VC Dimension

PAC and Agnostic Learning

- Finite H, assume target function c ∈ H

$$Pr(\exists h \in H,\ s.t.\ (error_{train}(h) = 0) \wedge (error_{true}(h) > \epsilon)\ ) \leq |H|e^{-\epsilon m}$$

  - Suppose we want this to be at most δ. Then *m* examples suffice:

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

- Finite H, agnostic learning: perhaps c *not* in H

$$P(\exists h \in H,\ |\epsilon(h) - \hat{\epsilon}(h)| > \gamma) \leq 2k\exp(-2\gamma^2 m)$$

  - ➔ $$m \geq \frac{1}{2\gamma^2}\log\frac{2k}{\delta}$$

  - with probability at least (1-δ) every h in H satisfies

$$\epsilon(\hat{h}) \leq \left(\min_{h \in H} \epsilon(h)\right) + 2\sqrt{\frac{1}{m}\log\frac{2k}{\delta}}$$

What if H is not finite?

- Can’t use our result for infinite H

- Need some other measure of complexity for H
  - Vapnik-Chervonenkis (VC) dimension!

Shattering a Set of Instances

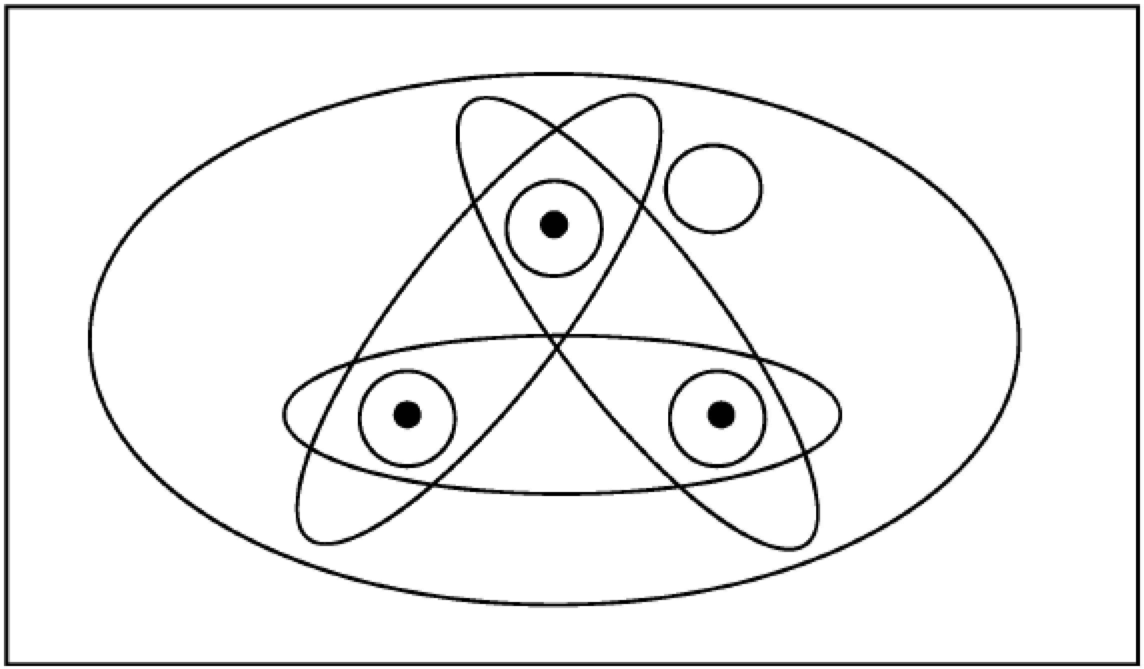
- *Definition*: Given a set $\mathcal{S} = \{x^{(1)}, \ldots, x^{(m)}\}$ (no relation to the training set) of points $x^{(i)} \in$ X, we say that $\mathcal{H}$ shatters $\mathcal{S}$ if $\mathcal{H}$ can realize any labeling on $\mathcal{S}$.

  I.e., if for any set of labels $\{y^{(1)}, \ldots, y^{(d)}\}$, there exists some $h \in \mathcal{H}$ so that $h(x^{(i)}) = y^{(i)}$ for all i = 1, …, m.

- There are 2^m different ways to separate the sample into two sub-samples (a dichotomy)

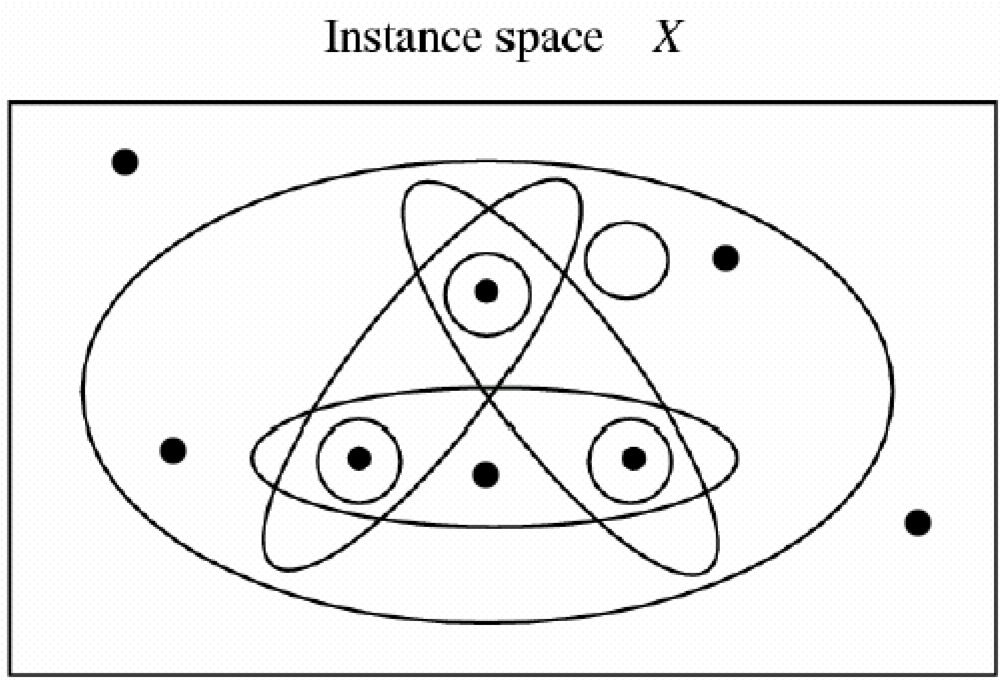
Three Instances Shattered

Instance space X



The Vapnik-Chervonenkis Dimension

- *Definition*: The **Vapnik-Chervonenkis dimension**, $VC(H)$, of hypothesis space H defined over instance space X is the size of the *largest finite subset* of X shattered by H . If arbitrarily large finite sets of X can be shattered by H , then $VC(H) \equiv \infty$.



VC dimension: examples

Consider $X = \mathbb{R}$, want to learn c: $X \rightarrow \{0,1\}$

What is VC dimension of

- Open intervals:

  H1: if $x>a$, then $y=1$ else $y=0$

- Closed intervals:

  H2: if $a<x<b$, then $y=1$ else $y=0$

VC dimension: examples

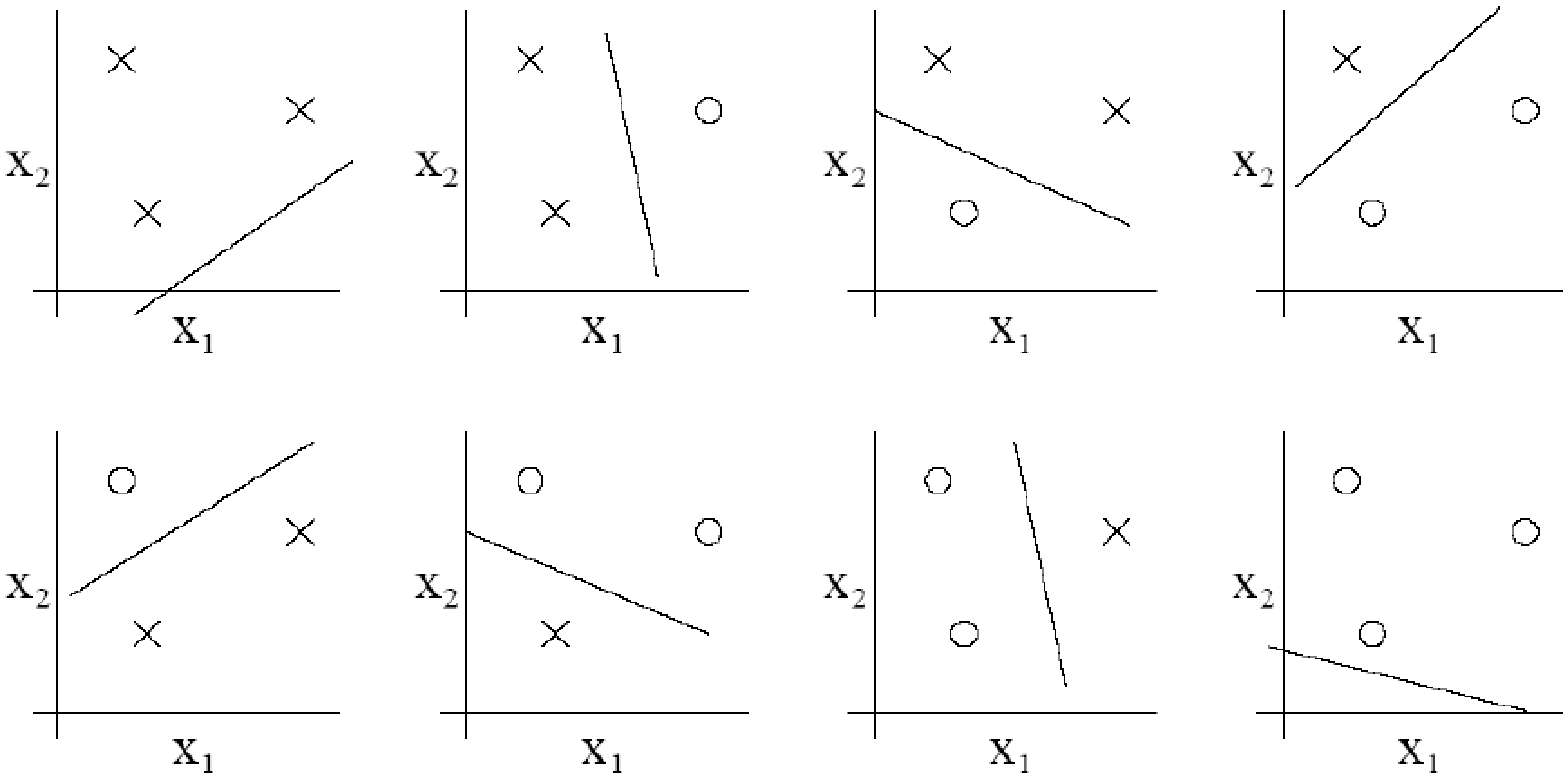
Consider X = $\mathbb{R}^2$, want to learn c: X→{0,1}

- What is VC dimension of lines in a plane?

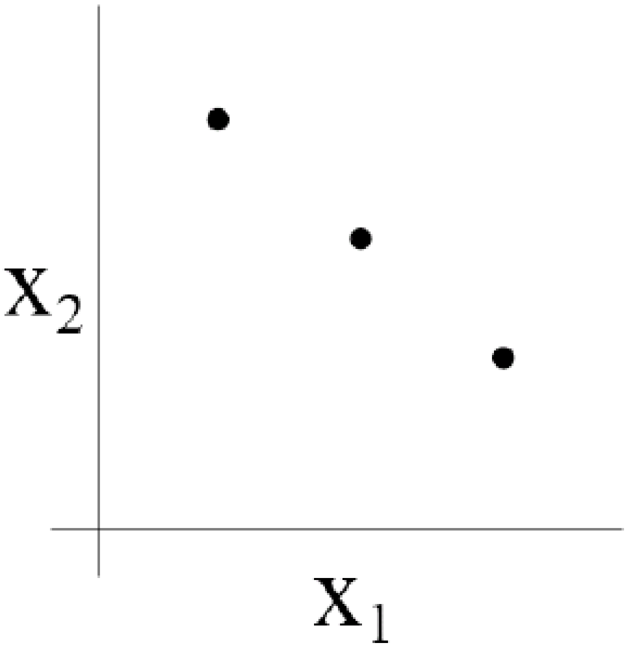
  H= { ((wx+b)>0 → y=1) }

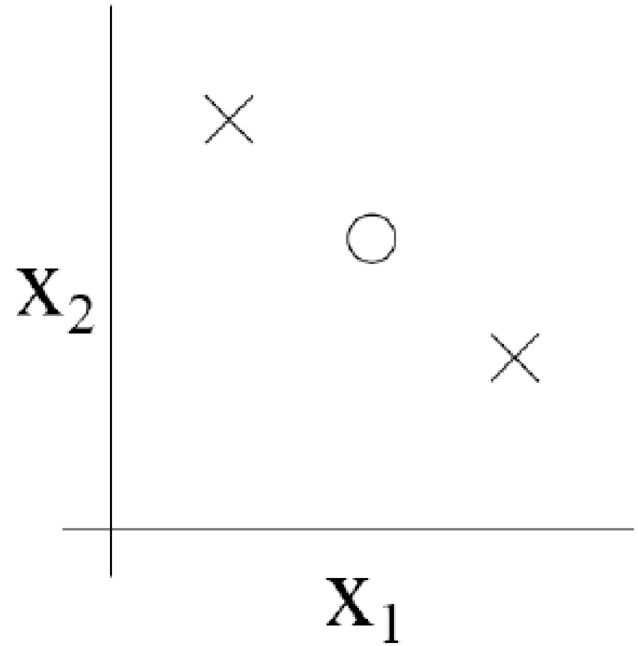
(a)

(b)



- For any of the eight possible labelings of these points, we can find a linear classier that obtains "zero training error" on them.
- Moreover, it is possible to show that there is no set of 4 points that this hypothesis class can shatter.





- The VC dimension of H here is 3 even though there may be sets of size 3 that it cannot shatter.
- under the definition of the VC dimension, in order to prove that VC(H) is at least *d*, we need to show only that there's at least one set of size d that H can shatter.

- **Theorem** Consider some set of m points in $\mathbb{R}^n$. Choose any one of the points as origin. Then the m points can be shattered by oriented hyperplanes if and only if the position vectors of the remaining points are linearly independent.

- **Corollary**: The VC dimension of the set of oriented hyperplanes in $\mathbb{R}^n$ is n+1.

  Proof: we can always choose n + 1 points, and then choose one of the points as origin, such that the position vectors of the remaining n points are linearly independent, but can never choose n + 2 such points (since no n + 1 vectors in $\mathbb{R}^n$ can be linearly independent).

Sample Complexity from VC Dimension

- How many randomly drawn examples suffice to ε-exhaust $VS_{H,S}$ with probability at least (1 - δ)?

ie., to guarantee that any hypothesis that perfectly fits the training data is probably (1-δ) approximately (ε) correct on testing data from the same distribution

$$m \geq \frac{1}{\varepsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\varepsilon))$$

Compare to our earlier results based on |*H*|:

$$m \geq \frac{1}{2\varepsilon^2}(\ln|H| + \ln(1/\delta))$$

The Vapnik-Chervonenkis dimension

- The Vapnik-Chervonenkis dimension of H is d if there exists such an S, |S|=d which it can shatter, but it cannot shatter any S for |S|=d+1 (If it can shatter any finite S then VCD=∞.)
- Theorem: Let **C** be a concept class,a and H a representation set for which VCD(H)=d. Let L be a learning algorithm that learns c∈**C** by getting a set S of training samples with |S|=m, and it outputs a hypothesis h∈H which is consistent with S. The learning of **C** over H is PAC learning if

$$m \geq c_0 \frac{1}{\varepsilon} \left(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right)$$

(where c_0 is a proper constant)

- Remark: In contrary to the finite case, the bound obtained here is tight (that is, m samples are not only sufficient, but in certain cases they are necessary as well).

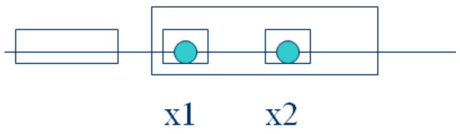
The Vapnik-Chervonenkis dimension

- Let's compare the bounds obtained for the finite and infinite cases:
- Finite case: $m \geq \frac{1}{\varepsilon}(\ln|H| + \ln(\frac{1}{\delta}))$
- Infinite case: $m \geq c_0 \frac{1}{\varepsilon}\left(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right)$

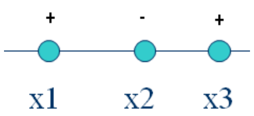
- The two formulas look quite similar, but the role of |H| is taken by the Vapnik-Chervonenkis dimension in the infinite case
  - Both formulas increase relatively slowly as a function of ε and δ, so in this sense these are not bad boundaries…

Examples of VC-dimension

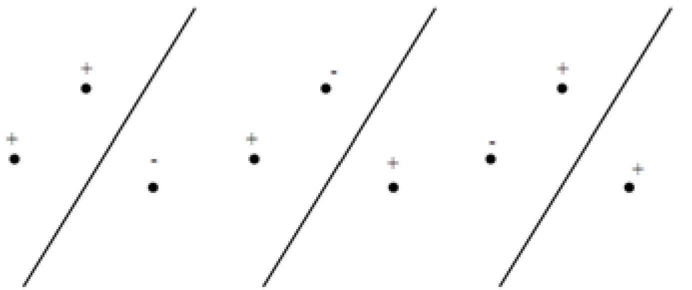
- Finite intervals over the line: VCD=2
  - VCD≥2, as **these** two points can be shattered: (=separated for all label configurations)



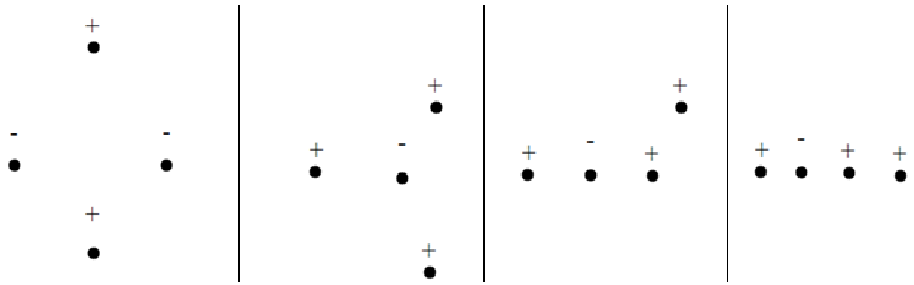
  - VCD<3, as **no** 3 points can be shattered:



- Separating the two classes by lines on the plane: VCD=3 (in d-dimensional space: VCD=d+1)
  - VCD ≥3, as these 3 points can be shattered: (all labeling configurations should be tried!)

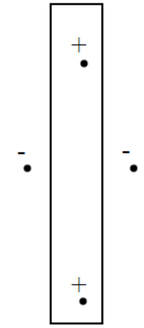


  - VCD<4, as no 4 points can be shattered: (all point arrangements should be tried!)



Examples of VC-dimension

- Axis-aligned rectangles on the plane: VCD=4
  - VCD≥4, as **these** 4 points can be shattered:



  - VCD<5, as **no** 5 points can be shattered:

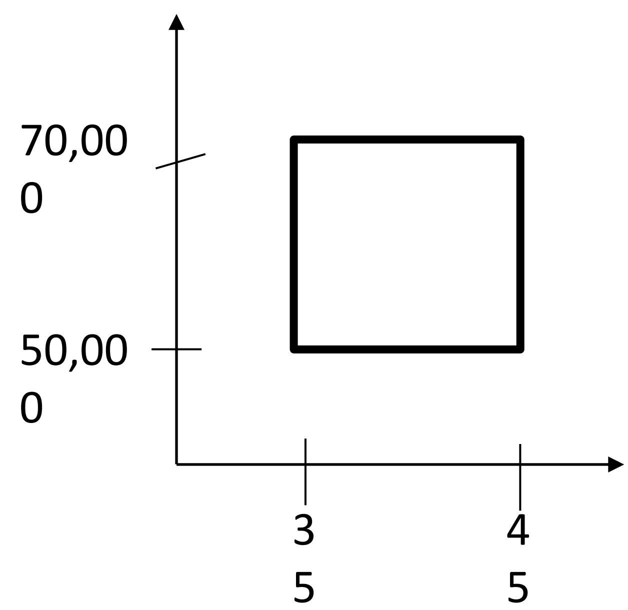
- Convex poligons on the plane: VCD=2d+1 (d is the number of vertices) (the book proofs only one of the directions)
- Conjunctions of literals over $\{0,1\}^n$: VCD=n (See Mitchell's book, only one direction is proved)

Examples

- Intervals of the real axis:
  - $Vcdim = 2, H[n] = O(n^2)$
- Rectangle with axis-parallel edges:
  - $Vcdim = 4, H[n] = O(n^4)$
- Union of 2 intervals of the real axis (Divide an orders set of numbers into two different intervals)
  - $Vcdim = 4, H[n] = O(n^4)$
- Convex polygons:
  - $Vcdim \to \infty, H[n] = 2^n$

Example

- Consider a database consisting of the salary and age for a random sample of the adult population in the United States.
- We are interested in using the database to answer the question:
- What fraction of the adult population in the US has:
- - age between 35 and 45
- - salary between 50,000$ and 70,000$?



Axis Aligned Rectangles

Let $\mathcal{H}$ be the class of axis aligned rectangles, formally:

$$\mathcal{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 \leq a_2 \text{ and } b_1 \leq b_2\}$$

where

$$h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq a_2 \text{ and } b_1 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

The Natarajan Dimension

Natarajan dimension, which is a generalization of the VC dimension to classes of multiclass predictors.

let $\mathcal{H}$ be a hypothesis class of multiclass predictors; namely, each $h \in \mathcal{H}$ is a function from $\mathcal{X}$ to $[k]$.

To define the Natarajan dimension, we first generalize the definition of shattering.

DEFINITION 29.1 (Shattering (Multiclass Version)) We say that a set $C \subset \mathcal{X}$ is shattered by $\mathcal{H}$ if there exist two functions $f_0, f_1 : C \to [k]$ such that

- For every $x \in C$, $f_0(x) \neq f_1(x)$.
- For every $B \subset C$, there exists a function $h \in \mathcal{H}$ such that

$$\forall x \in B, h(x) = f_0(x) \text{ and } \forall x \in C \setminus B, h(x) = f_1(x).$$

DEFINITION 29.2 (Natarajan Dimension) The Natarajan dimension of $\mathcal{H}$, denoted $\mathrm{Ndim}(\mathcal{H})$, is the maximal size of a shattered set $C \subset \mathcal{X}$.

It is not hard to see that in the case that there are exactly two classes, $\mathrm{Ndim}(\mathcal{H}) = \mathrm{VCdim}(\mathcal{H})$. Therefore, the Natarajan dimension generalizes the VC dimension. We next show that the Natarajan dimension allows us to generalize the fundamental theorem of statistical learning from binary classification to multiclass classification.

Thanks