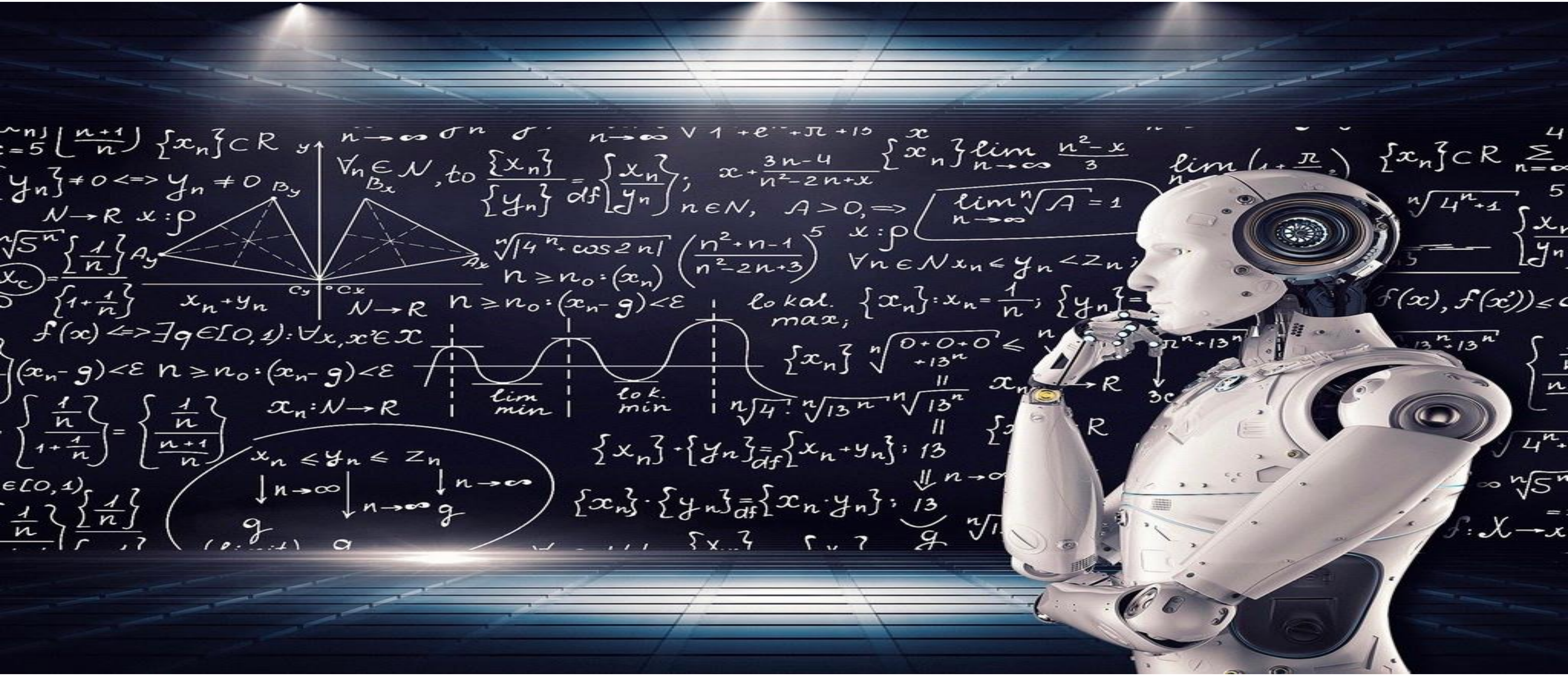
Learning Theory



IDC HERZLIYA

Agenda

- No Free Lunch
- Learning complexity
- VC dimension

No Free Lunch – definition

- $Acc_G(L, c)$ = Generalization accuracy of learner L on concept c

  = Accuracy of L on non-training examples
- The accuracy on the training, without taking into consideration the generalization accuracy is not interesting – we can easily make it 100%
- C = Set of all possible concepts, y=c(x)
  - Concept is a map from the data to label

No Free Lunch

- **Theorem:** For any learner L,

$$\frac{1}{|C|}\sum_{c\in C} Acc_G(L,c) = \frac{1}{2}$$

- The average generalization accuracy over all concepts in C is ½
  - For any given distribution D on X and training set size n
- Why?

No Free Lunch

- **Theorem:** For any learner L,

$$\frac{1}{|C|}\sum_{c\in C} Acc_G(L,c) = \frac{1}{2}$$

- **Proof:** Given any training set S:

For every concept c where $Acc_G(L,c) = \frac{1}{2} + \delta,$

there is a concept c’ where $Acc_G(L,c') = \frac{1}{2} - \delta$

$$\forall x \in S, c'(x) = c(x) = y \quad \forall x \notin S, c'(x) = \neg c(x)$$

No Free Lunch

- **Corollary:**

  - For any two learner L1, L2

    If ∃ learning problem c s.t $Acc_G(L_1, c) > Acc_G(L_2, c)$

    Then ∃ learning problem c' s.t $Acc_G(L_2, c') > Acc_G(L_1, c')$

No Free Lunch – simple example

L1=

x1	x2	x3	y
0	0	0	0
0	0	1	0
1	1	0	1
0	1	0	1
1	1	1	0
0	1	1	1
1	0	0	0
1	0	1	1

Training

Test

L2=

x1	x2	x3	y
0	0	0	0
0	0	1	0
1	1	0	1
0	1	0	1
1	1	1	1
0	1	1	0
1	0	0	1
1	0	1	0

If the concept is (0,0,1,1,0,1,0,0) then L1 is more accurate with 75% and L2 has 25%

If the concept is (0,0,1,1,1,0,1,1) then L2 is more accurate with 75% and L1 has 25%

No Free Lunch – conclusions

- Don’t expect your favorite learner to always be the best
- Simple algorithm can be better sometimes (the complex ones will over fit)
- Try different approaches

Learning Complexity

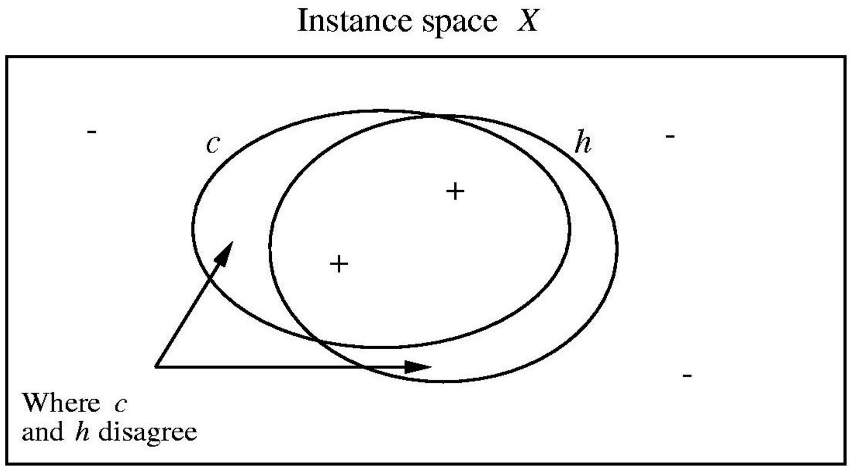
- What is the problem with the training data?
  - We can’t measure our algorithm on the data that we used for learning – it will be pretty easy to get 0% error
  - Do we have enough data?
- We want to infer the true (generalization) error from the training error
- We want to know how many training example are sufficient

Learning Complexity

- Given:
  - Set of instances *X*
  - Set of hypotheses *H*
  - Set of possible target concepts *C*
  - Training instances generated by fixed, unknown probability distribution $\mathcal{D}$ over *X* in an independent manner
- Learner observes a sequence of training examples $\langle x, c(x) \rangle$, for some target concept $c \in C$
- Learner outputs a hypothesis $h \in H$ best estimating *c*
  - *h* should be evaluated by its performance on subsequent instances drawn according to $\mathcal{D}$

Learning Complexity

True Error of a Hypothesis



Definition: The **true error** (denoted $error_{\mathcal{D}}(h)$) of hypothesis h with respect to target concept c and distribution $\mathcal{D}$ is the probability that h will misclassify an instance drawn at random according to $\mathcal{D}$.

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)]$$

Learning Complexity

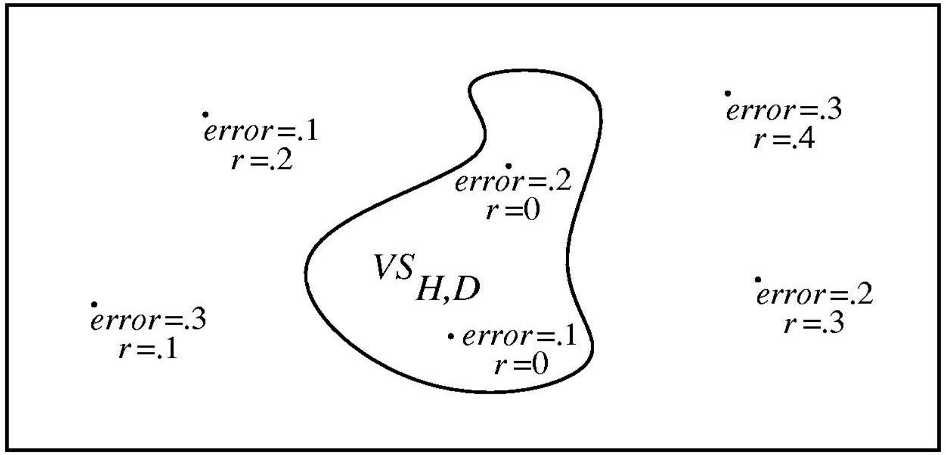
- **Training error:**
  - How often $h(x) \neq c(x)$ over training instances
- **True error:**
  - How often $h(x) \neq c(x)$ over future random instances
- We want to bound the true error given the training error
- First consider when training error of h is zero

Learning Complexity

Version Spaces

Version Space $VS_{H,D}$:
Subset of hypotheses in H consistent with training data D

Hypothesis space H



$(r = \text{training error}, \; error = \text{true error})$

How many examples are enough?

- **Theorem:**

If the hypothesis space *H* is finite, and *D* is a sequence of $m \geq 1$ independent random examples of some target concept c, then for any $0 \leq \varepsilon \leq 1$, the probability that there exists $\mathrm{h} \in VS_{H,D}$ with $error_D(h) > \varepsilon$ is less than:

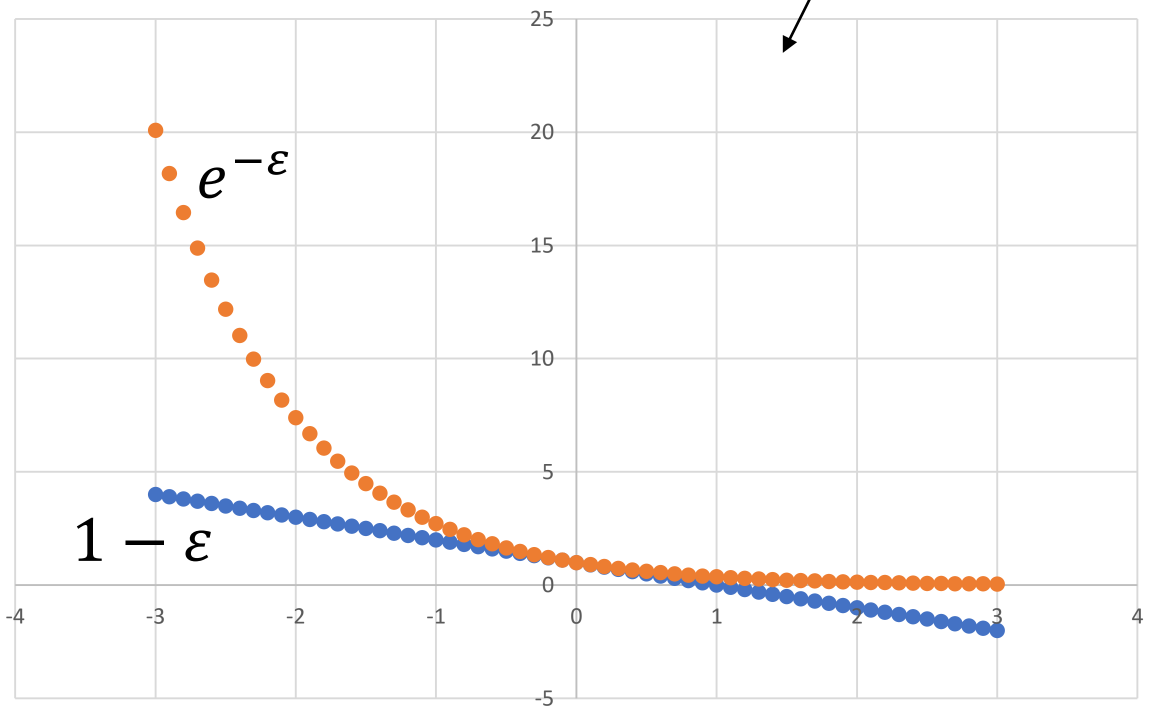
$$|H|e^{-\varepsilon m}$$

How many examples are enough?

- **Proof:**

By definition

P(1 hyp. w/ error > ε consistent w/ 1 ex.) $< 1-\varepsilon \leq e^{-\varepsilon}$



How many examples are enough?

- **Proof:**

P(1 hyp. w/ error > ε consistent w/ 1 ex.) $< 1 - \varepsilon \leq e^{-\varepsilon}$

P(1 hyp. w/ error > ε consistent w/ m ex.) $< (e^{-\varepsilon})^m = e^{-m\varepsilon}$

* D is a sequence of $m \geq 1$ independent random examples

How many examples are enough?

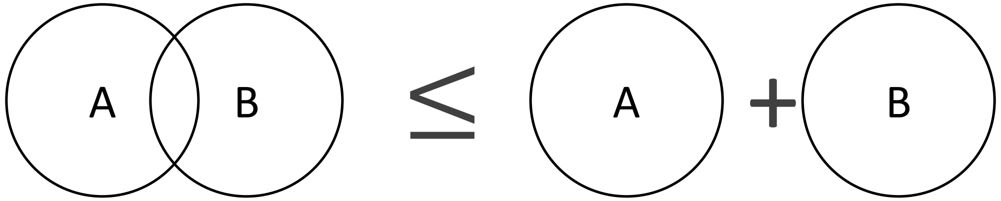
- **Proof:**

P(1 hyp. w/ error > ε consistent w/ 1 ex.) $< 1 - \varepsilon \leq e^{-\varepsilon}$

P(1 hyp. w/ error > ε consistent w/ m ex.) $< (e^{-\varepsilon})^{m} = e^{-m\varepsilon}$

P(1 of $|H|$ hyps. w/ error > ε consistent w/ m ex.) $\leq |H| e^{-m\varepsilon}$

* Because of Union Bound



How many examples are enough?

- This bounds the probability that any consistent learner will output a hypothesis *h* with $error_D(h) \geq \varepsilon$

- We want this probability to be at most δ

$$|H|e^{-\varepsilon m} \leq \delta$$
$$\ln(|H|e^{-\varepsilon m}) \leq \ln(\delta)$$
$$\ln(|H|) + \ln(e^{-\varepsilon m}) \leq \ln(\delta)$$
$$-\varepsilon m \leq \ln(\delta) - \ln(|H|)$$
$$m \geq \frac{1}{\varepsilon}(\ln(|H|) - \ln(\delta))$$
$$m \geq \frac{1}{\varepsilon}\left(\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right)$$

How many examples are enough?

- Now, we can know how many examples are sufficient to ensure with probability at least $(1-\delta)$ that every *h* in $VS_{H,D}$ satisfies *error(h)* $\leq \varepsilon$ (true error)
- We use the formula from the theorem:

$$m \geq \frac{1}{\varepsilon}\left(\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right)$$

How many examples are enough? Example I

- Suppose *H* contains conjunctions of constraints on up to *n=13* Boolean attributes. Then $|H| = 3^{13} = 1594323$
- We want to ensure in 95% that our hypothesis will have error < 5%

$$m \geq \frac{1}{0.05}\left(\ln(1594323) + \ln\left(\frac{1}{0.05}\right)\right) = 346$$

How many examples are enough? Example II

- 1 attribute with 3 values
- 9 attributes with 2 values

$$|X| = 3 \times 2^9$$

- H contains conjunctions of attributes, then

$$|H| = 4 \times 3^9 = 78733$$

- We want to ensure in 95% that our hypothesis will have error < 10%

$$m \geq \frac{1}{0.1}\left(\ln(78733) + \ln\left(\frac{1}{0.05}\right)\right) = 143$$

VC Dimension

- The VC dimension (for Vapnik–Chervonenkis dimension) is a measure of the capacity (complexity, expressive power, richness, or flexibility) of a statistical classification algorithm, defined as the cardinality of the largest set of points that the algorithm can shatter

Shattering

- Definition :
  An hypothesis class H **shatters** a set of points $\mathrm{X} = \{x_1, x_2, \ldots, x_m\} \in U$ iff
  for **<u>every</u>** assignment $\mathrm{Y} = \{y_1, y_2, \ldots, y_m\} \in \{-1, 1\}^m$,
  there exists $h \in H$ s.t $\forall i$: $h\,(x_i) = y_i$

- Let $S(H, X) \;= \begin{cases} T & H\ Shatters\ X \\ F & H\ Can't\ shatter\ X \end{cases}$

- If $S(H, X) = F$ this means there is a specific assignment $y_1, y_2, \ldots, y_m$ for which $\forall h \in H\ \exists i\ h(x_i) \neq y_i$

Shattering

- Let U be some universe and let $X = \{x_1, x_2\}$. how many possible assignments Y does X have?

Y_1	Y_2	Y_3	Y_4
$X_1 = -1$	$X_1 = 1$	$X_1 = -1$	$X_1 = 1$
$X_2 = -1$	$X_2 = -1$	$X_2 = 1$	$X_2 = 1$

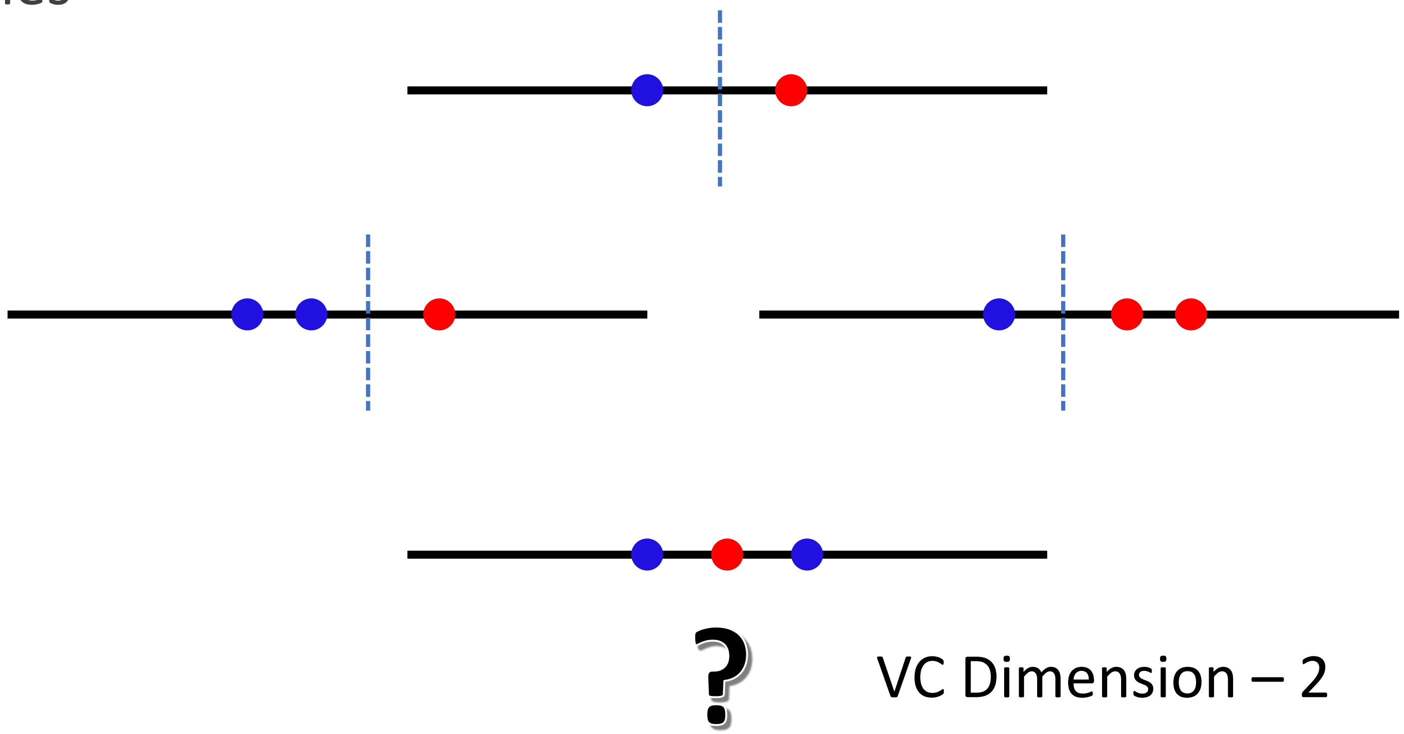
- Let H by some hypothesis space.
  - Can $S(H, X) = True$ if $|\boldsymbol{H}| < \boldsymbol{4}$?
  - Can $S(H, X) = True$ if $\boldsymbol{h}(\boldsymbol{x_2}) = -\boldsymbol{1}\ \forall \boldsymbol{h} \in \boldsymbol{H}$?
  - Can $S(H, X) = True$ if $\boldsymbol{h}(\boldsymbol{x_1}) = \boldsymbol{h}(\boldsymbol{x_2}) \forall \boldsymbol{h} \in \boldsymbol{H}$?

VC Dimension

- *VC(H)*, VC dimension of *H*, defined over instance space *U,* is the size of the largest finite subset of *X* shattered by hypothesis space *H*
- *Note: it's enough to find one subset of a given size that H can shatter*!
- If arbitrarily large finite sets of *U* can be shattered by *H*, then $VC(H) = \infty$
- This is a measure for the hypothesis space *H*

Shattering – example I

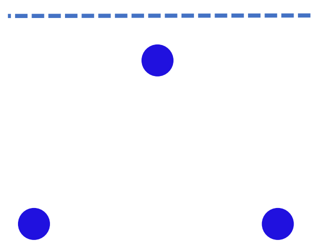
- 1-dimension space
- *H* – linear lines

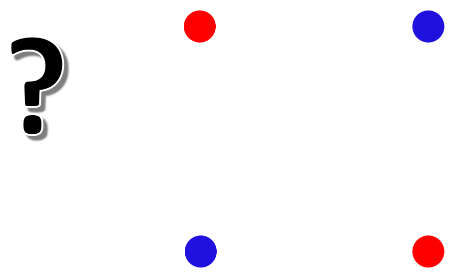


VC Dimension – 2

Shattering – example II

- 2-dimension space
- *H* – linear lines





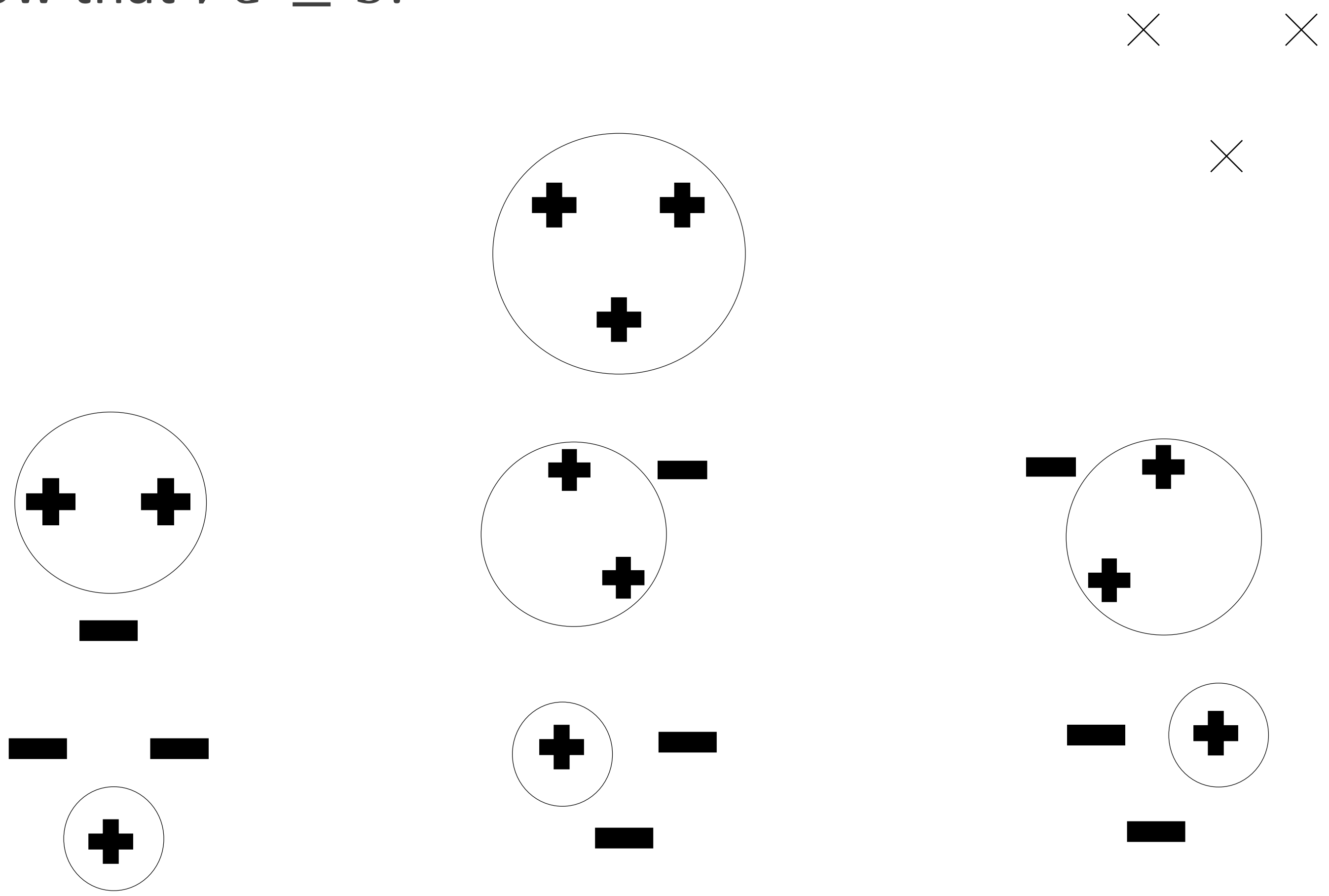
- VC Dimension – 3
- We need to find only one subset of instances – not all!

VC Dimension - example

- Consider U, the instances space, to be the set of all points in the 2-D plane, i.e., $(x, y) \in \mathbb{R}^2$
- Give the VC dimension where the hypothesis space is the set of all circles (the internal part of each circle is classify as positive)

VC Dimension - example

- First, we'll show that $VC \geq 3$:

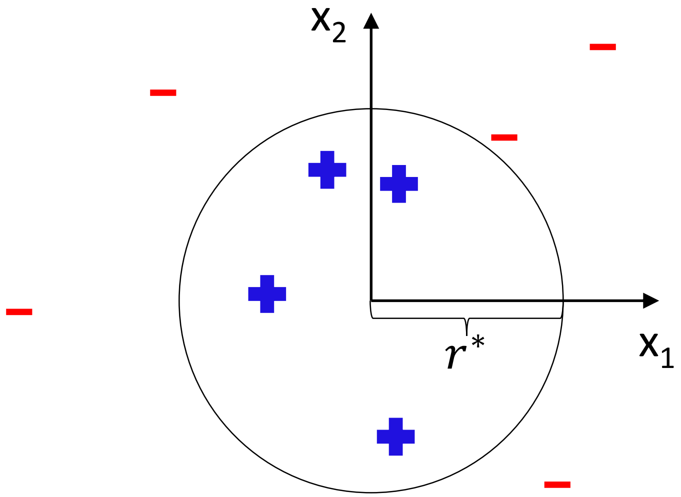


VC Dimension - example

- Second, we'll show that $VC < 4$:
- We show this by constructing a counterexample in several cases
  - If the four points are collinear, the labeling +-+- (going along the line) is impossible, among numerous others
  - If the convex hull of the four points is a triangle, then the labeling with + (the three points of the triangle) and - (the interior point) is not possible
  - If the convex hull of the four points is a quadrilateral, then let (a1, a2) be the points separated by the long diagonal and (b1, b2) be the points separated by the short diagonal. At least one of the labelings +(a1, a2), −(b1, b2) or +(b1, b2), −(a1, a2) must be impossible:
    - If they were both possible, then there would be some satisfying circle c1 for the first labeling and some other circle c2 satisfying the second labeling, and the symmetric difference of these circles ((c1 \ c2) ∪ (c2 \ c1)) would consist of four disjoint regions, which is impossible for circles
- Since some set of 3 points is shattered by the class of circles, and no set of 4 points is, the VC dimension of the class of circles is 3

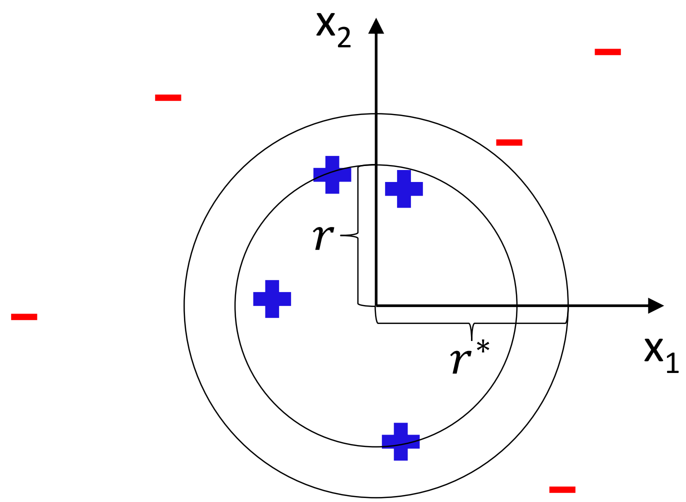
Direct calculation of sample complexity

- Consider a game to learn an unknown concentric circle in the Euclidean plane with 2 dimensions
- Let $r*$ be the radius of the target circle
- Each instance viewed in the sample is drawn from an unknown distribution $\mathscr{D}$ and comprises of 2 features (position of the instance, (x_1, x_2)) and a target value ($+1$ if it's inside the circle and -1 otherwise)



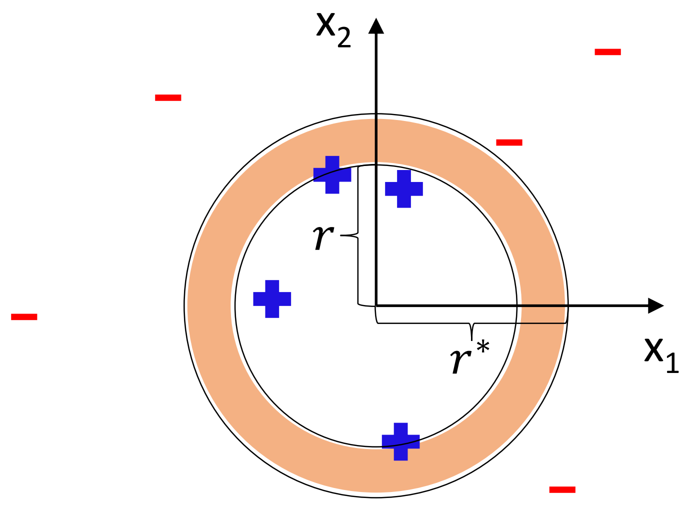
Direct calculation of sample complexity

- **Theorem** – The concept class of concentric circles is efficiently PAC-learnable
- **Proof** – Note that there is a simple and efficient way to come up with a hypothesis r by taking a large number of examples and fitting the tightest circle around the positive ones so that all the given positive points lie inside it



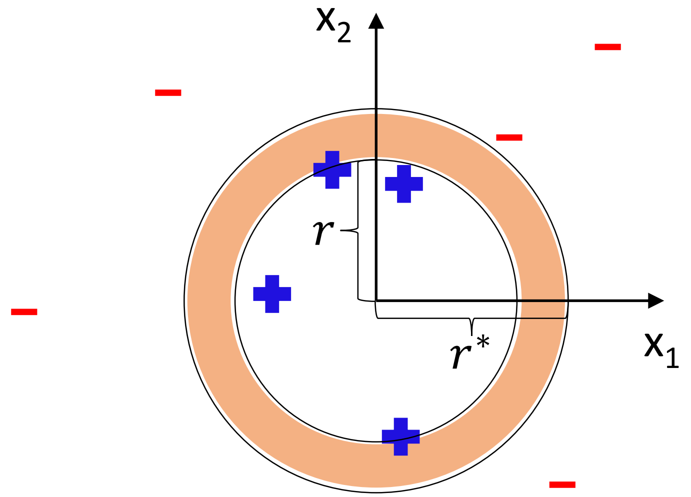
Direct calculation of sample complexity

- Note that $r \leq r^*$ always
- The error annulus is given in solid color
- Intuition – the more examples we see, the closer r gets to r^*, and the smaller the annulus becomes



Direct calculation of sample complexity

- Possible problem – what if the distribution $\mathcal{D}$ is such that observing a point in the annulus A_r is a very unlikely event?
- In this case, the r will converge to r^* very slowly, or not at all! Is this really a problem?
- If we want error ε, then if $\Pr[(x_1, x_2) \in A_r] \leq$ ε we have no problem since the error will be small enough



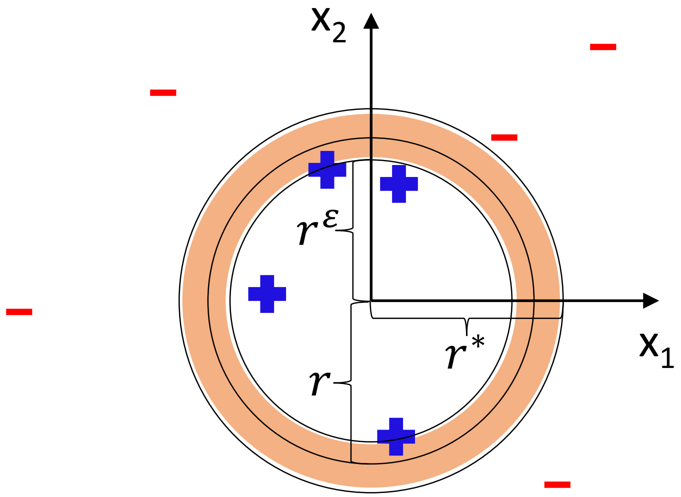
Direct calculation of sample complexity

- We split the proof in two cases
- Define

$$r^{\varepsilon} = \text{arginf}_r \Pr[(x_1, x_2) \in A_r] \leq \varepsilon$$

i.e. the largest annulus with probability at most ε

- Case 1: If $r^{\varepsilon} \leq r$ then the probability of the annulus is less than ε



Direct calculation of sample complexity

- Case 2: Otherwise, what is the probability of missing the annulus of radii r^{ε}, r^{*} with m traning examples?

$$(1-\varepsilon)^{m} \leq \exp(-\varepsilon m)$$

- With sample size $\mathrm{m} \geq \frac{\ln\left(\frac{1}{\delta}\right)}{\epsilon}$, we get

$$\exp(-\varepsilon m) \leq \exp\left(-\ln\left(\frac{1}{\delta}\right)\right) = \exp(\ln(\delta)) = \delta$$

- So if the probability of the annulus is very small, the error it incurs is also small
- With enough examples, it is very unlikely to miss the annulus

Questions

?