Learning Theory

Aarti Singh

Machine Learning 10-315
Dec 2, 2020

Slides courtesy: Carlos Guestrin

Summary of PAC bounds for finite model class

With probability $\geq 1-\delta$,

1) For all $h \in H$ s.t. $\text{error}_{train}(h) = 0$,

$$\text{error}_{true}(h) \leq \varepsilon = \frac{\ln|H| + \ln\frac{1}{\delta}}{m}$$

Haussler's bound

2) For all $h \in H$

$$|\text{error}_{true}(h) - \text{error}_{train}(h)| \leq \varepsilon = \sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}$$

Hoeffding's bound

What about continuous hypothesis spaces?

With probability ≥ 1-δ,

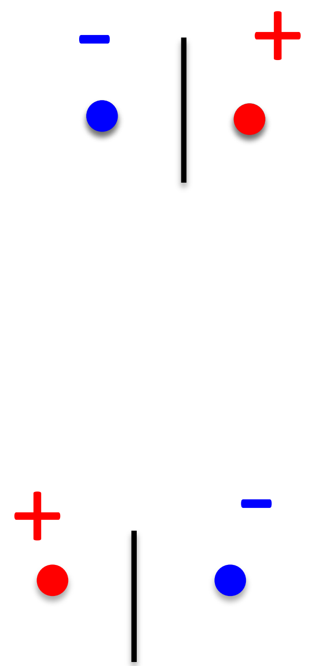
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}$$

- Continuous model class (e.g. linear classifiers):
  - $|H| = \infty$
  - Infinite gap???

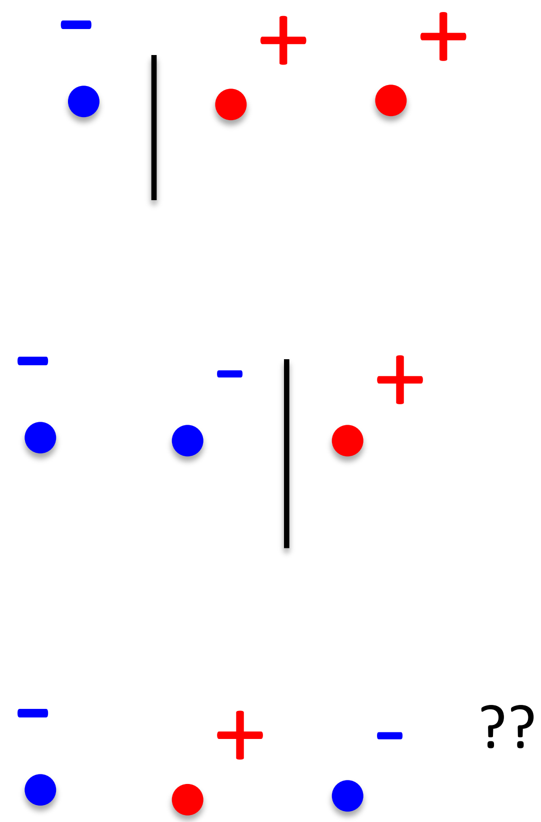
- **As with decision trees, complexity of model class only depends on maximum number of points that can be classified exactly (and not necessarily its size)!**

How many points can a linear boundary classify exactly? (1-D)

2 pts



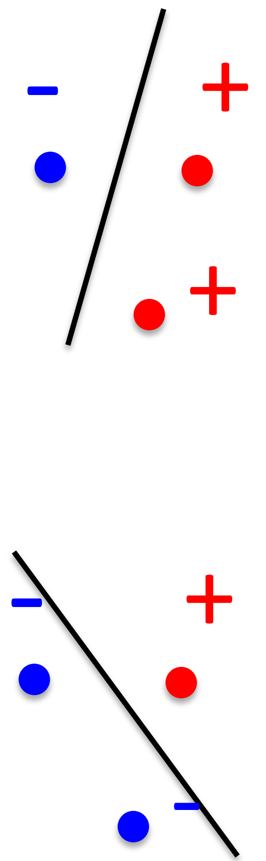
3 pts



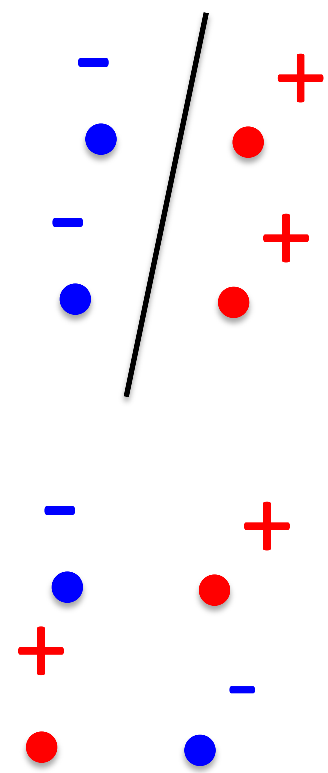
There exists placement s.t. all labelings can be classified

How many points can a linear boundary classify exactly? (2-D)

3 pts



4 pts

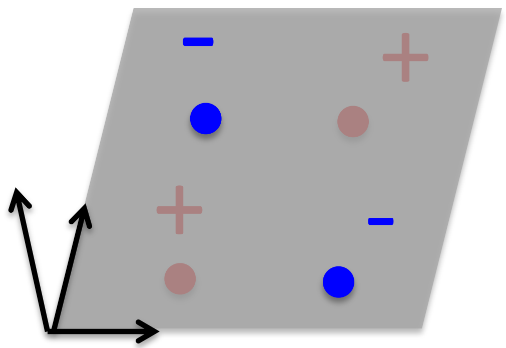


??

There exists placement s.t. all labelings can be classified

How many points can a linear boundary classify exactly? (d-D)

d+1 pts



How many parameters in linear Classifier in d-Dimensions?

$$w_0 + \sum_{i=1}^{d} w_i x_i$$

d+1

There exists placement s.t. all labelings can be classified

PAC bound using VC dimension

- Number of training points that can be classified exactly is VC dimension!!!
  - Measures relevant size of hypothesis space, as with decision trees with k leaves

With probability ≥ 1-δ,

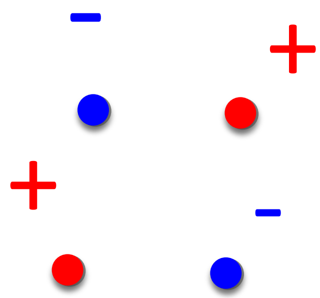
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + 8\sqrt{\frac{VC(H)\left(\ln \frac{m}{VC(H)} + 1\right) + \ln \frac{8}{\delta}}{2m}}$$

Instead of ln|H|

VC dimension

<u>Definition</u>: VC dimension of a hypothesis space H is the maximum number of points such that there exists a hypothesis in H that is consistent with (can correctly classify) any labeling of the points.

- You pick set of points
- Adversary assigns labels
- You find a hypothesis in H consistent with the labels



If VC(H) = k, then for all k+1 points, there exists a labeling that cannot be shattered (can’t find a hypothesis in H consistent with it)

PAC bound using VC dimension

- Number of training points that can be classified exactly is VC dimension!!!
  - Measures relevant size of hypothesis space, as with decision trees with k leaves
  - Bound for infinite dimension hypothesis spaces:

w.p. ≥ 1-δ

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + 8\sqrt{\frac{VC(H)\left(\ln \frac{m}{VC(H)} + 1\right) + \ln \frac{8}{\delta}}{2m}}$$

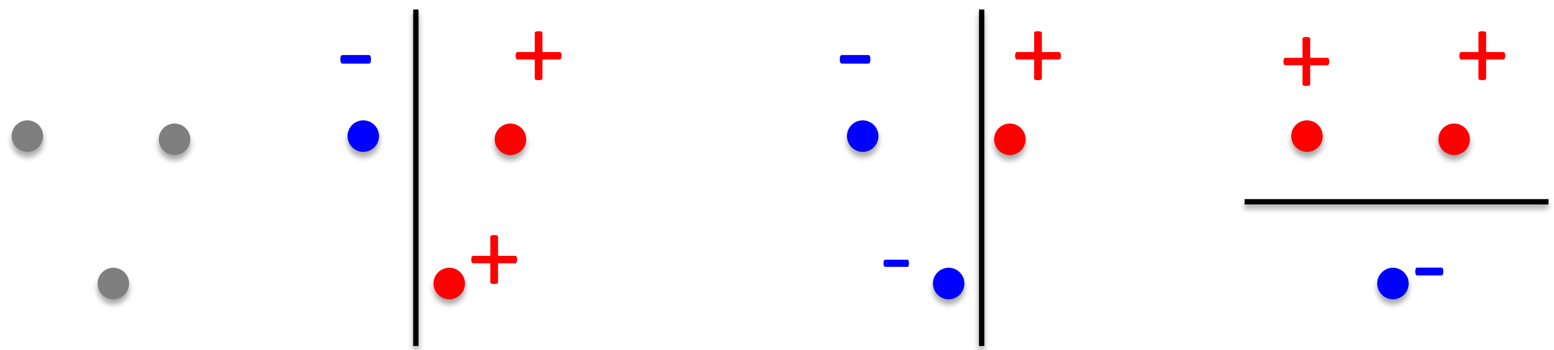
linear classifiers		
2D	large	small
10,000 D	small	large

Examples of VC dimension

- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term

Another VC dim. example - What can we shatter?

- What's the VC dim. of decision stumps in 2D?



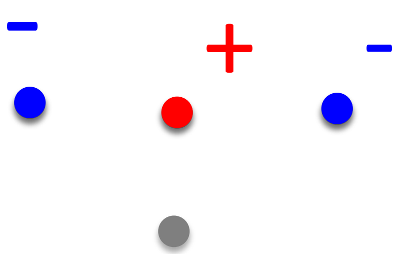
$VC(H) \geq 3$

Another VC dim. example - What can't we shatter?

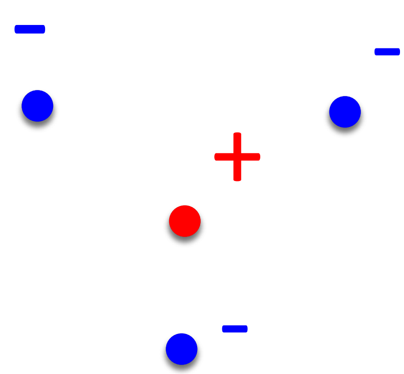
- What's the VC dim. of decision stumps in 2D?

If VC(H) = 3, then for all placements of 4 pts, there exists a labeling that can't be shattered

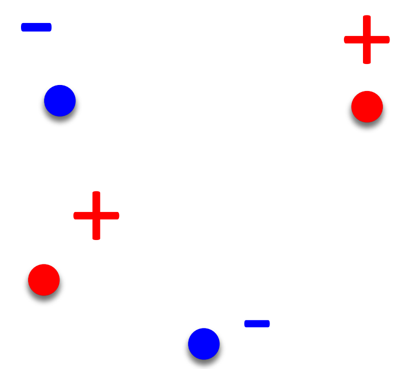
3 collinear



1 in convex hull of other 3



quadrilateral



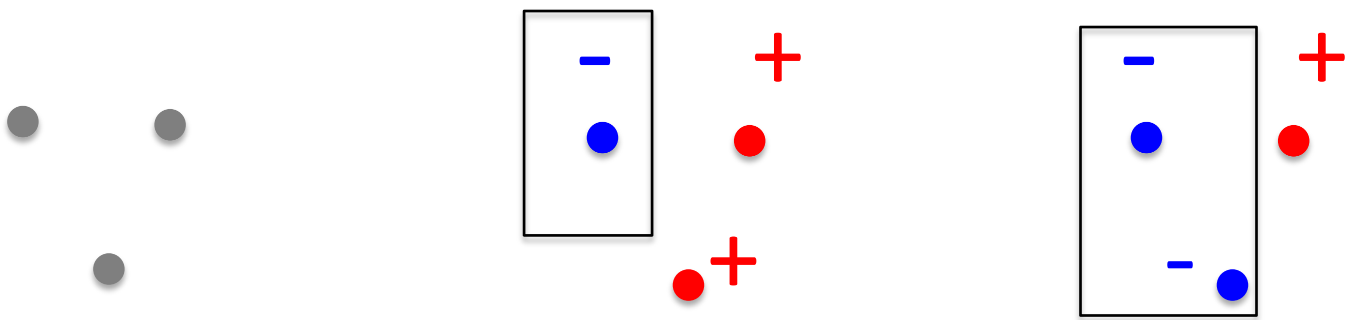
Examples of VC dimension

- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term

- Decision stumps: VC(H) = d+1 (3 if d=2)

Another VC dim. example - What can we shatter?

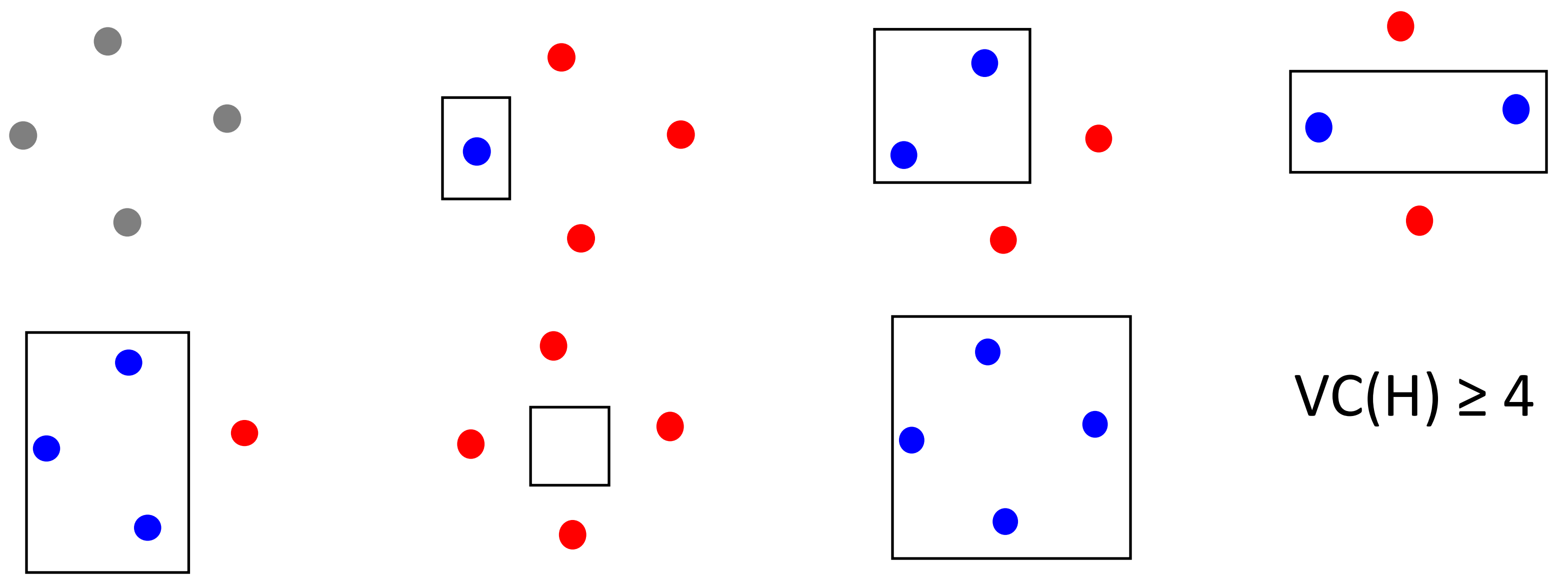
- What's the VC dim. of axis parallel rectangles in 2D? $\mathrm{sign}(1 - 2 * \mathbf{1}_{x \in \text{rectangle}})$



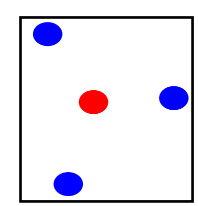
VC(H) ≥ 3

Another VC dim. example - What can't we shatter?

- What's the VC dim. of axis parallel rectangles in 2D? sign(1- 2*$\mathbf{1}_{x \in \text{rectangle}}$)



VC(H) ≥ 4

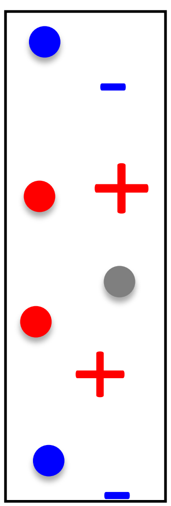
- Some placement of 4 pts can't be shattered 

Another VC dim. example - What can't we shatter?

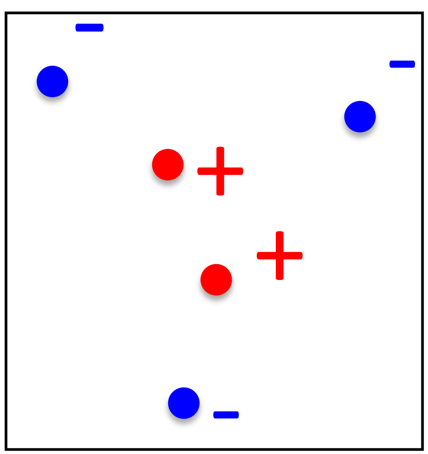
- What's the VC dim. of axis parallel rectangles in 2D? $\quad \text{sign}(1- 2*\mathbf{1}_{x \in \text{rectangle}})$

If VC(H) = 4, then for all placements of 5 pts, there exists a labeling that can't be shattered

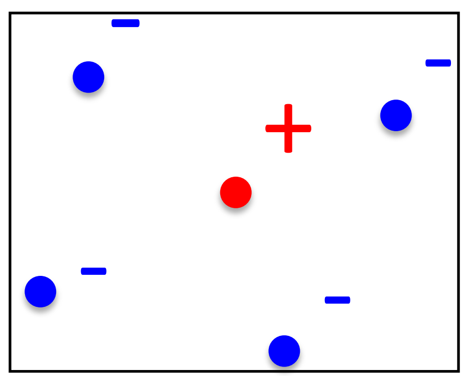
4 collinear



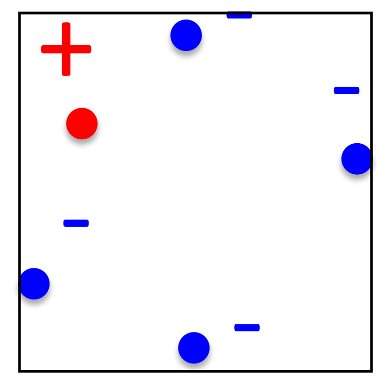
2 in convex hull of other 3



1 in convex hull of other 4



pentagon



Examples of VC dimension

- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term

- Decision stumps: VC(H) = d+1

- Axis parallel rectangles: VC(H) = 2d (4 if d=2)

- 1 Nearest Neighbor: VC(H) = ∞

VC dimension and size of hypothesis space

- To be able to shatter m points, how many hypothesis do we need?

2^m labelings ⇨ $|H| \geq 2^m$

Given $|H|$ hypothesis can hope to shatter max $m = \log_2|H|$ points

$$VC(H) \leq \log_2|H|$$

So VC bound is tighter.

Summary of PAC bounds

With probability ≥ 1-δ,

1) for all h $\in$ H s.t. $\text{error}_{\text{train}}(h) = 0$,

$$\text{error}_{\text{true}}(h) \leq \varepsilon = \frac{\ln|H| + \ln\frac{1}{\delta}}{m}$$

2) for all h $\in$ H,

$$|\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h)| \leq \varepsilon = \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}$$

Finite hypothesis space

3) for all h $\in$ H,

$$|\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h)| \leq \varepsilon = 8\sqrt{\frac{VC(H)\left(\ln\frac{m}{VC(H)} + 1\right) + \ln\frac{8}{\delta}}{2m}}$$

Infinite hypothesis space

Limitation of VC dimension

- Hard to compute for many hypothesis spaces

  VC(H) ≥ lower bound (easy)

  VC(H) = …   (HARD!)

  For all placements of VC(H)+1 points, there exists a labeling that can't be shattered

- Too loose for many hypothesis spaces

  linear SVMs, VC dim = d+1 (d features)

  kernel SVMs, VC dim = ??

  = ∞ (Gaussian kernels)

  Deep Neural nets, VC dim = very large

Suggests Gaussian kernels and deep nets are really BAD!! But contradicts practice!

What you need to know

- PAC bounds on true error in terms of empirical/training error and complexity of hypothesis space
- Complexity of the classifier depends on number of points that can be classified exactly
  - Finite case – Number of hypothesis
  - Infinite case – VC dimension

Other bounds – Rademacher complexity (data dependent), Margin based (complexity low if margin achieved high), Mistake bounds, …