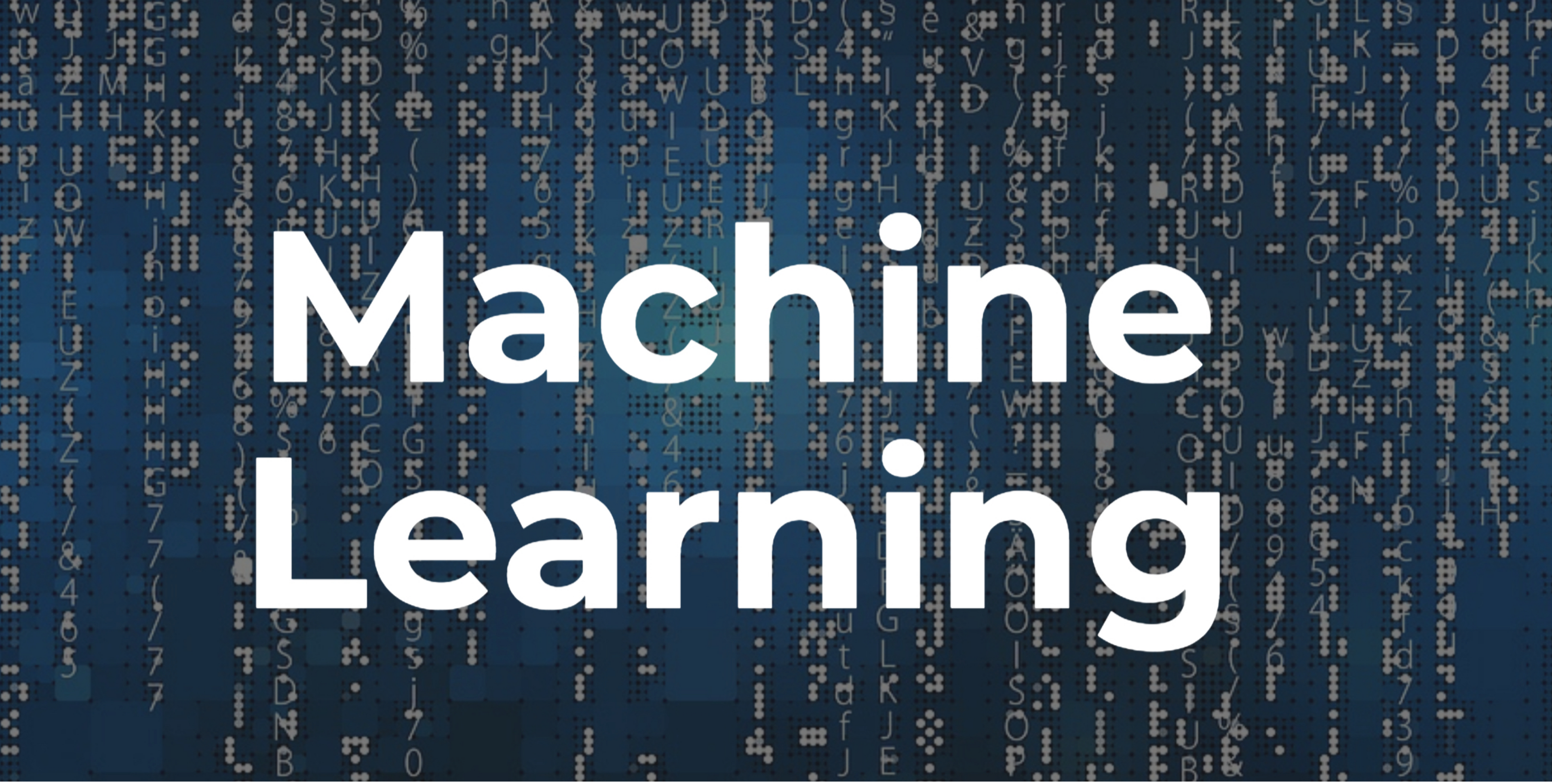


CSCI 5622 Fall 2020

Prof. Claire Monteleoni

Today

- Intro. to Learning Theory
  - Complexity of classifiers
    - VC dimension
    - Margin
    - Sauer's lemma (if time)

Complexity of classifiers

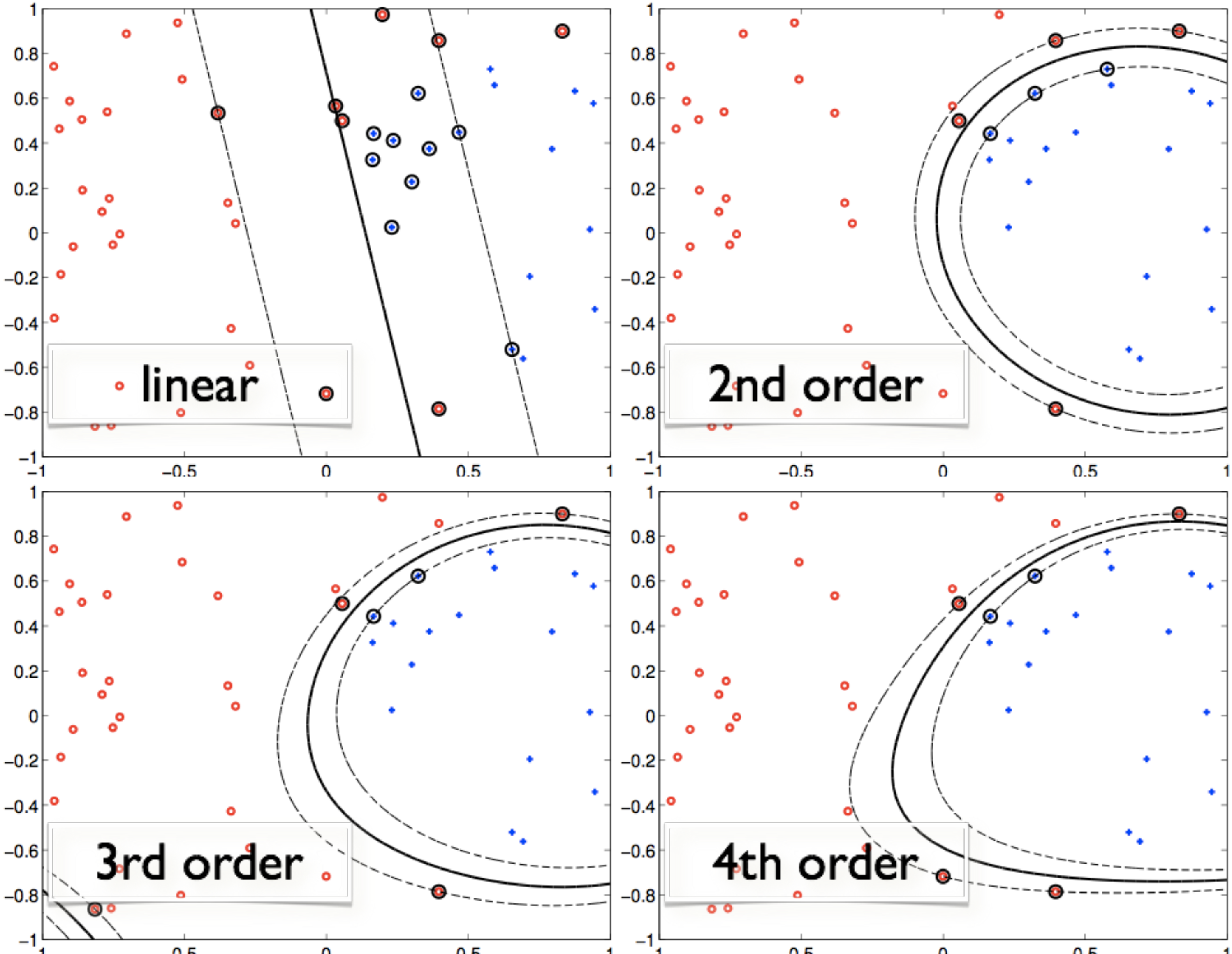


Image credit: T. Jaakkola

VC dimension

We have seen various different hypothesis classes, e.g. linear separators, polynomial separators, decision trees, nearest-neighbor classifiers, etc.

How do we compare them? One way is to define a notion of the level of complexity of a hypothesis class H.

Complexity can be defined in various ways.

Complexity and the ability to generalize to future data are inversely related (as we have mentioned repeatedly!)

→ i.e. as complexity of the classifier increases, it can lead to overfitting.

VC dimension

Definition: Vapnik-Chervonenkis (VC) dimension.

The VC-dimension of a hypothesis class H is the maximum number of points that a classifier can shatter.

This is a measure of the complexity of H.

To prove that the VC dimension of a class H is V, you must prove both of the following:

- H can shatter V points.
- H cannot shatter V+1 points.

Shattering

Definition: A hypothesis class H can shatter n points if there exists a set of n points such that for every possible labeling of the points, there exists some h in H that can attain that labeling.

To prove that a hypothesis class H can shatter n points, you must:

$\exists$ give a point set S of size n

$\forall$ and show that for all possible labelings of S

$\exists$ there exists some h in H which achieves that labeling.

Shattering

To prove that a hypothesis class H cannot shatter n points you need to you need to show that:

- $\forall$ for any set of n points
  - $\exists$ there exists some labeling of that configuration
    - $\forall$ such that no h in H can achieve that labeling of that configuration (i.e. for all h in H, h cannot)

VC dimension

- Example: what is the VC dimension of decision stumps in R^1?

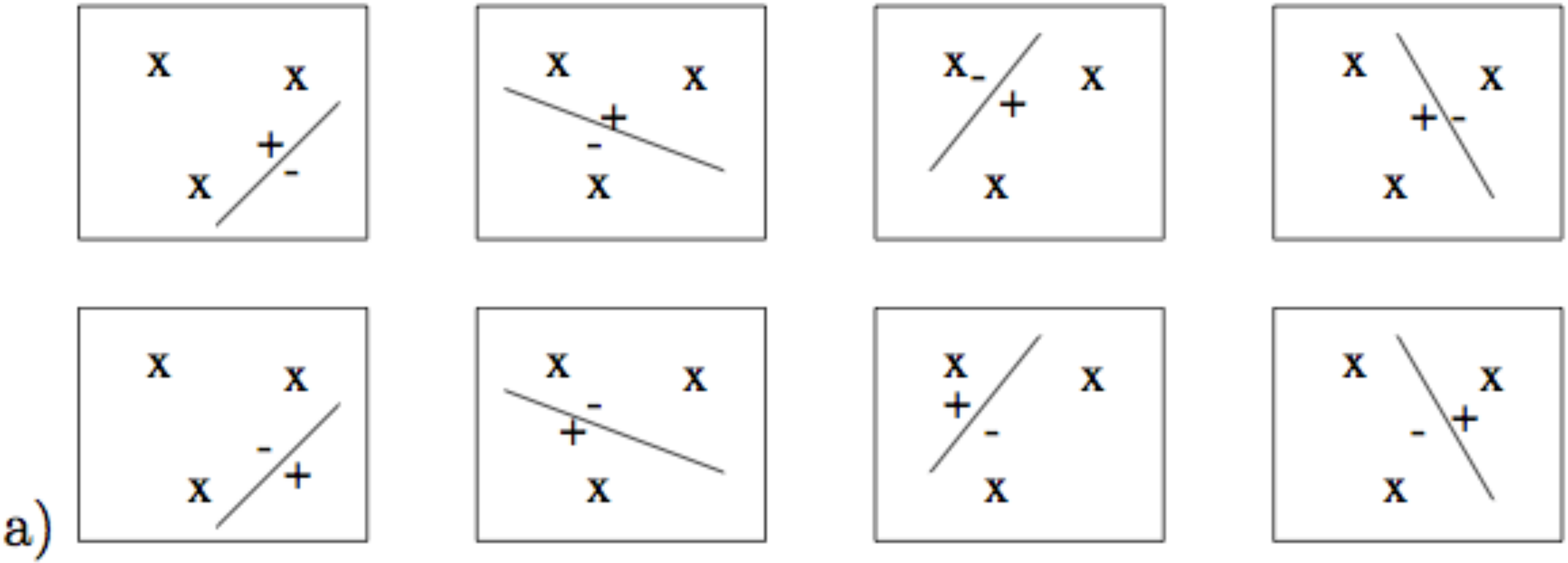
- Example: what is the VC dimension of decision stumps in R^2?

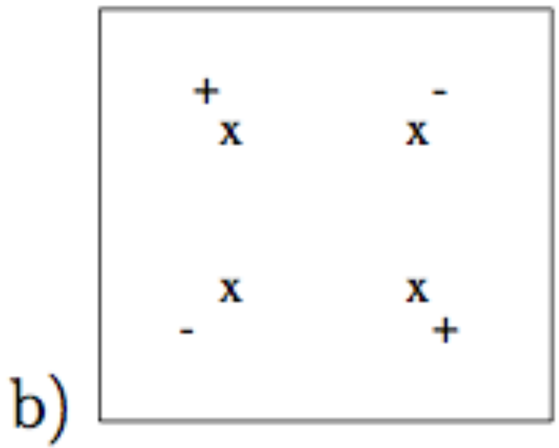
- Example: what is the VC dimension of 1-nearest-neighbor classifiers?

VC dimension

- Example: What is the VC dimension of linear classifiers in the plane (R^2)?

- Example: What is the VC dimension of linear classifiers in R^d?





For b, see proof for stumps in R^2. Same arguments hold.

Image credit: T. Jaakkola

VC dimension

- The VC dimension of Linear Classifiers in R^d is: d+1

- If the data has margin r, and B upper bounds the norm of all points, then the VC dimension of Linear Classifiers in R^d is

  $V \leq (B/r)^2$ [like perceptron mistake bound!]

- This means even if d = ∞ we can use linear separators and not suffer from high complexity, if there's a margin!

- Margin is therefore another measure of complexity.

Sauer's Lemma

Suppose you are given m (unlabeled) points.

Before we introduce Sauer's lemma, what's an upper bound on the number of possible (binary) labelings of the m points?

If we know that the VC dimension, V, of a class H is finite, then we can apply Sauer's lemma which says that the number of possible labelings of the m points that can be achieved by classifiers in H is $O(m^V)$.

As long as we have more points than the VC dimension (i.e. $m > V$), this is a much tighter bound, i.e. $O(m^V) \leq O(2^m)$

Sauer's Lemma

Formally, given a hypothesis class H define H(m) as the maximum number of ways to label any set of m points using hypotheses in H. Let V = VC-dimension(H) < ∞.

Sauer's Lemma:

$$H(m) \leq \sum_{i=1}^{V} \binom{m}{i} = O(m^V)$$

Effective size of H

Given a hypothesis class H, the size of H, |H|, is the number of classifiers in H.

This can be infinite, for example if H = {Linear Classifiers in R^d}

However, as soon as we fix a set of m (unlabeled) data points, M, the "effective" size of H becomes finite.

> Group the hypotheses into equivalence classes, where each class contains all hypotheses in H that output the same labeling on M.

A (potentially loose) upper bound on the effective size of H is:

If H has finite VC dimension, V, what's a tighter upper bound on the effective size of H?