University of Colorado **Boulder**

Department of Computer Science

CSCI 5622: Machine Learning

Chris Ketelsen

Lecture 11: Learning Theory Part 2
VC Dimension

Learning Objectives

- [ ] Learn about the VC dimension for infinite dimensional H
- [ ] See some examples of computing the VC dimension
- [ ] See some generalization bounds in terms of VC dimension

Assumptions and Definitions

Assumptions:

- Data comes from distribution $\mathcal{D}$
- Concept c comes from concept class C
- Hypothesis h comes from hypothesis class H

Def: Generalization Error

$$R(h) = Pr_{x \sim D}\left[h(x) \neq c(x)\right]$$

Def: Training Error

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} I[h(x_i) \neq c(x_i)]$$

Generalization Bound for Finite H

Theorem: Let H be a finite hypothesis set. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln |H| + \ln(2/\delta)}{2m}}$$

- Larger m is, better training error predicts gen. error

If we make H more complex:

- Training error would go down
- Bound term would go up ...

Generalization Bound for Infinite H

Note that when H is infinite this bound is completely uninformative

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln |H| + \ln(2/\delta)}{2m}}$$

Questions:

- Is efficient learning possible when $|H| = \infty$?
- Are there useful measures of complexity when $|H| = \infty$?

Answers:

- Must be b/c we showed it for intervals (Frank the Alien)
- Yes! We'll explore some today

Complexity Measures - Counting Bits

First Pass Attempt: Consider linear classifiers in 2D

$$h_{\mathbf{w}}(\mathbf{x}) = I\left(w_0 + w_1 x_1 + w_2 x_2 \geq 0\right)$$

Hypotheses are defined by 3 real parameters w_0, w_1, and w_2

Usually w's represented on the computer by double precision variables which are defined by 64 bits.

Thus we have $3 \cdot 64 = 192$ degrees of freedom in the hypothesis

For binary classification H then consists of at most $|H| = 2^{3 \cdot 64} = 2^{192}$ different hypothesis

Complexity Measures - Counting Bits

First Pass Attempt: Consider linear classifiers in 2D

$$h_{\mathbf{w}}(\mathbf{x}) = I\left(w_0 + w_1 x_1 + w_2 x_2 \geq 0\right)$$

This doesn't seem particularly helpful

Especially since depends on the number of parameters of the model

But could parameterize same H as

$$h_{\mathbf{u},\mathbf{v}}(\mathbf{x}) = I\left((u_0^2 - v_0^2) + (u_1^2 - v_1^2)x_1 + (u_2^2 - v_2^2)x_2 \geq 0\right)$$

Same H but analysis suggests this is worse somehow

Complexity Measures - Growth Function

Second Pass Attempt:

What is the maximum number of ways that m point can be classified by H?

Definition: The **Growth Function** of H is the maximum number of ways that m points can be classified by hypotheses in H

$$\Pi_H(m) = \max_{S \subset \mathcal{X}} |\{(h(x_1), \ldots, h(x_m) \ : \ h \in H\}|$$

Question: What is the smallest possible growth function?

Complexity Measures - Growth Function

Second Pass Attempt:

What is the maximum number of ways that m point can be classified by H?

Definition: The **Growth Function** of H is the maximum number of ways that m points can be classified by hypotheses in H

$$\Pi_H(m) = \max_{S \subset \mathcal{X}} |\{(h(x_1), \ldots, h(x_m) \ : \ h \in H\}|$$

Question: What is the smallest possible growth function?

Answer: $\Pi_H(m) = 1$, Hypotheses predict 1 no matter what

Complexity Measures - Growth Function

Second Pass Attempt:

What is the maximum number of ways that m point can be classified by H?

Definition: The **Growth Function** of H is the maximum number of ways that m points can be classified by hypotheses in H

$$\Pi_H(m) = \max_{S \subset \mathcal{X}} |\{(h(x_1), \ldots, h(x_m) \ : \ h \in H\}|$$

Question: What is the largest possible growth function?

Complexity Measures - Growth Function

Second Pass Attempt:

What is the maximum number of ways that m point can be classified by H?

Definition: The **Growth Function** of H is the maximum number of ways that m points can be classified by hypotheses in H

$$\Pi_H(m) = \max_{S \subset \mathcal{X}} |\{(h(x_1), \ldots, h(x_m) \ : \ h \in H\}|$$

Question: What is the largest possible growth function?

Answer: $\Pi_H(m) = 2^m$, H can classify all m points correctly

Complexity Measures - Growth Function

This seems like a reasonable measure of complexity

Can update our generalization error bound in terms of $\Pi_H(m)$

Theorem: Let H be a hypothesis set. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{2 \ln \Pi_H(m)}{m}} + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

Note: In worst-case scenario first complexity term is constant and can't be reduced by increasing m

Complexity Measures - Growth Function

Example: Find $\Pi_H(m)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

We need a function of m that counts the number of possible hypotheses that can classify m points perfectly

What do **you** think?

Complexity Measures - Growth Function

Example: Find $\Pi_H(m)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

We need a function of m that counts the number of possible hypotheses that can classify m points perfectly

Suppose that there is at least one positive point followed by at least one negative point

Complexity Measures - Growth Function

Example: Find $\Pi_H(m)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

We need a function of m that counts the number of possible hypotheses that can classify m points perfectly

Suppose that there is at least one positive point followed by at least one negative point

Since we need to pick the first positive point and the first negative point, this gives $\binom{m}{2}$ possible hypotheses

Complexity Measures - Growth Function

Example: Find $\Pi_H(m)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

We need a function of m that counts the number of possible hypotheses that can classify m points perfectly

Suppose that there is at least one positive point with no trailing negative points

Complexity Measures - Growth Function

Example: Find $\Pi_H(m)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

We need a function of m that counts the number of possible hypotheses that can classify m points perfectly

Suppose that there is at least one positive point with no trailing negative points

This time we only have to count the first positive point, which gives us m possibly hypotheses

Complexity Measures - Growth Function

Example: Find $\Pi_H(m)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

We need a function of m that counts the number of possible hypotheses that can classify m points perfectly

Finally, consider the case where there are no positive points

Clearly there is only 1 hypothesis associated with this. Thus,

$$\Pi_H(m) = 1 + m + \binom{m}{2} = \sum_{k=0}^{2} \binom{m}{k}$$

Complexity Measures - VC Dimension

OK, so we can compute the growth function

But it is unsatisfactory for a couple of reasons

- It is different for different values of m
- It's kind of a pain to compute

Instead we introduce the VC Dimension of a hypothesis class

$$\mathrm{VC}dim(H) = \max\{m \ : \ \Pi_H(m) = 2^m\}$$

In other words, $\mathrm{VC}dim(H)$ is the largest possible m such that H can accurately classify every possible combinations of labels

Complexity Measures - VC Dimension

OK, so we can compute the growth function

But it is unsatisfactory for a couple of reasons

- It is different for different values of m
- It's kind of a pain to compute

Instead we introduce the VC Dimension of a hypothesis class

$$\mathrm{VC}dim(H) = \max\{m \ : \ \Pi_H(m) = 2^m\}$$

Invented by Vapnik and Chervonenkis in early 60's. Formalized for learning theory in late 90's.

Complexity Measures - VC Dimension

Alternate formulation

Definition: A set of points S is **shattered** by hypothesis class H if H can correctly classify all possible labellings of the points in S

Definition: The VC dimension can then be defined as

$$\text{VC}dim(H) = \max \{|S| \ : \ H \text{ shatters } S\} \quad \text{for some } S$$

Notes:

- Just need to find one S that works
- Can be pathological choices of S that don't work

Complexity Measures - VC Dimension

Alternate formulation

Definition: A set of points S is **shattered** by hypothesis class H if H can correctly classify all possible labellings of the points in S

Definition: The VC dimension can then be defined as

$$\text{VC}dim(H) = \max \{|S| \ : \ H \text{ shatters } S\}$$

Notes:

- Need lower bound on m points that can be shattered
- Need upper bound on m points that can be shattered

Complexity Measures - VC Dimension

Example: Find $\mathrm{VC}dim(H)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

First see if we can shatter a set of 2 points

Complexity Measures - VC Dimension

Example: Find $\mathrm{VC}dim(H)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

First see if we can shatter a set of 2 points

Complexity Measures - VC Dimension

Example: Find $\mathrm{VC}dim(H)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

First see if we can shatter a set of 2 points

Complexity Measures - VC Dimension

Example: Find $\mathrm{VC}dim(H)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

First see if we can shatter a set of 2 points

So we've found $\mathrm{VC}dim(H) \geq 2$

Complexity Measures - VC Dimension

Example: Find $\mathrm{VC}dim(H)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

Upper bound is usually harder because we need to show that there is **NO** set of three points that can be shattered by H.

Handwavy: Consider the following case

Complexity Measures - VC Dimension

Example: Find $\mathrm{VC}dim(H)$ for intervals on the real line:

$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

Upper bound is usually harder because we need to show that there is **NO** set of three points that can be shattered by H.

So we must have $\mathrm{VC}dim(H) < 3$

Together with the lower bound this gives us

$$\mathrm{VC}dim(H) = 2$$

Complexity Measures - VC Dimension

Technically we need to prove lower bound for **all** possible S

Proof: Want to prove that for all S s.t. $|S| = 3$ there exists some labeling such that S cannot be captured by H

Let x_1, x_2, x_3 be arbitrary points s.t. $x_1 \leq x_2 \leq x_3$ (WLOG)

Let $y = (+1, -1, +1)$ and assume some $h = [a, b]$ works

Since $y_1 = +1$ must have $a \leq x_1$

Since $y_3 = +1$ must have $b \geq x_3$

Then $a \leq x_1 \leq x_2 \leq x_3 \leq b$

But $y_2 = -1$ which is a contradiction

Complexity Measures - VC Dimension

Can update our generalization error bound in terms of $\mathrm{VC}dim(H)$

Theorem: Let H be a hypothesis set and let $\mathrm{VC}dim(H) = d$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{2d \ln(em/d)}{m}} + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

For fixed δ and $m > d$ we eventually have

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \mathcal{O}\left(\sqrt{\frac{\ln(m/d)}{m/d}}\right)$$

Training error is good indication of classification error if $m \gg d$

Complexity Measures - VC Dimension

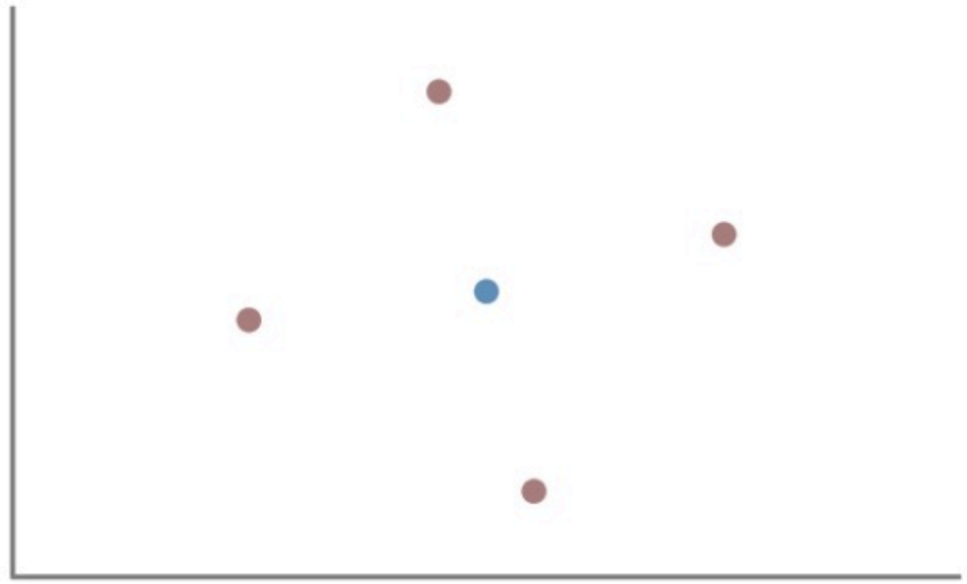
Example: Let $\mathbf{x}$ be 2D points and find $\mathrm{VC}dim(H)$ where H is the class of all minimal axis-aligned rectangles

Can we shatter some set of $m = 4$ points?

Complexity Measures - VC Dimension

Example: Let $\mathbf{x}$ be 2D points and find $\mathrm{VC}dim(H)$ where H is the class of all minimal axis-aligned rectangles

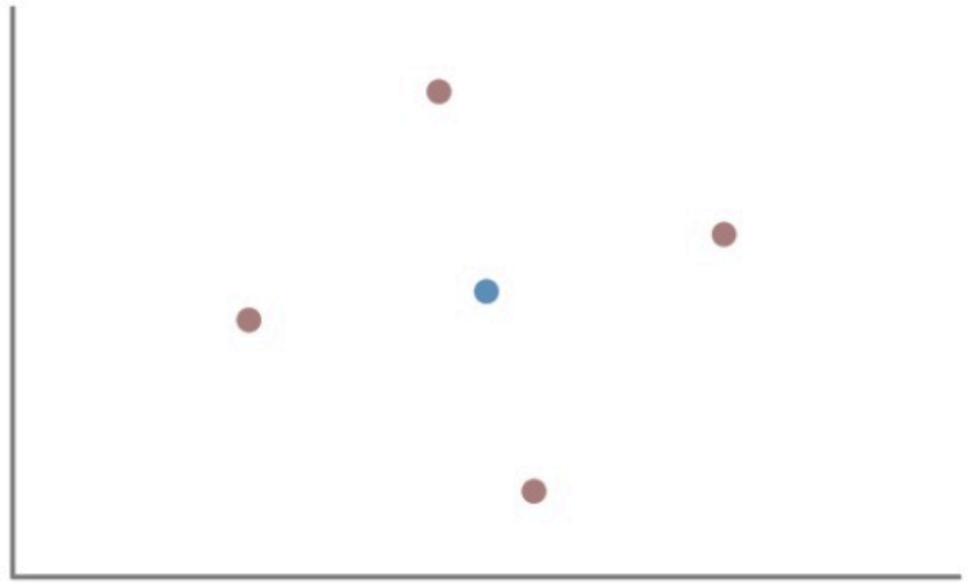
Yes we shatter some set of $m = 4$ points? What about 5 points?



Complexity Measures - VC Dimension

Example: Let $\mathbf{x}$ be 2D points and find $\mathrm{VC}dim(H)$ where H is the class of all minimal axis-aligned rectangles

Yes we shatter some set of $m = 4$ points? What about 5 points?

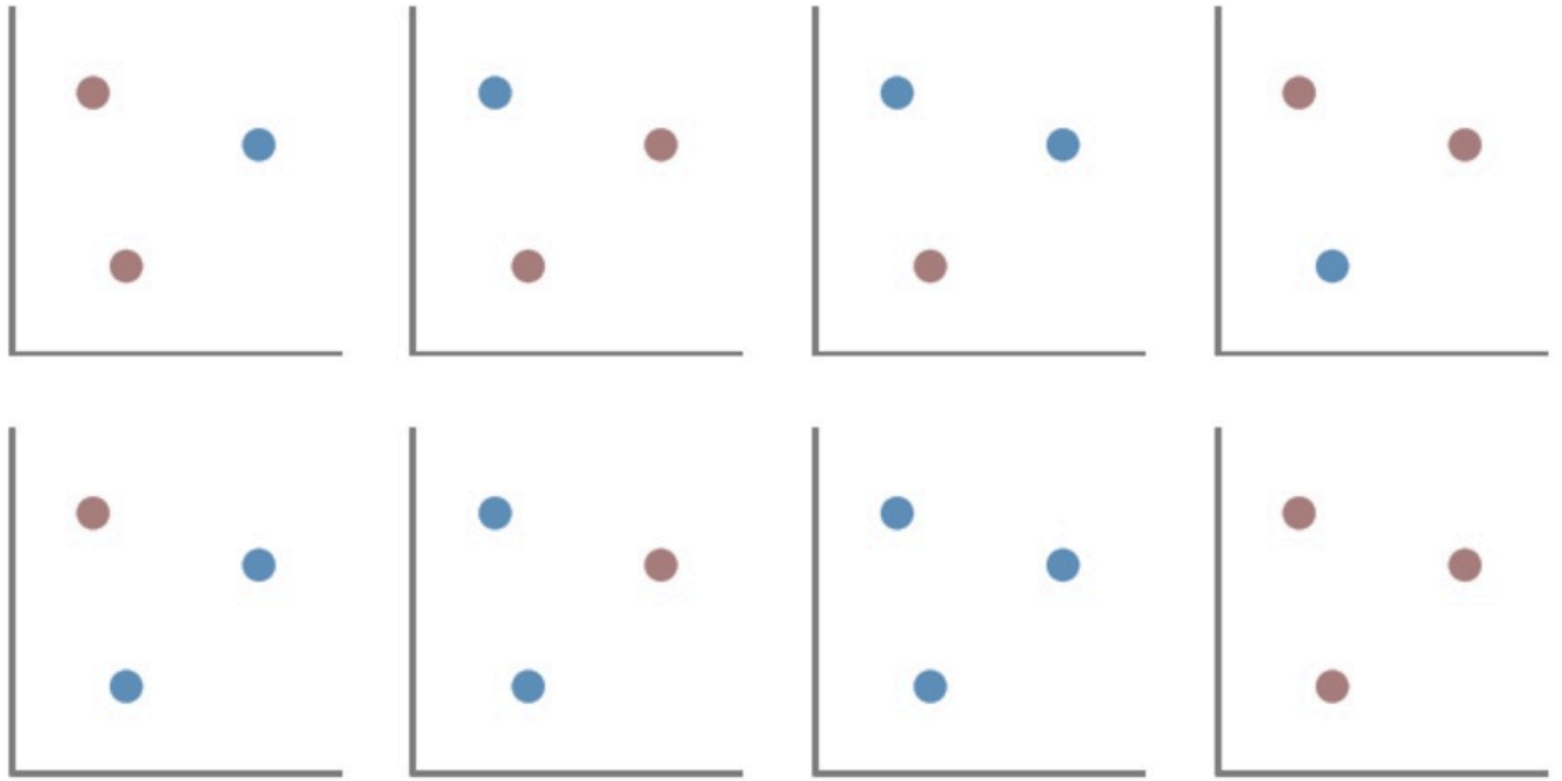


Can't do it, so $\mathrm{VC}dim(H) = 4$

Complexity Measures - VC Dimension

Example: Let $\mathbf{x}$ be 2D points and find $\mathrm{VC}dim(H)$ where H is the class of all linear classifiers

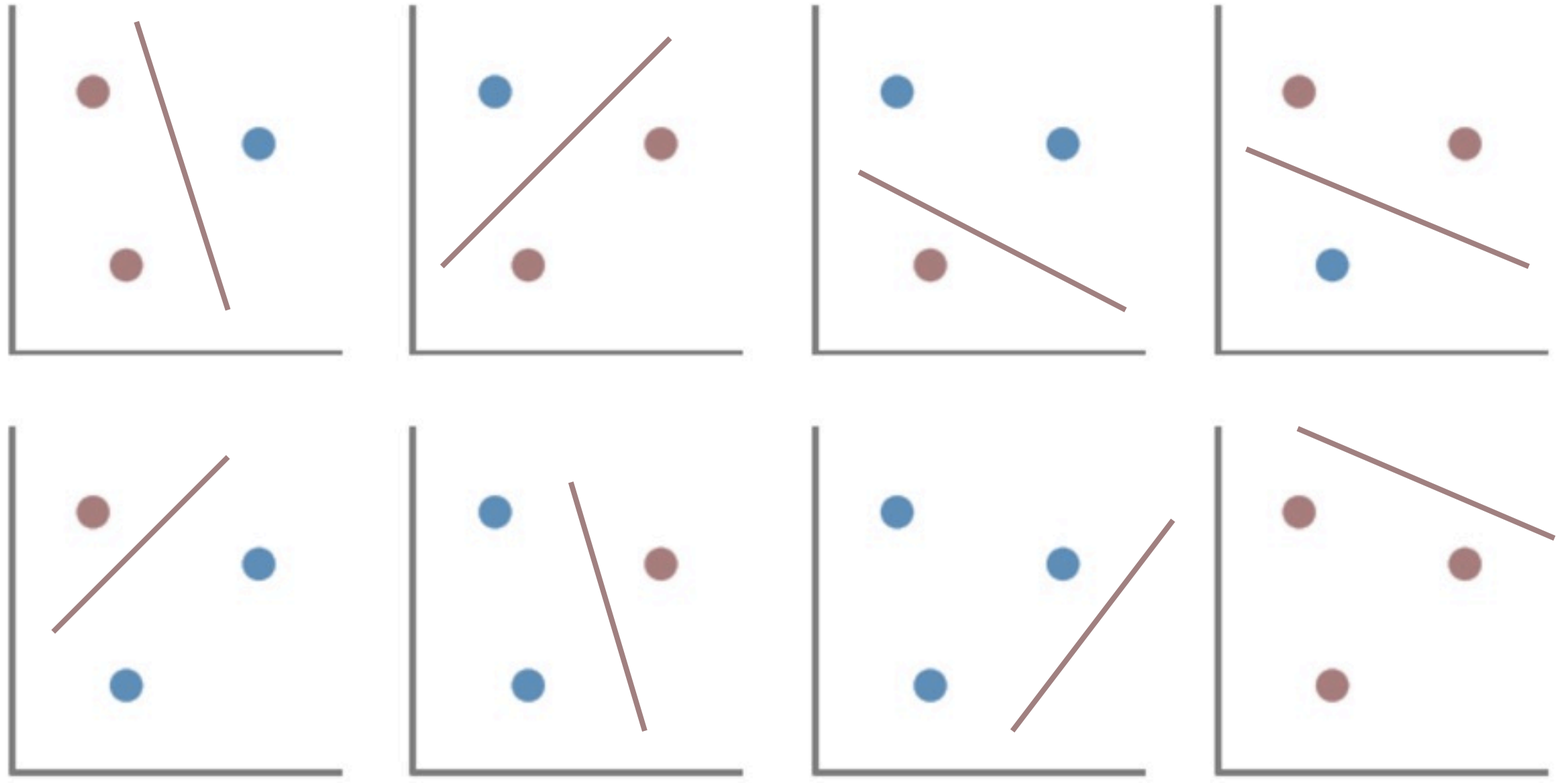
Can we shatter a set of 3 points?



Complexity Measures - VC Dimension

Example: Let $\mathbf{x}$ be 2D points and find $\mathrm{VC}dim(H)$ where H is the class of all linear classifiers

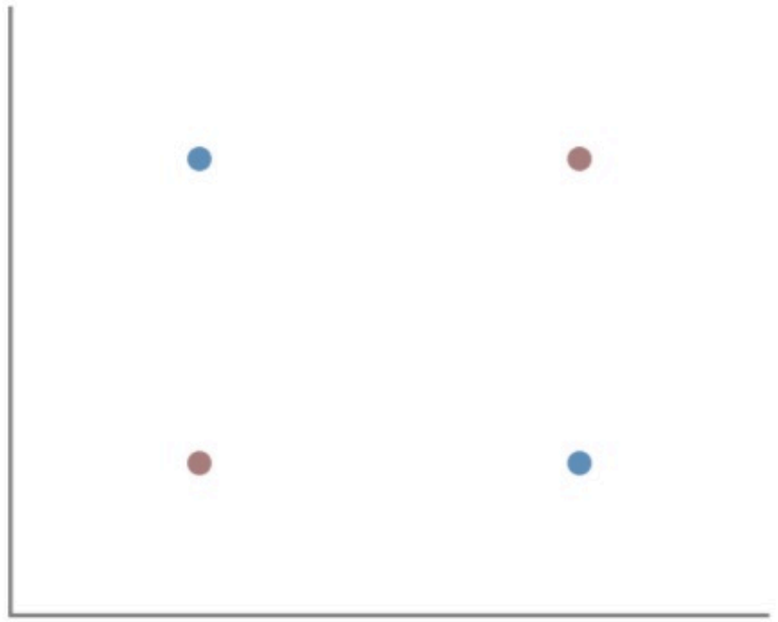
Can we shatter a set of 3 points?



Complexity Measures - VC Dimension

Example: Let $\mathbf{x}$ be 2D points and find $\mathrm{VC}dim(H)$ where H is the class of all linear classifiers

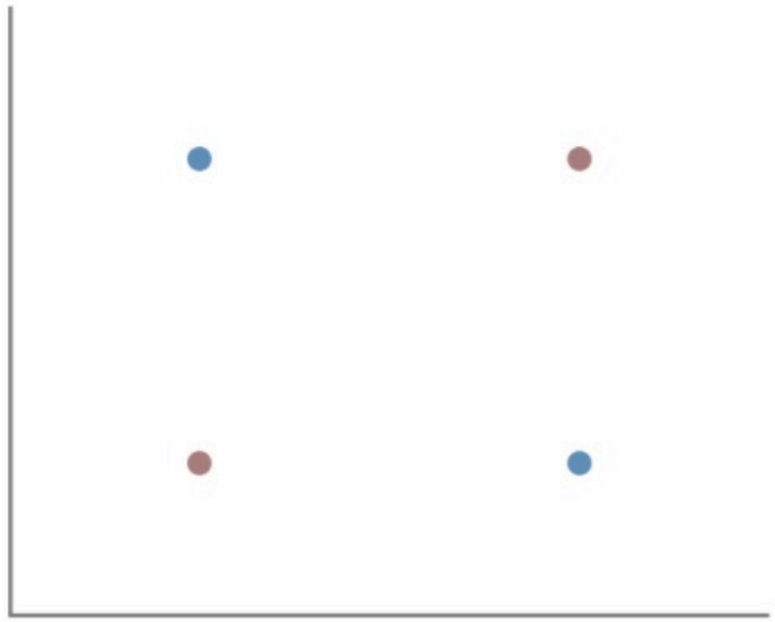
Can we shatter a set of 4 points?



Complexity Measures - VC Dimension

Example: Let $\mathbf{x}$ be 2D points and find $\mathrm{VC}dim(H)$ where H is the class of all linear classifiers

Can we shatter a set of 4 points?

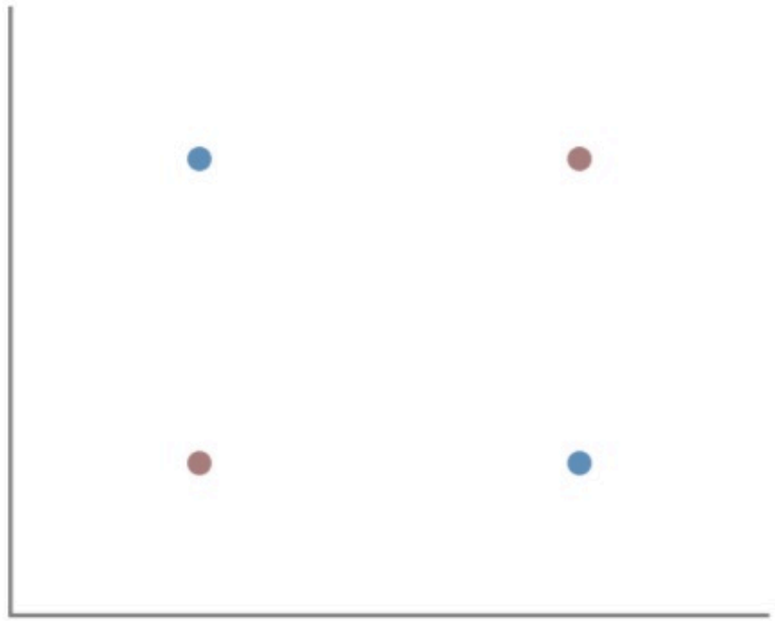


Nope! So $\mathrm{VC}dim(H) = 3$

Complexity Measures - VC Dimension

Example: Let $\mathbf{x}$ be 2D points and find $\mathrm{VC}dim(H)$ where H is the class of all linear classifiers

Can we shatter a set of 4 points?



Fact: If $\mathbf{x} \in \mathbb{R}^n$ then lin. classifiers have $\mathrm{VC}dim(H) = n + 1$

Complexity Measures - VC Dimension

Theorem: Let H be a finite hypothesis set and let $\mathrm{VC}dim(H) = d$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \mathcal{O}\left(\sqrt{\frac{\ln(m/d)}{m/d}}\right)$$

OK, This is great!: All of examples we've looked at have relatively small $\mathrm{VC}dim(H)$

Complexity Measures - VC Dimension

Theorem: Let H be a finite hypothesis set and let $\mathrm{VC}dim(H) = d$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \mathcal{O}\left(\sqrt{\frac{\ln(m/d)}{m/d}}\right)$$

OK, This is great!: All of examples we've looked at have relatively small $\mathrm{VC}dim(H)$

Question: What's the VC Dimension of KNN with $K = 1$?

Complexity Measures - VC Dimension

Theorem: Let H be a finite hypothesis set and let $\mathrm{VC}dim(H) = d$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \mathcal{O}\left(\sqrt{\frac{\ln(m/d)}{m/d}}\right)$$

OK, This is great!: All of examples we've looked at have relatively small $\mathrm{VC}dim(H)$

Question: What's the VC Dimension of KNN with $K = 1$?

Answer: $\mathrm{VC}dim(H) = \infty$

Wrap-Up

- The VC Dimension is a convenient measure of the complexity of infinite hypothesis classes
- Nice generalization bounds that depend on the VCdim

In Class

- Answer **your question**!
- Work some simple examples
- Examples of rigorous proofs of VC Dimensions

Acknowledgments

Many of these slides were adopted from Jordan Boyd-Graber

Some of the figures in this presentation were adopted from *Foundations of Machine Learning* by Mohri, et. al.

In Class

In Class

In Class