VC Dimension

Lecture 12

Last Time

- ***Agnostic probably approximately correct (PAC) learning*** is a property of a hypothesis class $\mathcal{H}$. If it holds, there's a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and an algorithm such that if we have m i.i.d. examples where $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, then with probability at least $1 - \delta$ the algorithm returns h such that

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

- We've shown ***any finite hypothesis class is agnostic PAC learnable*** via ERM with respect to a loss function with range [0,1], with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2 \log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

  Textbook: chapter 4

This Class

- Our final tool of learning theory: what makes a hypothesis class learnable? Can infinite hypothesis classes ever be learnable?

- Textbook: chapters 6.0, 6.1, 6.2, 6.3, 6.4, 9.1.3

An Infinite Example

Example: Can Cats Have Salami?

- Let’s say we randomly sample cats and measure
    - What fraction of their diet is salami?
    - Does their vet say they’re healthy?

- What fraction can be salami and the cat still be healthy? 10%? 20%? 30%?

Example: 1-D Threshold Functions

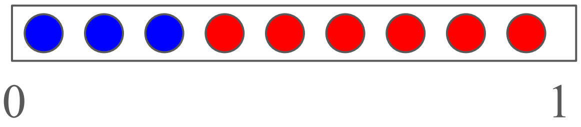
- Like a 1-D halfspace, except assume that lower values get positive label:

- Assume we're in the realizable setting and $a^\star$ is the true threshold

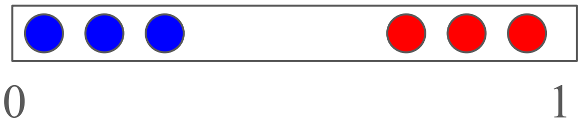
- But, as usual in distribution-free learning, no additional assumptions on $\mathcal{D}_x$, the marginal distribution over the attribute x

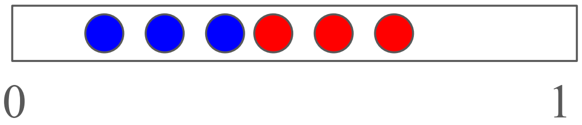
ERM for 1-D Threshold Functions

How do we choose a for a sample S?



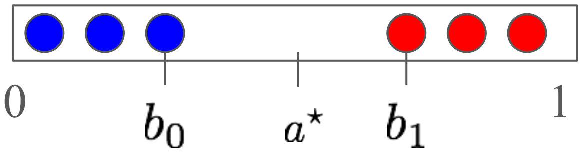
Which sample below would you rather have?



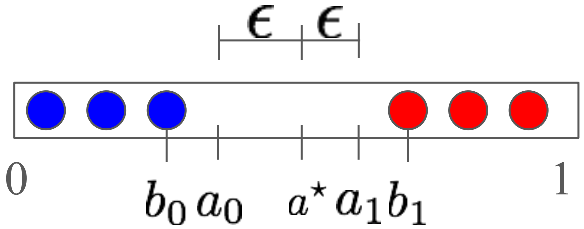


Analyzing ERM for 1-D Threshold Functions

Call the largest x observed to be positive b_0 and the smallest observed negative b_1



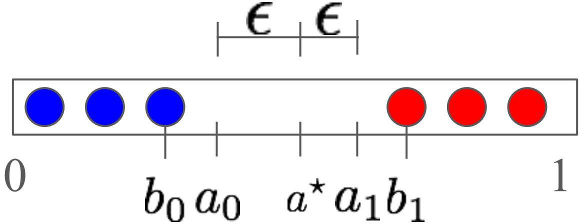
Denote the endpoint of the ϵ probability mass to the left of $a^\star$ as a_0 and the endpoint of another ϵ mass to the right as a_1, where the mass is according to $\mathcal{D}_x$



Question

What is the Probability of Failure?

We've defined our learning problem as follows:



Which of the following is an upper bound on the probability that learning fails, i.e., $\mathcal{D}^m(\{S : L_{\mathcal{D}}(h_S) > \epsilon\})$?

A: $\mathcal{D}^m(\{S : b_0 < a_0 \vee b_1 > a_1\})$

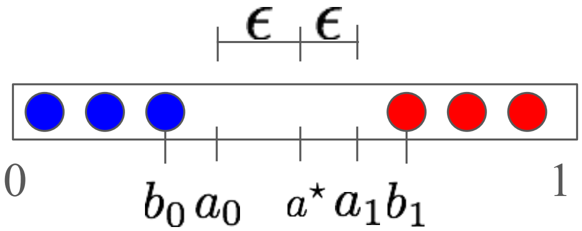
B: $\mathcal{D}^m(\{S : b_0 > a_0 \wedge b_1 < a_1\})$

C: $\mathcal{D}^m(\{S : b_0 > a_0 \vee b_1 < a_1\})$

D: $\mathcal{D}^m(\{S : b_0 < a_0 \wedge b_1 > a_1\})$

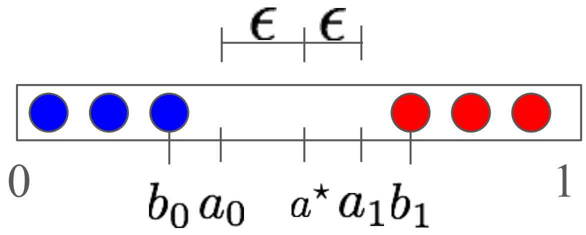
Answer

Answer: probability that b_0 or b_1 is too far from a* (A)



- If we choose a between a_0 and a_1 then the maximum error is ϵ
- We will choose such an a if $b_0 > a_0$ and $b_1 < a_1$
- Therefore $\mathcal{D}^m(\{S : L_{\mathcal{D}}(h_S) > \epsilon\}) \leq \mathcal{D}^m(\{S : b_0 < a_0 \vee b_1 > a_1\})$

Finishing our Analysis



$$\mathcal{D}^m(\{S : L_{\mathcal{D}}(h_S) > \epsilon\})$$

$$\leq \mathcal{D}^m(\{S : b_0 < a_0 \vee b_1 > a_1\})$$ (previous slide)

$$\leq \mathcal{D}^m(\{S : b_0 < a_0\}) + \mathcal{D}^m(\{S : b_1 > a_1\})$$ (union bound)

$$= \mathcal{D}^m(\{S : \nexists x_i \in (a_0, a^\star)\}) + \mathcal{D}^m(\{S : \nexists x_i \in (a^\star, a_1)\})$$ (definition of b_0 and b_1)

$$= (1-\epsilon)^m + (1-\epsilon)^m$$ (definition of ϵ)

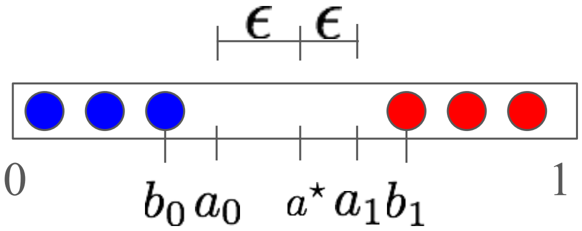
$$\leq 2e^{-\epsilon m}$$ ($1 - \epsilon \leq e^{-\epsilon}$)

Conclusions

- For any δ in (0, 1), if we choose

  $$m \geq \left\lceil \frac{\log(2/\delta)}{\epsilon} \right\rceil$$, then with probability

  at least $1 - \delta$, $L_{\mathcal{D}}(h_S) \leq \epsilon$



- It is not necessary for hypothesis classes to be finite in order to be learnable!

Intuition

- What made it so we could pick an ϵ-accurate hypothesis out of an infinitely large set?

- Could we use the same argument for an arbitrary function in 1-d?

- We're using the fact that the hypothesis class is highly structured. There's really one degree of freedom a and the data we get helps us narrow it down

- Intuitively, the *dimension* of the hypothesis space is important

Shattering

Shattering

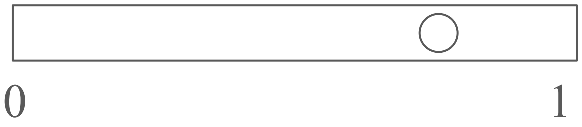
- Intuitively:
  Every possible true labeling of a set of points can be classified with zero training error

- Formally:
  A hypothesis class $\mathcal{H}$ *shatters* a finite set $C \subset \mathcal{X}$ if, for every possible assignment of outputs to the points in C, there's some $h \in \mathcal{H}$ that induces it

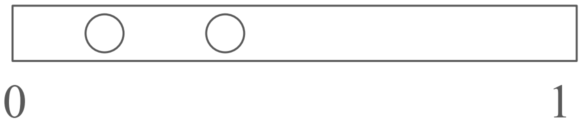
- Example for threshold hypothesis class:

Shattering Example: 1-d

Does our threshold hypothesis class shatter this data?

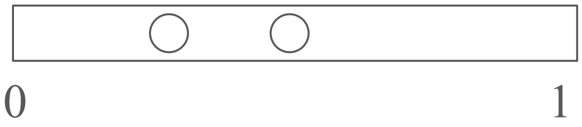


What about this one?

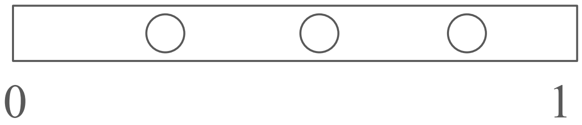


Shattering Example: 1-d

- What if we extend our hypothesis class to 1-d non-homogeneous halfspaces?

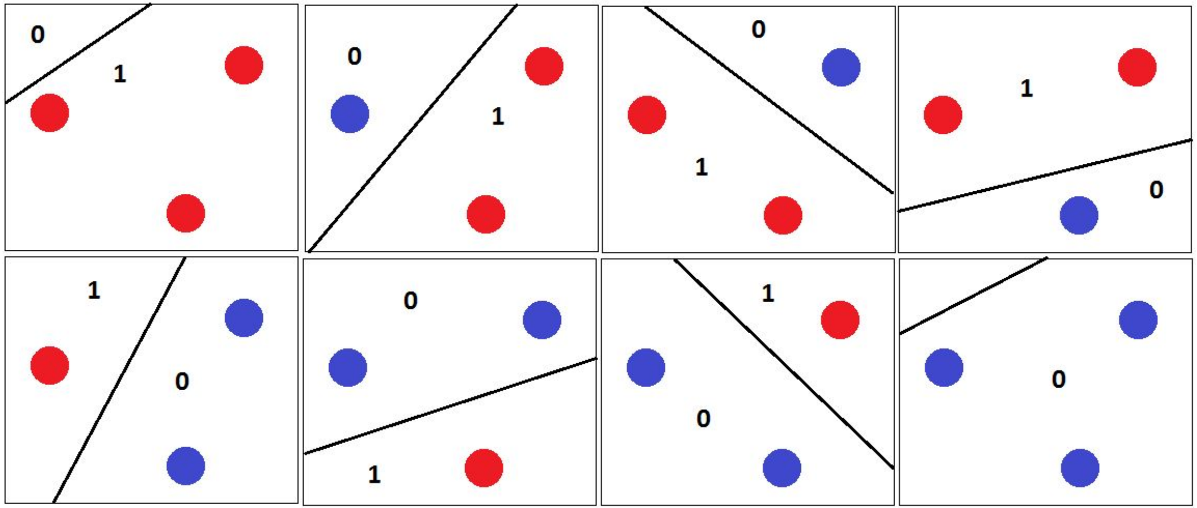


- And does the set of 1-d non-homogeneous halfspaces shatter this set?



Shattering Example: 2-d

- Any three noncollinear points in $\mathbb{R}^2$ are shattered by the class of 2-d non-homogeneous halfspaces:



VC Dimension

Vapnik-Chervonenkis Dimension

$\mathrm{VCdim}(\mathcal{H})$: The maximal size set C such that $\mathcal{H}$ shatters C.

Intuitively:
The largest number of distinct points such that every possible labeling of the points can be classified with zero error.

Proving VC Dimension is some integer d:

1. There exists a set C of size *d* that is shattered by $\mathcal{H}$
2. Every set C of size $d+1$ is not shattered by $\mathcal{H}$

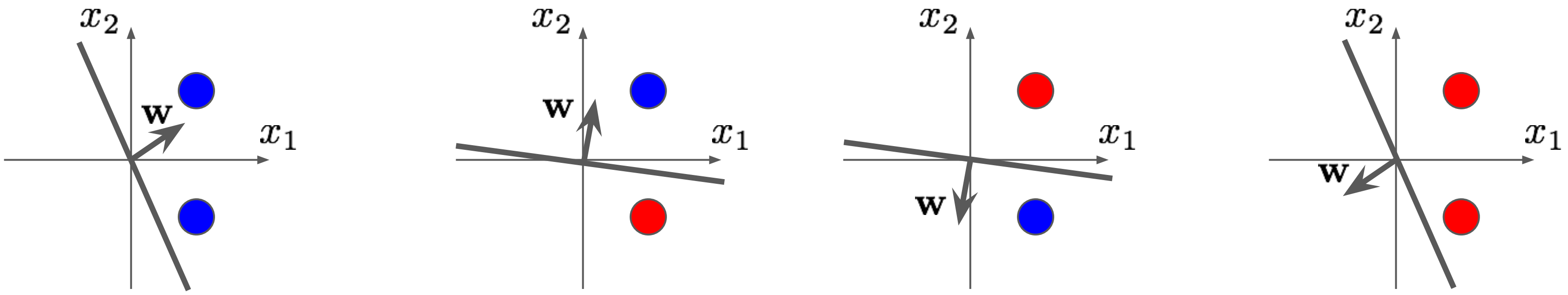
VC Dimension of Halfspaces

- The VC dimension of the class of homogenous halfspaces in $\mathbb{R}^d$ is d

- The VC dimension of the class of non-homogeneous halfspaces in $\mathbb{R}^d$ is $d + 1$

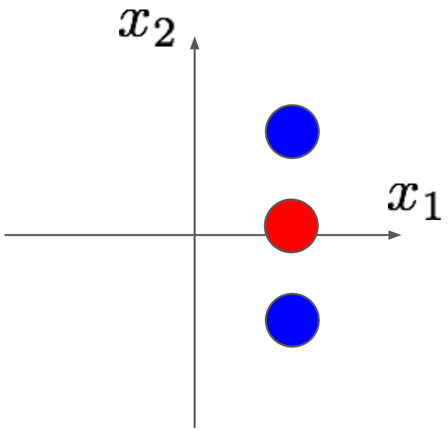
Example: Homogeneous Halfspaces in 2D

- Want to show: the VC dimension of homogenous halfspaces in $\mathbb{R}^2$ is 2
- Step 1: there exists 2 points in $\mathbb{R}^2$ that are shattered



Example: Homogeneous Halfspaces in 2D

- Want to show: the VC dimension of homogenous halfspaces in $\mathbb{R}^2$ is 2
- Step 2: there **do not exist** 3 points in $\mathbb{R}^2$ that are shattered
- Intuition:



Not linearly separable with **homogeneous** halfspace!

Step 2: Formally

- Need to make an argument about **any** three points

- Sufficient rigor for homework/exams:
    - Pick any two points and their labels
    - Learn a halfspace
    - Add a third point anywhere
    - There is a labeling of the third point that is not realized by that halfspace

- Fully formal:
  theorems 9.2 and 9.3

Question

VC Dimension?

Consider the hypothesis class for 1-d data with parameter a:

$$\mathcal{H} : h_a(x) = \begin{cases} 1 & \text{if } |x - a| < 0.2 \\ 0 & \text{otherwise} \end{cases}$$

Recall that VC dimension is integer d such that:

1. There exists a set C of size *d* that is shattered by $\mathcal{H}$.
2. Every set $d+1$ of size C is not shattered by $\mathcal{H}$.

What is VCdim($\mathcal{H}$)?

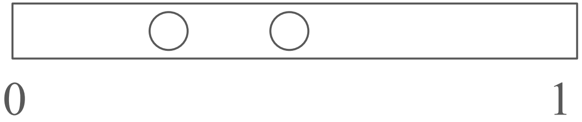
A. 1
B. 2
C. 3
D. 4

Answer

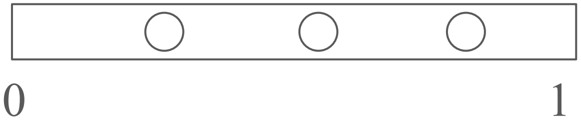
Answer: 2 (B)

$$\mathcal{H} : h_a(x) = \begin{cases} 1 & \text{if } |x - a| < 0.2 \\ 0 & \text{otherwise} \end{cases}$$

We can shatter this set of size 2:



But not any set of size 3 (always can set middle to opposite label of others):



The Fundamental Theorem of Statistical Learning

Let H be a hypothesis class of functions from a domain X to {0, 1} and let the loss function be the 0−1 loss. Then, the following are equivalent:

1. H has the uniform convergence property
2. Any ERM rule is a successful agnostic PAC learner for H
3. H is agnostic PAC learnable
4. H is PAC learnable
5. An ERM algorithm is a successful PAC learner for H
6. H has a finite VC-dimension

VC Dimension also Determines Sample Complexity

- Sample complexity of uniform convergence is linear in d
- Sample complexity of agnostic PAC learning is linear in d
- Sample complexity of PAC learning is linear in d

Full details in Theorem 6.8

Other Hypothesis Classes

VC Dimension of Boosting

LEMMA 10.3 *Let B be a base class and let $L(B,T)$ be as defined in Equation (10.4). Assume that both T and* $\mathrm{VCdim}(B)$ *are at least* 3. *Then,*

$$\mathrm{VCdim}(L(B,T)) \leq T\,(\mathrm{VCdim}(B)+1)\,(3\log(T\,(\mathrm{VCdim}(B)+1))+2).$$

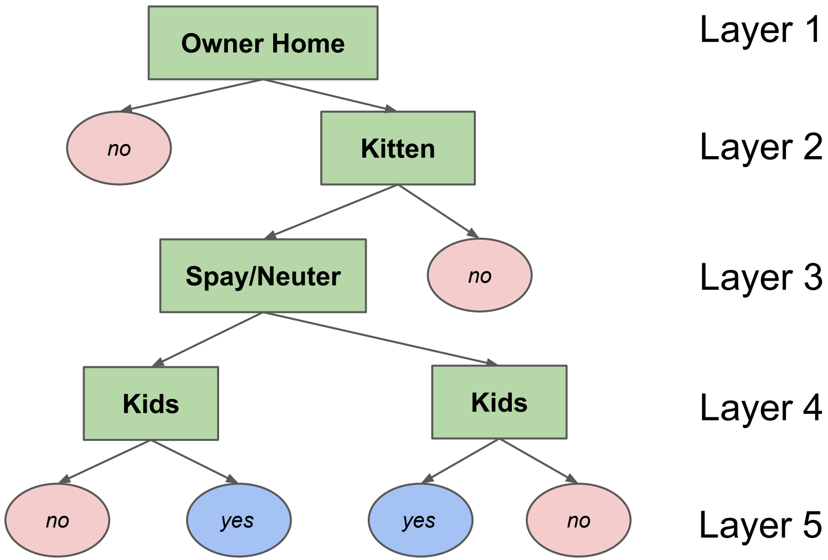
Understanding Machine Learning. Shalev-Shwartz and Ben-David, 2014.

Decision Trees of Depth T

$$\mathcal{X} = \{0,1\}^d$$

$$\mathcal{Y} = \{0,1\}$$

$$\mathcal{H} = \{h : h \text{ is a decision tree with layers } \leq T\}$$



Implications

VC dim = 2^T

The Most Important Things

- Some infinite hypothesis classes are learnable!

- A necessary and sufficient condition for PAC learnability is finite ***VC dimension***, which is the maximal size of a set shattered by $\mathcal{H}$

- We can prove that a hypothesis class has a VC dimension of *d* by showing:
  - There exists a set C of size *d* that is shattered by $\mathcal{H}$
  - Every set C of size $d+1$ is not shattered by $\mathcal{H}$

- Textbook: chapters 6.0, 6.1, 6.2, 6.3, 6.4, 9.1.3

Next Time

- What if we add assumptions about the distribution $\mathcal{D}$?
  - A.K.A. What does all this have to do with statistical inference?

- Textbook: chapter 24.0, 24.1