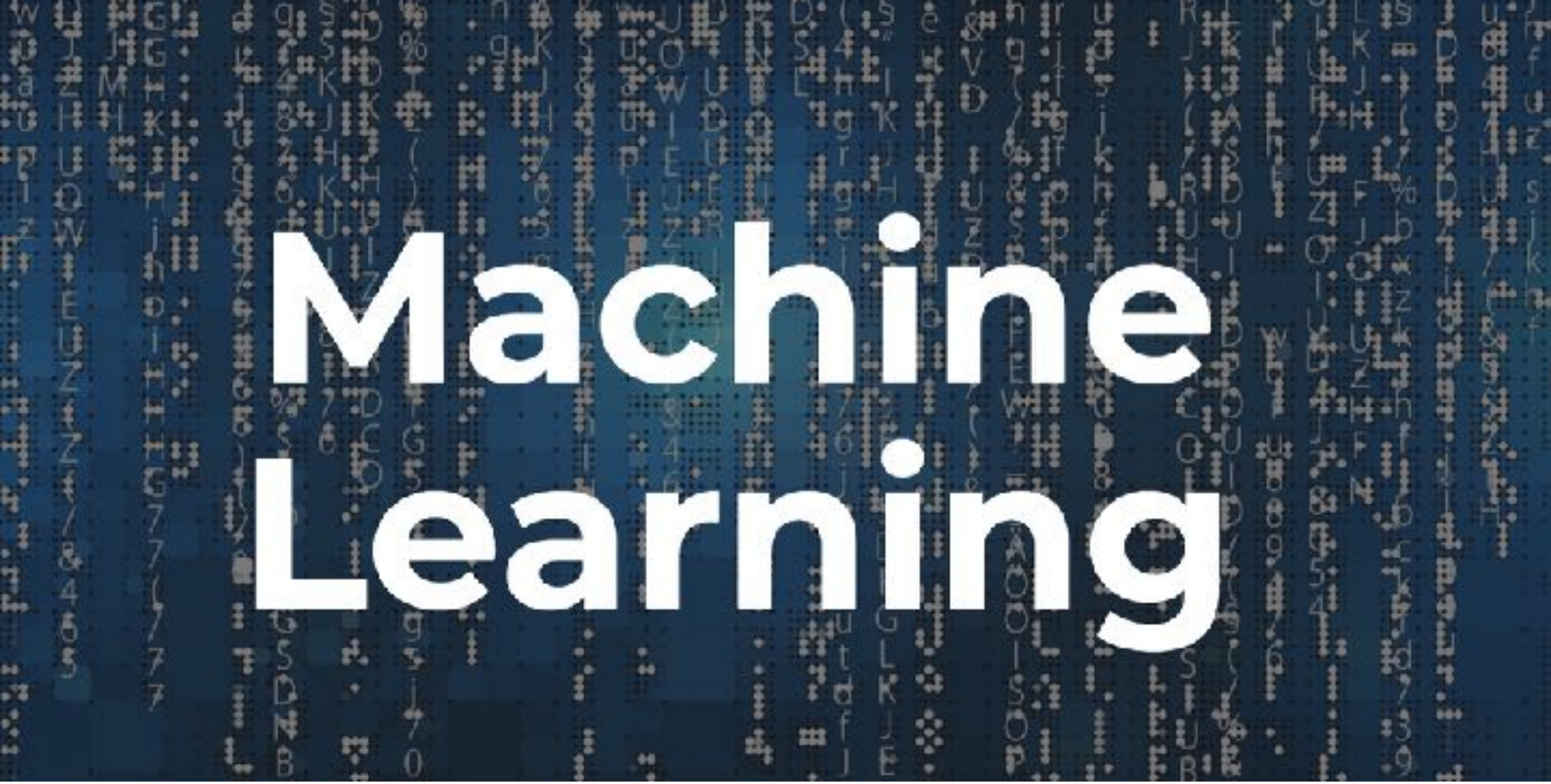


CSCI 4622 Fall 2019

Prof. Claire Monteleoni

Today: Lecture 9

- Intro. to Learning Theory
  - Complexity of classifiers
    - VC dimension (continued)
    - Margin
    - Sauer's lemma
  - Standard analysis tools in learning theory

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?

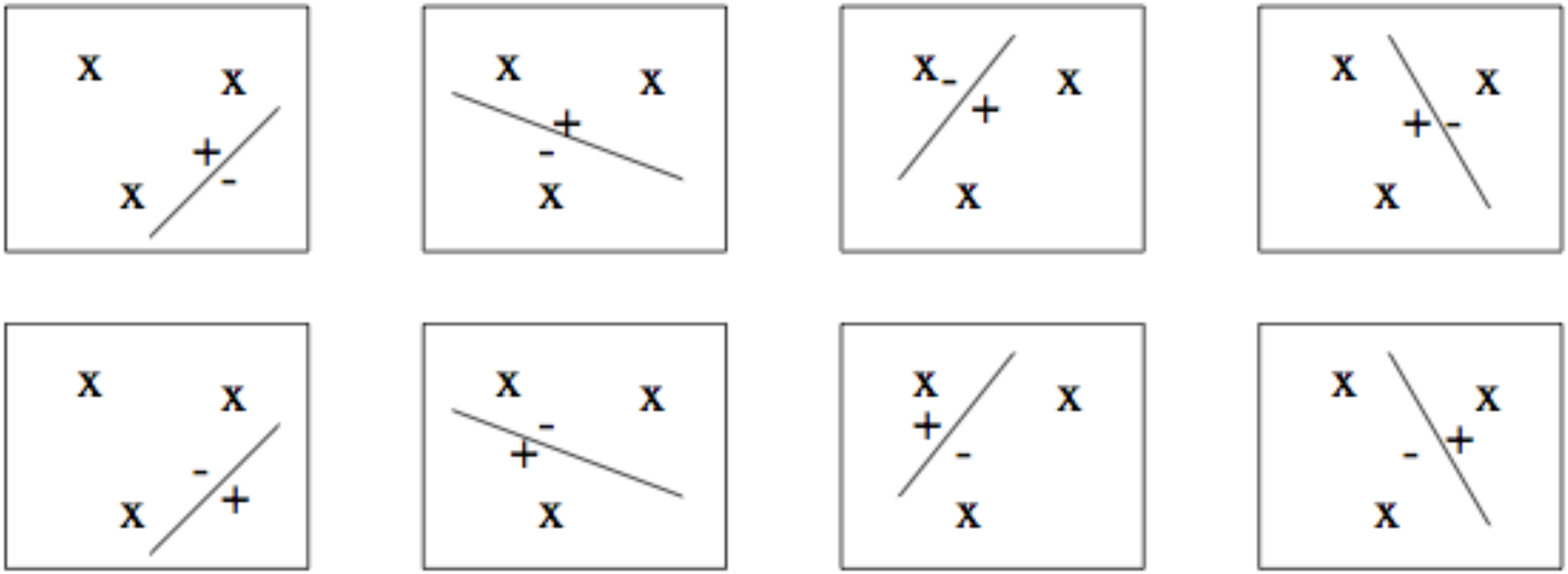
- Example: what is the VC dimension of 1-nearest-neighbor classifiers?

VC dimension

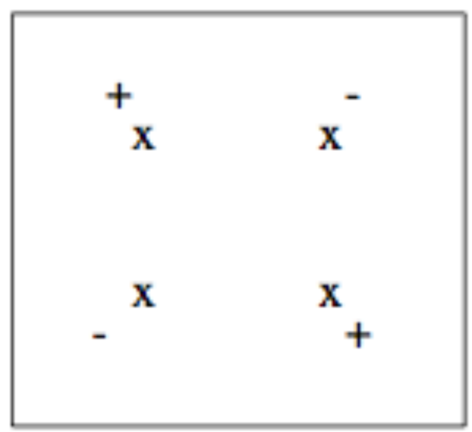
- Example: What is the VC dimension of linear classifiers in the plane (R^2)?

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

a)



b)



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

Image credit: T. Jaakkola

VC-dimension, examples

- The VC-dimension of the set of linear classifiers through origin in $\mathcal{R}^d$ is d.
- The VC-dimension of the set of linear classifiers with offset parameter in $\mathcal{R}^d$ is $d+1$.
- The VC-dimension of an ensemble with m decision stumps is at least $m/2$
- The VC-dimension of a kernel classifier with the Radial Basis kernel is ∞ !

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

  $V \leq (B/r)^2$ [No dependence on d!]

  [Like perceptron mistake bound!]

- This means even if $d = \infty$ 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?

Bounding Random Variables

- Markov Inequality
- Chebyshev Inequality

Use the mean, and possibly the variance, to provide bounds on the probabilities of certain events.

Markov Inequality

If X is a random variable that can only take nonnegative values, then:

$$\boxed{P(X \geq a) \leq \frac{E[X]}{a}}$$

Interpretation: If a nonnegative random variable has a small mean, then the probability it takes a large value is small.

Chebyshev Inequality

If X is a random variable with mean μ and var σ^2, then for all $c > 0$:

$$\boxed{P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}}$$

Interpretation: If a random variable has small variance, then the probability it takes a value far from its mean is small.

Probability amplification

- We saw methods to take a statement that holds in expectation and get a statement that holds with high probability: i.e., with probability $\geq 1 - \delta$.
- There are a variety of inequalities for doing so, known as concentration inequalities, e.g.
  - Markov Inequality
  - Chebyshev Inequality
  - Chernoff bounds, etc.
- This is an example of probability amplification. More generally, there are techniques in which the original statement just needs to holds with constant probability.

The Union Bound

Given any probability space, and any events $A_1, \ldots, A_n$ defined on the space:

$$\boxed{P(A_1 \cup A_n \cup \cdots \cup A_n) = P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)}$$

Interpretation: a (possibly loose) upper bound on the probability of **any** of the events, A_i, occurring.

- Often used to bound the probability of any “bad” event occurring.
- The events, A_i, need not be independent!

PAC Learning

- Probably Approximately Correct (PAC) Learning
- If a Hypothesis Class (a.k.a. Concept Class) has finite VC dimension, then when learning from labeled, i.i.d. data, there are performance guarantees for PAC learning
  - [On whiteboard]