Învățare Automată (Machine Learning)

Bogdan Alexe,

bogdan.alexe@fmi.unibuc.ro

Master Informatică, anul I, 2018-2019, cursul 6

Recap - Shattering

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:

$$\mathcal{H}_C = \{h\colon C \to \{0, 1\} \mid h \in \mathcal{H}\}$$

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 $|\mathcal{H}_C| = 2^{|C|}$.

Recap - 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 $C \subseteq X$ 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.

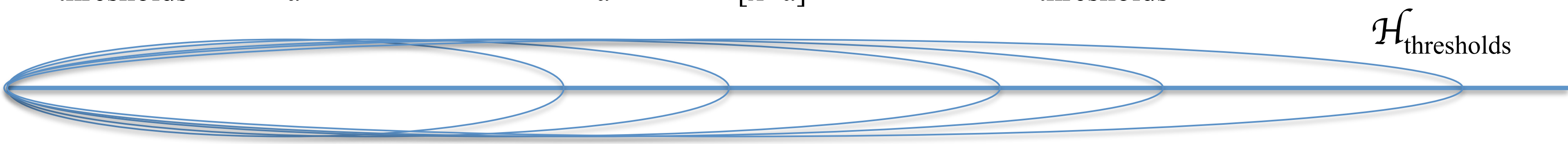
In order to show that the VC-dimension of a hypothesis class $\mathcal{H}$ is d, we need to show that:

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)**

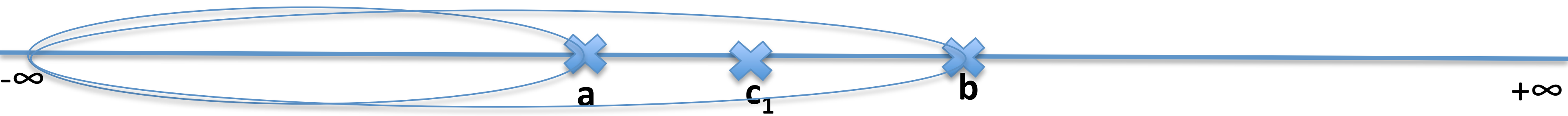
We will see in this lecture that the converse is also true: ***a finite VC- dimension guarantees learnability***. ***Hence, the VC-dimension characterizes PAC learnability. VC-dimension is a combinatorial measure, does not imply computing probabilities.***

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

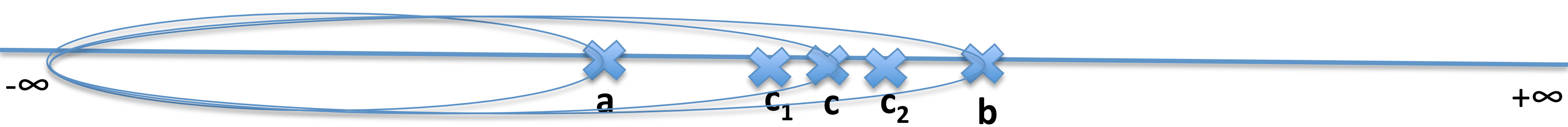
Consider $\mathcal{H} = \mathcal{H}_{\text{thersholds}}$ be the set of threshold functions over the real line.
$\mathcal{H}_{\text{thresholds}} = \{h_a\text{: R} \rightarrow \{0, 1\}, h_a(\text{x}) = \mathbf{1}_{[x<a]}, \text{a} \in \mathbf{R}\}, |\mathcal{H}_{\text{thresholds}}| = \infty.$



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

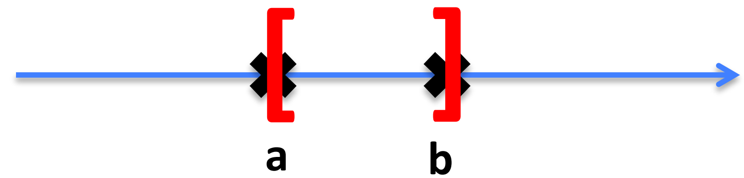


So, **VCdim($\mathcal{H}_{\text{thresholds}}$) = 1**

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

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

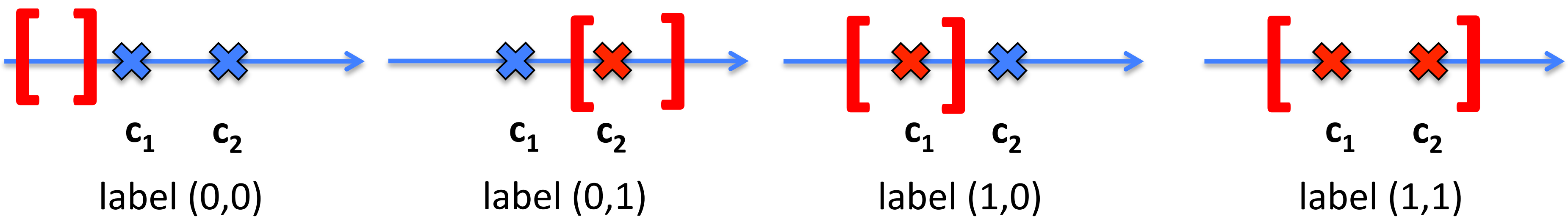
$\mathcal{H}_{\text{intervals}} = \{[a,b] \mid a \leq b,\ a,\ b \in \mathbf{R}\ \}$



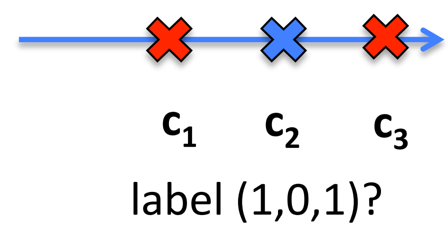
Can also view $\mathcal{H}_{\text{intervals}}$ as:

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

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



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



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

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

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

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

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

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

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

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

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

$\mathcal{H}_{rec}{}^2$ shatters the set A of 4 points arranged as a diamond. So VCdim($\mathcal{H}_{rec}{}^2$) ≥ 4

$\mathcal{H}_{rec}{}^2$ doesn't shatter any set of A of five pints in $\mathbf{R^2}$ (geometric argument).

So, **VCdim($\mathcal{H}_{rec}{}^2$) = 4**

Today's lecture: Overview

- Computing the VC-dimension for some particular $\mathcal{H}$
- Assignment 1
- The fundamental theorem of statistical learning

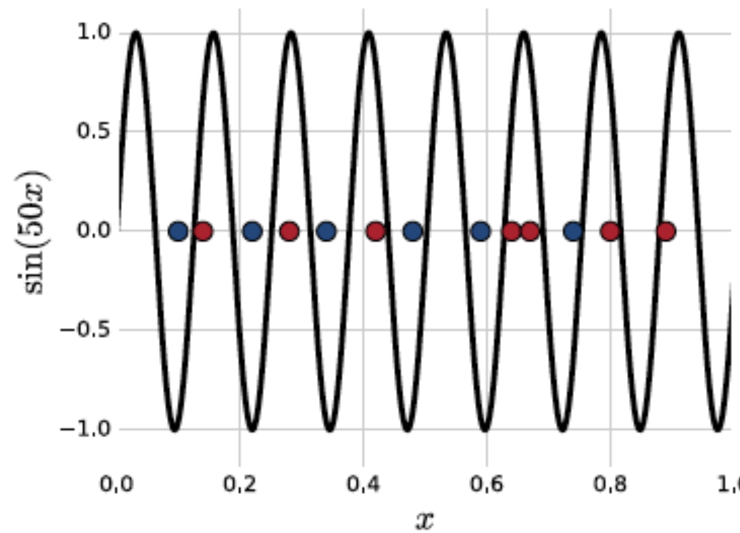
VCdim($\mathcal{H}_{\text{sin}}$)

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

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

We will show that VCdim($\mathcal{H}_{\text{sin}}$) = ∞



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

Lemma

Let x $\in (0,1)$ and let $0.x_1x_2x_3\ldots$ be the binary representation of x. Then, for any natural number m, provided that there exist $k \geq m$ such that $x_k = 1$, we have:

$$\left\lceil \sin(2^m \pi x) \right\rceil = 1 - x_m$$

Example of binary representation:

$x = (0.x_1x_2x_3\ldots)_2 = x_1 \times 2^{-1} + x_2 \times 2^{-2} + x_3 \times 2^{-3} + \ldots$

$x = 0.75 = ½ + ¼ = (0.110000\ldots.)_2$

$x = 0.3 = 0 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} + 0 \times 2^{-4} + 1 \times 2^{-5} + \ldots = (0.01001\ldots)_2$

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

Lemma

Let x $\in (0, 1)$ and let $0.x_1x_2x_3\ldots$ be the binary representation of x. Then, for any natural number m, provided that there exist $k \geq m$ such that $x_k = 1$, we have:

$$\lceil \sin(2^m \pi x) \rceil = 1 - x_m$$

Proof

$\sin(2^m\pi x) = \sin(2^m\pi(0.x_1x_2x_3\ldots)) = \sin(2\pi * 2^{m-1}(0.x_1x_2x_3\ldots)) =$

$= \sin(2\pi*(x_1x_2x_3\ldots x_{m-1}.x_m x_{m+1}\ldots))$ (left shift with m-1 position)

$= \sin(2\pi*(x_1x_2x_3\ldots x_{m-1}.x_m x_{m+1}\ldots) - 2\pi*(x_1x_2x_3\ldots x_{m-1}.0))$ (sin has period 2π)

$= \sin(2\pi*(0.x_m x_{m+1}\ldots))$

Note that $0.x_m x_{m+1}\ldots > 0$ as there exist $k \geq m$ such that $x_k = 1$

Case 1: $x_m = 0$, then $0 < 2\pi*(0.x_m x_{m+1}\ldots) < 2\pi*1/2 = \pi$. So $0 < \sin(2^m\pi x) < 1$, and from here it results that: $\lceil \sin(2^m \pi x) \rceil = 1 = 1 - 0 = 1 - x_m$

Case 2: $x_m = 1$, then $2\pi > 2\pi*(0.x_m x_{m+1}\ldots) \geq 2\pi*1/2 = \pi$. So $-1 \leq \sin(2^m\pi x) \leq 0$, and from here it results that: $\lceil \sin(2^m \pi x) \rceil = 0 = 1 - 1 = 1 - x_m$ (we consider $\lceil -1 \rceil = 0$)

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

To prove VCdim($\mathcal{H}_{\sin}$) = ∞, we need to pick *n* points which are shattered by $\mathcal{H}_{\sin}$, for any *n*. To do so, we construct *n* points $x_1, x_2, \ldots, x_n \in [0, 1]$, such that the set of the *m*-th bits in the binary expansion, as m ranges from 1 to 2^n, ranges over all possible labelings of $x_1, x_2, \ldots, x_n$.

x_1 = 0. 0 0 0 0 ... 1 1

x_2 = 0. 0 0 0 0 ... 1 1

...

x_{n-1} = 0 0 0 1 1 ... 1 1

x_n = 0. 0 1 0 1 ... 0 1

m=1

For example, to give the labeling 1 for all instances, we just pick m=1:

$$h(x) = \left\lceil \sin(2^1 \pi x) \right\rceil$$

which returns the first bit (column) in the binary expansion.

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

To prove VCdim($\mathcal{H}_{\sin}$) = ∞, we need to pick *n* points which are shattered by $\mathcal{H}_{\sin}$, for any *n*. To do so, we construct *n* points $x_1, x_2, \ldots, x_n \in [0, 1]$, such that the set of the *m*-th bits in the binary expansion, as m ranges from 1 to 2^n, ranges over all possible labelings of $x_1, x_2, \ldots, x_n$.

x_1 = 0. 0 0 0 0 ... 1 1

x_2 = 0. 0 0 0 0 ... 1 1

...

x_{n-1} = 0. 0 0 1 1 ... 1 1

x_n = 0. 0 1 0 1 ... 0 1

m=2

If we wish to give the labeling 1 for $x_1, x_2, \ldots, x_{n-1}$, and the labeling 0 for x_n, we pick m=2:

$$h(x) = \lceil \sin(2^2 \pi x) \rceil$$

which returns the second bit (column) in the binary expansion.

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

To prove VCdim($\mathcal{H}_{\sin}$) = ∞, we need to pick *n* points which are shattered by $\mathcal{H}_{\sin}$, for any *n*. To do so, we construct *n* points $x_1, x_2, \ldots, x_n \in [0, 1]$, such that the set of the *m*-th bits in the binary expansion, as m ranges from 1 to 2^n, ranges over all possible labelings of $x_1, x_2, \ldots, x_n$.

x_1 = 0. 0 0 0 0 ... 1 1

x_2 = 0. 0 0 0 0 ... 1 1

...

x_{n-1} = 0. 0 0 1 1 ... 1 1

x_n = 0. 0 1 0 1 ... 0 1

$m=2^n$

If we wish to give the labeling 0 for $x_1, x_2, \ldots, x_{n-1}, x_n$ we pick $m=2^n$:

$$h(x) = \left\lceil \sin(2^{2^n} \pi x) \right\rceil$$

which returns the bit on position 2^n (column) in the binary expansion.

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

To prove VCdim($\mathcal{H}_{\sin}$) = ∞, we need to pick *n* points which are shattered by $\mathcal{H}_{\sin}$, for any *n*. To do so, we construct *n* points $x_1, x_2, \ldots, x_n \in [0, 1]$, such that the set of the *m*-th bits in the binary expansion, as m ranges from 1 to 2^n, ranges over all possible labelings of $x_1, x_2, \ldots, x_n$.

x_1 = 0. 0 0 0 0 ... 1 1

x_2 = 0. 0 0 0 0 ... 1 1

...

x_{n-1} = 0. 0 0 1 1 ... 1 1

x_n = 0. 0 1 0 1 ... 0 1

We conclude that $x_1, x_2, \ldots, x_{n-1}, x_n$ can be given any labeling by some $h \in \mathcal{H}_{\sin}$, so it is shattered. This can be done for any *n*, so VCdim($\mathcal{H}_{\sin}$) = ∞.

VCdim($\mathcal{HS}_0^n$)

Consider $\mathcal{H} = \mathcal{HS}^n$ be the set of halfspaces (linear classifiers) in R^n

$\mathcal{H} = \mathcal{HS}^n = \{h_{w,b}: \mathbf{R}^n \to \{-1, 1\},\ h_{w,b}(x) = sign\left(\sum_{i=1}^{n} w_i x_i + b\right) \mid w \in \mathbf{R}^n, b \in \mathbf{R}\}$

Consider label -1 to correspond to label 0, they are basically the same.

For n = 2 we have:

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

Let us restrict our attention to "homogenous" linear classifiers, the ones that go through origin, b = 0.

$\mathcal{HS}_0^n = \{h_{w,0}: \mathbf{R}^n \to \{-1, 1\},\ h_{w,0}(x) = sign\left(\sum_{i=1}^{n} w_i x_i\right) \mid w \in \mathbf{R}^n\}$

What is the VCdim($\mathcal{HS}_0^n$)?

VCdim($\mathcal{H}S_0^n$)

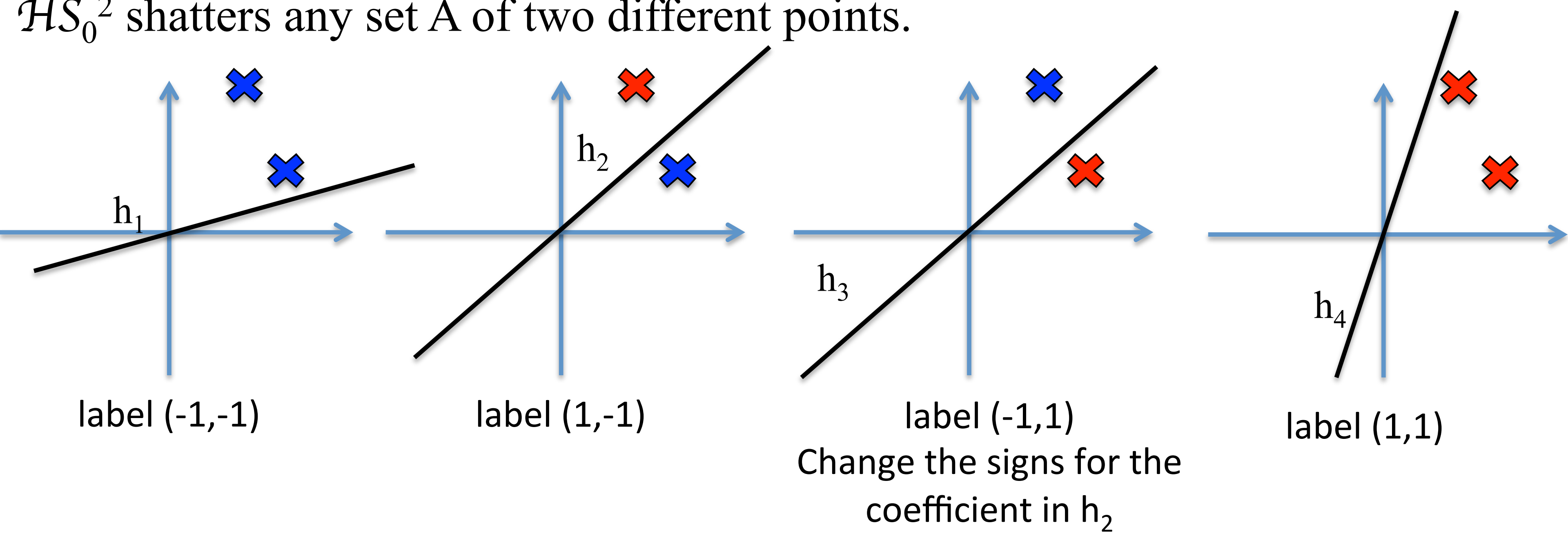
$\mathcal{H}S_0^n = \{h_{w,0}: \mathbf{R}^n \rightarrow \{-1, 1\},\ h_{w,0}(x) = sign\left(\sum_{i=1}^{n} w_i x_i\right) \mid w \in \mathbf{R}^n\}$

For n = 2 we have:

$\mathcal{H}S_0^2 = \{h_{w1,w2}: \mathbf{R}^2 \rightarrow \{-1, 1\},\ h_{w1,w2}(x) = sign(w_1 x_1 + w_2 x_2) \mid (w_1, w_2) \in \mathbf{R}^2\}$

What is the VCdim(HS_0^2) ?

$\mathcal{H}S_0^2$ shatters any set A of two different points.



Does HS_0^2 shatter a set A of three points?

Difficult to reason geometrically… choose the algebraic proof.

VCdim($\mathcal{H}S_0{}^n$)

We will show that VCdim($\mathcal{H}S_0{}^n$) = n.

Proof: 1st part

We first show that VCdim($\mathcal{H}S_0{}^n$) $\geq n$.

We find a set A consisting of n points in $\mathbf{R}^n$ that is shattered by $\mathcal{H}S_0{}^n$.

Take A = $\{e_1, e_2, \ldots, e_n\}$ to be the orthonormal basis of $\mathbf{R}^n$.

$e_1 = (1, 0, 0, \ldots, 0)$; $e_2 = (0, 1, 0, \ldots, 0)$; …. ; $e_n = (0, 0, 0, \ldots, 1)$

We want to proof that $\mathcal{H}S_0{}^n$ shatters A, so that VCdim($\mathcal{H}S_0{}^n$) $\geq n$. This is equivalent to proof that for every B $\subseteq$ A, there is a function $h_B \in \mathcal{H}S_0{}^n$ such that h_B gives label +1 to all elements in B and label -1 to all elements of A \ B.

Pick B subset of A, B $\subseteq \{e_1, e_2, \ldots, e_n\}$. Choose $w = (w_1, w_2, \ldots, w_n)$ such that:

$$w_i = \begin{cases} 1, \text{ if } e_i \in B \\ \\ -1, \text{ if } e_i \notin B \end{cases}$$

Then, $h_B(e_i) = \text{sign}(\langle w, e_i \rangle) = w_i$ will generate the labels +1 for elements in B, -1 for elements not in B

VCdim($\mathcal{H}S_0{}^n$)

Proof: 2nd part

We now show that VCdim($\mathcal{H}S_0{}^n$) < n + 1.

We will prove that given any set A = $\{x_1, x_2, \ldots, x_{n+1}\}$ of n + 1 points in $\mathbf{R}^n$, A cannot be shattered by $\mathcal{H}S_0{}^n$.

The points $\{x_1, x_2, \ldots, x_{n+1}\}$ "live" in $\mathbf{R}^n$, a vector space with dimension n. So, $\{x_1, x_2, \ldots, x_{n+1}\}$ are linearly dependent and there exist coefficients $a_1, a_2, \ldots$ a_{n+1} not all of them 0 such that:

$$\sum_{i=1}^{n+1} a_i x_i = 0$$

Take P ⊆ { 1, 2, …, n +1} the set of positive coefficients a_i and N ⊆ { 1, 2, …, n +1} the set of negative coefficients of a_i. So we have:

$$\sum_{i \in P} a_i x_i = \sum_{j \in N} |a_j| x_j$$

VCdim($\mathcal{H}S_0^n$)

Take $P \subseteq \{1, 2, \ldots, n+1\}$ the set of positive coefficients a_i and $N \subseteq \{1, 2, \ldots, n+1\}$ the set of negative coefficients of a_i. So we have:

$$\sum_{i \in P} a_i x_i = \sum_{j \in N} |a_j| x_j$$

Assume that A is shattered by $\mathcal{H}S_0^n$ and take $B = \{x_i \mid i \in P\}$. In particular, there exist h_B such that it realizes the label consisting of +1 for all $x_i \in B$ and -1 for all $x_i \notin B$.

$$h_B\left(\sum_{i \in P} a_i x_i\right) = \sum_{i \in P} a_i h_B(x_i) > 0$$

$$h_B\left(\sum_{i \in P} a_i x_i\right) = h_B\left(\sum_{j \in N} |a_j| x_j\right) = \sum_{j \in N} |a_j| h_B(x_j) < 0$$

So, this is a contradiction.

Assignment 1

Assignment 1 – good to know

- 5 problems = 3.5 points
- 1 bonus problem = 1 point
- deadline: Sunday, 21 April 2019, 23:59
  - late submission policy: maximum 3 days allowed, -10% (= 0.35 points) for each day
  - submit hard copy

  OR
  - send a pdf with a scan of your solution to bogdan.alexe@fmi.unibuc.ro

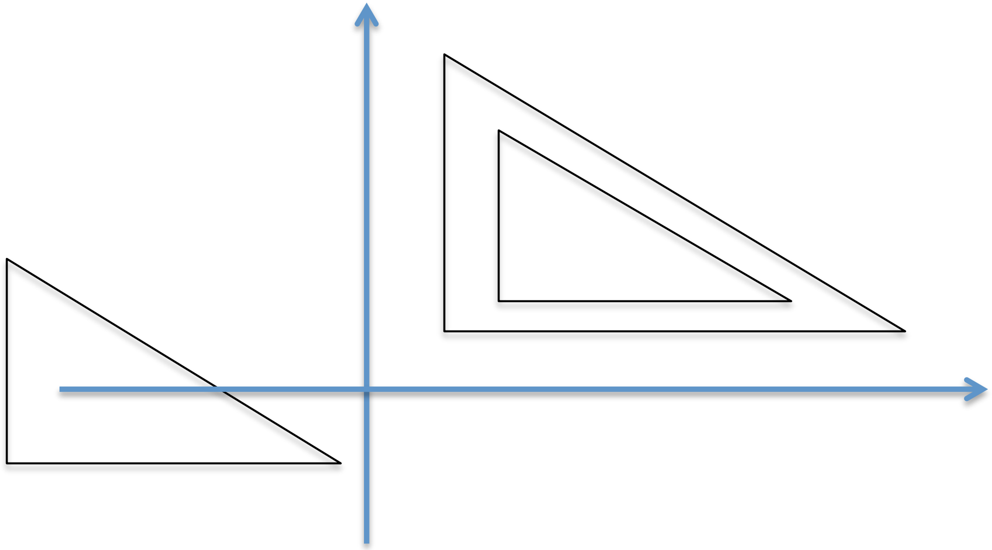
- for every problem write clear explanations, proofs to justify your answer (if you write just some indications you will not get too many points)

- do not share/copy the solution with/from your colleagues: you + your colleague/s will get 0 points

Problem 1

Problem 1 (0.5 points) - problem 2.5 in the book "Foundation of the Machine Learning", 2nd Edition

Let $X = \mathbf{R}^2$ and consider the set of concepts defined by the area inside a right triangle ABC with the two catheti AB and AC parallel to the axes and with AB/AC = α (fixed constant > 0). Show, using similar methods to those used in the seminar class for the axis-aligned rectangles, that this class can be PAC-learned from training data of size $m \geq (3/\varepsilon)\ln(3/\delta)$.



Problem 2

Problem 2 (0.5 points) - problem from lecture 5.

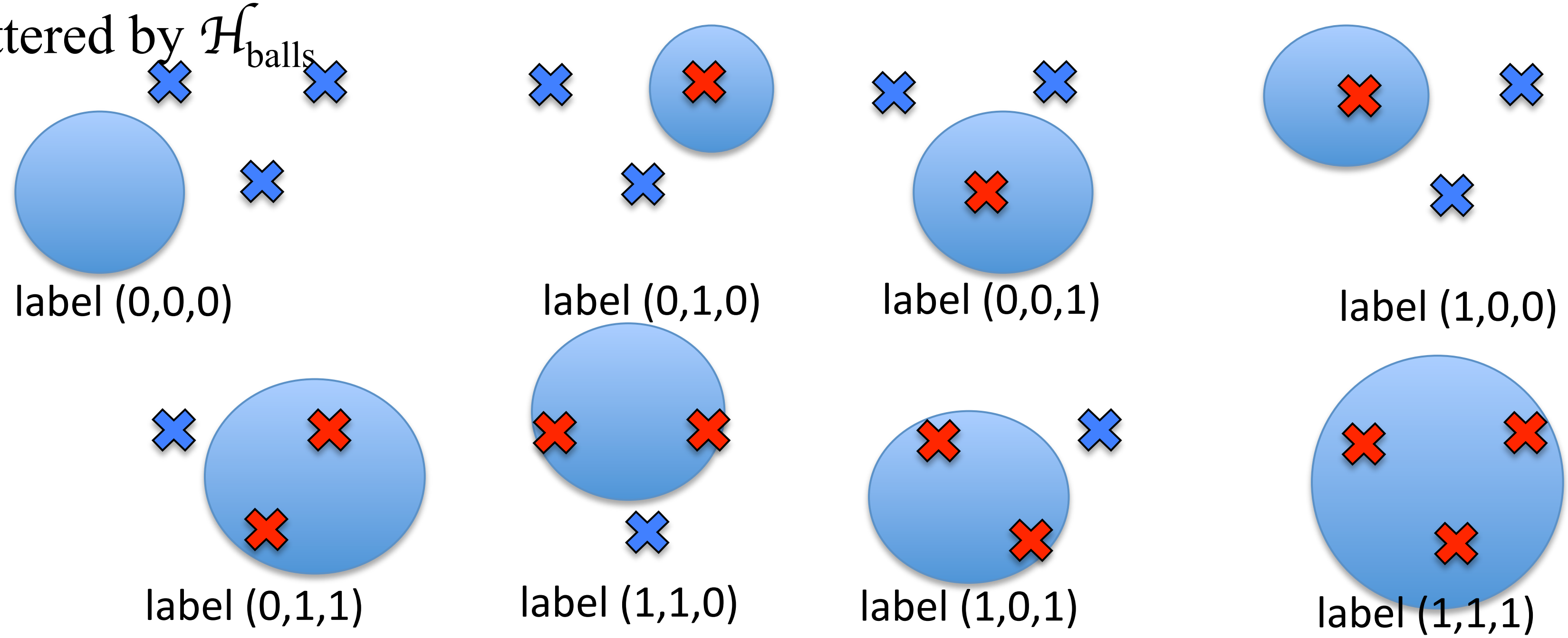
Consider $\mathcal{H}_{\text{balls}}$ to be the set of all balls in $\mathbf{R}^2$: $\mathcal{H}_{\text{balls}} = \{B(x,r),\ x \in \mathbf{R}^2,\ r \geq 0\}$,

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

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

Justify your answer.

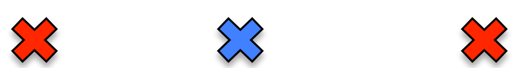
Remember: any set A of three distinct points in $\mathbf{R}^2$ that are not collinear is shattered by $\mathcal{H}_{\text{balls}}$



Problem 2

Problem 2 (0.5 points) - problem from lecture 5.
Consider $\mathcal{H}_{\text{balls}}$ to be the set of all balls in $\mathbf{R}^2$: $\mathcal{H}_{\text{balls}} = \{B(\mathbf{x},r), \mathbf{x} \in \mathbf{R}^2, r \geq 0\}$,
where $B(x,r) = \{y \in \mathbf{R}^2 \mid \| y - x \|_2 \leq r\}$
What are the conditions for which a set A in $\mathbf{R}^2$ of size 4 is shattered by $\mathcal{H}_{\text{balls}}$?
Justify your answer.

Remember: 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)

Problem 3

Problem 3 (0.75 points) – problem 5.2 in the book "Understanding Machine Learning: From Theory to Algorithms"

Assume you are asked to design a learning algorithm to predict whether patients are going to suffer a heart attack. Relevant patient features the algorithm may have access to include blood pressure (BP), body-mass index (BMI), age (A), level of physical activity (P), and income (I). You have to choose between two algorithms; the first picks an axis aligned rectangle in the two dimensional space spanned by the features BP and BMI and the other picks an axis aligned rectangle in the five dimensional space spanned by all the preceding features.

1. Explain the pros and cons of each choice.
2. Explain how the number of available labeled training samples will affect your choice

Problem 4

Problem 4 (0.75 points) - problem 6.3 in the book "Understanding Machine Learning: From Theory to Algorithms"

Let $\mathcal{X}$ be the Boolean hypercube $\{0,1\}^n$. For a set $I \subseteq \{1, 2, \ldots, n\}$ we define a parity function h_I as follows. On a binary vector $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$,

$$h_I(x) = \left(\sum_{i \in I} x_i\right) \bmod 2$$

(That is, h_I computes parity of bits in I .) What is the VC-dimension of the class of all such parity functions, $\mathcal{H}_{n\text{-parity}} = \{h_I : I \subseteq \{1, 2, \ldots, n\}\}$? Prove your claim.

Problem 5

Problem 5 (1 point) - problem 6.2 in the book "Understanding Machine Learning: From Theory to Algorithms"

Given some finite domain set, $\mathcal{X}$, and a number $k \leq |\mathcal{X}|$, figure out the VC-dimension of each of the following classes (and prove your claims):

1. $\mathcal{H}^{\mathcal{X}}_{=k} = \{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| = k\}$: that is, the set of all functions that assign the value 1 to exactly k elements of $\mathcal{X}$.

2. $H_{\text{at-most-k}} = \{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| \leq k \text{ or } |\{x : h(x) = 0\}| \leq k\}$.

Bonus Problem

Bonus Problem (1 point)

Compute the VC-dimension of the class of convex d-gons (convex polygons with exactly d sides) in the plane. Provide a detailed proof of your result.

d = 4