Învățare Automată (Machine Learning)

Bogdan Alexe,

bogdan.alexe@fmi.unibuc.ro

Master Informatică, anul I, 2018-2019, cursul 5 https://www.eeml.eu

ML

Home | Advisory board | Program | Application | Registration | Venue

Eastern European Machine Learning Summer School

(previously TMLSS)

1-6 July 2019, Bucharest, Romania

Deep Learning and Reinforcement Learning

Important dates

The application period is now open! Go to Application page.

Application deadline: March 29, 2019

Notification of acceptance: first round April 19, 2019; second round April 26, 2019

Registration open: early May, 2019

The No-Free-Lunch theorem

Theorem (No-Free-Lunch)

Let A be any learning algorithm for the task of binary classification with respect to the 0−1 loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size.

Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

1. there exists a function $f: \mathcal{X} \rightarrow \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$.
2. with probability of at least 1/7 over the choice of $S \sim \mathcal{D}^m$ we have that

$L_{\mathcal{D}}(A(S)) \geq 1/8$.

In other words, for every learning algorithm A there are cases for which this algorithm will fail whereas there is another learner (e.g. a trivial successful learner in this case would be an ERM learner with the hypothesis class $\mathcal{H} = \{ f \}$, or more generally, ERM with respect to any finite hypothesis class that contains f and whose size satisfies the equation $m \geq 8\log(7|H|/6)$) that solves the task. It simply means that an adversary can use the fact that A has no clue what happens on the other half of the domain. We cannot learn perfectly without the proper background knowledge.

The error decomposition

Decompose the error of an $\mathrm{ERM}_{\mathcal{H}}$ predictor that chooses h_S from a restricted class $\mathcal{H}$ into two components:

$$L_{\mathcal{D}}(h_S) = \epsilon_{\mathrm{app}} + \epsilon_{\mathrm{est}}$$

$$\text{where}: \ \epsilon_{\mathrm{app}} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h), \quad \epsilon_{\mathrm{est}} = L_{\mathcal{D}}(h_S) - \epsilon_{\mathrm{app}}$$

The approximation error ($\varepsilon_{\mathrm{app}}$)

- the minimum risk achievable by a predictor in the hypothesis class $\mathcal{H}$
- it measures how well our hypothesis class $\mathcal{H}$ fits the distribution
- it is determined only by $\mathcal{H}$, enlarging it decreases the approximation error

The estimation error ($\varepsilon_{\mathrm{est}}$)

- it measures how well our particular sample let us estimate the best classifier
- the empirical risk is only an estimate of the true risk
- the quality of this estimation depends on the training set size and on the size (complexity) of $\mathcal{H}$
- it varies with samples

Uniform convergence for agnostic PAC learning?

Definition (*uniform convergence*)

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* wrt a domain $\mathcal{Z}$, loss function ℓ if:

- there exists a function $m_H^{UC} : (0,1)^2 \to \mathrm{N}$
- such that for all $(\varepsilon, \delta) \in (0,1)^2$
- and for any probability distribution $\mathcal{D}$ over $\mathcal{Z}$

if S is a sample of m $\geq m_H^{UC}(\varepsilon,\delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1 - \delta$, S is ε-representative.

Definition (ε – representative sample)

A sample S is called ε – representative wrt domain $\mathcal{Z}$, hypothesis class $\mathcal{H}$, loss function ℓ and distribution $\mathcal{D}$ if:

$$\forall h \in \mathcal{H}, \ \ |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$

Lemma

Let S be a sample that is ε/2 – representative wrt domain $\mathcal{Z}$, hypothesis class $\mathcal{H}$, loss function ℓ and distribution $\mathcal{D}$. Then any output of $\mathrm{ERM}_{\mathcal{H}}(S)$ i.e any $h_S \in \mathrm{argmin}_h L_S(h)$ satisfies:

$$L_{\mathcal{D}}(A(S)) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

Finite classes are agnostic PAC learnable

Theorem
Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain and let ℓ: $\mathcal{H} \times Z \rightarrow [0,1]$ be a loss function. Then $\mathcal{H}$ has the uniform convergence property with sample complexity:

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

Moreover, the class $\mathcal{H}$ is agnostically PAC learnable using the ERM paradigm with sample complexity:

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

Today’s lecture: Overview

- Learnability - what do we know so far

- Shattering

- VC-dimension

Learnability - what do we know so far?

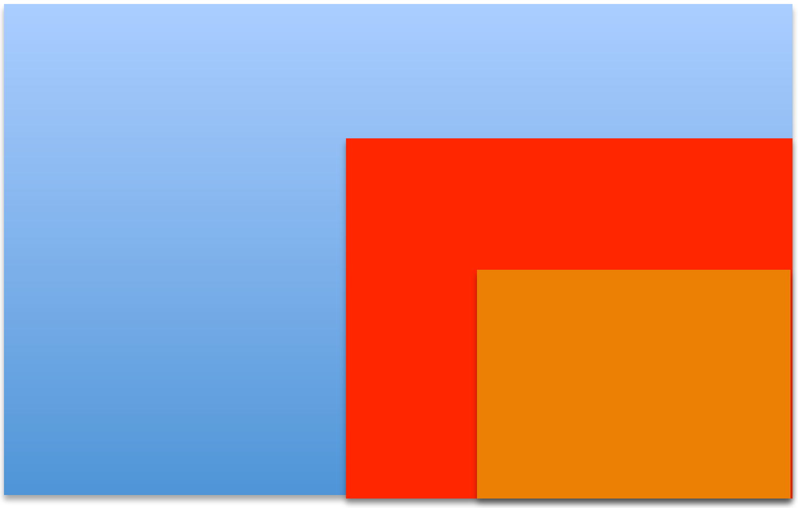
Let $\mathcal{H}$ a hypothesis class.

- is it PAC learnable?
- is it agnostic PAC learnable?
  - we do know that agnostic PAC learnability → PAC learnability

all hypothesis classes

PAC learnable

agnostic PAC learnable



Definition 3.1 (PAC Learnability). A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$, $\boxed{\text{if the realizable assumption holds with respect to } \mathcal{H}, \mathcal{D}, f,}$ then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

Definition 3.4 (Agnostic PAC Learnability for General Loss Functions). A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set Z and a loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$, if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over Z, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns $h \in \mathcal{H}$ such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon,$$

where $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$.

Learnability - what do we know so far?

Let $\mathcal{H}$ a hypothesis class.

- is it PAC learnable?
- is it agnostic PAC learnable?
  - we do know that agnostic PAC learnability $\rightarrow$ PAC learnability
  - we don't know that PAC learnability $\rightarrow$ agnostic PAC learnability

Size of class $\mathcal{H}$:

- finite
- infinite

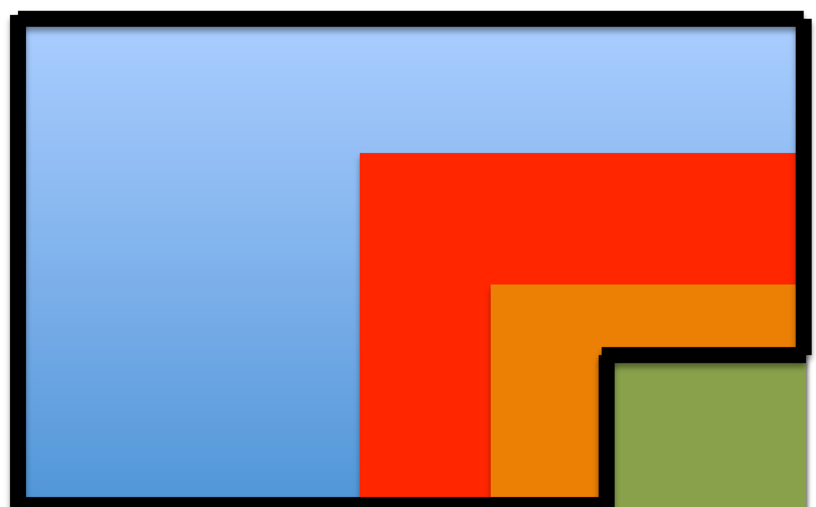
all hypothesis classes

PAC learnable

agnostic PAC learnable

finite size classes

infinite size classes



Learnability - what do we know so far?

Let $\mathcal{H}$ a hypothesis class.

- is it PAC learnable?
- is it agnostic PAC learnable?
  - we do know that agnostic PAC learnability → PAC learnability
  - we don't know that PAC learnability → agnostic PAC learnability

Size of class $\mathcal{H}$:

- finite
- infinite

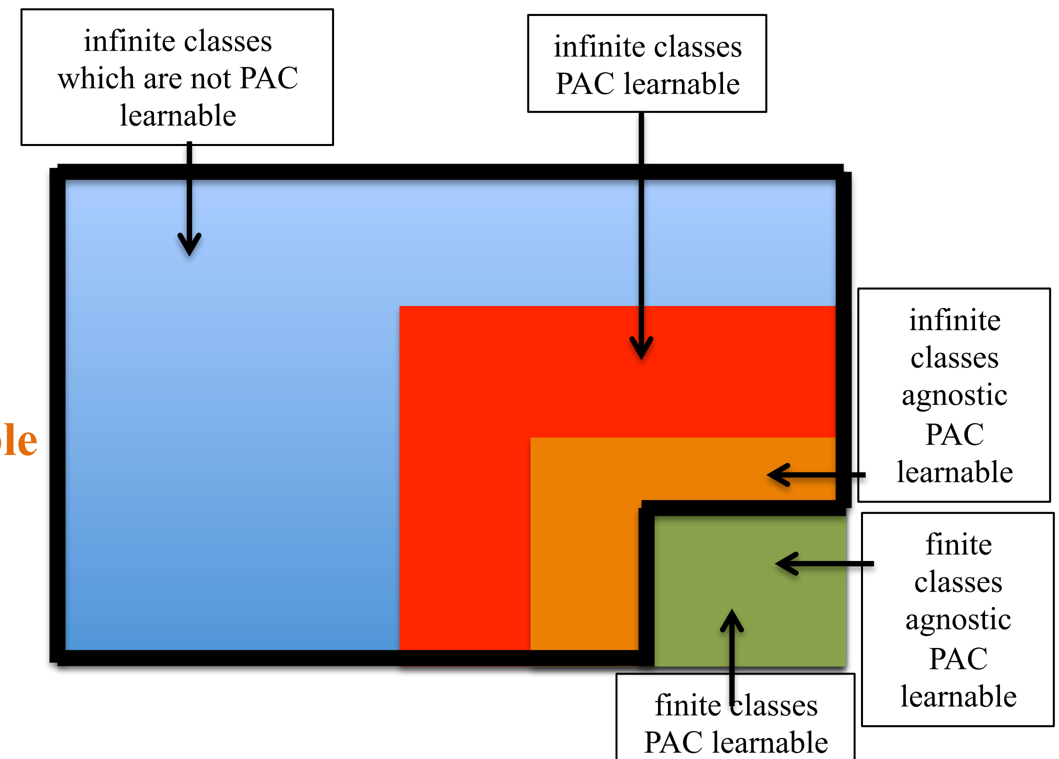
all hypothesis classes

PAC learnable

agnostic PAC learnable

finite size classes

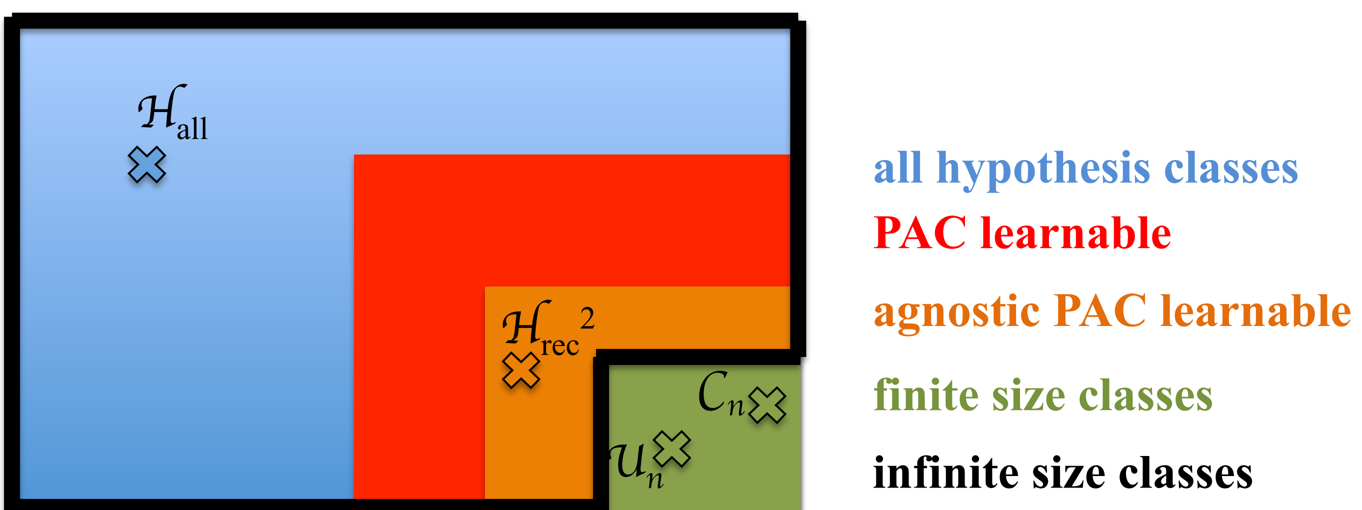
infinite size classes



Learnability - what do we know so far?

Hypothesis classes $\mathcal{H}$ encountered until now:

- finite $\mathcal{H}$
  - C_n concept class of conjunctions of at most n Boolean literals $x_1, \ldots, x_n$
  - $\mathcal{U}_n$ universal concept class
- infinite $\mathcal{H}$
  - $\mathcal{H}_{rec}{}^2$ set of all axis-aligned rectangle lying in R^2 (with positive labels unaffected – PAC or affected by noise – agnostic PAC)
  - $\mathcal{H}_{all}$ all functions from X to {0,1}(No Free-Lunch theorem)

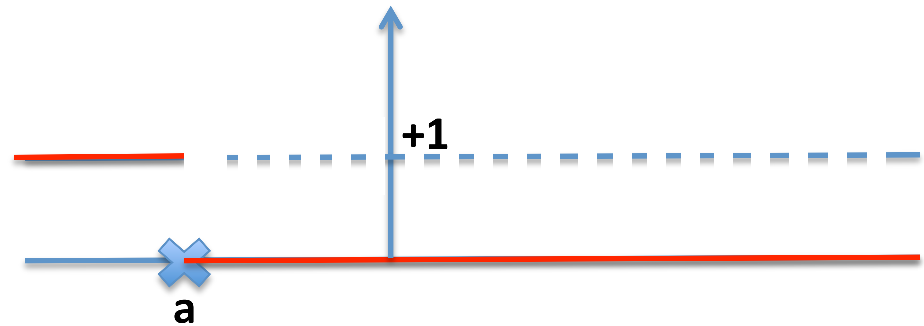


Another class example - $\mathcal{H}_{\text{thresholds}}$

Consider $\mathcal{H}_{\text{thresholds}}$ be the set of threshold functions over the real line

$\mathcal{H}_{\text{thresholds}} = \{h_a\colon \mathrm{R} \longrightarrow \{0, 1\},\ h_a(x) = \mathbf{1}_{[x<a]},\ a \in \mathbf{R}\},\ |\mathcal{H}_{\text{thresholds}}| = \infty$

$$\mathbf{1}_{[x<a]} = \begin{cases} 1, & x < a \\ 0, & x \geq a \end{cases},$$



is the indicator function of the set $\{x \in \mathbf{R} \mid x < a\}$

Lemma

$\mathcal{H}_{\text{thresholds}}$ is PAC learnable, using the ERM learning rule, with sample complexity of

$$m_{H_{thresholds}} \leq \left\lceil \frac{1}{\varepsilon} \log \frac{1}{\delta} \right\rceil$$

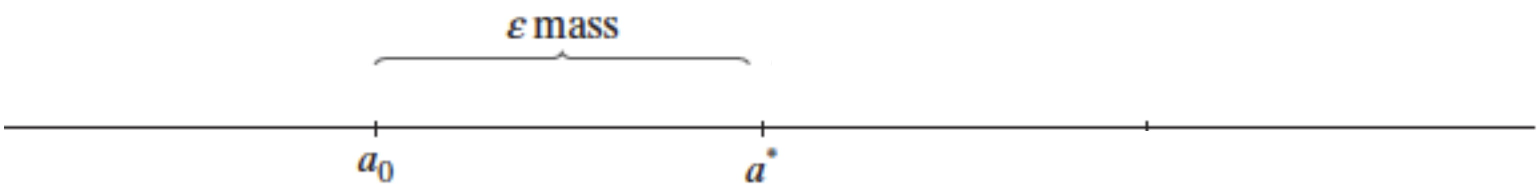
Another class example - $\mathcal{H}_{\text{thresholds}}$

Proof

Let $a^* \in \mathbf{R}$ such that $h_{a^*}(x) = \mathbf{1}_{[x<a^*]}$ achieves $L(h_{a^*}) = 0$ (realizability assumption in the PAC learning scenario).

Consider $\mathcal{D}_x$ a distribution over $\mathcal{X} = \mathbf{R}$ and take $a_0 < a^* \in \mathbf{R}$ such that:

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \epsilon.$$



If $D_x(-\infty, a^*) \leq \varepsilon$ take $a_0 = -\infty$.

Consider the training set $S = ((x_1,y_1), \ldots, (x_m,y_m))$ and the following algorithm A:

- take $b = \max\{x_i | (x_i,1) \in S\}$ (if no positive example appears in S take $b = -\infty$)
- output $A(S) = h_S = h_b$

Another class example - $\mathcal{H}_{\text{thresholds}}$

Then $L_D(h_S) > \varepsilon$ means that $b < a_0$. We have that:

$$\underset{S \sim D^m}{P}(L_D(h_S) > \varepsilon) = P(b < a_0) = (1-\varepsilon)^m \le e^{-\varepsilon m}$$

Take $$m = \left\lceil \frac{1}{\varepsilon} \log \frac{1}{\delta} \right\rceil$$

So, we have that:

$$\underset{S \sim D^m}{P}(L_D(h_S) > \varepsilon) < \delta$$

which shows that $\mathcal{H}_{\text{thresholds}}$ is PAC learnable, using the ERM learning rule.

No Free Lunches vs. $\mathcal{H}_{\text{thresholds}}$

Why is $\mathcal{H}_{\text{thresholds}}$ = set of threshold classifiers not a victim of the No Free Lunch theorem? (we can PAC learn them)

The reason is simple:

- the class of threshold classifiers is so simple that an adversary has no room to create an adversarial distribution

In fact, as our discussion above shows:

- if two threshold classifiers agree on a large enough sample
- their respective thresholds will be close to each other
- there is no way you can force them to behave completely differently on unseen examples.

If that would have been possible then:

- we would have been able to create an adversarial distribution.

So, it seems necessary for PAC learnability that the general class $\mathcal{H}$ considered isn't too expressive

Shattering

How expressive is $\mathcal{H}$?

In our binary classification context, a hypothesis is a function h: $\mathcal{X} \rightarrow \{0, 1\}$

Hence the expressiveness of $\mathcal{H}$:

- is necessary a measure of how many functions $\mathcal{H}$ can express
- in the light of the No Free Lunch theorem, not only functions on $\mathcal{X}$, but also functions on (finite) subsets C of $\mathcal{X}$

Definition (restriction of $\mathcal{H}$ to C)

Let $\mathcal{H}$ be a set hypothesis, i.e., set of functions from $\mathcal{X}$ to $\{0, 1\}$, and let C be a (finite) subset of $\mathcal{X}$, C = $\{c_1, c_2, \ldots, c_m\}$. The restriction of $\mathcal{H}$ to C, denoted by $\mathcal{H}_C$, is the set of functions from C to $\{0, 1\}$ that can be derived from $\mathcal{H}$. That is:

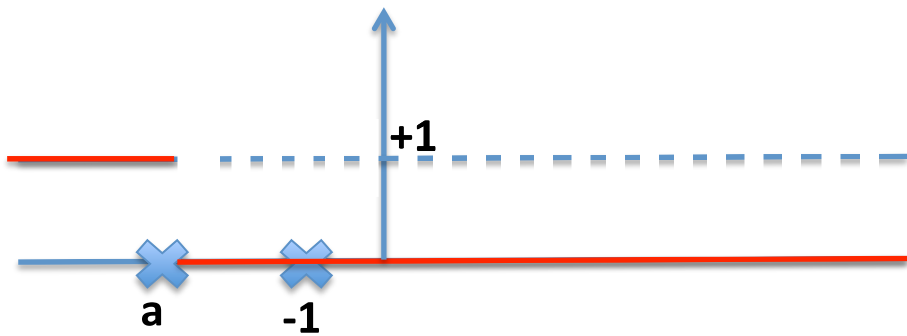
$$\mathcal{H}_C = \{h: C \rightarrow \{0, 1\} \mid h \in \mathcal{H}\}$$

Example: restriction of $\mathcal{H}$ to C

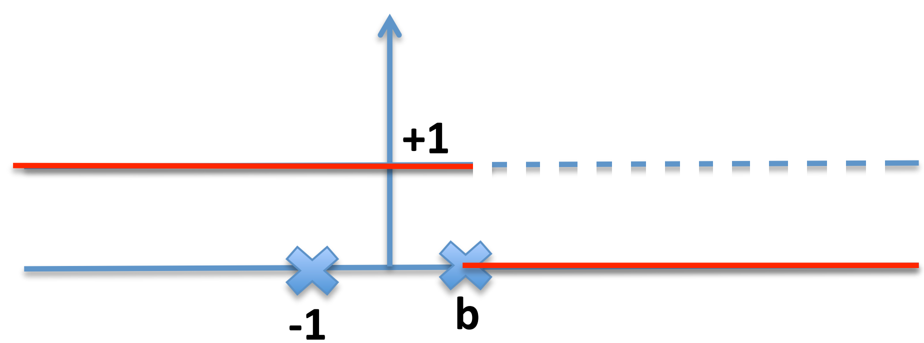
Consider $\mathcal{H} = \mathcal{H}_{\text{thresholds}}$ be the set of threshold functions over the real line $\mathcal{H}_{\text{thresholds}} = \{h_a\!: \mathrm{R} \to \{0, 1\},\ h_a(x) = \mathbf{1}_{[x<a]},\ a \in \mathbf{R}\}$, $|\mathcal{H}_{\text{thresholds}}| = \infty$.

Consider $C = \{-1\}$. Then $\mathcal{H}_C = \{h\!: C \to \{0, 1\} \mid h \in \mathcal{H}\} = \{h\!:\{-1\} \to \{0, 1\} \mid h \in \mathcal{H}\}$ has 2 elements $h_{a,}$ h_b where:

$h_a(-1) = 0$, $a \leq -1$



$h_b(-1) = 1$, $b > -1$

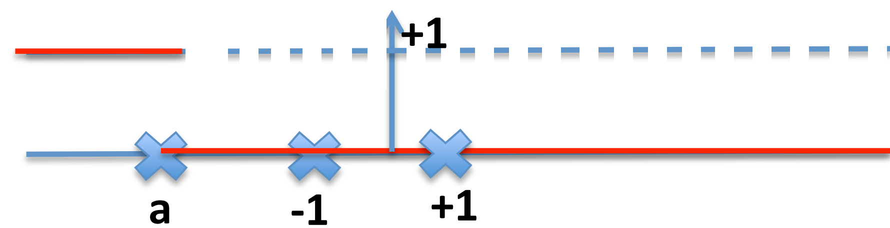


Example: restriction of $\mathcal{H}$ to C

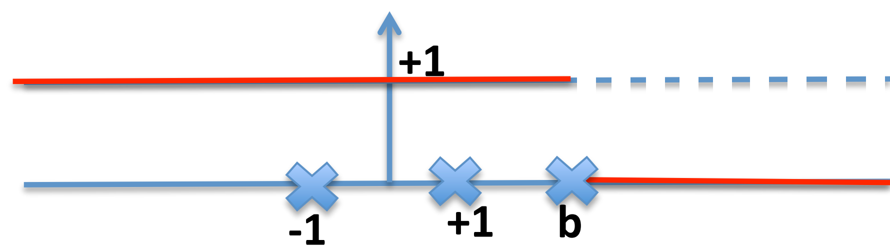
Consider $\mathcal{H} = \mathcal{H}_{\text{thresholds}}$ be the set of threshold functions over the real line $\mathcal{H}_{\text{thresholds}} = \{h_a\colon \mathrm{R} \rightarrow \{0, 1\},\ h_a(x) = \mathbf{1}_{[x<a]},\ a \in \mathbf{R}\}$, $|\mathcal{H}_{\text{thresholds}}| = \infty$.

Consider $C = \{-1, 1\}$. Then $\mathcal{H}_C = \{h\colon C \rightarrow \{0, 1\} \mid h \in \mathcal{H}\} = \{h\colon \{-1, 1\} \rightarrow \{0, 1\} \mid h \in \mathcal{H}\}$ has 3 elements h_a, h_b, h_c where:

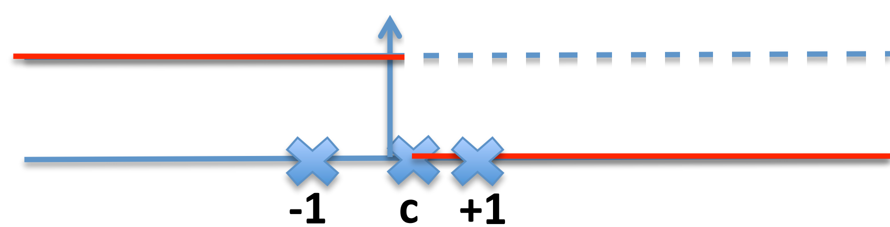
$h_a(-1) = 0$, $h_a(1) = 0$, $a \leq -1$



$h_b(-1) = 1$, $h_b(1) = 1$, $b > 1$



$h_c(-1) = 1$, $h_c(1) = 0$, $-1 < c \leq 1$



Example: restriction of $\mathcal{H}$ to C

Consider $\mathcal{H} = \mathcal{H}_{\text{thresholds}}$ be the set of threshold functions over the real line $\mathcal{H}_{\text{thresholds}} = \{h_a: \mathrm{R} \rightarrow \{0, 1\}, h_a(x) = \mathbf{1}_{[x<a]}, a \in \mathbf{R}\}$, $|\mathcal{H}_{\text{thresholds}}| = \infty$.

Consider $C = \{-1, 1\}$. Then $\mathcal{H}_C = \{h: C \rightarrow \{0, 1\} | h \in \mathcal{H}\} = \{h: \{-1, 1\} \rightarrow \{0, 1\} | h \in \mathcal{H}\}$ has 3 elements $h_{a,}$ h_b, h_c where:

$h_a(-1) = 0$, $h_a(1) = 0$, $a \leq -1$

$h_b(-1) = 1$, $h_b(1) = 1$, $b > 1$

$h_c(-1) = 1$, $h_c(1) = 0$, $-1 < c \leq -1$

There is no function h_d in $\mathcal{H}_C$ such that $h_d(-1) = 0$ and $h_d(1) = 1$ (as we work with the threshold functions)

Alternative view of functions

There is equivalence between functions from $\mathcal{X}$ to $\{0,1\}$ and subsets C of $\mathcal{X}$

- given h: $\mathcal{X} \rightarrow \{0, 1\}$, we can define $C = \{x \in \mathcal{X}|$ such that $h(x) = 1\}$
- given C subset of $\mathcal{X}$, we can define h: $\mathcal{X} \rightarrow \{0,1\}$, $h(x) = \mathbf{1}_C(x)$ indicator function of C

Can represent a subset C of $\mathcal{X}$ as a vector, put 1 for element in C, 0 otherwise.

Can represent each function from C to $\{0,1\}$ as a vector in $\{0,1\}^{|C|}$.

The restriction of $\mathcal{H}$ to C, denoted by $\mathcal{H}_C$, is the set of functions from C to $\{0, 1\}$ that can be derived from $\mathcal{H}$. It can be also seen as the set of all possible vectors that can be generated by $h \in \mathcal{H}$ with elements from C. That is:

$$\mathcal{H}_C = \{h: C \rightarrow \{0, 1\}|\ h \in \mathcal{H}\} = \{(h(c_1), h(c_2), \ldots, h(c_m))|\ h \in \mathcal{H}\}$$

Consider $C = \{-1, 1\}$. Then $\mathcal{H}_{thresholds} = \mathcal{H}_C = \{h: C \rightarrow \{0, 1\}|\ h \in \mathcal{H}\} = \{h: \{-1, 1\} \rightarrow \{0, 1\}|\ h \in \mathcal{H}\} = \{(0,0), (1,0),(1,1)\}$. The vector (0,1) is not realizable by $\mathcal{H}_C$.

Shattering

Definition (Shattering)
A hypothesis class $\mathcal{H}$ *shatters* a finite set C of $\mathcal{X}$, if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0, 1\}$. That is $|\mathcal{H}_C| = 2^{|C|}$.

Examples:
Consider $\mathcal{H} = \mathcal{H}_{\text{thresholds}}$ be the set of threshold functions over the real line.

Consider $C = \{c_1\}$. Then $\mathcal{H}_C = \{h: C \rightarrow \{0, 1\} | h \in \mathcal{H}\}$ has two elements $\{h_a, h_b\}$ with $a \leq c_1$ and $b > c_1$ so $\mathcal{H}$ shatters C. $\mathcal{H}_C = \{(0),(1)\}$, $|\mathcal{H}_C| = 2^{|C|} = 2^1$

Consider $C = \{c_1, c_2 | c_1 \leq c_2\}$. Then $\mathcal{H}_C = \{h: C \rightarrow \{0, 1\} | h \in \mathcal{H}\}$ has at most three elements, there is no function that realizes the labeling (0,1) and so $\mathcal{H}$ does not shatter C.

Alternative view of Shattering

Definition (Shattering)

A hypothesis class $\mathcal{H}$ *shatters* a finite set C of $\mathcal{X}$, if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0, 1\}$. That is $|\mathcal{H}_C| = 2^{|C|}$.

Similar Definition (Shattering)

Let $\mathcal{H}$ be a collection of subsets of $\mathcal{X}$ and C a finite subset of $\mathcal{X}$.

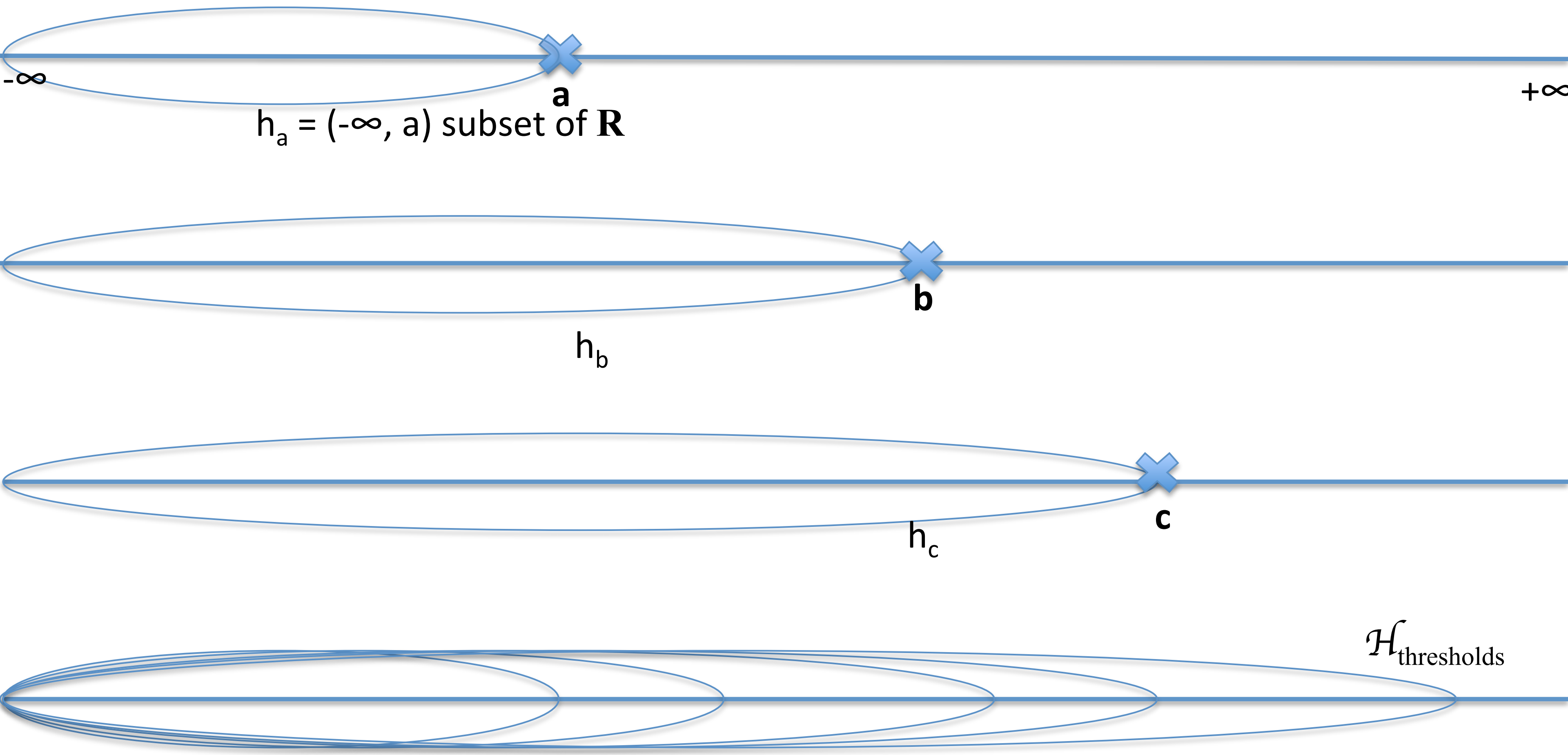
$\mathcal{H}$ shatters C if for every subset B of C there exist some subset $h_B \in \mathcal{H}$ such that $B = h_b \cap C$

In other words, using the elements of $\mathcal{H}$, we can cut C in every possible way.

Shattering – graphical representation

Examples:

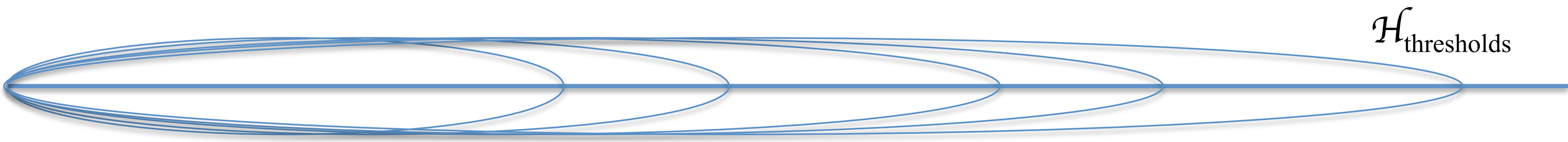
Consider $\mathcal{H} = \mathcal{H}_{\text{thresholds}}$ be the set of threshold functions over the real line.



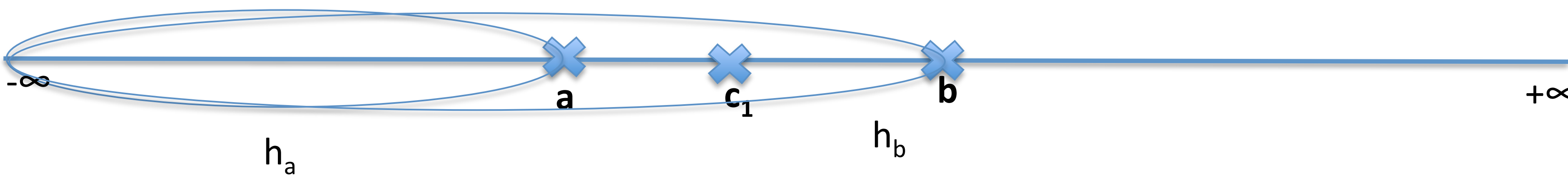
Shattering – graphical representation

Examples:

Consider $\mathcal{H} = \mathcal{H}_{\text{thersholds}}$ be the set of threshold functions over the real line.



Consider $C = \{c_1\}$. Then $\mathcal{H}_C = \{h: C \rightarrow \{0, 1\} \mid h \in \mathcal{H}\}$ has two elements $\{h_a, h_b\}$ with $a \leq c_1$ and $b > c_1$ so $\mathcal{H}$ shatters C. $\mathcal{H}_C = \{(0),(1)\}$, $|\mathcal{H}_C| = 2^{|C|} = 2^1$



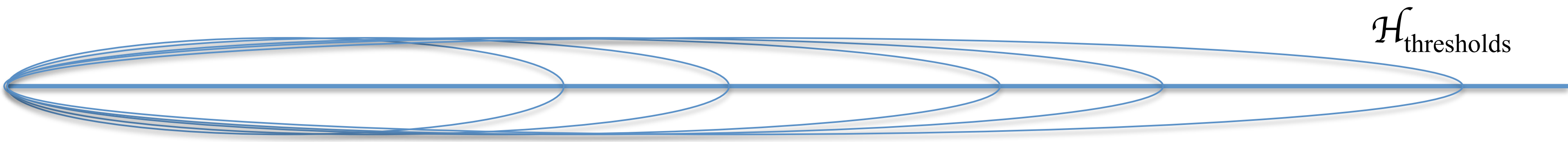
h_a generates label 0

h_b generates label 1

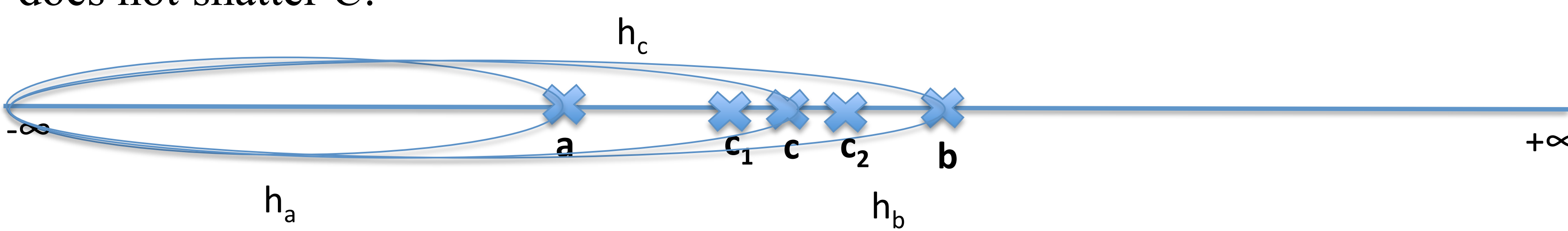
Shattering – graphical representation

Examples:

Consider $\mathcal{H} = \mathcal{H}_{\text{thersholds}}$ be the set of threshold functions over the real line.



Consider $C = \{c_1, c_2 | c_1 \leq c_2\}$. Then $\mathcal{H}_C = \{h: C \rightarrow \{0, 1\} | h \in \mathcal{H}\}$ has at most three elements, there is no function that realizes the labeling (0,1) and so $\mathcal{H}$ does not shatter C.



h_a generates label (0,0)

h_b generates label (1,1)

h_c generates label (1,0)

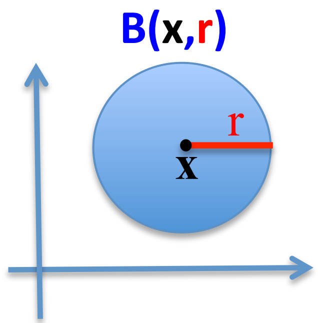
cannot generate the label (0,1)

Shattering – example $\mathcal{H}_{\text{balls}}$

Consider $\mathcal{H} = \mathcal{H}_{\text{balls}}$ be the set of all balls in $\mathbf{R}^2$:

$\mathcal{H}_{\text{balls}} = \{B(x,r),\ x \in \mathbf{R}^2,\ r \geq 0\}$,

$B(x,r) = \{y \in \mathbf{R}^2 \mid \| y - x \|_2 \leq r\}$

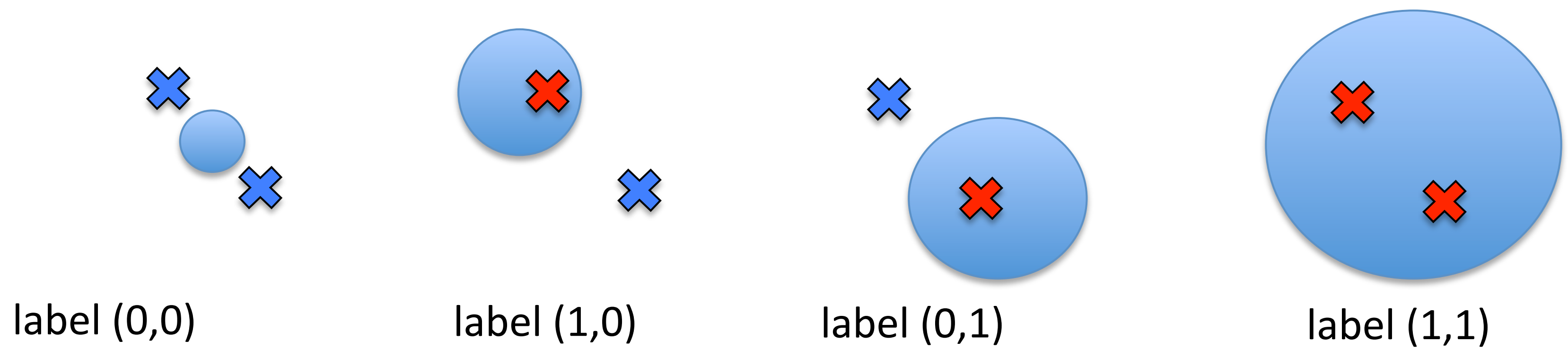


Can also view $\mathcal{H}_{\text{balls}}$ as:

$\mathcal{H}_{\text{balls}} = \{h_{x,r}: \mathbf{R}^2 \rightarrow \{0, 1\},\ h_{x,r} = \mathbf{1}_{B(x,r)},\ x \in \mathbf{R}^2,\ r \geq 0\}$

Is there a set A in $\mathbf{R}^2$ of size 2 shattered by $\mathcal{H}_{\text{balls}}$?

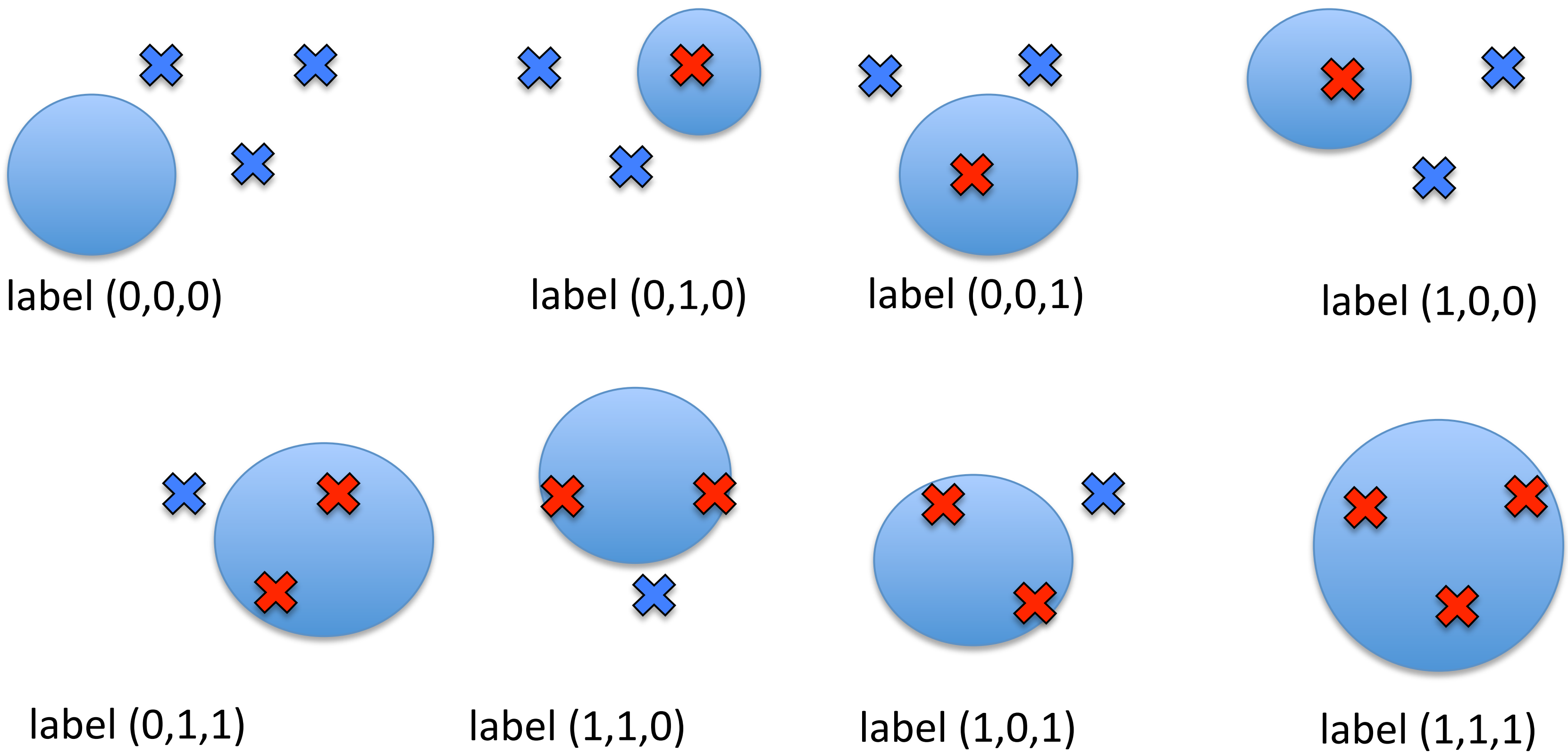
Any set A of two distinct points in $\mathbf{R}^2$ is shattered by $\mathcal{H}_{\text{balls}}$



Shattering – example $\mathcal{H}_{\text{balls}}$

Is there a set A in $\mathbf{R}^2$ of size 3 shattered by $\mathcal{H}_{\text{balls}}$?

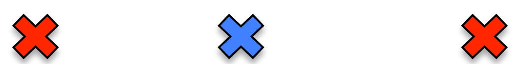
Any set A of three distinct points in $\mathbf{R}^2$ that are not collinear is shattered by $\mathcal{H}_{\text{balls}}$



Shattering – example $\mathcal{H}_{\text{balls}}$

Is there a set A in $\mathbf{R}^2$ of size 3 shattered by $\mathcal{H}_{\text{balls}}$?

Any set A of three distinct points in $\mathbf{R}^2$ that are collinear is not shattered by $\mathcal{H}_{\text{balls}}$



cannot realize the label (1, 0, 1)

What are the conditions for which a set A in $\mathbf{R}^2$ of size 4 is shattered by $\mathcal{H}_{\text{balls}}$?

No Free lunches - revisited

- from the proof of the No Free Lunch theorem, we saw that we can create an adversarial distribution if $\mathcal{H}$ shatters a too large class.

- let $\mathcal{H}$ be a hypothesis class of functions h: X → {0, 1} and S a training set of size m. If there exists a set $C \subseteq \mathcal{X}$ of size $2m$ that is shattered by $\mathcal{H}$, then for any learning algorithms A there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ such that:
  - there exists a function $f : \mathcal{X} \rightarrow \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$
  - with probability of at least 1/7 over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$

- the labels of the m instances give us no information about the labels of the rest of the instances in C – every possible labeling of the rest of the instances can be explained by some hypothesis in $\mathcal{H}$.

- *"If someone can explain every phenomenon, his explanations are worthless"*

- shattering is good, but don't shatter too much.

VC-dimension

The VC-dimension

Definition (VC-dimension)

The VC - dimension of a hypothesis class $\mathcal{H}$, denoted VCdim($\mathcal{H}$), is the maximal size of a set $C \subset X$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite VC-dimension.

Theorem

Let $\mathcal{H}$ be a class of infinite VC-dimension. Then, $\mathcal{H}$ is not PAC learnable.

Proof: Since $\mathcal{H}$ has an infinite VC-dimension, for any training set S of size *m*, there exists a shattered set of size *2m*, and the claim follows for the No Free Lunch theorem.

We will see in the next lecture that the converse is also true: ***a finite VC-dimension guarantees learnability***. ***Hence, the VC-dimension characterizes PAC learnability. VC-dimension is a combinatorial measure, does not imply computing probabilities.***

Determining the VC-dimension of $\mathcal{H}$

Definition (VC-dimension)

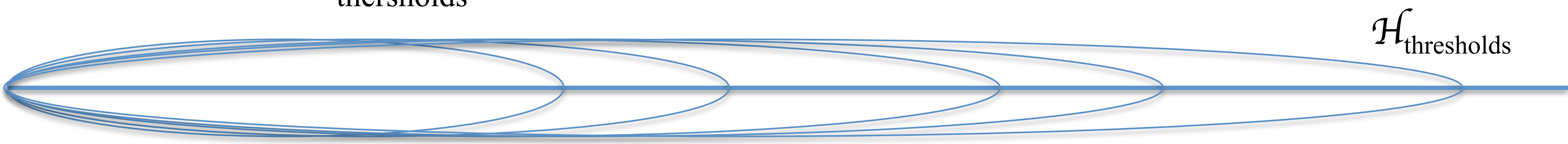
The VC - dimension of a hypothesis class $\mathcal{H}$, denoted VCdim($\mathcal{H}$), is the maximal size of a set $C \subset X$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite VC-dimension.

In order to show that the VC-dimension of a hypothesis class $\mathcal{H}$ is d, we need to show that:

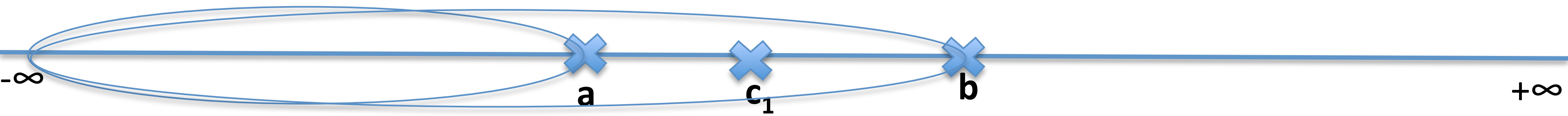
1. **There exists a set C of size d that is shattered by $\mathcal{H}$. (VCdim($\mathcal{H}$) $\geq$ d)**
2. **Every set C of size $d + 1$ is not shattered by $\mathcal{H}$. (VCdim($\mathcal{H}$) $<$ d+1)**

VCdim($\mathcal{H}_{\text{thresholds}}$)

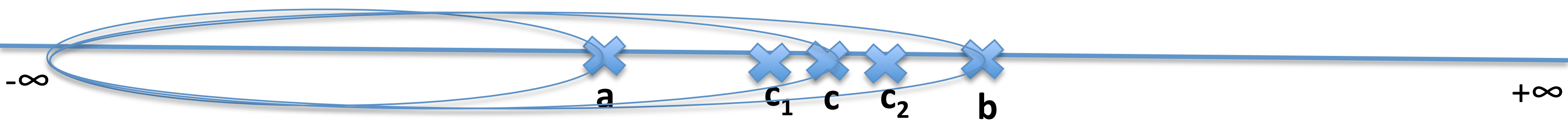
Consider $\mathcal{H} = \mathcal{H}_{\text{thersholds}}$ be the set of threshold functions over the real line.



Consider $C = \{c_1\}$. Then $\mathcal{H}_C = \{h\!: C \to \{0, 1\} \mid h \in \mathcal{H}\}$ has two elements $\{h_a, h_b\}$ with $a \leq c_1$ and $b > c_1$ so $\mathcal{H}$ shatters C. $\mathcal{H}_C = \{(0),(1)\}$, $|\mathcal{H}_C| = 2^{|C|} = 2^1$



Consider $C = \{c_1, c_2 \mid c_1 \leq c_2\}$. Then $\mathcal{H}_C = \{h\!: C \to \{0, 1\} \mid h \in \mathcal{H}\}$ has at most three elements, there is no function that realizes the labeling (0,1) and so $\mathcal{H}$ does not shatter C.

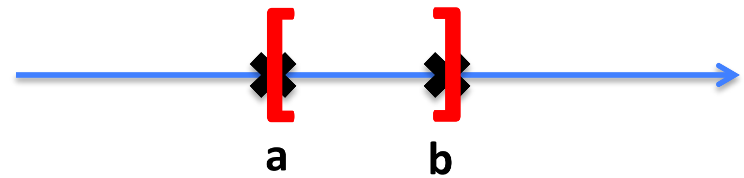


So, **VCdim($\mathcal{H}_{\text{thresholds}}$) = 1**

VCdim($\mathcal{H}_{\text{intervals}}$)

Consider $\mathcal{H} = \mathcal{H}_{\text{intervals}}$ be the set of intervals over the real line.

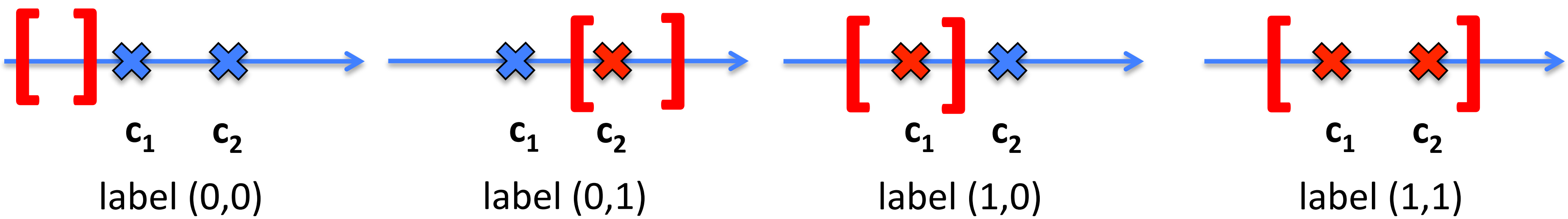
$\mathcal{H}_{\text{intervals}} = \{[a,b] \mid a \leq b,\ a,\ b \in \mathbf{R}\ \}$



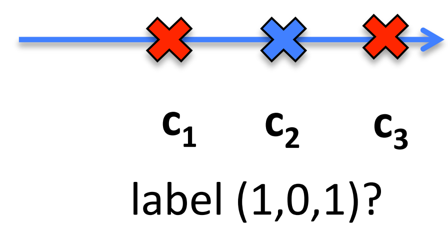
Can also view $\mathcal{H}_{\text{intervals}}$ as:

$\mathcal{H}_{\text{intervals}} = \{h_{a,b}\colon \mathbf{R} \to \{0,\ 1\},\ h_{a,b} = \mathbf{1}_{[a,b]},\ a \leq b,\ a,\ b \in \mathbf{R}\}$

$\mathcal{H}_{\text{intervals}}$ shatters any set A of two different points in $\mathbf{R}$.



$\mathcal{H}_{\text{intervals}}$ cannot shatter any set A of three points in $\mathbf{R}$.



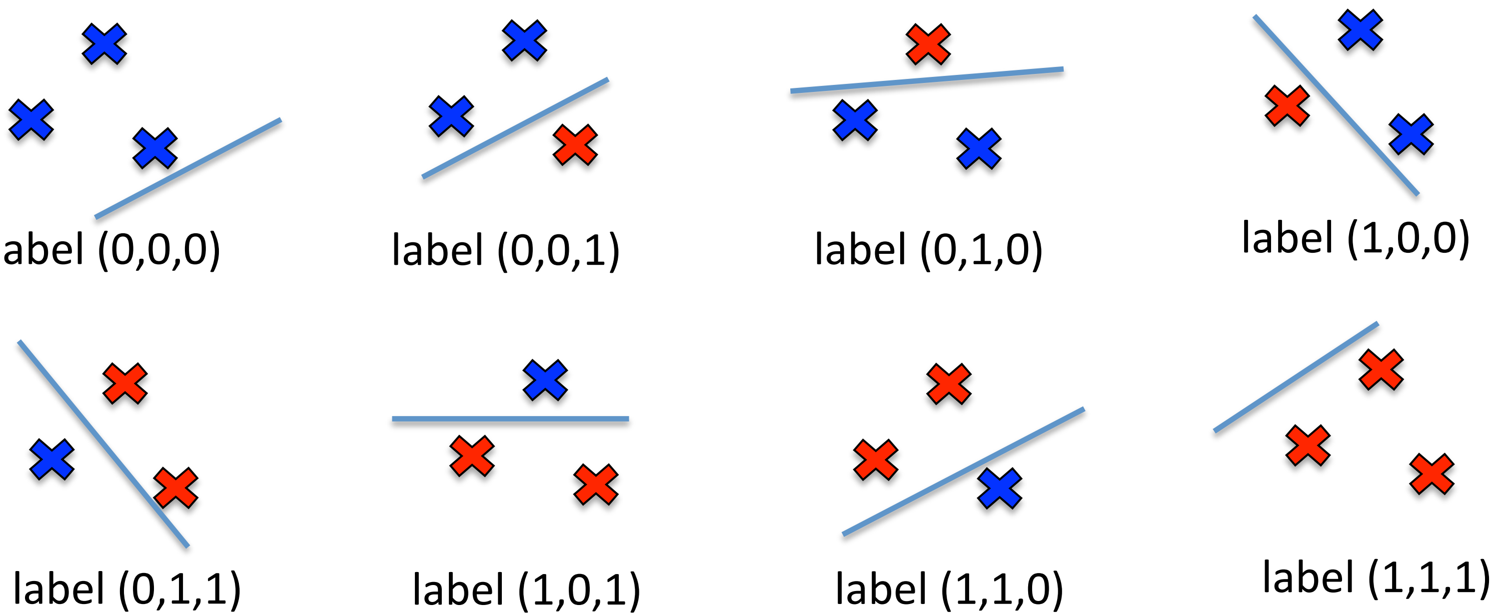
So, **VCdim($\mathrm{H}_{\text{intervals}}$) = 2**

VCdim($\mathcal{H}_{\text{lines}}$)

Consider $\mathcal{H} = \mathcal{H}_{\text{lines}}$ be the set of lines in $\mathbf{R}^2$.

$\mathcal{H}_{\text{lines}} = \{h_{a,b,c}\colon \mathbf{R^2} \rightarrow \{0, 1\}, h_{a,b,c}((x,y)) = \mathbf{1}_{[ax+by+c>0]}((x,y))$, a, b, c$\in \mathbf{R}\}$

$\mathcal{H}_{\text{lines}}$ shatters any set A of three non-colinnear points in $\mathbf{R^2}$.



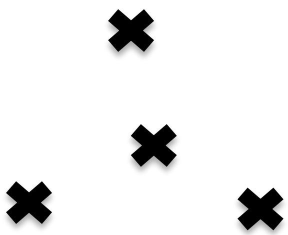
VCdim($\mathcal{H}_{\text{lines}}$)

Consider $\mathcal{H} = \mathcal{H}_{\text{lines}}$ be the set of lines in $\mathbf{R}^2$.

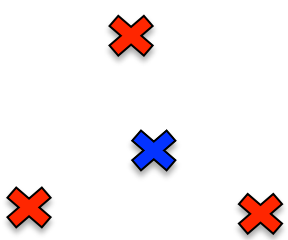
$\mathcal{H}_{\text{lines}} = \{h_{a,b,c}: \mathbf{R}^2 \to \{0, 1\}, h_{a,b,c}((x,y)) = \mathbf{1}_{[ax+by+c>0]}((x,y))$, a, b, c $\in \mathbf{R}\}$

$\mathcal{H}_{\text{lines}}$ doesn't shatter any set A of four points in $\mathbf{R}^2$.

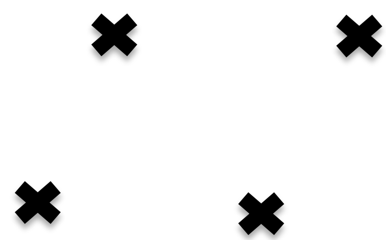
Case a: one point is interior to the convex hull of the other 3



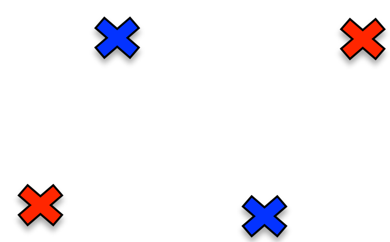
label (1,1,1,0) is un-realizable



Case b: no point is interior to the convex hull of the other 3



label (1,0,1,0) is un-realizable

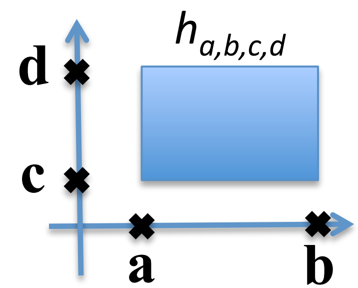


So, **VCdim(H_{lines}) = 3**

VCdim($\mathcal{H}_{rec}{}^2$)

Consider $\mathcal{H} = \mathcal{H}_{rec}{}^2$ be the set of axis aligned rectangles in the $\mathbf{R}^2$.

$\mathcal{H}_{rec}{}^2 = \{[a,b] \times [c,d] \mid a \leq b,\ c \leq d,\ a, b, c, d \in \mathbf{R}\ \}$

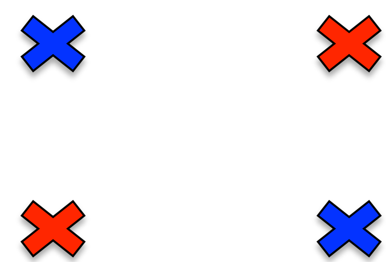


Can also view $\mathcal{H}_{rec}{}^2$ as:

$\mathcal{H}_{rec}{}^2 = \{h_{a,b,c,d}\colon \mathbf{R}^2 \rightarrow \{0, 1\},\ h_{a,b,c,d} = \mathbf{1}_{[a,b] \times [c,d]},\ a \leq b,\ c \leq d,\ a, b, c, d \in \mathbf{R}\ \}$

Does $\mathcal{H}_{rec}{}^2$ shatters any set A of four different points in $\mathbf{R}$?

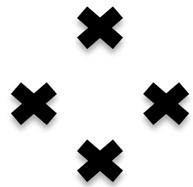
Take A the set of 4 vertices of a rectangle with axis aligned in $\mathbf{R}^2$. Then $\mathcal{H}_{rec}{}^2$ doesn't shatter A (the (0,1,0,1) labeling is not realizable).

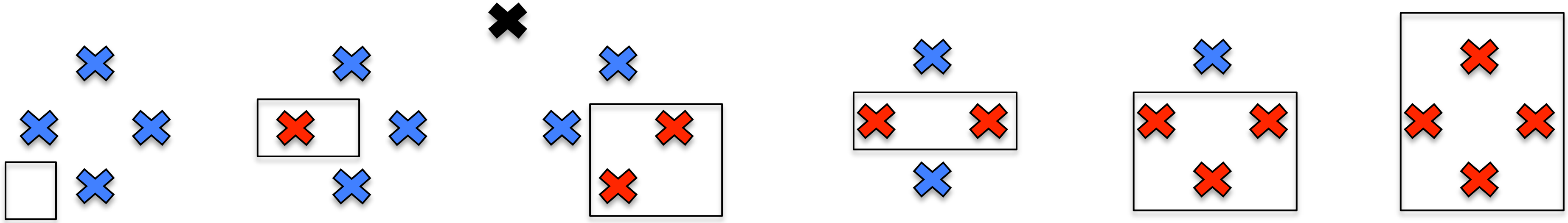


Does this mean that VCdim($\mathcal{H}_{rec}{}^2$) < 4?

No! Either we show that all sets A of size 4 are not shattered by $\mathcal{H}_{rec}{}^2$ (VCdim($\mathcal{H}_{rec}{}^2$) < 4) or find a set A of size 4 that is shattered (VCdim($\mathcal{H}_{rec}{}^2$) $\geq$ 4) .

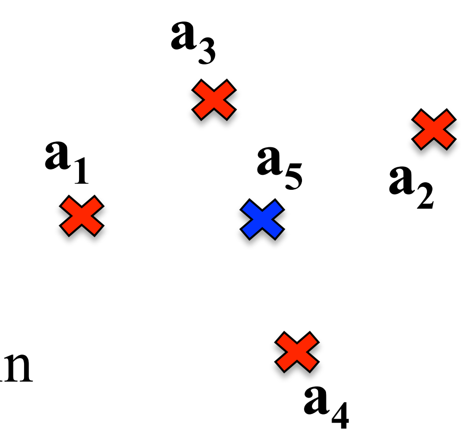
VCdim($\mathcal{H}_{rec}{}^2$)

Consider the set A:  Then A is shattered by $\mathcal{H}_{rec}{}^2$.



label (0,0,0,0) | label (1,0,0,0) | label (0,1,1,0) | label (1,0,1,0) | label (1,1,1,0) | label (1,1,1,1)

Can realize all 16 possible labels. So VCdim($\mathcal{H}_{rec}{}^2$) ≥ 4
Show now that all sets A of 5 points are not shattered by $\mathcal{H}_{rec}{}^2$.
$A = \{a_1, a_2, a_3, a_4, a_5\}$. Consider a_1 – the leftmost point (smaller x) , a_2 – the rightmost point (larger x), a_3 – the lowest point (smaller y), a_4 – the highest point (larger y). Then every rectangle containing a_1, a_2, a_3, a_4 will also contain a_5, so (1,1,1,1,0) is not realizable. So, **VCdim($\mathcal{H}_{rec}{}^2$) = 4**

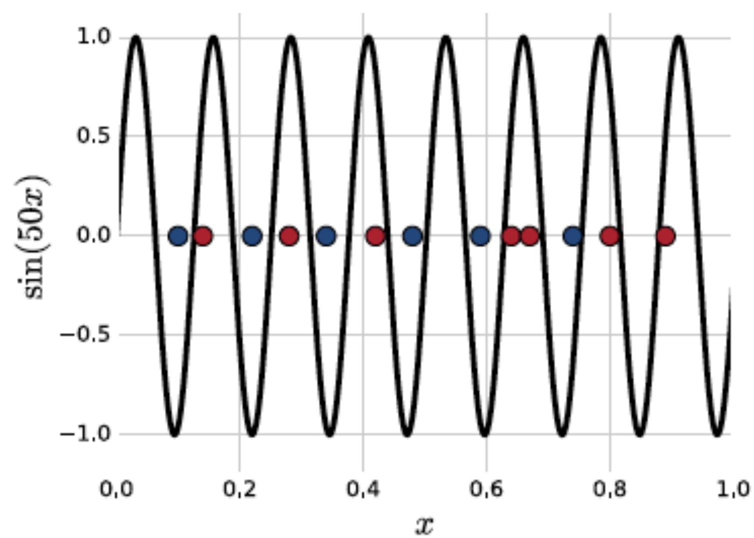


VCdim($\mathcal{H}_{\sin}$)

VCdim($\mathcal{H}_{\text{thresholds}}$) = 1, VCdim($\mathcal{H}_{\text{intervals}}$) = 2, VCdim($\mathcal{H}_{\text{lines}}$) = 3
VCdim($\mathcal{H}_{\text{rec}^2}$) = 4

Consider $\mathcal{H} = \mathcal{H}_{\sin}$ be the set of sin functions:
$\mathcal{H}_{\sin} = \{h_\theta : \mathbf{R} \rightarrow \{0,1\} \mid h_\theta(x) = \lceil \sin(\theta x) \rceil, \theta \in \mathrm{R}\}$, $\lceil -1 \rceil = 0$

VCdim($\mathcal{H}_{\sin}$) = ?



It is possible to prove that VCdim($\mathcal{H}$)=∞, namely, for every d, one can find d points that are shattered by $\mathcal{H}$.

Some basic properties of the VCdim($\mathcal{H}$)

1. VCdim($\mathcal{H}$) $\leq \log_2|\mathcal{H}|$

2. If $\mathcal{H}_1 \subseteq \mathcal{H}_2$ then VCdim($\mathcal{H}_1$) $\leq$ VCdim($\mathcal{H}_2$)

3. If VCdim($\mathcal{H}$) $= \infty$ then $\mathcal{H}$ is not PAC learnable

Next time