Learning Theory

Pradeep Ravikumar
Co-instructor: Manuela Veloso

Machine Learning 10-701

Learning Theory

- We have explored **many** ways of learning from data
- But…
  - How good is our classifier, really?
  - How much data do I need to make it “good enough”?

A simple setting

- Classification
  - n i.i.d. data points (X_i,Y_i), i = 1,…,n
  - **finite** number of possible hypotheses (e.g., decision trees of depth d)
- A learner finds a hypothesis *h*
- We are interested in:

$$\text{error}_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(h(X_i) \neq Y_i)$$

$$\text{error}_{\text{true}} = \mathbb{P}(h(X) \neq Y)$$

A simple setting

- Classification
  - n i.i.d. data points (X_i,Y_i), $i = 1,\ldots,n$
  - **finite** number of possible hypotheses (e.g., decision trees of depth d)
- A learner finds a hypothesis *h* that is **consistent** with training data
  - Gets zero error in training, $\text{error}_{\text{train}}(h) = 0$
- What is the probability that *h* has more than ε true error?
  - $\text{error}_{\text{true}}(h) \geq \varepsilon$

Even if *h* makes zero errors in training data, may make errors in test

How likely is a bad hypothesis to get m data points right?

- Consider a bad hypothesis *h* i.e. $error_{true}(h) \geq \varepsilon$

- Probability that *h* gets one data point right (i.e. does not make an error)

  $\leq 1 - \varepsilon$

- Probability that *h* gets *m* data points right

  $\leq (1 - \varepsilon)^m$

How likely is a learner to pick a bad hypothesis?

- Usually there are many (say k) bad hypotheses in the class

$$h_1, h_2, \ldots, h_k \quad \text{s.t. error}(h_i) \geq \varepsilon \qquad i = 1, \ldots, k$$

- Probability that learner picks a bad hypothesis = Probability that some bad hypothesis is consistent with m data points

Prob(h_1 consistent with m data points OR
h_2 consistent with m data points OR … OR
h_k consistent with m data points)

$\leq$ Prob(h_1 consistent with m data points) +
Prob(h_2 consistent with m data points) + … +
Prob(h_k consistent with m data points)

Union bound

Loose but works

$\leq \; k\,(1-\varepsilon)^m$

How likely is a learner to pick a bad hypothesis?

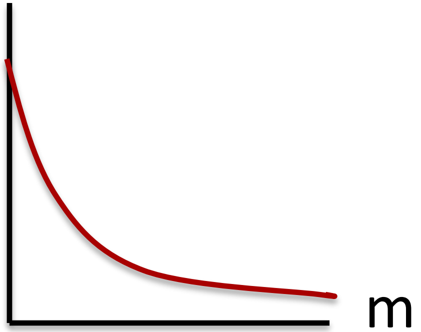
- Usually there are many many (say k) bad hypotheses in the class

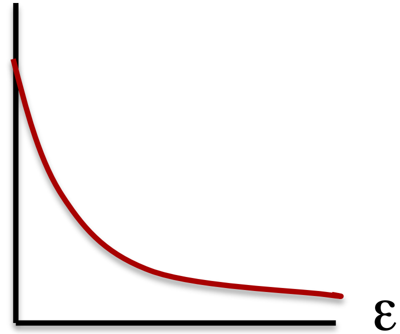
$$h_1, h_2, \ldots, h_k \quad \text{s.t. } \mathrm{error}(h_i) \geq \varepsilon \qquad i = 1, \ldots, k$$

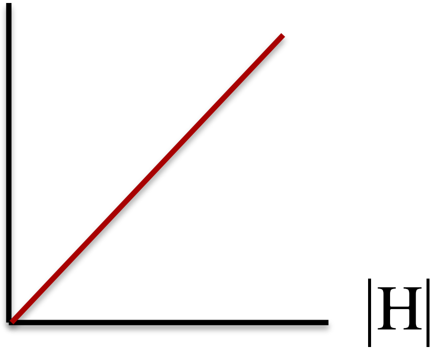
- Probability that learner picks a bad hypothesis

$$\leq \; k\,(1-\varepsilon)^m \quad \leq \; |H|\,(1-\varepsilon)^m \leq \; |H|\, e^{-\varepsilon m}$$

$\hookrightarrow$ Size of hypothesis class







Probability of Error

$|H|e^{-m\epsilon} \leq \delta$ .... Probability of error

- Given ε and δ, yields sample complexity

#training data, $m \geq \dfrac{\ln |H| + \ln \frac{1}{\delta}}{\epsilon}$

PAC (Probably Approximately Correct) bound

- ***Theorem [Haussler'88]***: Hypothesis space *H* finite, dataset *D* with *m* i.i.d. samples, 0 < ε < 1 : for any learned hypothesis *h* that is consistent on the training data, for sufficiently large m:

$$P(\text{error}_{true}(h) \geq \epsilon) \leq |H|e^{-m\epsilon} \leq \delta$$

- Equivalently, with probability $\geq 1 - \delta$

$$\text{error}_{true}(h) \leq \epsilon$$

What if our classifier does not have zero error on the training data?

- Question: What about a learner with $error_{train}(h) \neq 0$ in training set?
- The error of a hypothesis is like estimating the parameter of a coin!

$$error_{true}(h) := P(h(X) \neq Y) \quad \equiv \quad P(H=1) =: \theta$$

$$error_{train}(h) := \frac{1}{m} \sum_i \mathbf{1}_{h(X_i) \neq Y_i} \quad \equiv \frac{1}{m} \sum_i Z_i =: \widehat{\theta}$$

Hoeffding's bound for a single hypothesis

- Consider *m* i.i.d. flips $x_1,\dots,x_m$, where $x_i \in \{0,1\}$ of a coin with parameter θ. For $0<\varepsilon<1$:

$$P\left(\left|\theta - \frac{1}{m}\sum_i x_i\right| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

- For a single hypothesis h

$$P\left(\left|\text{error}_{true}(h) - \text{error}_{train}(h)\right| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

Hoeffding's bound for |H| hypotheses

- For each hypothesis h_i:

$$P\left(|\text{error}_{true}(h_i) - \text{error}_{train}(h_i)| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

- What if we are comparing |H| hypotheses?

  Union bound

- ***Theorem***: Hypothesis space *H* finite, dataset *D* with *m* i.i.d. samples, $0 < \varepsilon < 1$: for any learned hypothesis $h \in H$, with sufficiently large number of samples m:

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

Summary of PAC bounds for finite hypothesis spaces

With probability ≥ 1-δ,

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

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

Haussler's bound

2) For all $h \in H$

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

Hoeffding's bound

PAC bound and Bias-Variance tradeoff

- with probability $\geq 1 - \delta$

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

- Fixed m

hypothesis space	$\downarrow$	$\downarrow$
complex	small	large
simple	large	small

What about continuous hypothesis spaces?

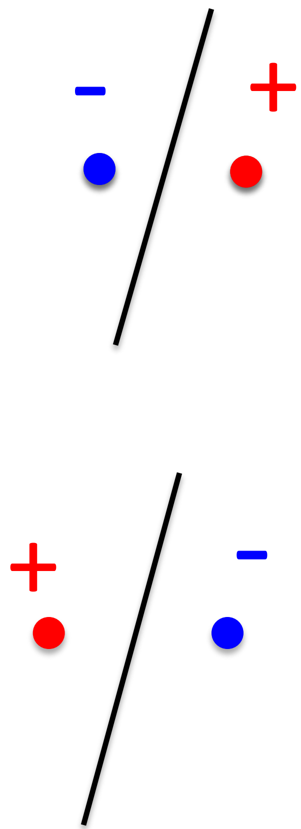
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}$$

- Continuous hypothesis space:
  - |H| = ∞
  - Infinite error ???

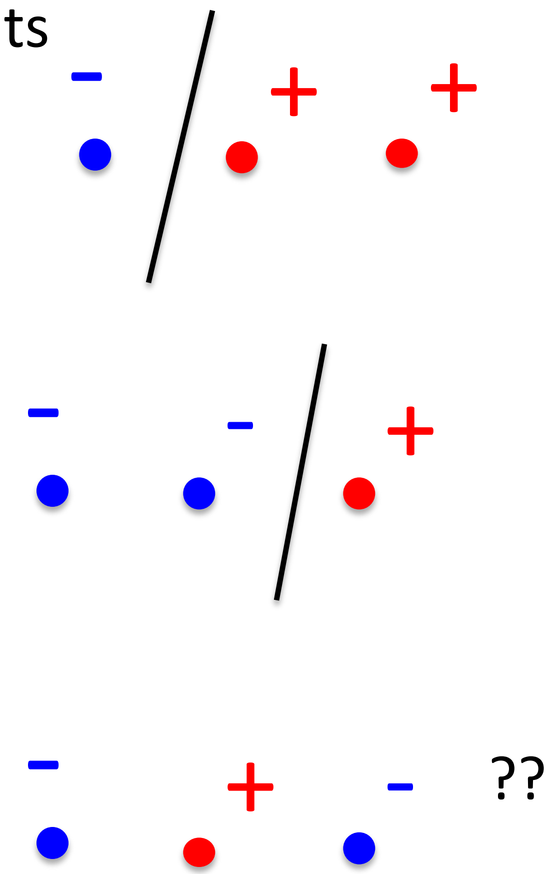
- **As with decision trees, complexity of hypothesis space only depends on maximum number of points that can be classified exactly (and not necessarily its size)!**

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

2 pts

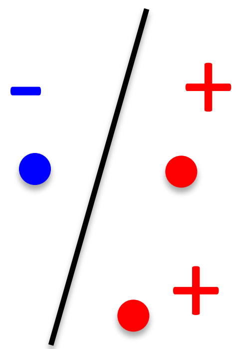


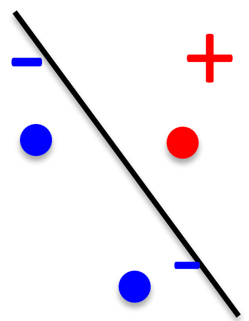
3 pts



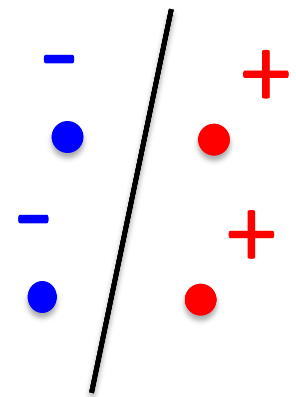
How many points can a linear boundary classify exactly? (2-D)

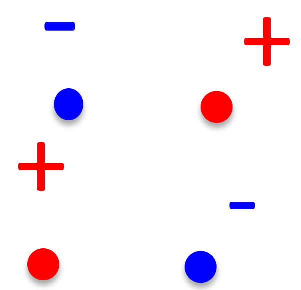
3 pts





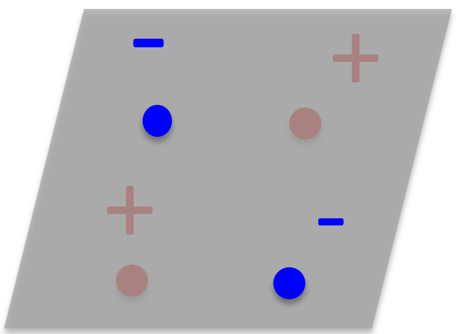
4 pts





??

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

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

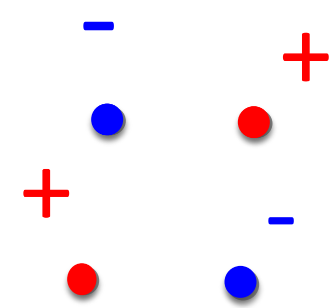
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + 8\sqrt{\frac{VC(H)\left(\ln\frac{m}{VC(H)}+1\right)+\ln\frac{8}{\delta}}{2m}}$$

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.

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



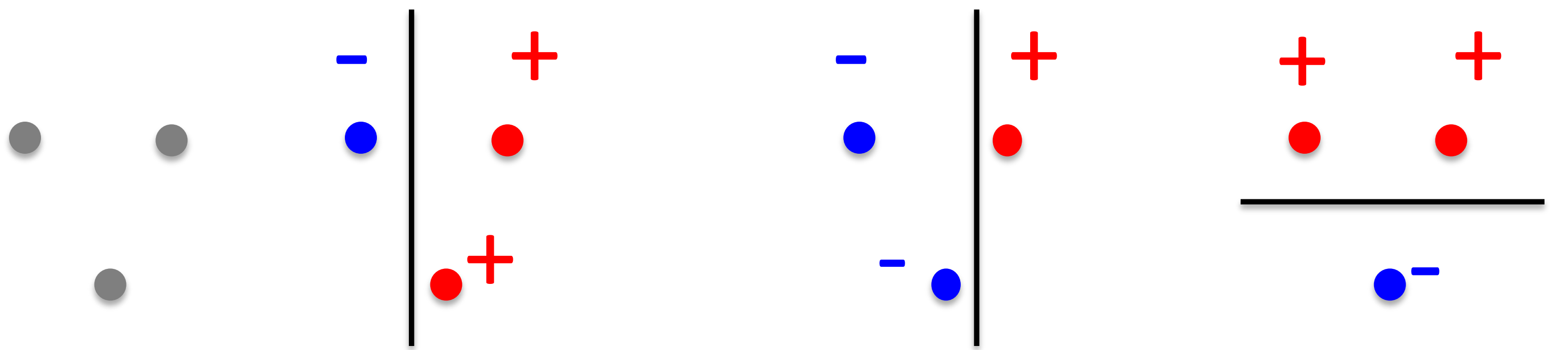
If VC(H) = k, then **for all** k+1 points, there exists a labeling such that can't find a hypothesis in H consistent with it

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 (axis parallel lines) in 2d?



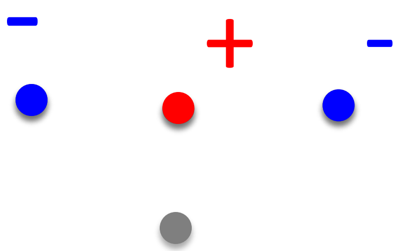
VC(H) ≥ 3

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

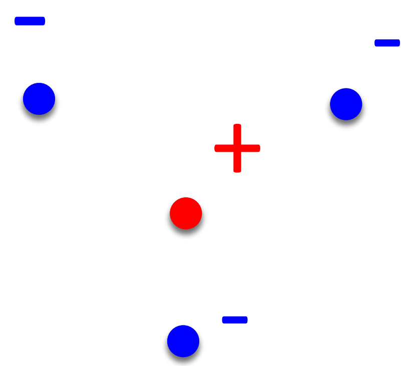
- What's the VC dim. of decision stumps in 2d?

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

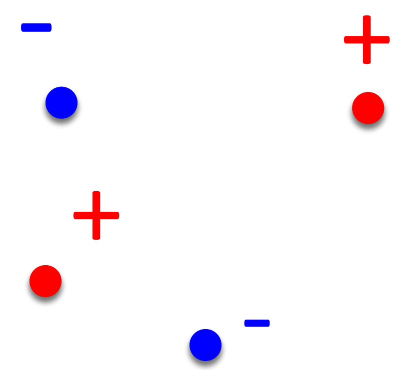
3 collinear



1 in convex hull of other 3



quadrilateral



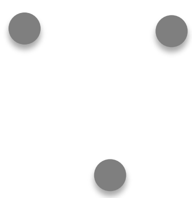
Examples of VC dimension

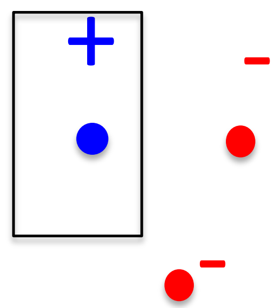
- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term

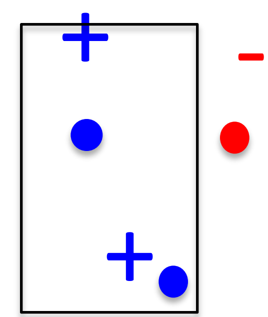
- Decision stumps: VC(H) = d+1 (3 if d=2)

Another VC dim. example - What can we shatter?

- What's the VC dim. of axis parallel rectangles in 2d?



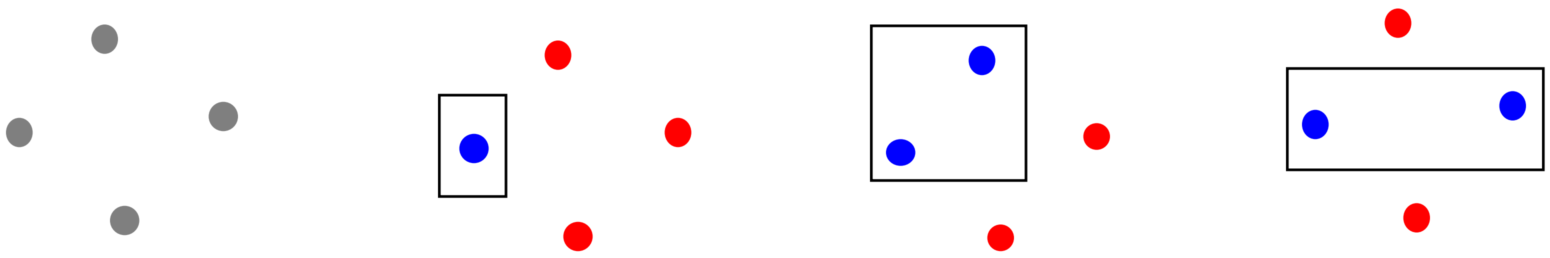


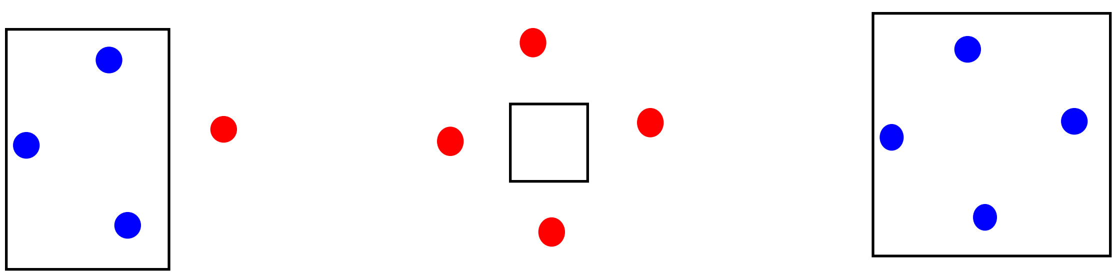


VC(H) $\geq$ 3

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

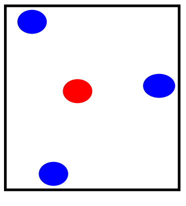
- What’s the VC dim. of axis parallel rectangles in 2d?





$VC(H) \geq 4$

- Some placement of 4 pts can’t be shattered

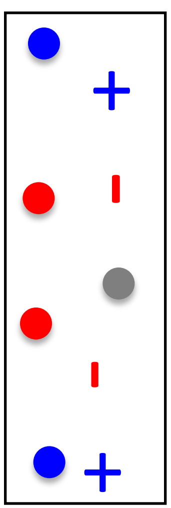


Another VC dim. example - What can't we shatter?

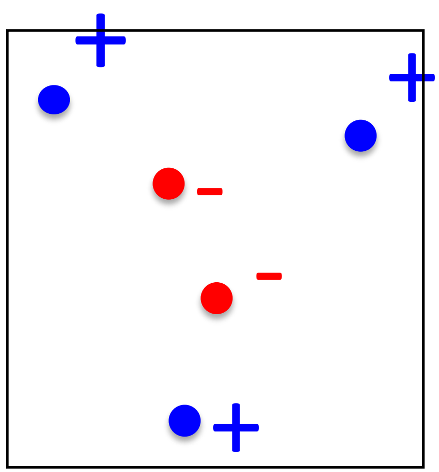
- What's the VC dim. of axis parallel rectangles in 2d?

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

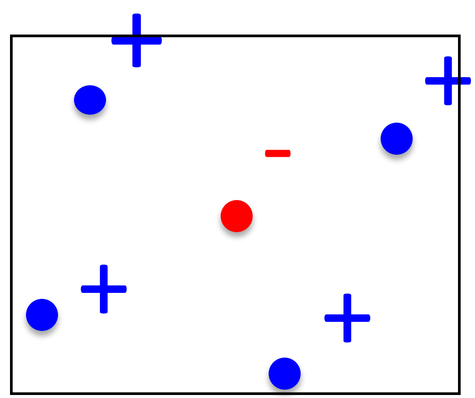
4 collinear



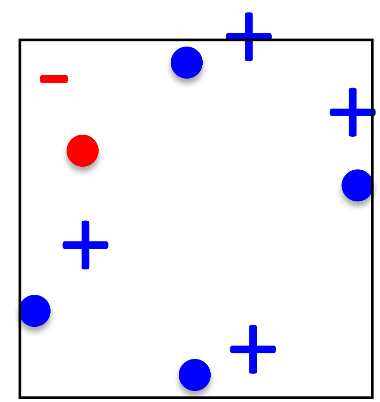
2 in convex hull of other 3



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, number of points we can shatter $m <= \log_2|H|$

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

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)

Suggests Gaussian kernels are really BAD!!

PAC Bounds

With probability $\geq 1-\delta$, for all $h \in H$,

$$|\text{error}_{true}(h) - \text{error}_{train}(h)| \leq \varepsilon(H)$$