Aalto University
School of Science

# CS-E4070 — Computational learning theory

## Slide set 04 : the Vapnik-Chervonenkis dimension

Cigdem Aslay and Aris Gionis

Aalto University

spring 2019

# reading material

- K&V, chapter 3
- SS&BD, chapter 6

# shattering a set of instances

- let $\mathcal{H}$ be a class of functions from $X$ to $\{0,1\}$
- let $A = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset X$ be a (finite) subset of $X$
- a dichotomy of a set is a partition of the set into two disjoint subsets
- a dichotomy of $A$ induced by $h \in \mathcal{H}$

$$h_A = \{h(\mathbf{x}_1), \ldots, h(\mathbf{x}_m)\} \in \{0,1\}^m$$

# shattering a set of instances

- definition: a set $A$ of instances is shattered by $\mathcal{H}$

  iff for every dichotomy of $A$, there exists some hypothesis in $\mathcal{H}$ consistent with this dichotomy

- let $\Pi_{\mathcal{H}}(A)$ be the set of all dichotomies on $A$ induced by $\mathcal{H}$

  (*a.k.a.,* restriction of $\mathcal{H}$ to $A$)

$$\Pi_{\mathcal{H}}(A) = \{(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_m)) : h \in \mathcal{H}\}$$

- $\mathcal{H}$ shatters $A$ iff

$$\Pi_{\mathcal{H}}(A) = \{0, 1\}^m$$

# the VC dimension

- definition: the VC dimension, $VCD(\mathcal{H})$, of a hypothesis class $\mathcal{H}$ is the cardinality of the largest finite subset of $X$ shattered by $\mathcal{H}$.

$$VCD(\mathcal{H}) = sup\{|A| : \mathcal{H} \text{ shatters } A\}$$

- If $\mathcal{H}$ can shatter arbitrarily large finite sets, then

$$VCD(\mathcal{H}) = \infty$$

# the VC dimension

- to show that $VCD(\mathcal{H})$ is $d$ we need to show that:
  - there exists a set of size $d$ which is shattered by $\mathcal{H}$
  - no set of size $d + 1$ can be shattered by $\mathcal{H}$

# example – threshold functions

- $X = \mathbb{R}$
- $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ where

$$h_a(x) = \mathbb{I}\left[x \leq a\right], \forall h_a \in \mathcal{H}$$

- claim: $VCD(\mathcal{H}) = 1$

# example – threshold functions

- first show that $d$ is at least $1$

  i.e., find a set of size $1$ that can be shattered

- let $A = \{x_1\}$

  for any $a \geq x_1$, we get $h_a(x_1) = 1$

  for any $a < x_1$, we get $h_a(x_1) = 0$

- $\exists A : \Pi_{\mathcal{H}}(A) = \{0, 1\} \implies d \geq 1$

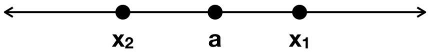
# example – threshold functions

- now show that $d < 2$

  i.e., show that no set of size 2 can be shattered

- let $A = \{x_1, x_2\}$ such that $x_1 \le x_2$

  no $h_a \in \mathcal{H}$ can induce a labeling $(0, 1)$




- $\forall A, \Pi_{\mathcal{H}}(A) \neq \{0, 1\}^2 \implies d < 2$

# example – intervals

- $X = \mathbb{R}$
- $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ where

$$h_{a,b}(x) = \mathbb{I}\left[x \in (a, b)\right], \forall h_{a,b} \in \mathcal{H}$$

- claim: $VCD(\mathcal{H}) = 2$

# example – intervals

- first show that $d$ is at least $2$

  i.e., find a set of size $2$ that can be shattered

- let $A = \{x_1, x_2\}$, $x_1 < x_2$

  $\exists (a, b) \in \mathbb{R}$ s.t. $h_{a,b}(x_1, x_2) = (1, 1)$

  $\exists (a, b) \in \mathbb{R}$ s.t. $h_{a,b}(x_1, x_2) = (1, 0)$

  $\exists (a, b) \in \mathbb{R}$ s.t. $h_{a,b}(x_1, x_2) = (0, 1)$

  $\exists (a, b) \in \mathbb{R}$ s.t. $h_{a,b}(x_1, x_2) = (0, 0)$

- $\exists A : \Pi_{\mathcal{H}}(A) = \{0, 1\}^2 \implies d \geq 2$

# example – intervals

- now show that $d < 3$

  i.e., show that no set of size 3 can be shattered

- let $A = \{x_1, x_2, x_3\}$ such that $x_1 \leq x_2 \leq x_3$

  no $h_{a,b} \in \mathcal{H}$ can induce a labeling $(1, 0, 1)$

  – whenever $x_1, x_3 \in (a, b)$, also $x_2 \in (a, b)$

- $\forall A, \Pi_{\mathcal{H}}(A) \neq \{0, 1\}^3 \implies d < 3$

# example – axis aligned rectangles

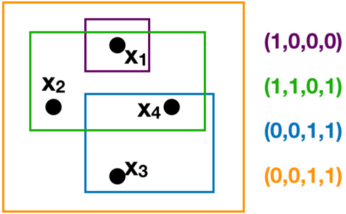
- $X = \mathbb{R}^2$
- $\mathcal{H} = \{h_{a_1,a_2,b_1,b_2} : a_1 \leq a_2, b_1 \leq b_2\}$
- claim: $VCD(\mathcal{H}) = 4$

# example – axis aligned rectangles

- first show that $d$ is at least 4

  i.e., find a set of size 4 that can be shattered

- let $A = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$




- $\exists A : \Pi_{\mathcal{H}}(A) = \{0, 1\}^4 \implies d \geq 4$

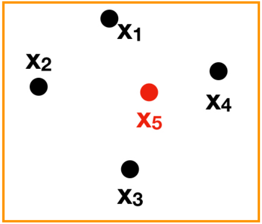
# example – axis aligned rectangles

- now show that $d < 5$

  i.e., show that no set of size 5 can be shattered

- let $A = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$

  no $h_{a_1,a_2,b_1,b_2} \in \mathcal{H}$ can induce a labeling $(1, 1, 1, 1, 0)$




- $\forall A, \Pi_{\mathcal{H}}(A) \neq \{0, 1\}^5 \implies d < 5$

# example – hyperplane classifiers

- hyperplane: let $\mathbf{x}, \mathbf{w} \in \mathbb{R}^n$, $b \in \mathbb{R}$, the equation

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

  specifies a hyperplane in $\mathbb{R}^n$

- a classifier is given by

$$h_{(\mathbf{w},b)}(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + b)$$

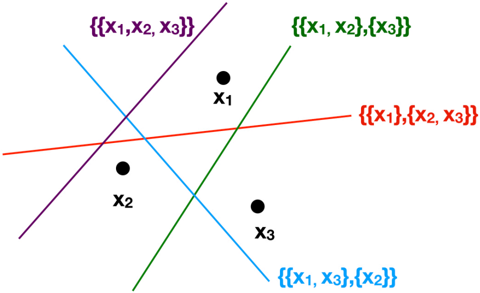
  (i.e., halfspaces define class membership)

- let $\mathcal{H}$ denote the set of hyperplanes defined on $X = \mathbb{R}^n$

$$\mathcal{H} = \{h_{(\mathbf{w},b)} : \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$$

# example – hyperplane classifiers

- claim: for hyperplanes in $\mathbb{R}^2$, $VCD(\mathcal{H}) = 3$
  - a hyperplane in $\mathbb{R}^2$ is a line
- let $A = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ be a set of non-collinear points in $\mathbb{R}^2$

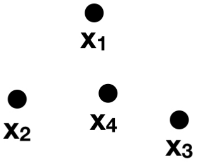



- $\exists A, \Pi_{\mathcal{H}}(A) = \{0,1\}^3 \implies d \geq 3$

# example – hyperplane classifiers

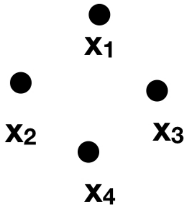
- now show that no set of size 4 can be shattered
- let $A = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ such that no 3 points of $A$ are collinear
- case 1: 3 of the 4 points define the convex hull of $A$

  (convex hull of $A$: smallest convex set that contains $A$)
- no $h_{(\mathbf{w},b)}(\mathbf{x}) \in \mathcal{H}$ can induce the labelings $(1, 1, 1, -1)$ and $(-1, -1, -1, 1)$



# example – hyperplane classifiers

- case 2: all 4 points define the convex hull of $A$
- any halfplane that contains 2 diagonally opposite points (e.g., $\mathbf{x}_1$ and $\mathbf{x}_4$) would also contain a third point from $A$ (e.g., $\mathbf{x}_2$ or $\mathbf{x}_3$)
- no $h_{(\mathbf{w},b)}(\mathbf{x}) \in \mathcal{H}$ can induce the labelings $(1, -1, -1, 1)$ and $(-1, 1, 1, -1)$




- $\forall A, \Pi_{\mathcal{H}}(A) \neq \{0, 1\}^4 \implies d < 4$

# example – hyperplane classifiers

- $\mathcal{H} = \{sign(\mathbf{w} \cdot \mathbf{x} + b) : \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$
- claim: for hyperplanes in $\mathbb{R}^n$, $VCD(\mathcal{H}) = n + 1$

# example – hyperplane classifiers

- let $A = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n\}$ where

$$\mathbf{x}_0 = \mathbf{0}_n \text{ and } \mathbf{x}_i = \mathbf{e}_i, 1 \leq i \leq n$$

- let $y_0, \ldots, y_n \in \{-1, 1\}$ and $b = y_0$
- let $\mathbf{w}$ be the vector with $w_i = y_i - b$ for $1 \leq i \leq n$
- we have $\mathbf{w} \cdot \mathbf{x}_0 + b = y_0$, and

  $\mathbf{w} \cdot \mathbf{x}_i + b = y_i$ for $1 \leq i \leq n$, which means

$$sign(\mathbf{w} \cdot \mathbf{x}_i + b) = y_i$$

- $A$ is shattered by $\mathcal{H}$, $VCD(\mathcal{H}) \geq n + 1$

# example – hyperplane classifiers

- to prove that $VCD(\mathcal{H}) < n + 2$, we need the following result
- Radon's lemma: let $A \subset \mathbb{R}^n$ be a set of size $n + 2$.
  then there exist two disjoint subsets $A_1$ and $A_2$ of $A$ such that the convex hulls of $A_1$ and $A_2$ intersect.
- given Radon's lemma, we need to show that for every $A \subset \mathbb{R}^n$ of size $n + 2$, there is a labelling that cannot be realized using hyperplanes

# example – hyperplane classifiers

- let $A \subset \mathbb{R}^n$ be any set of $n + 2$ points
- let $A_1$ and $A_2$ be two disjoint subsets of $A$
- consider a dichotomy of $A$ in which points in $A_1$ are labelled by $1$ and those in $A_2$ are labelled by $-1$
- fact: when two sets are separated by a hyperplane, their convex hulls are also separated by the hyperplane

# example – hyperplane classifiers

- if a hyperplane assigns a particular label to a set of points,

  then every point in their convex hulls is also assigned the same label

- assume there is a hyperplane consistent with such dichotomy
- from Radon's lemma, convex hulls of $A_1$ and $A_2$ has non-empty intersection, a contradiction
- $\mathcal{H}$ cannot shatter $A$ hence $VCD(\mathcal{H}) < n + 2$.

# the VC dimension – interpretation

- the VC dimension is the maximal size of a subset $A \subset X$ such that $\mathcal{H}$ gives no prior knowledge w.r.t. $A$
- it follows from the proof of no-free-lunch theorem that if

$$m \leq 2VCD(\mathcal{H})$$

  then it might be hard to find a good $h \in \mathcal{H}$ (verify!)
- in other words, a finite VC dimension tells us that we can distinguish between different hypothesis relatively quickly from a modest sample size

# growth function

- for any $m \in \mathbb{N}$, growth function is defined as

$$\Pi_{\mathcal{H}}(m) = max\{|\Pi_{\mathcal{H}}(A)| : |A| = m\}$$

- the growth function further characterizes complexity of $\mathcal{H}$:
  the faster growth, the more dichotomies with increasing $m$
- if $\mathcal{H}$ does not have finite VC dimension, then

$$\Pi_{\mathcal{H}}(m) = 2^m, \forall m$$

- if $VCD(\mathcal{H}) = d$, then

$$\Pi_{\mathcal{H}}(m) = 2^m, \forall m \leq d$$

- what about $m > d$? exponential growth?

# a polynomial bound on $\Pi_{\mathcal{H}}(m)$

- Sauer-Shelah-Perles lemma: let $\mathcal{H}$ be a hypothesis class with $VCD(\mathcal{H}) \leq d < \infty$. then, for all $m$

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

in particular, if $m > d + 1$ then

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^{d} = \mathcal{O}(m^{d})$$