\theta} \in \mathbb{R}^p$ such that

$$\begin{aligned} h(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}) = 1 \quad &\Leftrightarrow \quad \boldsymbol{\theta}^\top \mathbf{x}^{(i)} > 0 \quad \text{for all } i \in I \\ h(\mathbf{x}^{(j)} \mid \boldsymbol{\theta}) = -1 \quad &\Leftrightarrow \quad \boldsymbol{\theta}^\top \mathbf{x}^{(j)} < 0 \quad \text{for all } j \in J \end{aligned}$$

VC DIMENSION OF HYPERPLANES

- This implies

$$
\begin{aligned}
0 < \sum_{i \in I} a_i \cdot \boldsymbol{\theta}^\top \mathbf{x}^{(i)} &= \left(\sum_{i \in I} a_i \mathbf{x}^{(i)} \right)^\top \boldsymbol{\theta} \\
&= \left(\sum_{j \in J} |a_j| \, \mathbf{x}^{(j)} \right)^\top \boldsymbol{\theta} = \sum_{j \in J} |a_j| \cdot \boldsymbol{\theta}^\top \mathbf{x}^{(j)} < 0
\end{aligned}
$$

  which is a contradiction.

On the other hand, if we assume J (respectively, I) is empty, then the rightmost (respectively, leftmost) inequality should be replaced by an equality, which is still a contradiction.

□

VC DIMENSION OF HYPERPLANES

Theorem: The VC dimension of the class of non-homogeneous halfspaces, $\mathcal{H} = \{h : \mathbb{R}^p \to \{-1, 1\} \mid h(\mathbf{x} \mid \boldsymbol{\theta}) = \text{sign}(\mathbf{x}^T\boldsymbol{\theta} + \theta_0)\}$, in $\mathbb{R}^p$ is $p + 1$.

Proof: $p + 1$ as a lower bound: Similar to the proof of the previous theorem, the set of basis vectors and the origin, that is, $0, \mathbf{e}^{(1)}, \dots, \mathbf{e}^{(p)}$ can be shattered by non-homogenous halfspaces.

$p + 1$ as an upper bound:

- Assume that $p + 2$ vectors $\mathbf{x}^{(1)}, \dots \mathbf{x}^{(p+2)}$ are shattered.
- If we denote $\tilde{\boldsymbol{\theta}} = (\theta_0, \dots, \theta_p)^\top \in \mathbb{R}^{p+1}$, where θ_0 is the bias/intercept, and $\tilde{\boldsymbol{x}} = (1, x_1, \dots x_p)^\top \in \mathbb{R}^{p+1}$, then $h(\mathbf{x} \mid \boldsymbol{\theta}) = \text{sgn}\left(\mathbf{x}^\top\boldsymbol{\theta} + \theta_0\right) = \text{sgn}\left(\tilde{\boldsymbol{x}}^\top\tilde{\boldsymbol{\theta}}\right)$. Any affine function in $\mathbb{R}^p$ can be rewritten as a homogeneous linear function in $\mathbb{R}^{p+1}$.
- By the previous theorem, the set of homogeneous halfspaces in $\mathbb{R}^{p+1}$ cannot shatter any $p + 2$ points. Contradiction.

VC DIMENSION OF RECTANGLES

Example: Let $\mathcal{H}$ be the class of axis-aligned rectangles in $\mathbb{R}^2$

$$\mathcal{H} = \left\{ h_{(a_1, a_2, b_1, b_2)} : \mathbb{R}^2 \to \{0, 1\} : a_1 \leq a_2 \text{ and } b_1 \leq b_2 \right\}$$

where

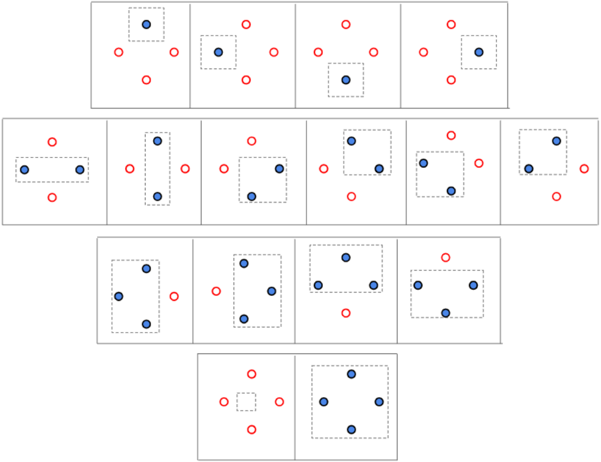
$$h_{(a_1, a_2, b_1, b_2)}(\boldsymbol{x}) = \begin{cases} 1 & a_1 \leq x_1 \leq a_2 \text{ and } b_1 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases}$$

Claim: $VC_2(\mathcal{H}) = 4$

Proof: (next slide)

VC DIMENSION OF RECTANGLES

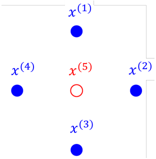
4 as a lower bound: There exists a set of 4 points that can be shattered.



VC DIMENSION OF RECTANGLES

4 as an upper bound: For any set of 5 points $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(5)} \in \mathbb{R}^2$:

- Assign the leftmost point (lowest x_1), rightmost point (highest x_1), highest point (highest x_2), and lowest point (lowest x_2) to class 1.
- The point not chosen, $\mathbf{x}^{(5)}$, is assigned to class 0.
- The rectangle must contain $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(4)}$.
- $\mathbf{x}^{(5)}$ is classified as 1 as well since its coordinates are within the intervals defined by the other 4.



Credit: Shalev-Shwartz, Ben-David. Understanding Machine Learning.

Therefore, the VC dimension of axis-aligned rectangles is 4. □

INFINITE VC DIMENSION

- We can show that if $\mathcal{H}$ has a VC dimension of $2n$, we cannot reliably learn $\mathcal{H}$ from only n examples. Similar to our first statement of the NFL theorem, we can now show that an adversarial data distribution exists, on which our learner fails. But also a function with risk 0 exists, but because of the shattering, this will be in $\mathcal{H}$.
- This directly implies that spaces of infinite VC dimension are not PAC learnable.

FUNDAMENTAL THEOREM OF PAC LEARNING

Assume hypothesis space $\mathcal{H}$, classification, and 0-1 loss. Then:

- $\mathcal{H}$ is agnostic PAC learnable if and only if $\mathcal{H}$ has finite VC dimension.
- Any ERM algorithm is a successful agnostic PAC learner for $\mathcal{H}$.
- For finite VC dimension d, the sample complexity is

$$C_1 \frac{d + \mathit{log}(1/\delta)}{\epsilon} \leq n_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \mathit{log}(1/\delta)}{\epsilon}$$

  with positive, absolute constants C_1 and C_2.

PROBABILISTIC BOUND ON TEST ERROR

Recall that the training error is an optimistic estimate of the generalization (or test) error. For a classification model with VC dimension d, 0-1-loss, and a training set of size n, the VC dimension can predict a probabilistic upper bound on the test error (with probability $1 - \delta$):

$$\mathcal{R}(f) \leq \mathcal{R}_{\text{emp}}(f) + \sqrt{\frac{1}{n}\left[d\left(\log\frac{2n}{d} + 1\right) - \log\frac{\delta}{4}\right]}$$

if the training data set is large enough ($d < n$ required).

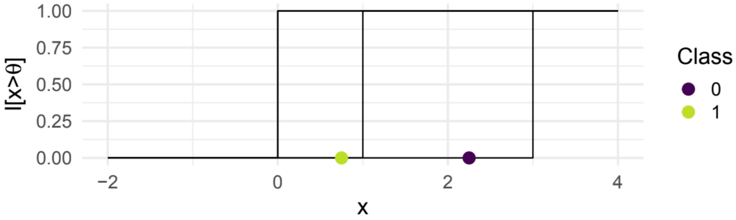
- So for finite VC dimension we could increase our sample size n so much, that the training error would give a close estimate of the test error, with high probability.
- Usually such a bound is too loose for practical relevance, and we would have to use an enormous amount of data.

VC DIMENSION VS NR OF PARAMETERS

Often, VC dimension of a hypothesis space increases with the number of learnable parameters. However, there are counterexamples.

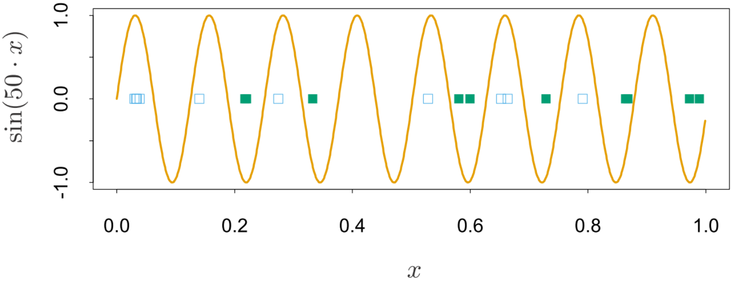
Example: A single-parametric threshold classifier ($h(x) = \mathbb{I}[x \geq \theta]$) has VC dimension 1:

- It can shatter a single point.
- It cannot shatter any set of 2 points (for every set of 2 numbers, if the smaller is labeled 1, the larger must also be labeled 1).



VC DIMENSION VS NR OF PARAMETERS

A single-parametric sine classifier $h(x) = \mathbb{I}[\sin(\theta x) > 0]$, for $x \in \mathbb{R}$, however, has infinite VC dimension, since it can shatter any set of points if the frequency θ is chosen large enough.



Credit: Trevor Hastie (2019). The Elements of Statistical Learning.