PAC LEARNING

Ronen Nir & Itay Evron

Based on slides by Shai Shalev-Schwarz

General classification setting

- Training data $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$
  - Sampled i.i.d. from an unknown distribution $\mathcal{D}$

- A learning algorithm $A(S)$ outputs a hypothesis $h_S\colon \mathcal{X} \to \mathcal{Y}$ from a hypothesis class $\mathcal{H}$.

Possible?

- Goal: minimize the generalization error $L_{\mathcal{D}}(h_S) \equiv \Pr_{(x,y)\sim\mathcal{D}}[h_S(x) \neq y]$

- Alternative: minimize the empirical error $L_S(h_S) \equiv \frac{1}{m} \sum_i \mathbb{I}\{h_S(x_i) \neq y_i\}$

ERM

- Our focus now:

What generalization guarantees can learning algorithms give?

Two assumptions for this tutorial

- An unknown labeling function $f\colon \mathcal{X} \to \mathcal{Y}$

$$\forall (x, y) \sim \mathcal{D}\colon\ y = f(x)$$

- The labeling function is realizable by the hypothesis class $\mathcal{H}$:

$$\exists h^* \in \mathcal{H} \text{ such that}$$

$$\forall x \sim \mathcal{D}\colon h^*(x) = f(x), \text{ or equivalently, } L_{\mathcal{D},f}(h^*) = 0$$

PAC LEARNABILITY

What generalization guarantees can learning algorithms give?

Probably Approximately Correct

- Consider for instance, $\mathcal{X} = \{x_1, x_2, x_3\}$, and $P(x_3) = 10^{-5}$.
- Cannot hope to find $h \in \mathcal{H}$ such that $L_{\mathcal{D}}(h) = 0$
  - With high probability, we never sample x_3.
  - Can't learn what you don't see!
- Approximately: Instead, we'd be happy with $L_{\mathcal{D}}(h) \leq \epsilon$
- Cannot guarantee even $L_{\mathcal{D}}(h) \leq \epsilon$
  - With some probability, we only sample x_3.
  - Again, can't learn what you don't see!
- Probably: Instead, we allow the algorithm to fail with probability $\delta \in (0,1)$

PAC learning

- A finite hypothesis class $\mathcal{H}$ is realizably PAC-learnable if there exist a function $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ and a learning algorithm A such that:
  - For every $\epsilon, \delta \in (0,1)$ and distribution $\mathcal{D}$ over $\mathcal{X}$
  - For every realizable labeling function $f: \mathcal{X} \to (0,1)$
  - The algorithm returns an (ϵ, δ)-probably approximately correct hypothesis

$$\Pr_{\substack{S\sim\mathcal{D}: \\ |S|\geq m_{\mathcal{H}}(\epsilon,\delta)}} \left[L_{\mathcal{D},f}\big(A(S)\big) \leq \epsilon\right] \geq 1-\delta$$

sample complexity

approximately

probably

PAC learning

- A finite hypothesis class $\mathcal{H}$ is realizably PAC-learnable if there exist a function $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ and a learning algorithm A such that:
  - For every $\epsilon, \delta \in (0,1)$ and distribution $\mathcal{D}$ over $\mathcal{X}$
  - For every realizable labeling function $f: \mathcal{X} \to (0,1)$
  - The algorithm returns an (ϵ, δ)-probably approximately correct hypothesis

$$\underbrace{\Pr_{\substack{S \sim \mathcal{D}: \\ |S| \geq m_{\mathcal{H}}(\epsilon,\delta)}}}_{\text{sample complexity}} \left[L_{\mathcal{D},f}\left(A(S)\right) \underbrace{\leq \epsilon}_{\text{approximately}} \right] \underbrace{\geq 1 - \delta}_{\text{probably}}$$

PAC learning

- A finite hypothesis class $\mathcal{H}$ is realizably PAC-learnable if there exist a function $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ and a learning algorithm A such that:
  - For every $\epsilon, \delta \in (0,1)$ and distribution $\mathcal{D}$ over $\mathcal{X}$
  - For every realizable labeling function $f: \mathcal{X} \to (0,1)$
  - The algorithm returns an (ϵ, δ)-probably approximately correct hypothesis

$$\Pr_{\substack{S \sim \mathcal{D}: \\ |S| \geq m_{\mathcal{H}}(\epsilon,\delta)}} \left[L_{\mathcal{D},f}\big(A(S)\big) \leq \epsilon \right] \geq 1 - \delta$$

Richer classes require more data!

- Using ERM:
  - Every finite hypothesis class is PAC learnable with $|S| \geq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$

PAC learning

- A finite hypothesis class $\mathcal{H}$ is realizably PAC-learnable if there exist a function $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ and a learning algorithm A such that:
  - For every $\epsilon, \delta \in (0,1)$ and distribution $\mathcal{D}$ over $\mathcal{X}$
  - For every realizable labeling function $f: \mathcal{X} \to (0,1)$
  - The algorithm returns an (ϵ, δ)-probably approximately correct hypothesis

$$\Pr_{\substack{S \sim \mathcal{D}: \\ |S| \geq m_{\mathcal{H}}(\epsilon,\delta)}} \left[L_{\mathcal{D},f}\big(A(S)\big) \leq \epsilon\right] \geq 1 - \delta$$

- Using ERM:
  - Every finite hypothesis class is PAC learnable with $|S| \geq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$
  - Example: Given a finite class with 100 hypotheses, how many samples are needed to $(0.1, 0.05)$-learn? A: $\left\lceil \frac{\ln(100/0.05)}{0.1} \right\rceil = \lceil 76.009 \rceil = 77$

PAC learning

- A finite hypothesis class $\mathcal{H}$ is realizably PAC-learnable if there exist a function $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ and a learning algorithm A such that:
  - For every $\epsilon, \delta \in (0,1)$ and distribution $\mathcal{D}$ over $\mathcal{X}$
  - For every realizable labeling function $f: \mathcal{X} \to (0,1)$
  - The algorithm returns an (ϵ, δ)-probably approximately correct hypothesis

$$\Pr_{\substack{S \sim \mathcal{D}: \\ |S| \geq m_{\mathcal{H}}(\epsilon, \delta)}} \left[L_{\mathcal{D},f}\big(A(S)\big) \leq \epsilon \right] \geq 1 - \delta$$

- Questions:
  - What if f is not realizable by $\mathcal{H}$ (agnostic case)? See the lecture
  - What if $\mathcal{H}$ is infinite, or is very large? Up next!

VC-DIMENSION

Infinite-size classes can be learnable!

Fundamental Theorem of Statistical Learning

- Let's start from the result (realizable case)
  - $\mathcal{H}$ is PAC learnable if and only if its VC dimension is finite.
  - Can learn with ERM using:

$$m_{\mathcal{H}}(\varepsilon, \delta) = \mathcal{O}\left(\frac{\mathrm{VCdim}(\mathcal{H}) \ln(1/\epsilon) + \log(1/\delta)}{\epsilon}\right)$$

- But what is the VC dimension?

VC Dimension: Idea

- The VC dimension of a hypothesis class $\mathcal{H}$ quantifies its capacity

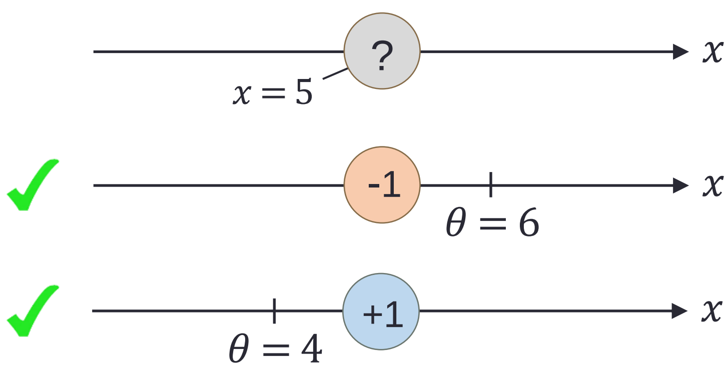
- VC dimension: the largest number of distinct data points, placed at positions of your choosing, such that every possible labeling of the points can be obtained by some hypothesis in $\mathcal{H}$
  - Larger number of points can be fitted $\equiv$ higher model complexity

Reminder: Threshold functions

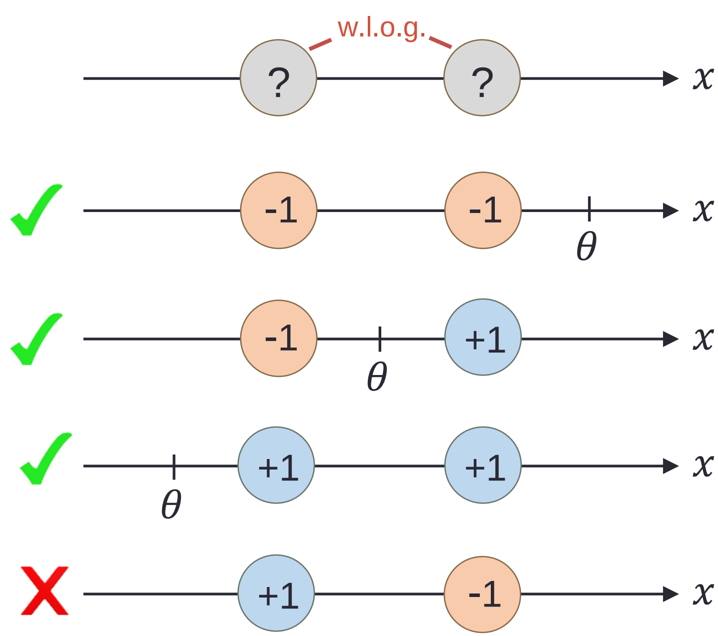
- In the lecture, we defined the following hypothesis class:

$$\mathcal{X} = \mathbb{R}, \qquad \mathcal{H} = \{x \mapsto \mathrm{sign}(x - \theta) : \theta \in \mathbb{R}\}$$

- There exists a single point which is shattered



- **Any** two points cannot be shattered



$$\boxed{\Longrightarrow \mathrm{VCdim}(\mathcal{H}) = 1}$$

VC Dimension: Formal Definition

- Let $C = \{x_1, \dots, x_{|C|}\} \subset \mathcal{X}$

$$\mathcal{H} \text{ shatters } C \quad \text{iff} \quad \forall y_1, \dots, y_{|C|} \in \mathcal{Y}\colon \exists h \in \mathcal{H}\colon \forall x_i \in C\colon h(x_i) = y_i$$

for any label assignment

there exists a hypothesis in $\mathcal{H}$

that completely agrees with it

- The VC dimension is the size of the largest set shattered by $\mathcal{H}$

$$\mathrm{VCdim}(\mathcal{H}) = \sup\{\, |C| \colon \mathcal{H} \text{ shatters } C \,\}$$

VC Dimension

To show that $\mathrm{VCdim}(\mathcal{H}) \triangleq \sup\{\, |C| : \mathcal{H} \text{ shatters } C \,\} = k$ we need to show:

1. There exists a set of k points that is shattered by $\mathcal{H}$

2. Any set of $k+1$ points cannot be shattered by $\mathcal{H}$

2. יהי S של $k+1$ נק' k', נראה כי עבור כל S כזה קיימת צביעה C של
$k+1$ הנק' בו כך שעבור כל $h \in H$ קיים $c \in C$ שאינו מסווג נכון.

usually harder to show

Exercise: Axis-aligned rectangles

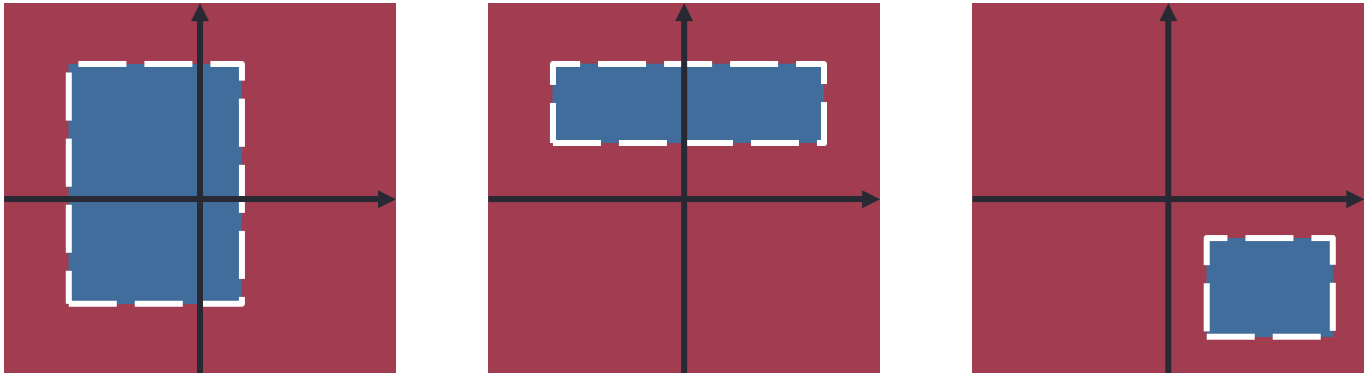
- Define the hypothesis class

$$\mathcal{X} = \mathbb{R}^2, \quad \mathcal{H}_{\text{rect}} = \{h_{(a_1,a_2,b_1,b_2)} : (a_1 < a_2) \wedge (b_1 < b_2)\},$$

where

$$h_{(a_1,a_2,b_1,b_2)}(\boldsymbol{x}) = 1 \text{ iff } x_1 \in [a_1, a_2] \text{ and } x_2 \in [b_1, b_2]$$

- Examples:



- Exercise: find $\mathrm{VCdim}(\mathcal{H}_{\text{rect}})$

Rule of thumb: in many cases (not all!), number of parameters = VC dimension

Exercise: Axis-aligned rectangles

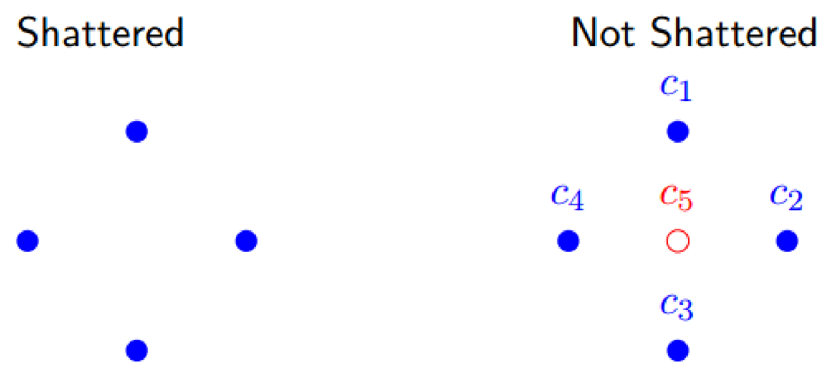
- Define the hypothesis class

$$\mathcal{X} = \mathbb{R}^2, \quad \mathcal{H}_{\text{rect}} = \{h_{(a_1,a_2,b_1,b_2)} : (a_1 < a_2) \wedge (b_1 < b_2)\},$$

where

$$h_{(a_1,a_2,b_1,b_2)}(\boldsymbol{x}) = 1 \text{ iff } x_1 \in [a_1, a_2] \text{ and } x_2 \in [b_1, b_2]$$

- Solution idea:



Exercise: Axis-aligned rectangles

- Define the hypothesis class

$$\mathcal{X} = \mathbb{R}^2, \quad \mathcal{H}_{\mathrm{rect}} = \{h_{(a_1,a_2,b_1,b_2)} : (a_1 < a_2) \wedge (b_1 < b_2)\},$$

where

$$h_{(a_1,a_2,b_1,b_2)}(\boldsymbol{x}) = 1 \text{ iff } x_1 \in [a_1, a_2] \text{ and } x_2 \in [b_1, b_2]$$

- Think: can the following set be shattered?

- Think: if so, how can $\mathrm{VCdim}(\mathcal{H}_{rect}) = 4$?

Exercise: Halfspaces

- Define the class of homogeneous halfspaces

$$\mathcal{X} = \mathbb{R}^d, \qquad \mathcal{H}_{\text{lin}}^d = \{\boldsymbol{x} \mapsto \operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x}) : \boldsymbol{w} \in \mathbb{R}^d\}$$

- Exercise:

1. Show that $\operatorname{VCdim}\left(\mathcal{H}_{\text{lin}}^d\right) \geq d$.

Exercise: Halfspaces

- Define the class of homogeneous halfspaces

$$\mathcal{X} = \mathbb{R}^d, \qquad \mathcal{H}_{\mathrm{lin}}^d = \{\boldsymbol{x} \mapsto \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x}) : \boldsymbol{w} \in \mathbb{R}^d\}$$

- Exercise:

1. Show that $\mathrm{VCdim}(\mathcal{H}_{\mathrm{lin}}^d) \geq d$.

   Solution idea: choose $C = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$

2. Extra (Q3b, Moed A, Spring 22):

   Show that any $d+1$ points cannot be shattered, i.e., $\mathrm{VCdim}(\mathcal{H}_{\mathrm{lin}}^d) = d$.

Exercise: Halfspaces

- Define the class of homogeneous halfspaces

$$\mathcal{X} = \mathbb{R}^d, \qquad \mathcal{H}_{\text{lin}}^d = \{\boldsymbol{x} \mapsto \text{sign}(\boldsymbol{w}^\top \boldsymbol{x}) : \boldsymbol{w} \in \mathbb{R}^d\}$$

- Exercise:

1. Show that $\text{VCdim}(\mathcal{H}_{\text{lin}}^d) \geq d$.

   Solution idea: choose $C = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$

2. Extra (Q3b, Moed A, Spring 22):

   Show that <u>any</u> $d+1$ points cannot be shattered, i.e., $\text{VCdim}(\mathcal{H}_{\text{lin}}^d) = d$.

Conclusion: for linear classifiers (on separable data), $m_{\mathcal{H}}(\varepsilon, \delta) = \Omega\left(\frac{d \ln(1/\epsilon) + \log(1/\delta)}{\epsilon}\right)$.

3. How does the number of features affect the sample complexity?

Summary

- Looked for guarantees of learning algorithms on the generalization error.
- Defined PAC learnability of $\mathcal{H}$ in the sense of:

$$\Pr_{\substack{S\sim\mathcal{D}:\\ |S|\geq m_{\mathcal{H}}(\epsilon,\delta)}} \left[L_{\mathcal{D},f}\Big(\underbrace{A(S)}_{h\in\mathcal{H}}\Big) \leq \epsilon\right] \geq 1-\delta$$

- In the finite realizable case, the required sample complexity is $\left\lceil\frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}\right\rceil$, which depends on $|\mathcal{H}|$.
- The VC dimension quantifies the capacity of infinite hypothesis classes.