Little Bit'O Theory

How Much Data Do We Need

We've talked a lot about the Bias-Variance Trade-Off

- The more complex/flexible a model, the more likely it is to overfit
- The more training data we have, the less likely a model is to overfit

Today:

- How can we measure how complex/flexible a model is?
- Given a measure of the complexity/flexibility of a model, how much data do we need?

Complexity Measures – Counting Bits

First-Pass Attempt: Consider a classifier in 2D of the form:

$$h_\beta(\mathbf{x}) = I(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq 0)$$

- We call a learned model a hypothesis
- The class of all hypothesis of a particular form is called the Hypothesis Space H
- Talk about the complexity of H, e.g. how complex is the set of all h of the given form

Note: No assumptions about the distribution of the data

Complexity Measures – Counting Bits

First-Pass Attempt: Consider a classifier in 2D of the form:

$$h_\beta(\mathbf{x}) = I(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq 0)$$

- Hypothesis are represented by 3 parameters
- Usually the parameters are represented on a computer by double-precision variables each defined by 64 bits
- Thus we have 3 x 64 = 192 degrees of freedom in the Hypothesis Space
- For binary classification, H then consists of at most $|H| = 2^{3\cdot 64} = 2^{192}$ different hypotheses

Complexity Measures – Counting Bits

First-Pass Attempt: Consider a classifier in 2D of the form:

$$h_{\beta}(\mathbf{x}) = I(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq 0)$$

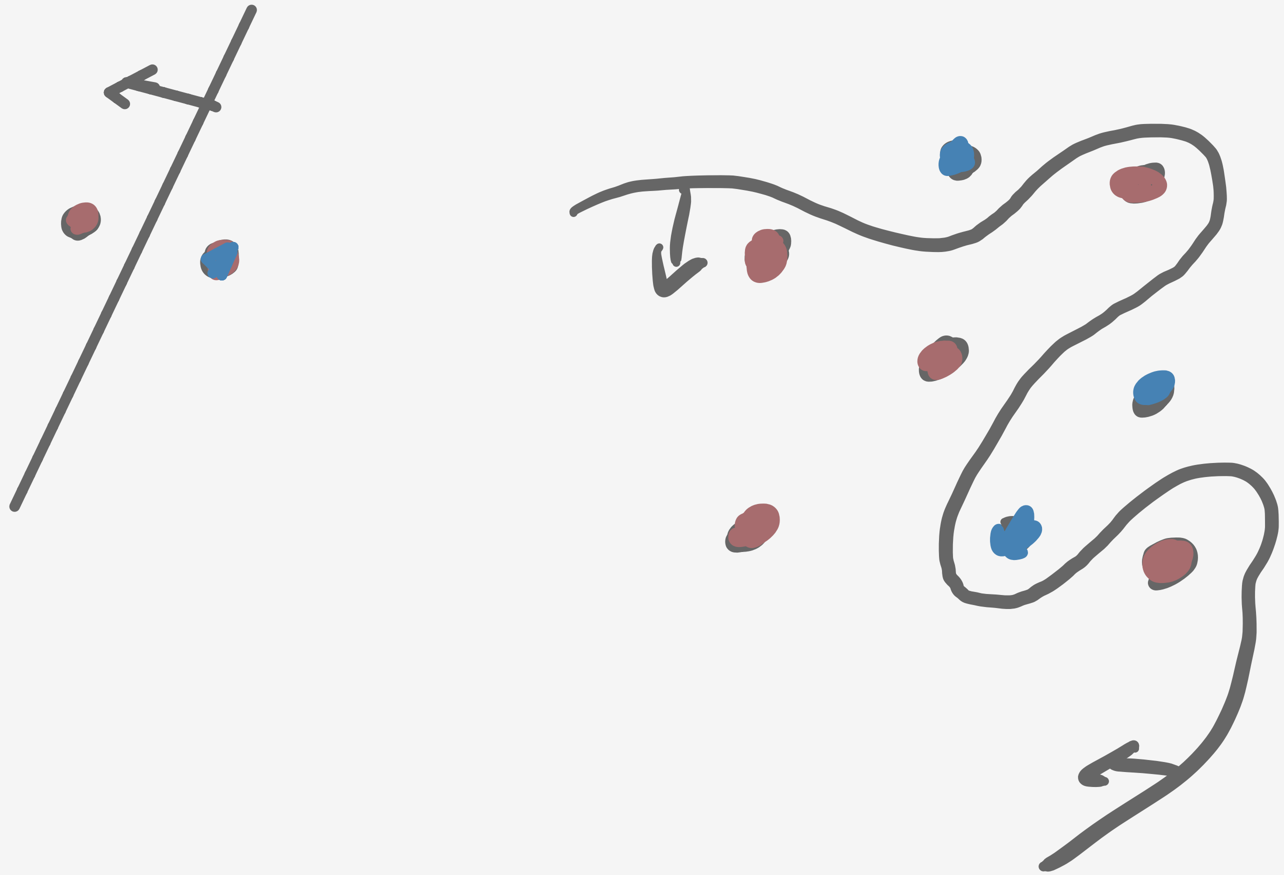
- For binary classification, H then consists of at most $|H| = 2^{3 \cdot 64} = 2^{192}$ different hypotheses
- This doesn't seem particularly helpful because it depends on number of parameters
- But we could alternatively write the model as

$$h_{\cancel{\beta}}(\mathbf{x}) = I((u_0^2 - v_0^2) + (u_1^2 - v_1^2)x_1 + (u_2^2 - v_2^2)x_2 \geq 0)$$

- Produces exactly the same hypotheses, but basing complexity of number of parameters then somehow suggests that this model is more complex …

A More Practical Complexity Measure

Second-Pass Attempt: Measure complexity/flexibility by what a model **can do**



A More Practical Complexity Measure

Second-Pass Attempt: Measure complexity/flexibility by what a model **can do**

Def: A **dichotomy** of a set S of points is a specific association of binary labels to the points in S

Def: A set of points S is **shattered** by Hypothesis Class H if H can correctly classify **ALL** dichotomies of S

Question: How many dichotomies must we consider if S contains n points?

A More Practical Complexity Measure

Second-Pass Attempt: Measure complexity/flexibility by what a model **can do**

Def: A **dichotomy** of a set S of points is a specific association of binary labels to the points in S

Def: A set of points S is **shattered** by Hypothesis Class H if H can correctly classify **ALL** dichotomies of S

Def: The VC Dimension of H is the size of the largest set S that can be shattered by H

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

VC Dimension

Def: The VC Dimension of H is the size of the largest set S that can be shattered by H

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

- Named for its inventors Vladimir Vapnik and Alexey Chervonenkis

Example: Suppose you have one feature. What is the VC Dimension of for intervals of the form

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

Need to:

- Find an example of the largest set of points that can be shattered
- Show that no bigger set of points can ever be shattered

VC Dimension

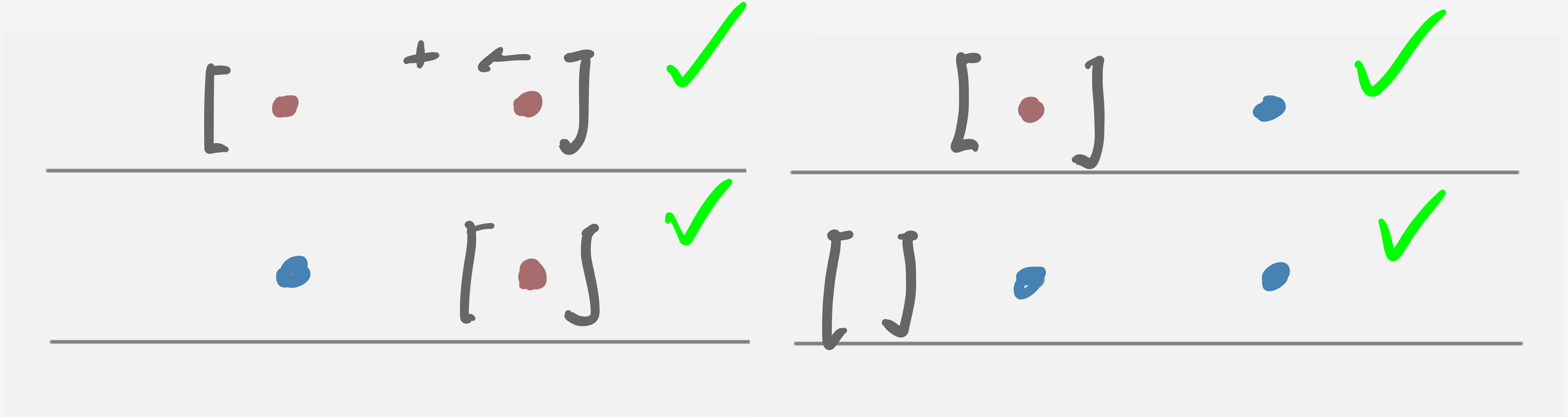
$x_1 \neq x_2$

Example: Suppose you have one feature. What is the VC Dimension of for intervals of the form

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

BOOM! SHATTERED!

Step 1: Pick a set of points and show that it can be shattered



VC Dimension

Example: Suppose you have one feature. What is the VC Dimension of for intervals of the form

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

Step 2: Argue that **NO** set of points of one size larger can ever be shattered

x_1, x_2, x_3 wolog

$x_1 \leq x_2 \leq x_3$

VC Dimension

Example: Suppose you have one feature. What is the VC Dimension of for intervals of the form

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

Framework:

- Step 1 proves that $\mathrm{VCdim}(H) \geq 2$
- Step 2 proves that $\mathrm{VCdim}(H) < 3$
- Combining Steps 1 and 2 proves that $\mathrm{VCdim}(H) = 2$

VC Dimension FAQs

Q: Am I allowed to use different hypothesis for different dichotomies?

A: Totally! It would be pretty hard to do otherwise!

Q: Does H have to shatter ALL sets of d points?

A: Nope! You're free to pick any convenient set that you like! Only need to find one!

Q: Why do I have to do Step 2? Can't I just find a set of d points that can be shattered?

A: Nope! Because there might be a set of d+1 points that could be shattered!

VC Dimension

Example: Suppose our data has 2 features. What is the VC Dim of the set of linear classifiers?

$$h_\beta(\mathbf{x}) = I(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq 0)$$

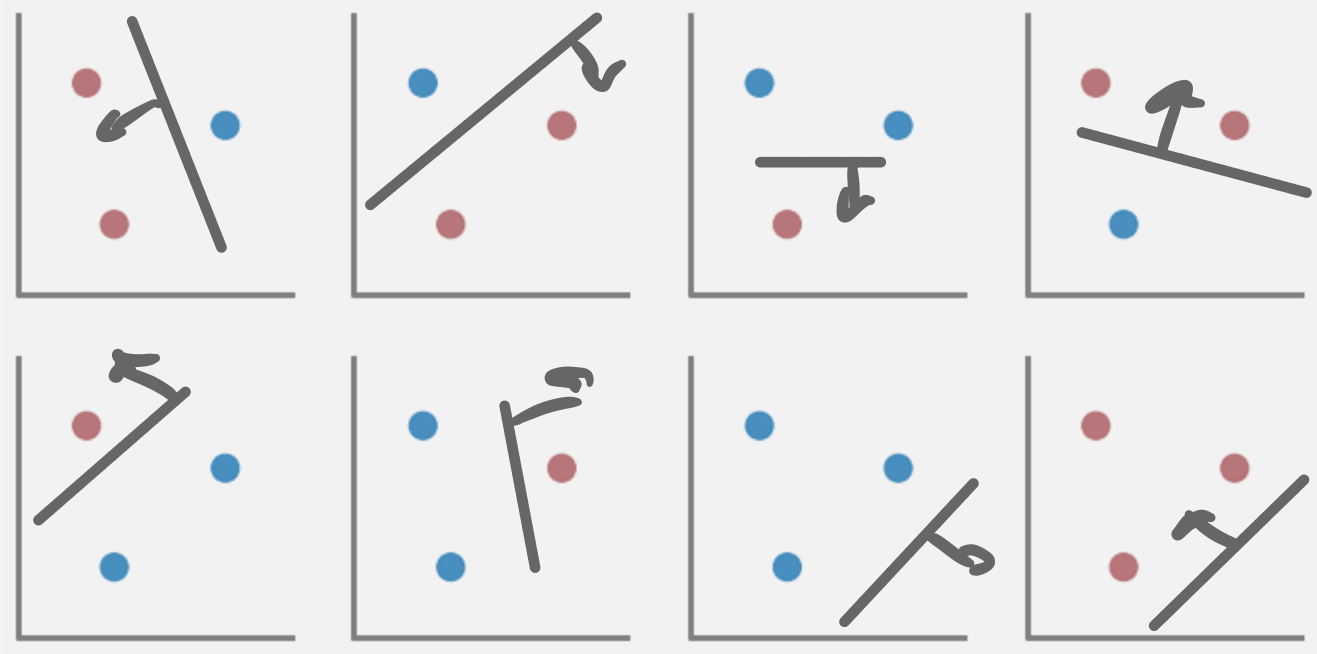
Question: If you had to guess, what would you say?

VC Dimension

Example: Suppose our data has 2 features. What is the VC Dim of the set of linear classifiers?

$$h_\beta(\mathbf{x}) = I(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq 0)$$

Step 1: Can we shatter a set of 3 points?

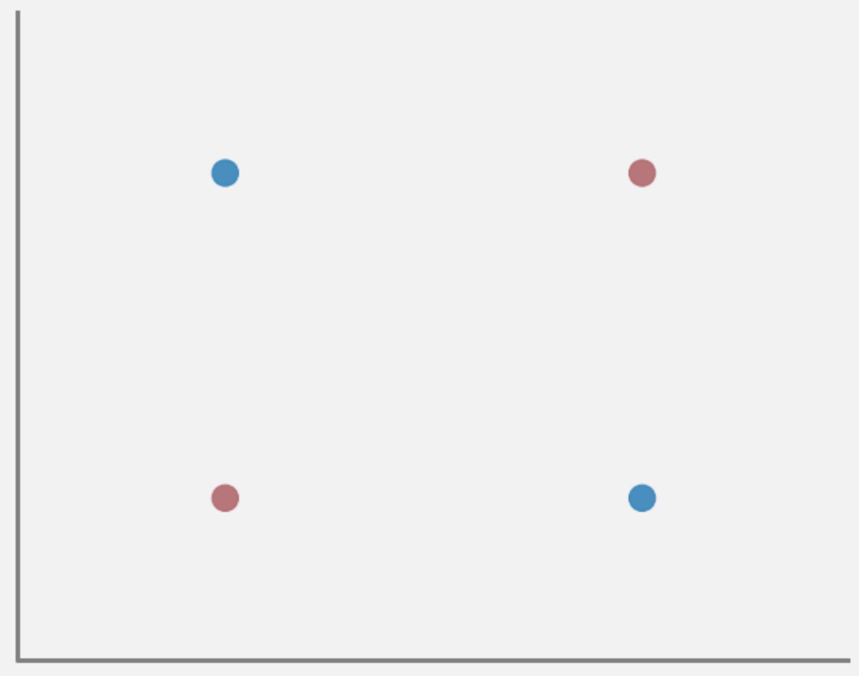


VC Dimension

Example: Suppose our data has 2 features. What is the VC Dim of the set of linear classifiers?

$$h_\beta(\mathbf{x}) = I(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq 0)$$

Step 2: Can we shatter **any** set of 4 points?

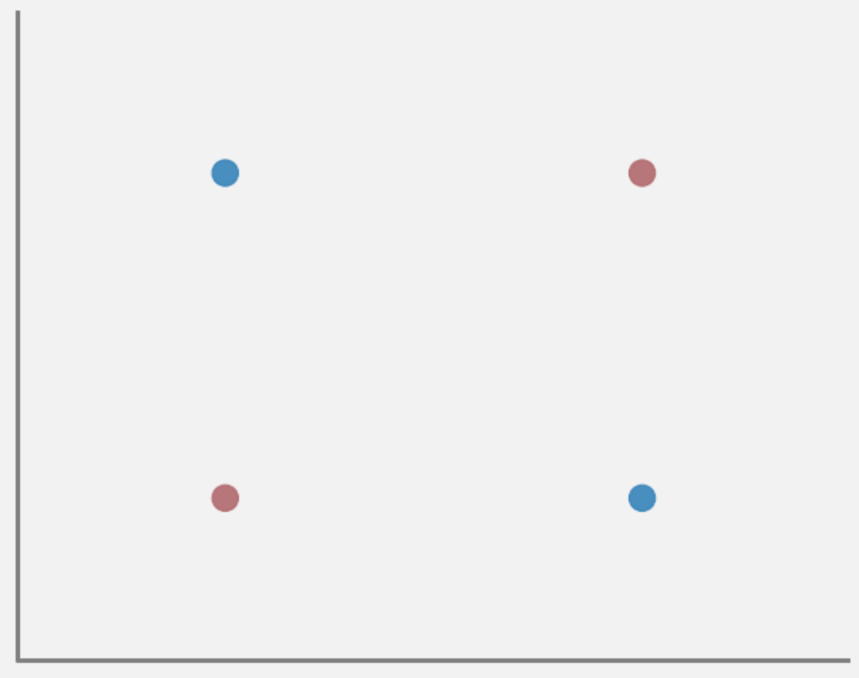


VC Dimension

Example: Suppose our data has 2 features. What is the VC Dim of the set of linear classifiers?

$$h_\beta(\mathbf{x}) = I(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq 0)$$

Step 2: Can we shatter **any** set of 4 points?



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

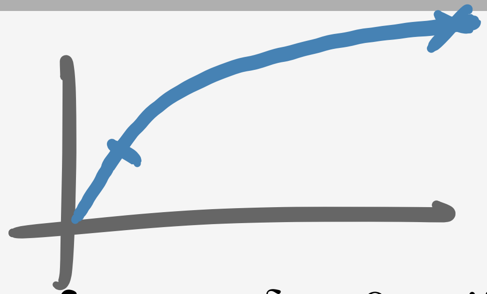
VC Dimension

Fact: If H is the set of all linear classifiers defined of data with p features, then

$$\text{VCdim}(H) = \underline{\underline{p + 1}}$$

Payoff: OK, so what does this tell us about flexibility, Bias-Variance, amount of data, etc?

VC Dimension



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

$$\text{Generalization Error} \leq \text{Training Error} + \underbrace{\sqrt{\frac{2d\log(em/d)}{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}}$$

where m is the number of training examples.

$$\boxed{x = \frac{m}{d}}$$

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

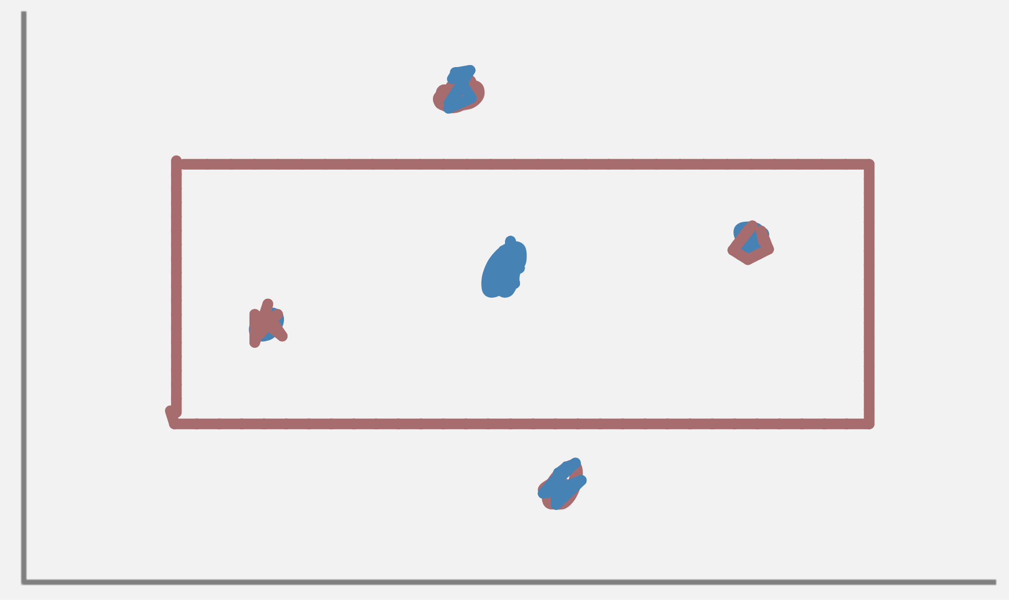
$$\text{Generalization Error} \leq \text{Training Error} + \mathcal{O}\left(\sqrt{\frac{\log(m/d)}{m/d}}\right)$$

Training error is a good indicator of Generalization error if $m \gg d$

VC Dimension

Example: Suppose our data has 2 features. What is the VC Dim of the hypothesis class H containing the set of all axis-aligned rectangles?

Step 1: Can we shatter **some** set of 4 points?

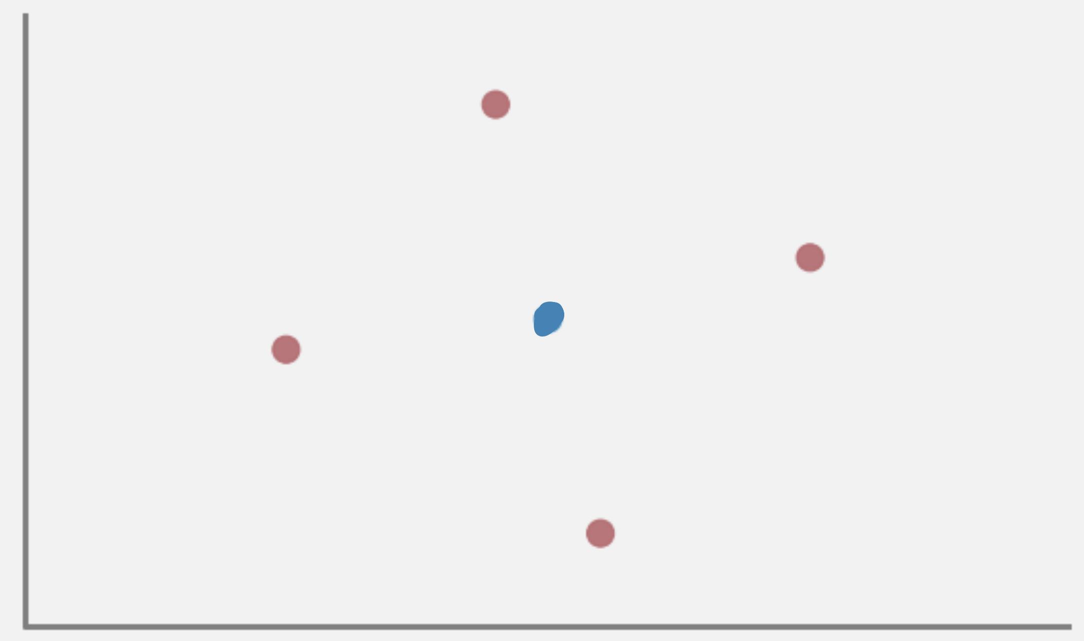


VC Dimension

Example: Suppose our data has 2 features. What is the VC Dim of the hypothesis class H containing the set of all axis-aligned rectangles?

Step 2: Can we shatter **any** set of 5 points?

Nope! So $\mathrm{VCdim}(H) = 4$



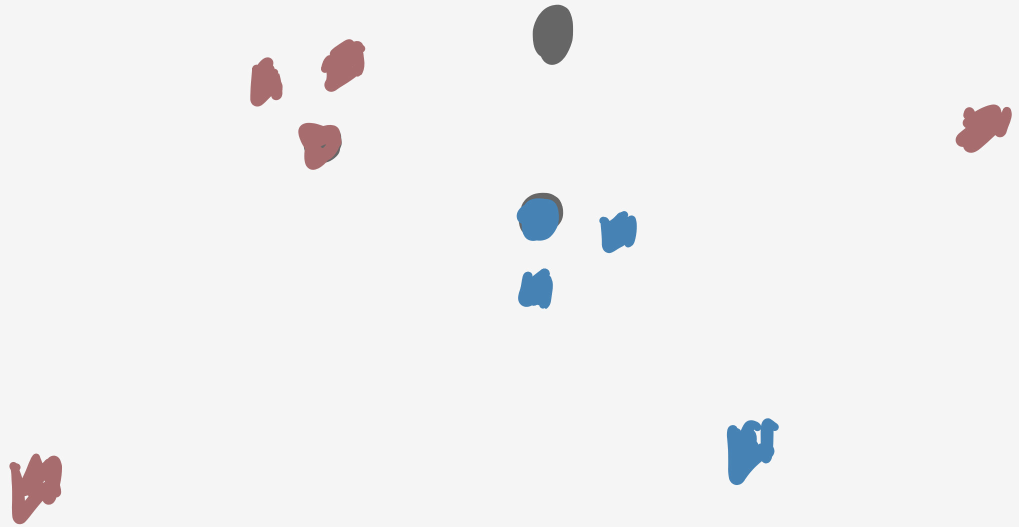
VC Dimension

2D FEATURS

OK, here's a **tough one**:

Example: What is the VC Dimension of K-Nearest Neighbors?

K = 2



VCDIM = ∞