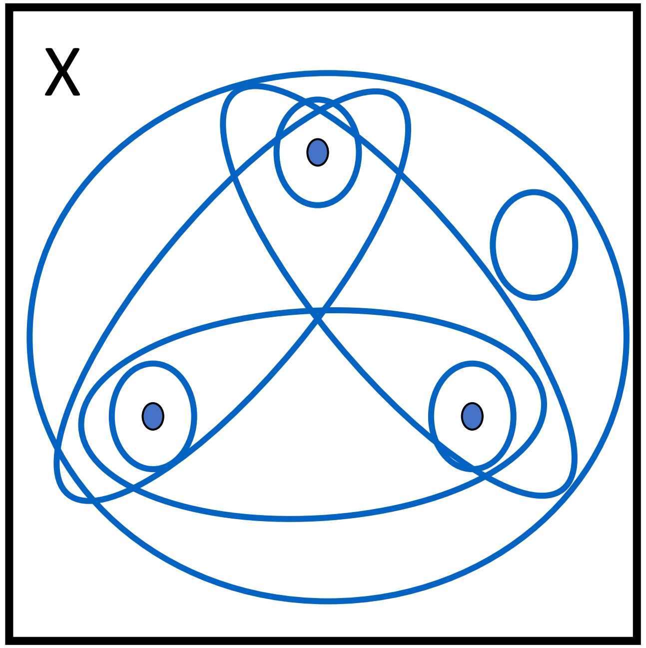
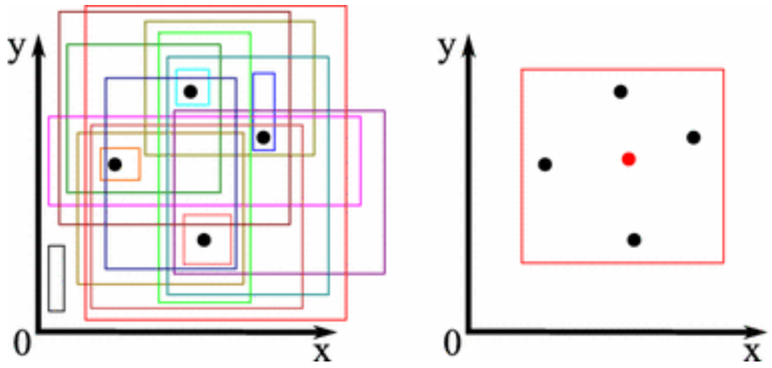
The VC Dimension of H

Ariel Shamir
Ben Galili
Zohar Yakhini





IDC HERZLIYA

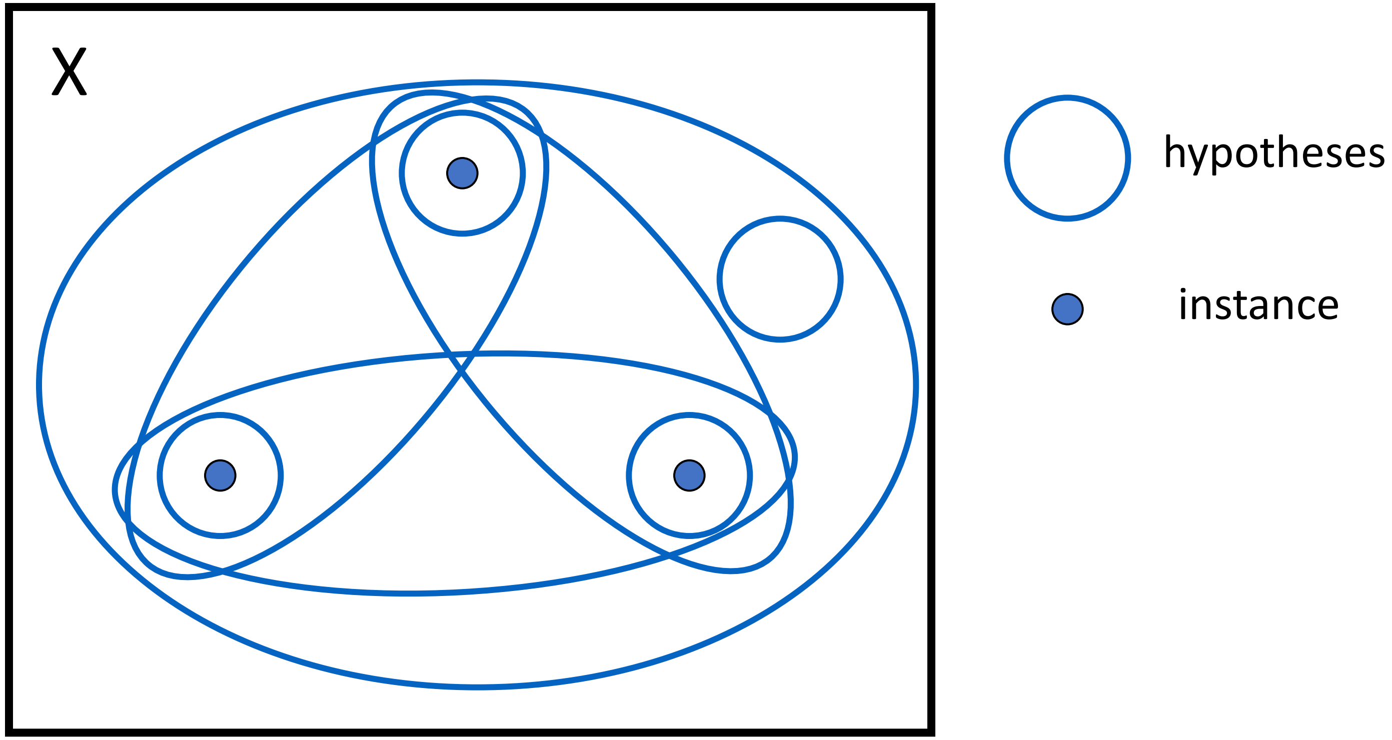
Infinite Hypothesis Spaces

- To analyze sample complexity in infinite Hs it is sometimes possible to use the geometry of H and C.
- Infinite Hs are also partially addressed by the Vapnik-Chervonenkis or VC-dimension.
- In some sense this takes the geometric considerations to the limit.

Shattering a Set of Instances

- A dichotomy of a set S is a partition of S into two disjoint subsets.
- A set of instances S is **shattered** by hypothesis space H if and only if **for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.**

Shattering Visual Example. H = Ellipses



Shattering & Expressiveness

- A hypotheses space that is capable of representing every possible concept (dichotomy) over an instance space X (i.e. an unbiased hypothesis space) is able to **shatter** the space X

- If this is not the case than the larger the subset S of X that H shatters, the more expressive H is.

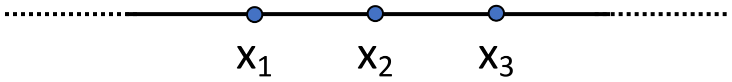
VC Dimension – a definition

- The Vapnik-Chervonenkis dimension, $VC(H)$, of a hypotheses space H, defined over an instance space X, is the size of the largest finite subset of X which is shattered by H.
- *Note: **it suffices** to find one subset of a given size that H can shatter*!
- If arbitrarily large finite sets of X can be shattered by H, then VC(H) = ∞.
- This is a measure for the expressiveness of the hypothesis space H

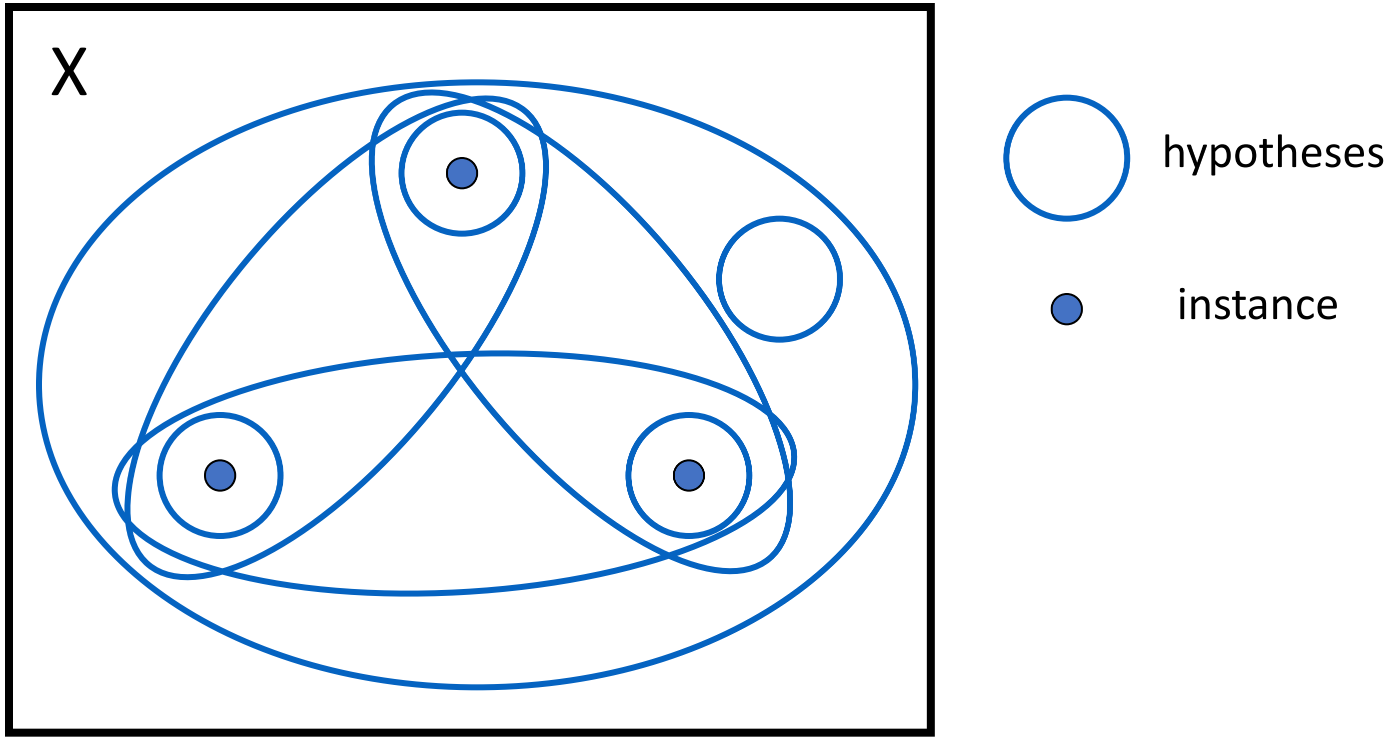
Example 1

- $X = \mathbb{R}$, $H = \{(a,b) \mid a,b \in \mathbb{R}\}$.
- $VC(H) \geq 2$ since:

- $VC(H) < 3$ since any hypothesis represented by an open segment that will include x_1 and x_3 must also include x_2:



Shattering Visual Example. H = Ellipses



Complexity bounds using VC dimension

- The VC dimension of H can be used to estimate sample complexity.
- The VC dimension of $C = H$ provides an upper bound on the required sample complexity of learning.

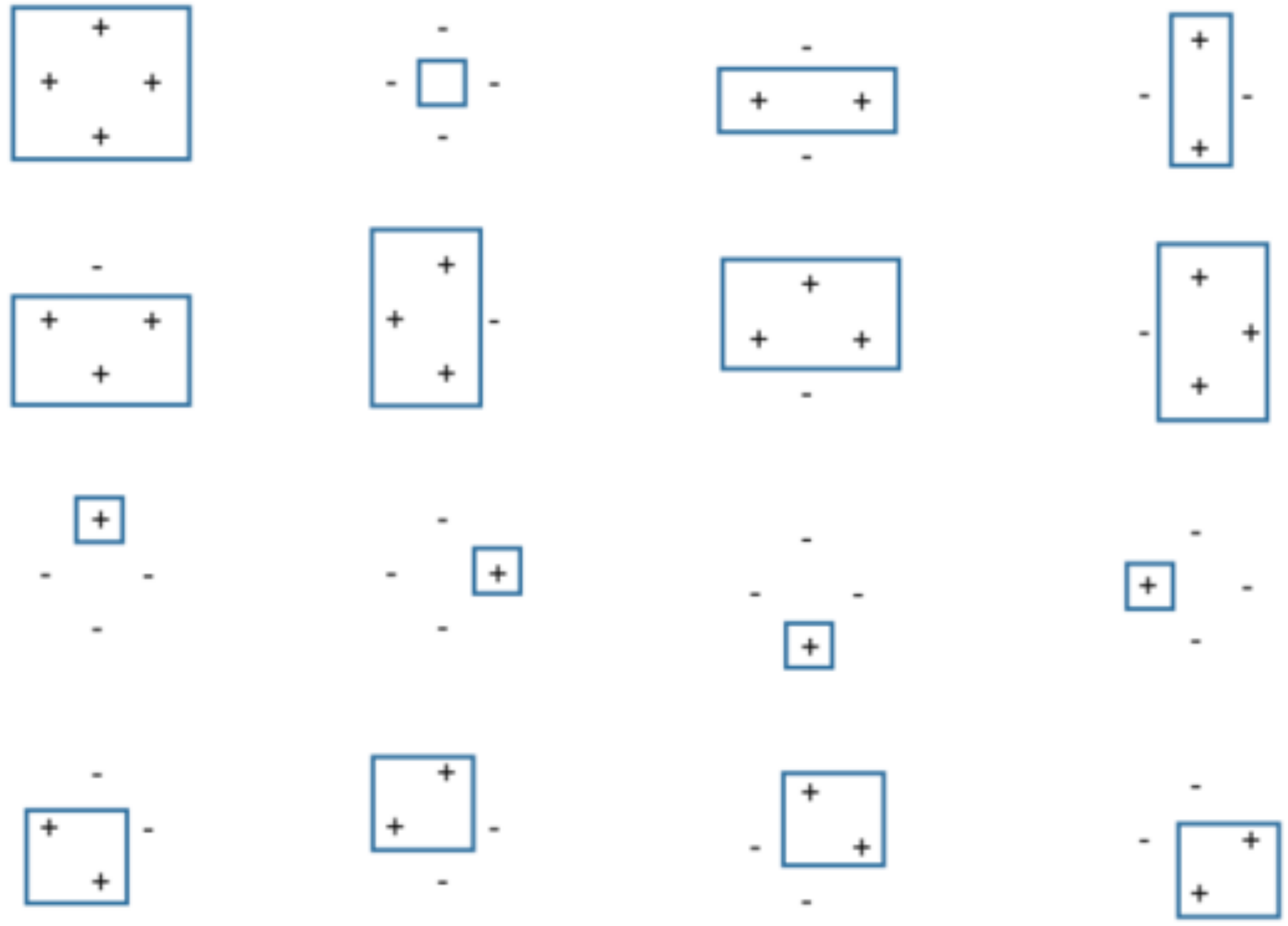
$$m(\varepsilon, \delta) \geq \frac{1}{\varepsilon}\left(4\, log_2\left(\frac{2}{\delta}\right) + 8VC(H)\, log_2\left(\frac{13}{\varepsilon}\right)\right) \quad \text{suffices}$$

Example – VC bound vs direct

- Let
  - $X = \mathbb{R}^2$
  - H be the set of axes aligned rectangles
- We now compare the sample complexity calculation obtained by using the VC bound to the one we directly calculated from te geometry
- First we will calculate the VC dimension of H

Example – VC bound vs direct

- $VC(H) \geq 4$:



Example – VC bound vs direct

- $VC(H) < 5$:
  - Consider any set of five distinct points $\{v_1, v_2, v_3, v_4, v_5\}$
  - Consider a rectangle that contains the points with maximum x-coordinate, minimum x-coordinate, maximum y-coordinate, and minimum y-coordinate. These points may not be distinct
  - However, there are at most four such points. Call this set of points
    $$S \subset \{v_1, v_2, v_3, v_4, v_5\}$$
  - Any axis-aligned rectangle that contains S must also contain all the points v_1, v_2, v_3, v_4, v_5
  - There is at least one v_i that was not used in S, but still must be in the rectangle
  - Therefore, the labeling that labels all points in S with + and v_i with − cannot be consistent with any axis-aligned rectangle
  - This means that there is no shattered set of size 5, and therefore $VC(H) < 5$
- Put together, we get $VC(H) = 4$

Example – VC bound vs direct, axes aligned rectangles

- Let $\varepsilon = 0.05$ and $\delta = 0.05$
- Using the VC bound we get:

$$m \geq \frac{1}{0.05}\left(4\,log_2\left(\frac{2}{0.05}\right) + 32\,log_2\left(\frac{13}{0.05}\right)\right) = 5560$$

- Using the direct calculation we get:

$$m \geq \frac{4}{\varepsilon}\left(\ln(4) + \ln\left(\frac{1}{\delta}\right)\right) = \frac{4}{0.05}(\ln(4) + \ln(20)) = 350$$

More examples of VC dimension in the recitation ….