Model Complexity and VC Dimension

Machine Learning
CSx824/ECEx242
Bert Huang
Virginia Tech

Outline

- Review form of bound
- VC dimension definition
- VC dimension of large-margin classifiers

Generalization Error Bound

true risk

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\mathrm{VC}(H)\left(\log \frac{2n}{\mathrm{VC}(H)} + 1\right) - \log \frac{\delta}{4}}{n}}$$

empirical risk

failure probability

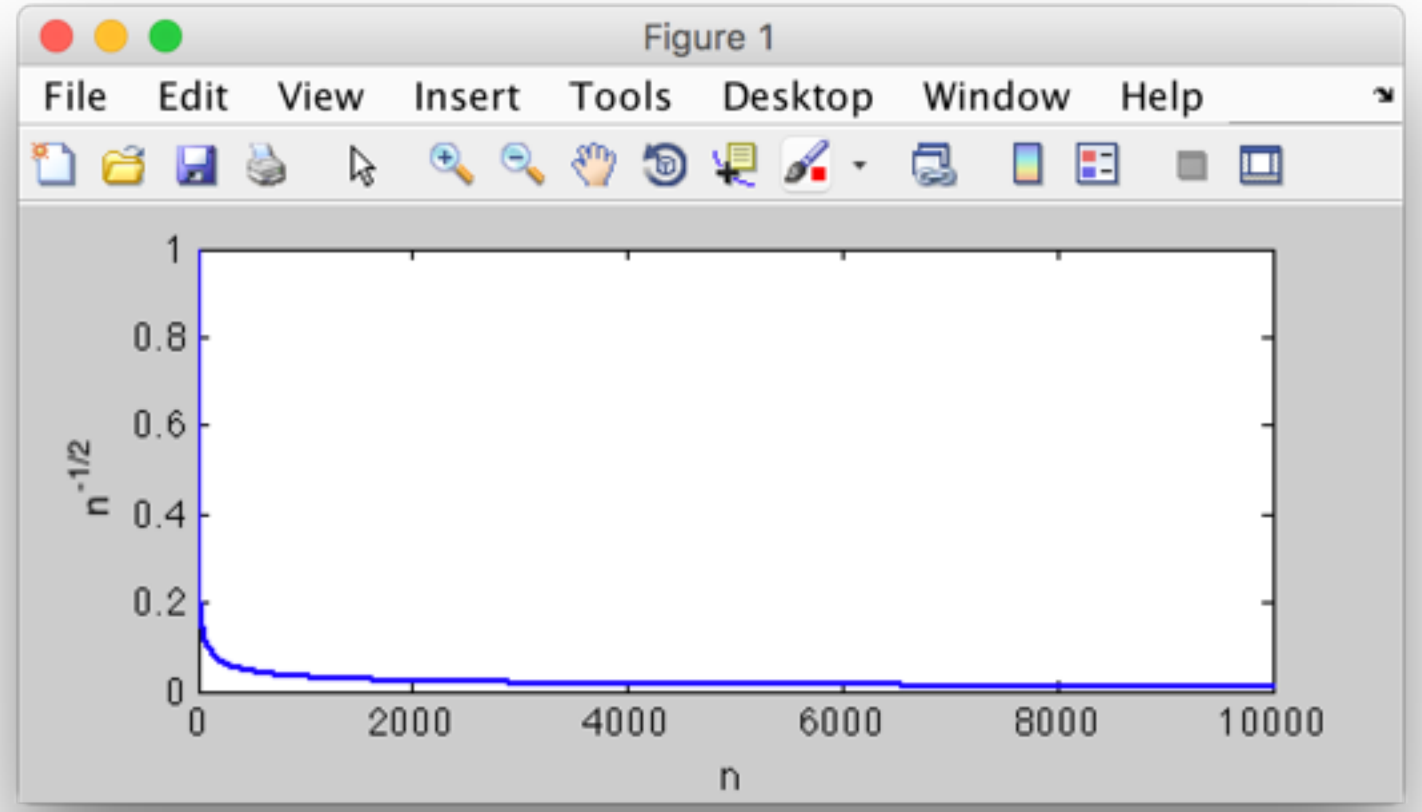
size of training data

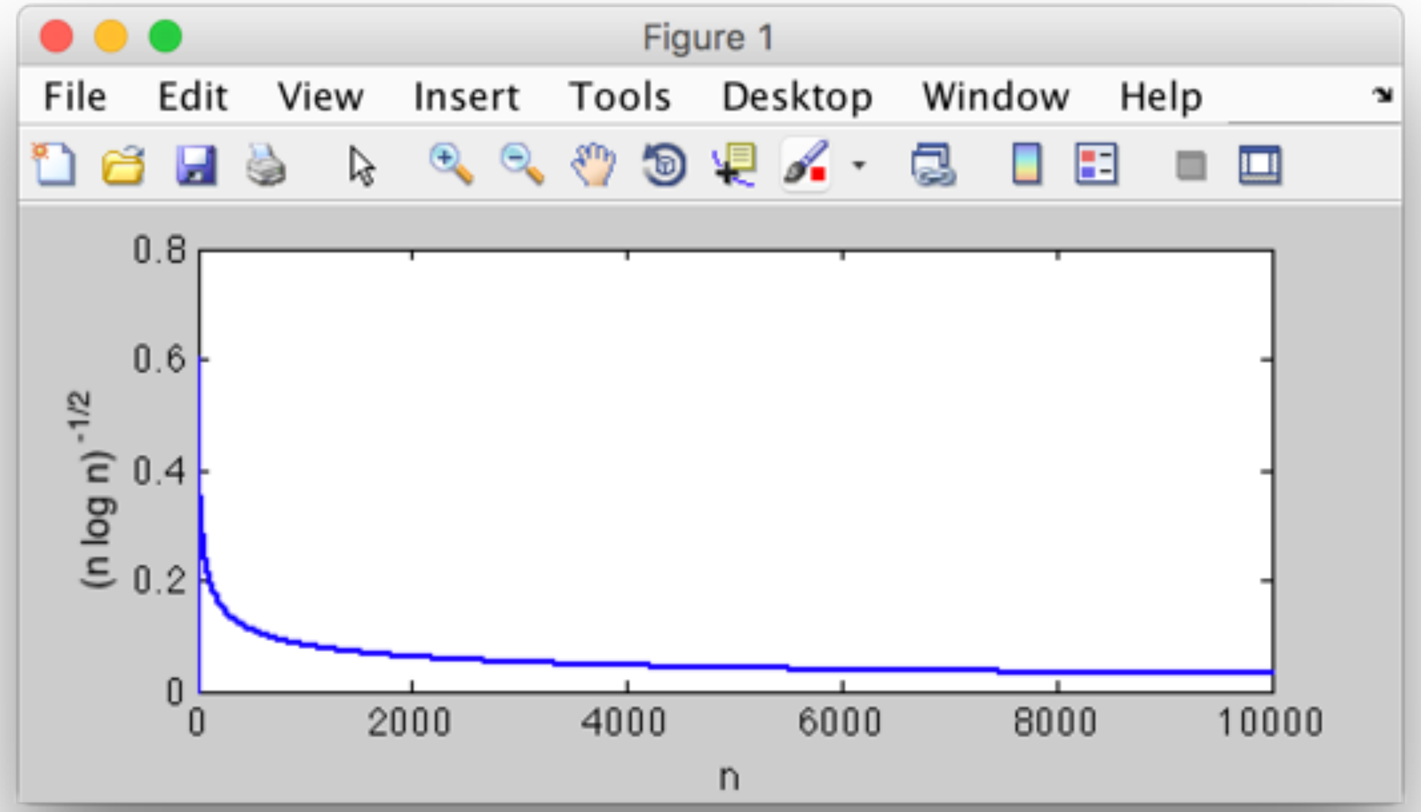
Vapnik-Chervonenkis dimension (model complexity)

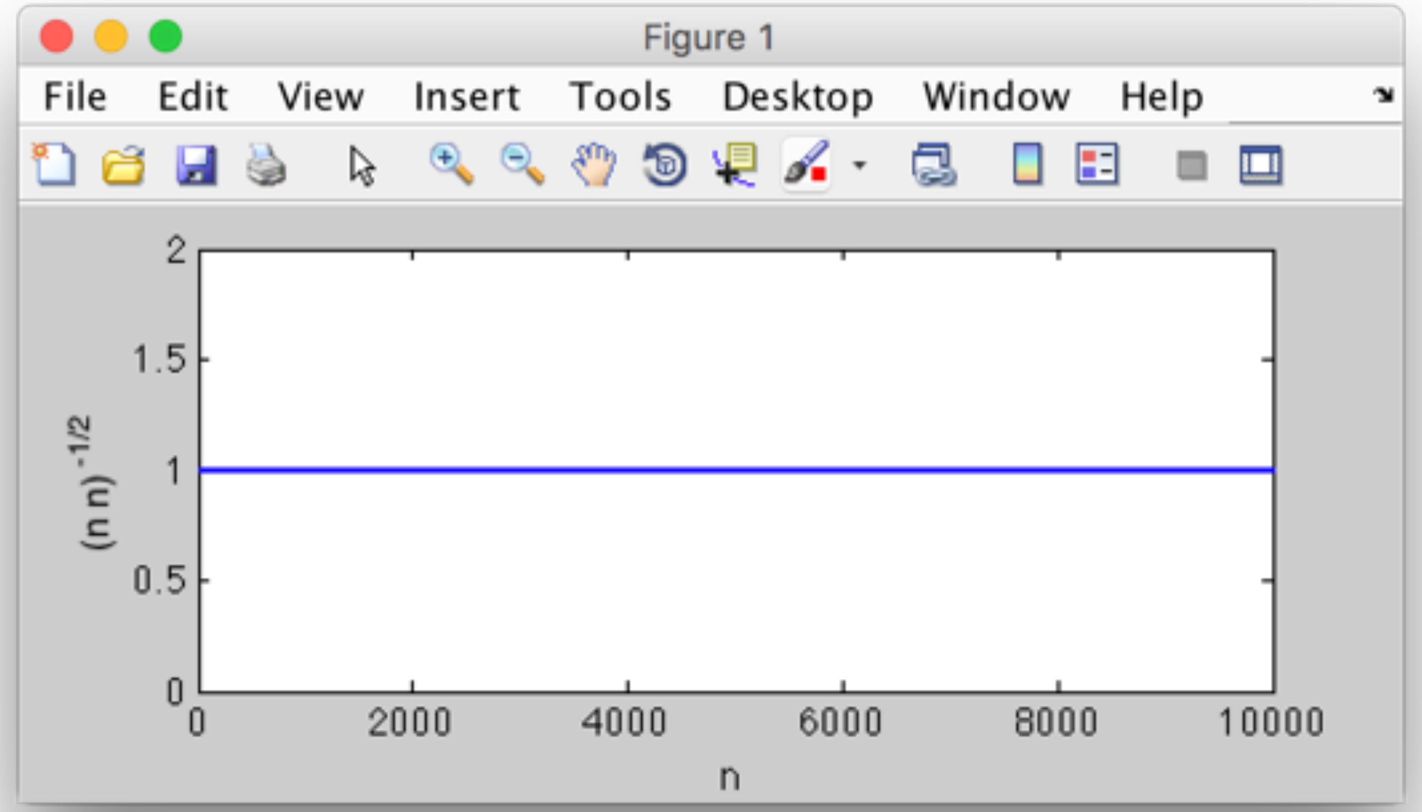
$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\mathrm{VC}(H)\left(\log \frac{2n}{\mathrm{VC}(H)} + 1\right) - \log \frac{\delta}{4}}{n}} \approx \sqrt{\frac{\text{complexity(H)}}{n}}$$

if complexity is fixed

if complexity is O(n)







Vapnik-Chervonenkis Dimension

- Expressive power, or **capacity**, of a **hypothesis class**
  - Linear classifiers in d-dimensional space
  - Degree k polynomial classifiers
  - Hierarchical axis-parallel classifiers (decision trees)
- Measured by ability of hypothesis class to **shatter** n points

Shattering

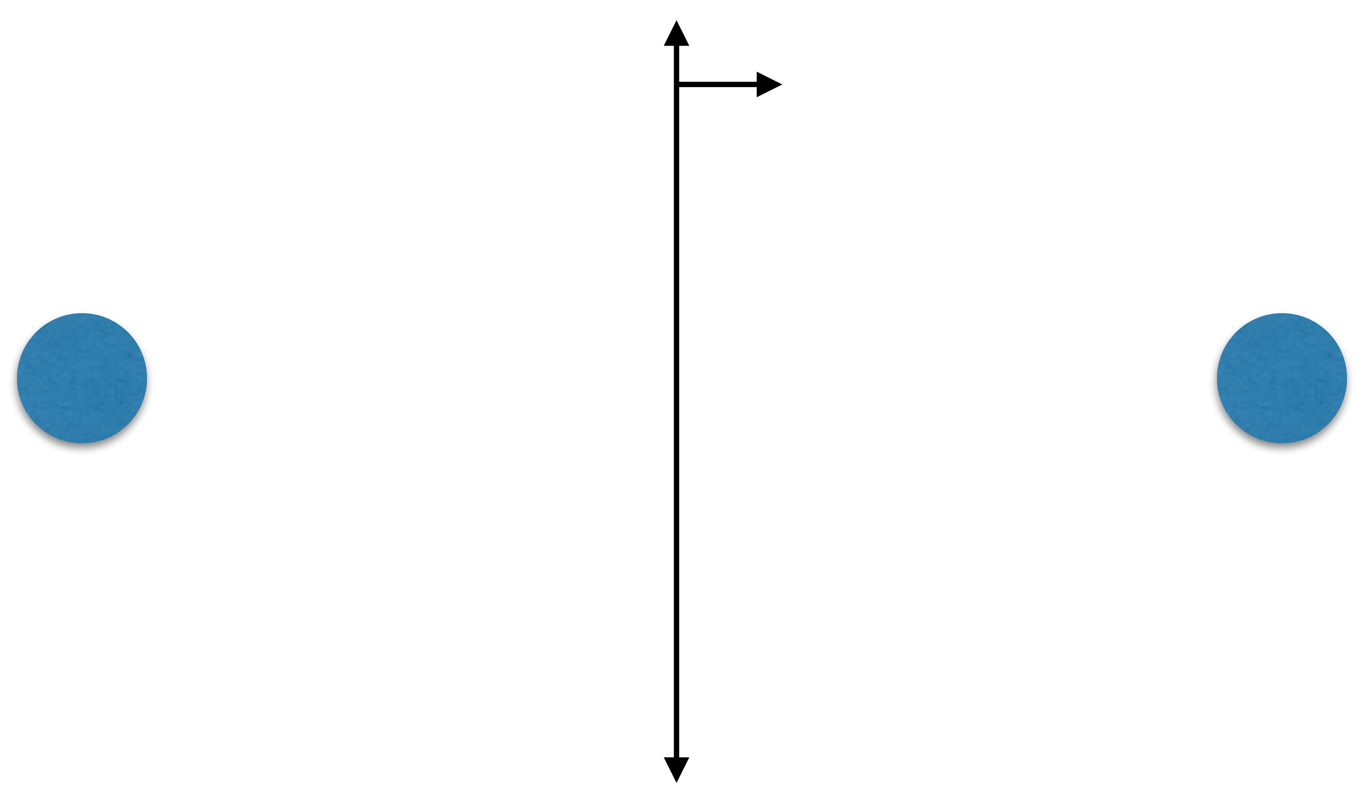
Classify points into all possible labels

✔
++, +-, -+, - -

Shattering

Classify points into all possible labels

✔ ✔
++, +-, -+, - -



Shattering

Classify points into all possible labels

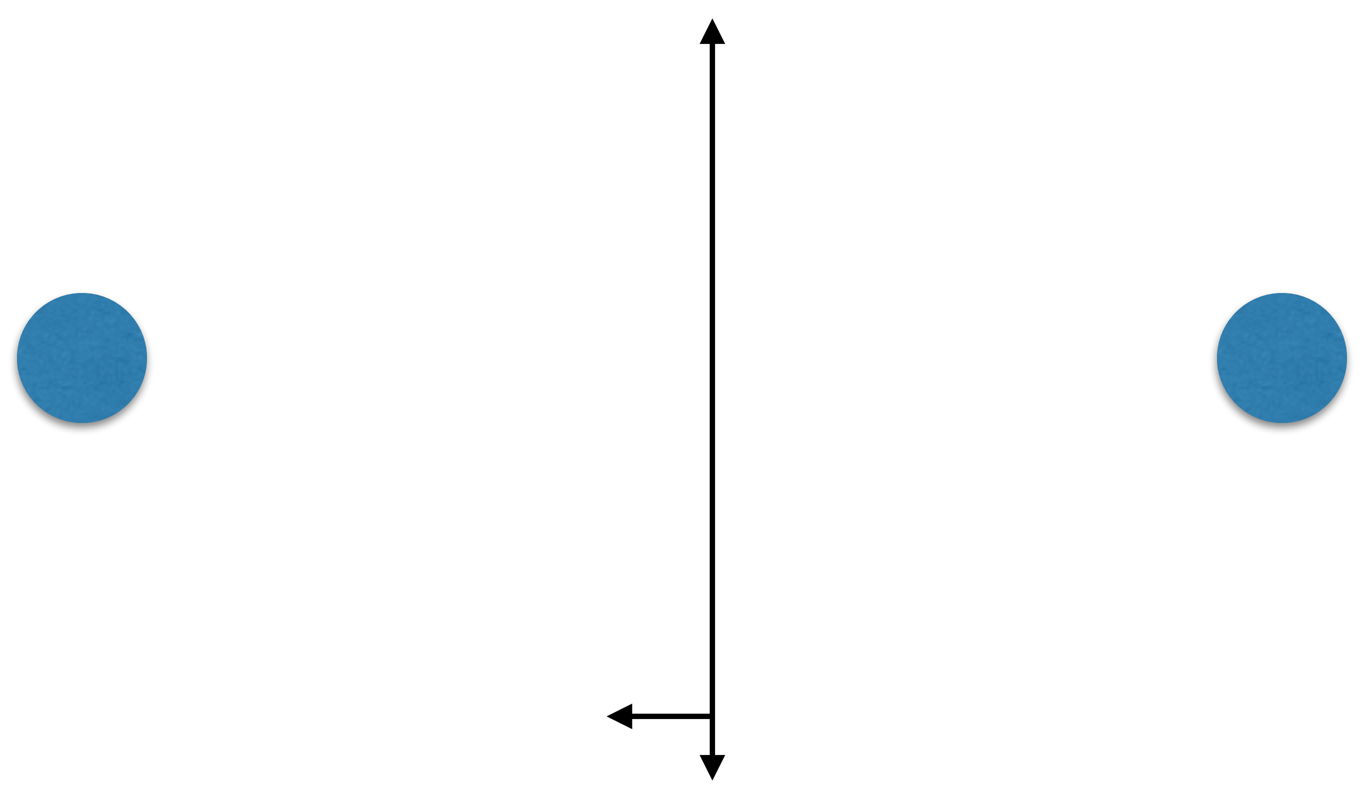
✔ ✔ ✔

++, +-, -+, - -

Shattering

Classify points into all possible labels

✔ ✔ ✔ ✔
++, +-, -+, - -

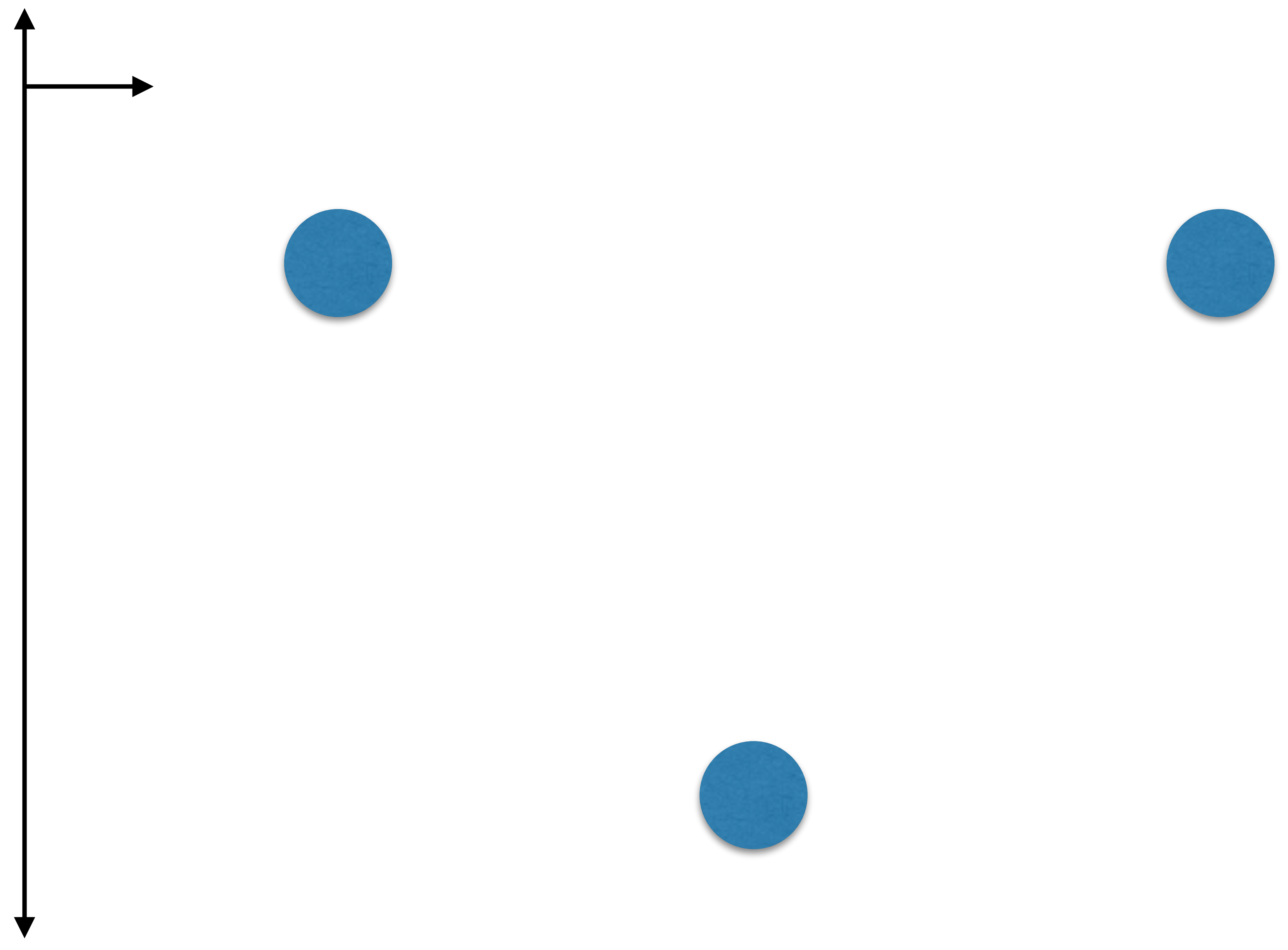


Shattering

Classify points into all possible labels

✔

+++, ++-, +-+, +- -, -++, - + -, - - -

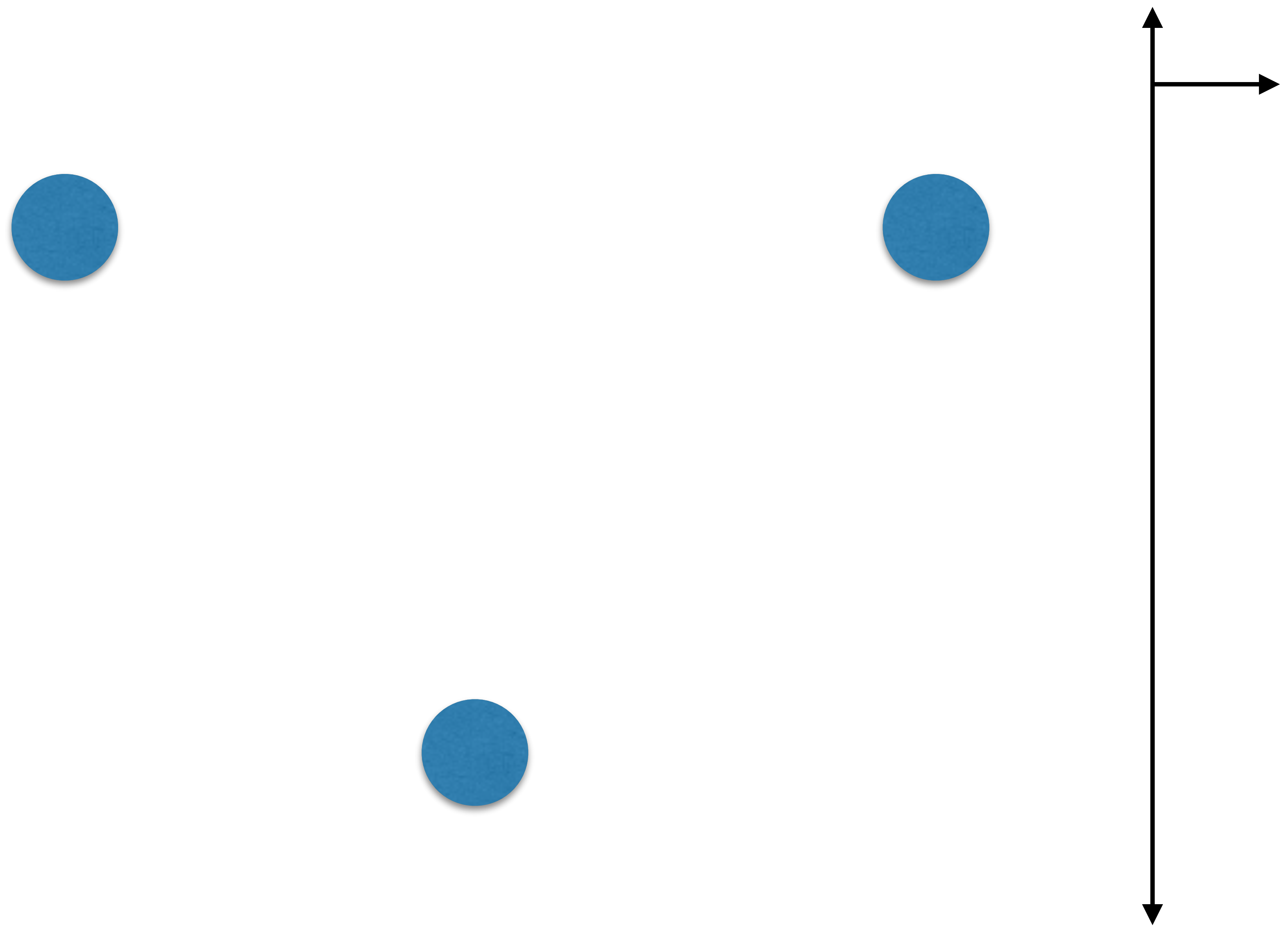


Shattering

Classify points into all possible labels

✔ ✔

+++, ++-, +-+, +- -, -++, - + -, - - -

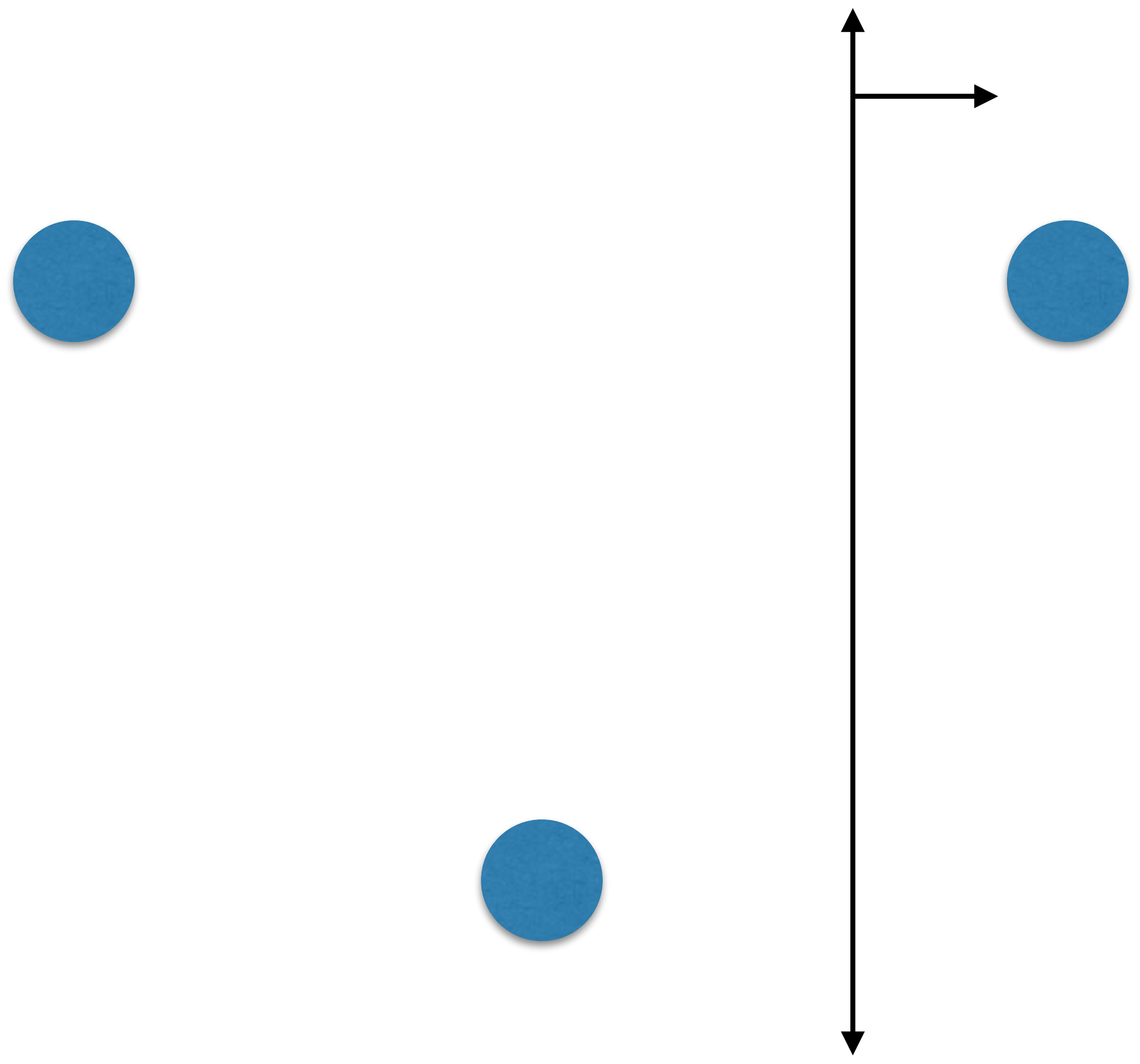


Shattering

Classify points into all possible labels

✔ ✔ ✔

+++, ++-, +-+, +- -, -++, - + -, - - -

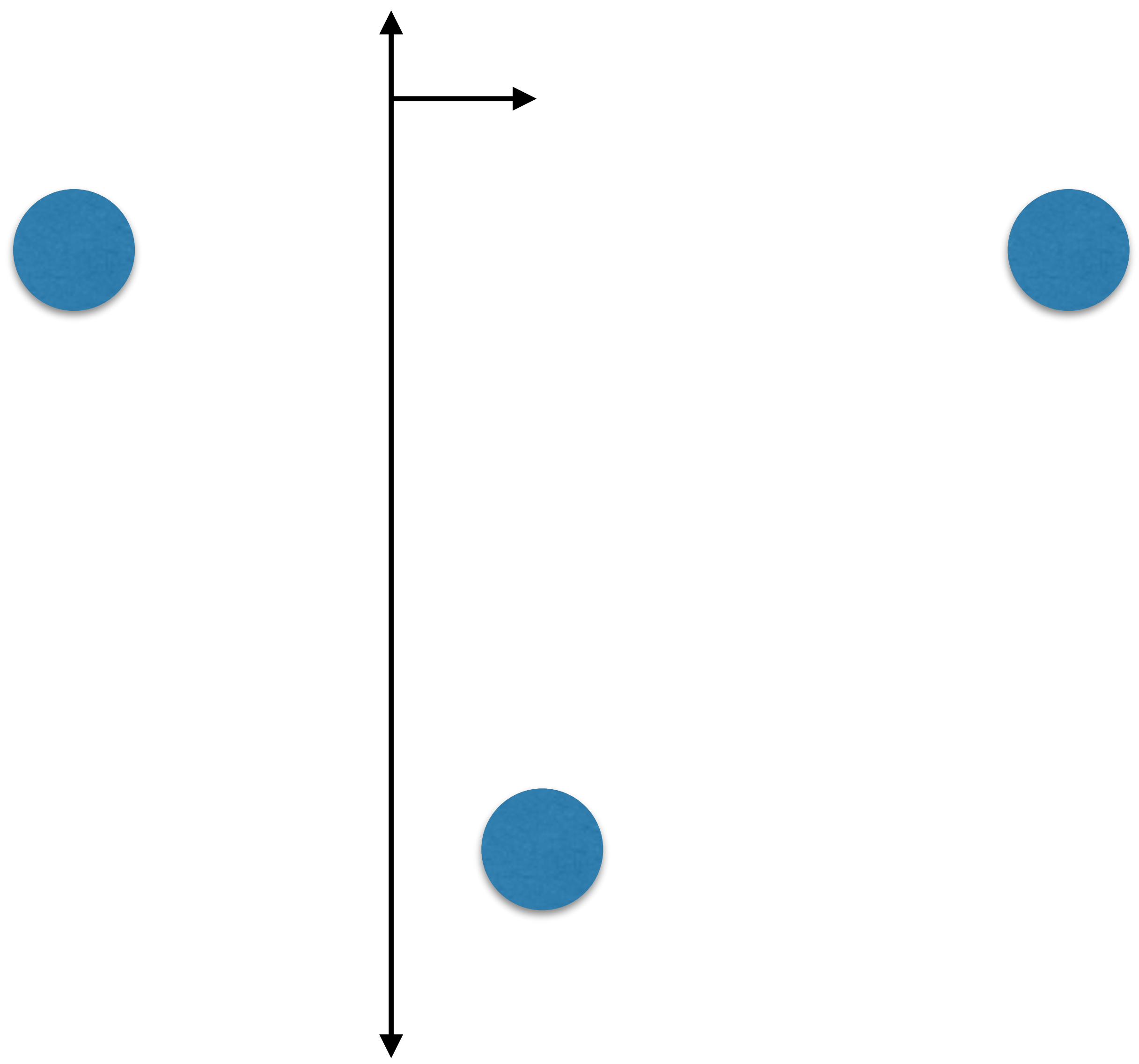


Shattering

Classify points into all possible labels

✔ ✔ ✔ ✔

+++, ++-, +-+, +- -, -++, - + -, - - -

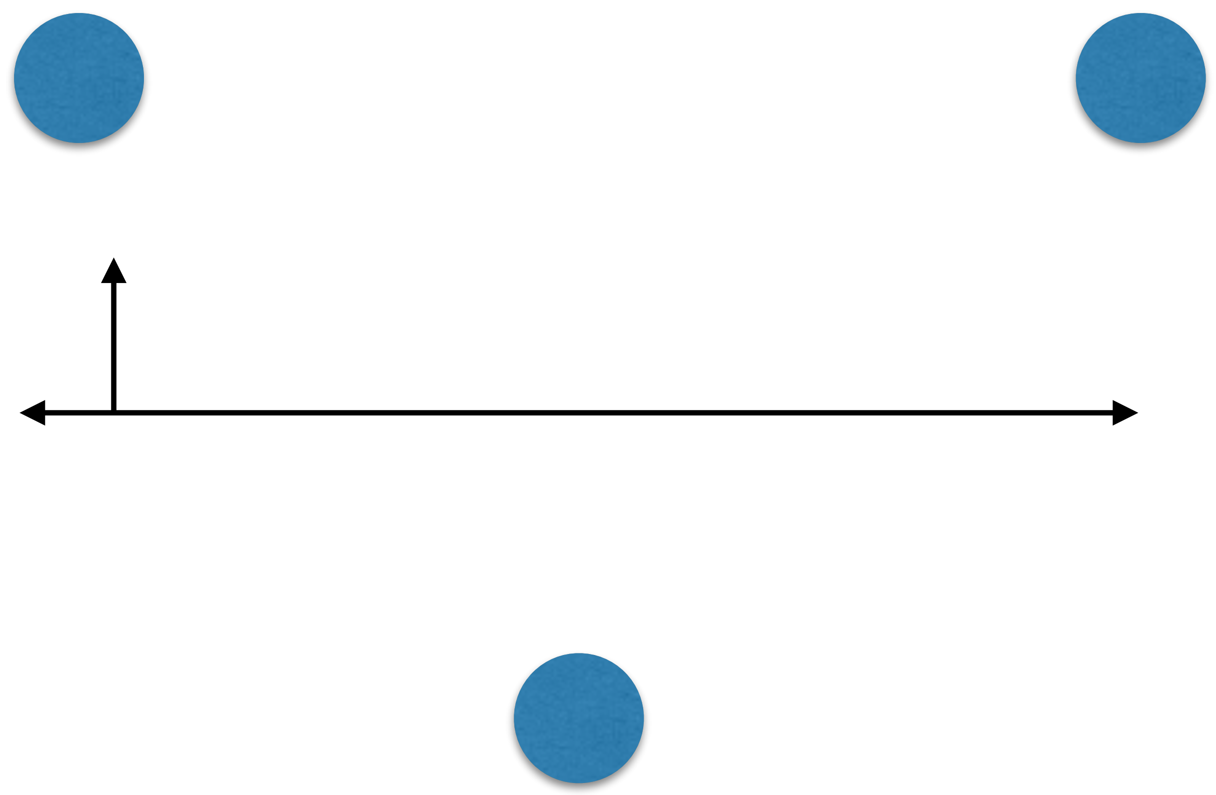


Shattering

Classify points into all possible labels

✔ ✔ ✔ ✔ ✔

+++, ++-, +-+, +- -, -++, - + -, - - -

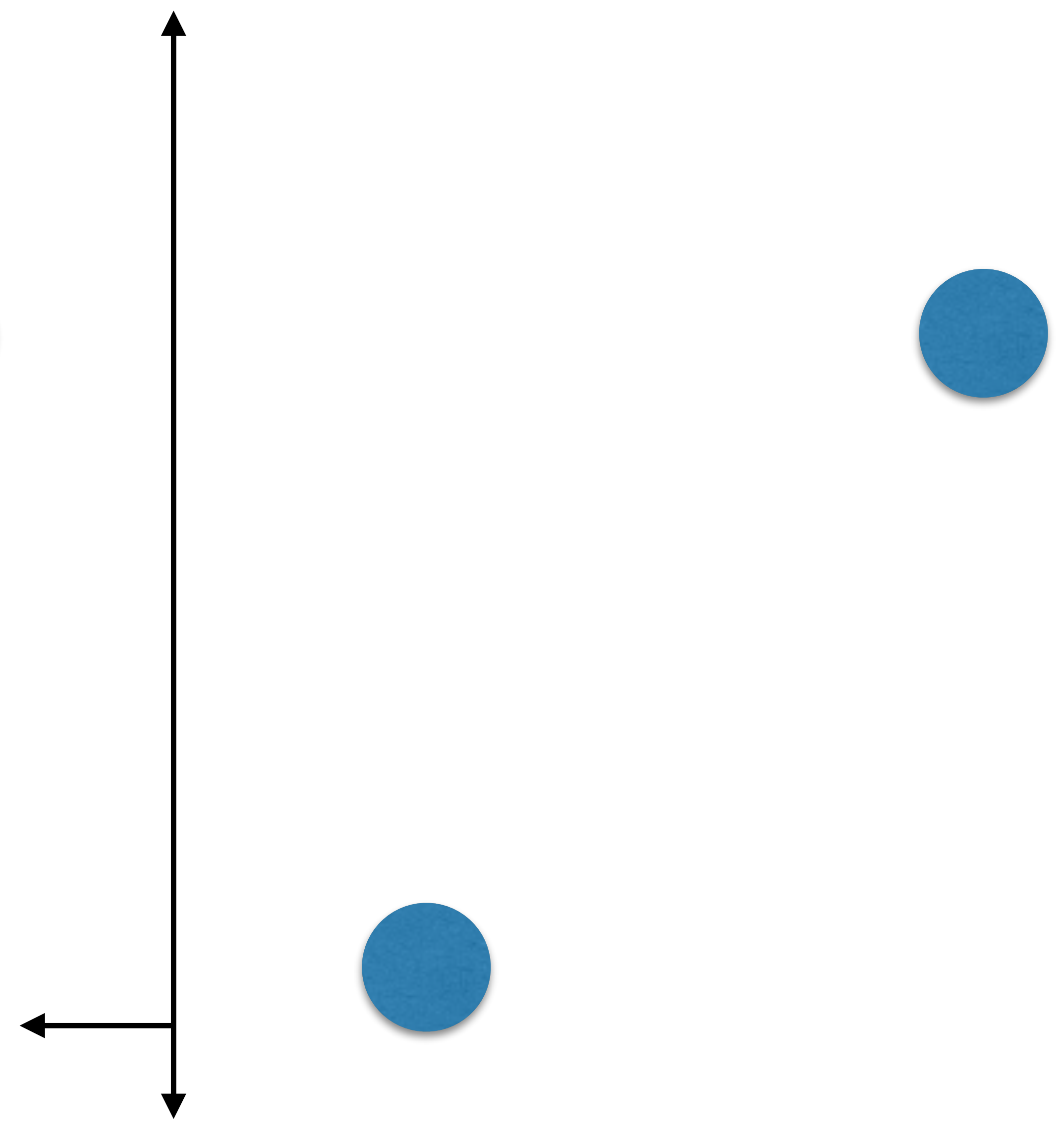


Shattering

Classify points into all possible labels

✔ ✔ ✔ ✔ ✔ ✔

+++, ++-, +-+, +- -, -++, - + -, - - -

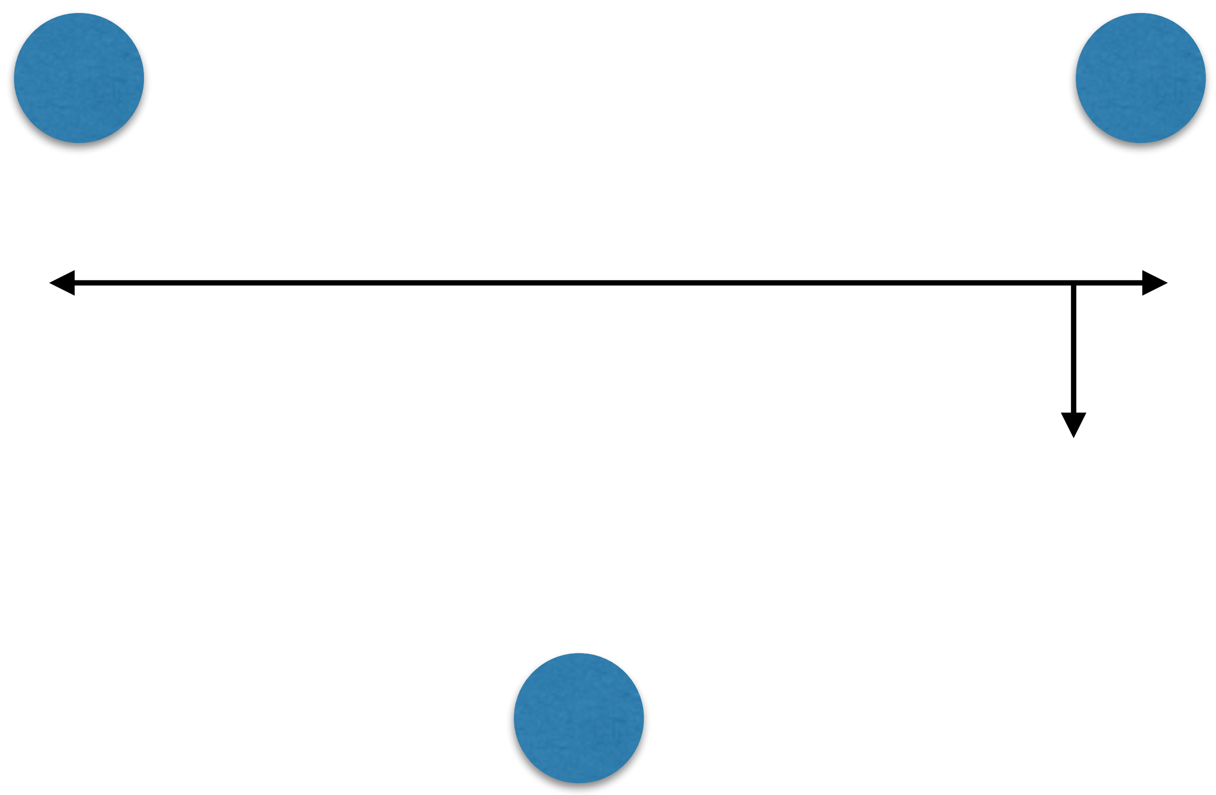


Shattering

Classify points into all possible labels

✔ ✔ ✔ ✔ ✔ ✔ ✔

+++, ++-, +-+, +- -, -++, - + -, - - -



Shattering

Classify points into all possible labels

4 points cannot be shattered by 2d linear classifier

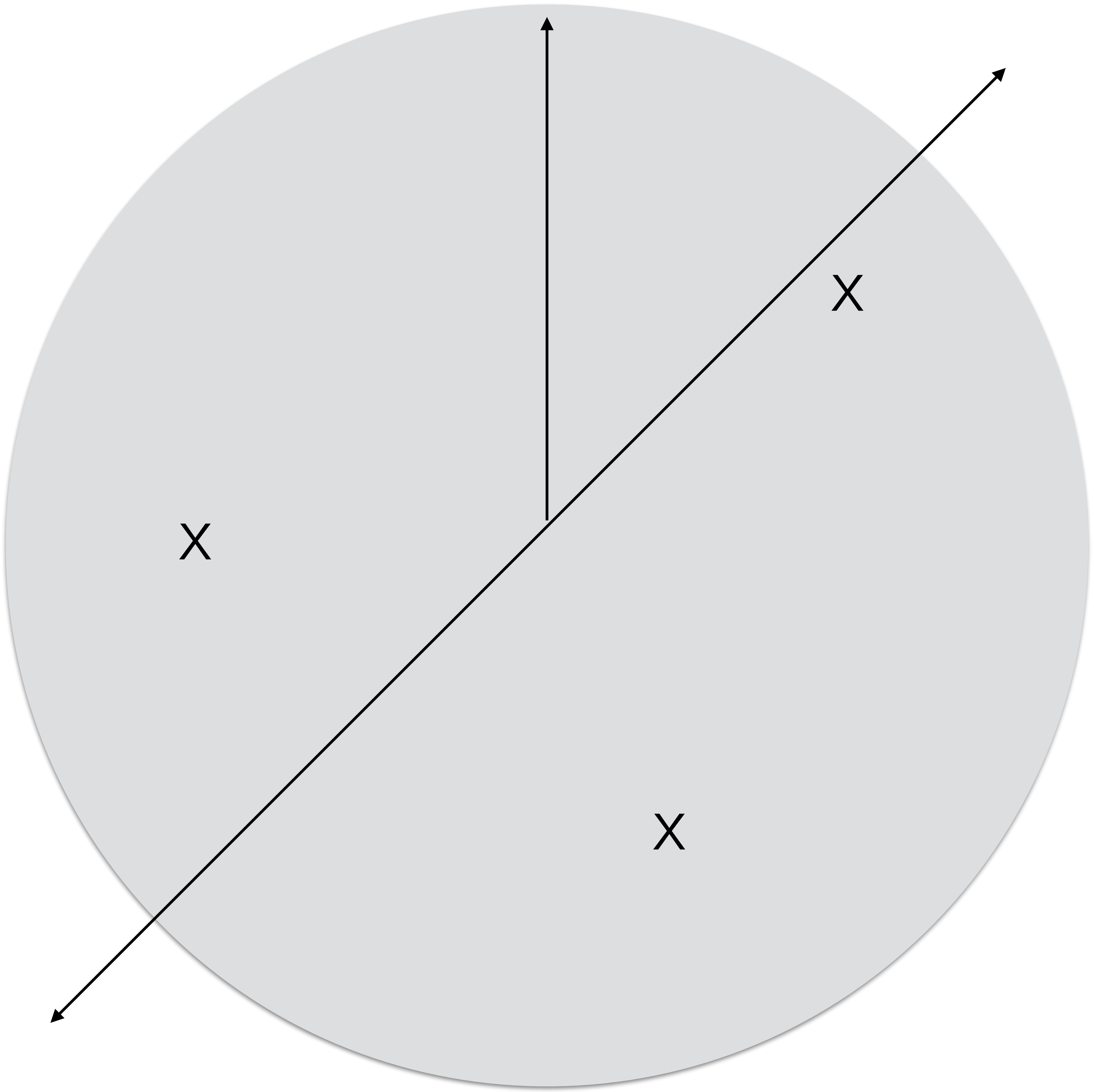
VC Dimension

- VC dimension of hypothesis class H:
- Maximum number of examples that can be shattered by H
- Examples can be arranged (feature values) in any way
- Must be shattered in same arrangement
- In general: linear classifier has VC dimension (d + 1)

VC Model Capacity Intuition

- How many points can this model class memorize?
- Game view:
    - We choose placement of points
    - Adversary chooses labeling
    - Can we classify labeling?
- Think of learning algorithm as function $A : \mathcal{X} \rightarrow \mathcal{H}$
  and hypothesis as a function $h : \mathcal{X} \rightarrow \mathcal{Y}$
- VC dimension |***y***| means ***A*** can output an ***h*** that can output any ***y***

Margin

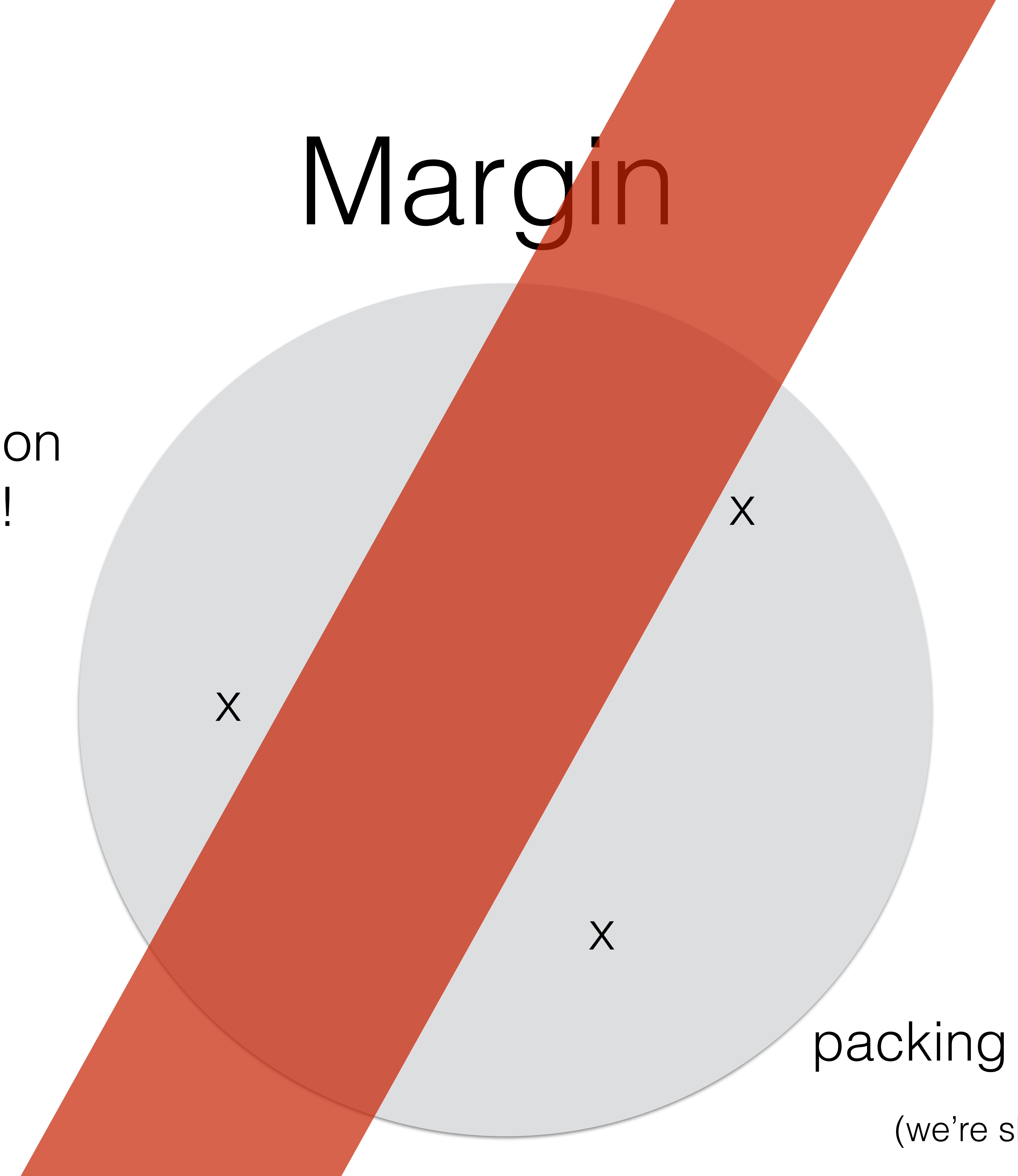


radius = R

Margin

$VC(H) = R^2 w^T w$

doesn't depend on dimensionality!



radius = R

packing points into a sphere

(we're skipping lots of details)

Summary and Thoughts

- From analysis, SVM appears to minimize VC dimension
  - but bound assumes VC dimension is fixed
- Generalization bounds tend to be loose for real data sizes
- Formally describe trend, but are they useful?
  - Better (tighter) bounds are certainly useful
  - But loose bounds help us formally understand properties of learning algorithms