Learning theory
Lecture 10

David Sontag
New York University

Slides adapted from Carlos Guestrin & Luke Zettlemoyer

What about continuous hypothesis spaces?

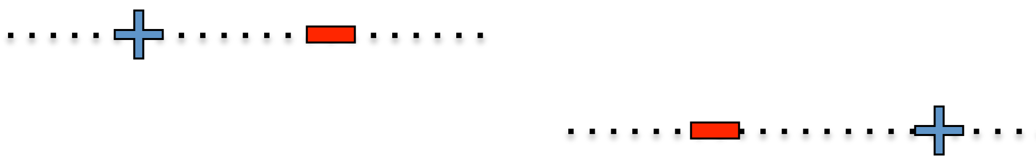
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}$$

- Continuous hypothesis space:
  - $|H| = \infty$
  - Infinite variance???

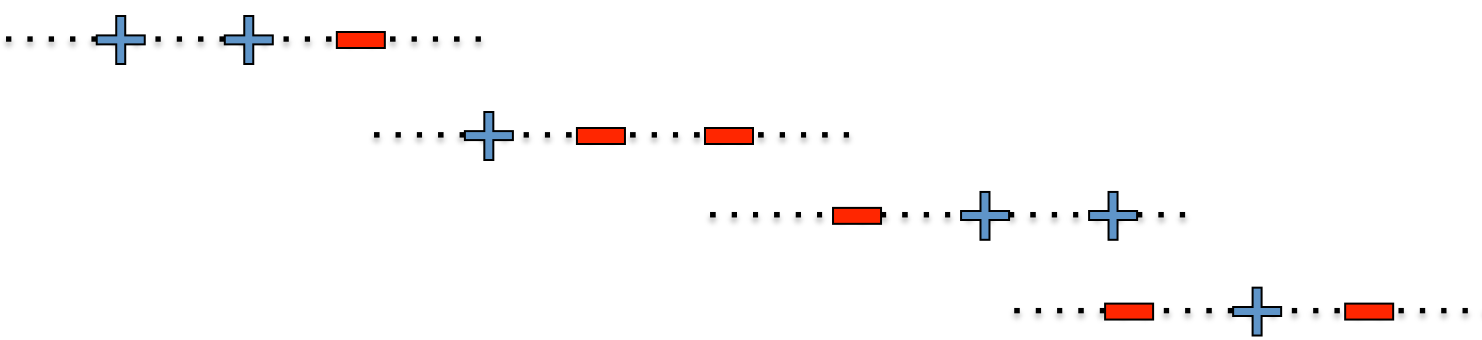
- **Only care about the maximum number of points that can be classified exactly!**

How many points can a linear boundary classify exactly? (1-D)

2 Points: Yes!!



3 Points: No…



etc (8 total)

Shattering and Vapnik–Chervonenkis Dimension

A **set of points** is *shattered* by a hypothesis space H iff:

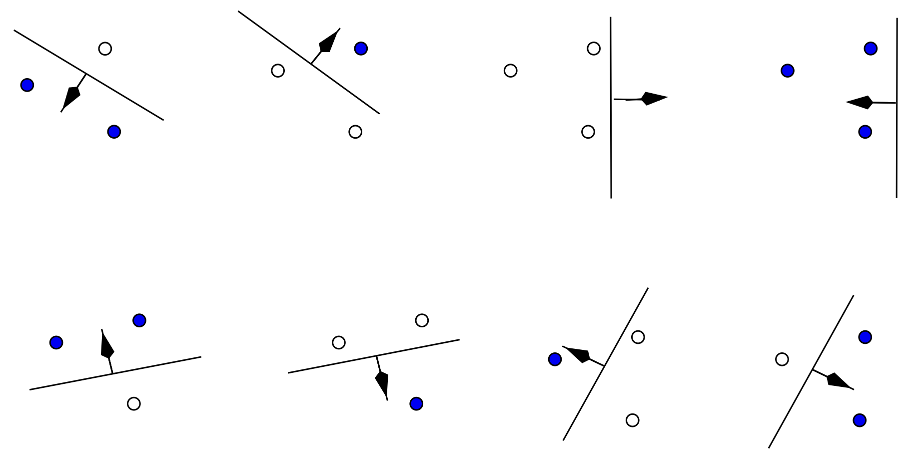
- For all ways of *splitting* the examples into positive and negative subsets
- There exists some *consistent* hypothesis h

The *VC Dimension* of H over input space X

- The size of the *largest* finite subset of X shattered by H

How many points can a linear boundary classify exactly? (2-D)

3 Points: Yes!!



4 Points: No…

etc.

[Figure from Chris Burges]

How many points can a linear boundary classify exactly? (d-D)

- A linear classifier $\sum_{j=1..d} w_j x_j + b$ can represent all assignments of possible labels to d+1 points
  - But not d+2!!
  - Thus, VC-dimension of d-dimensional linear classifiers is d+1
  - Bias term b required
  - Rule of Thumb: number of parameters in model often matches max number of points

- Question: Can we get a bound for error as a function of the number of points that can be completely labeled?

PAC bound using VC dimension

- VC dimension: number of training points that can be classified exactly (shattered) by hypothesis space H!!!
  – Measures relevant size of hypothesis space

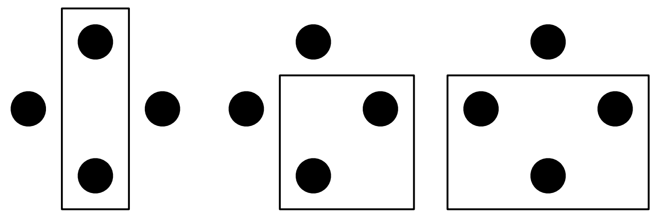
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln \frac{2m}{VC(H)} + 1\right) + \ln \frac{4}{\delta}}{m}}$$

- Same bias / variance tradeoff as always
  – Now, just a function of VC(H)

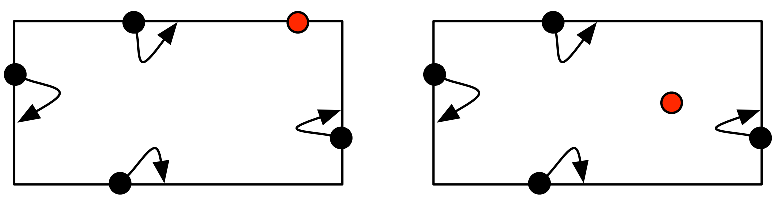
- Note: all of this theory is for **binary** classification
  – Can be generalized to multi-class and also regression

What is the VC-dimension of rectangle classifiers?

- First, show that there are 4 points that *can* be shattered:



- Then, show that no set of 5 points can be shattered:



[Figures from Anand Bhaskar, Ilya Sukhar]

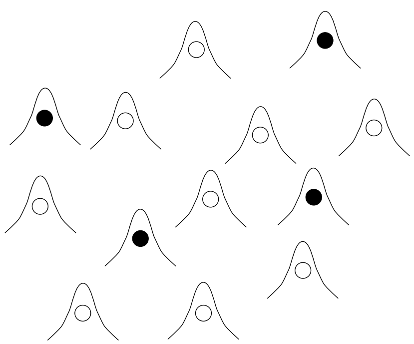
Generalization bounds using VC dimension

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln \frac{2m}{VC(H)} + 1\right) + \ln \frac{4}{\delta}}{m}}$$

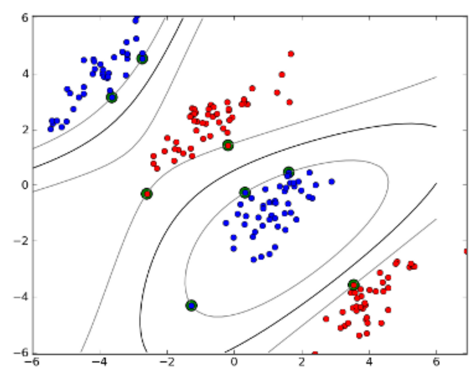
- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term *b*
- Classifiers using Gaussian Kernel
  - VC(H) = ∞

$$K(\vec{u}, \vec{v}) = \exp\left(-\frac{||\vec{u} - \vec{v}||_2^2}{2\sigma^2}\right)$$

Euclidean distance, squared



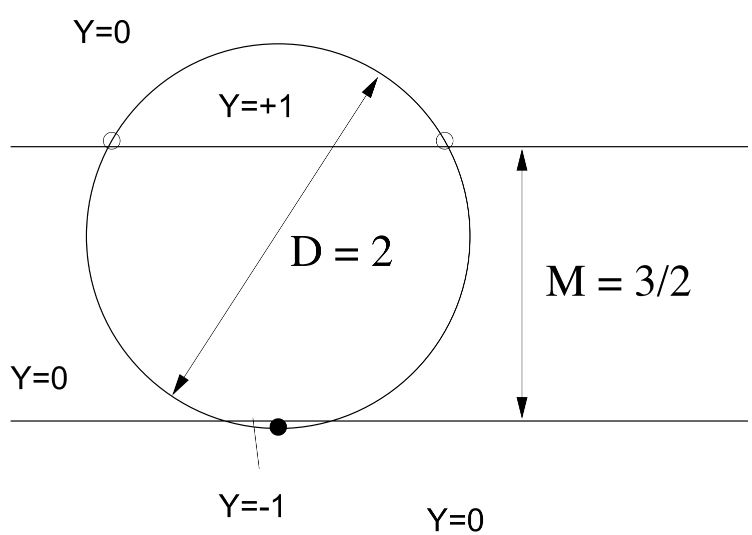
[Figure from Chris Burges]



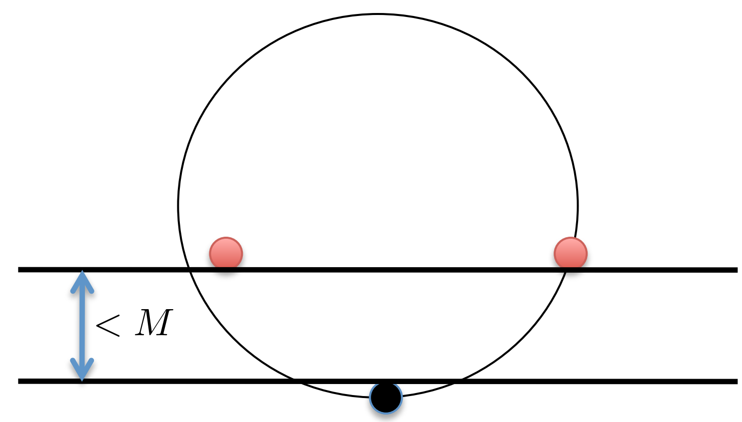
[Figure from mblondel.org]

Gap tolerant classifiers

- Suppose data lies in R^d in a ball of diameter **D**
- Consider a hypothesis class H of linear classifiers that can only classify point sets with margin at least **M**
- What is the largest set of points that H can shatter?



Cannot shatter these points:



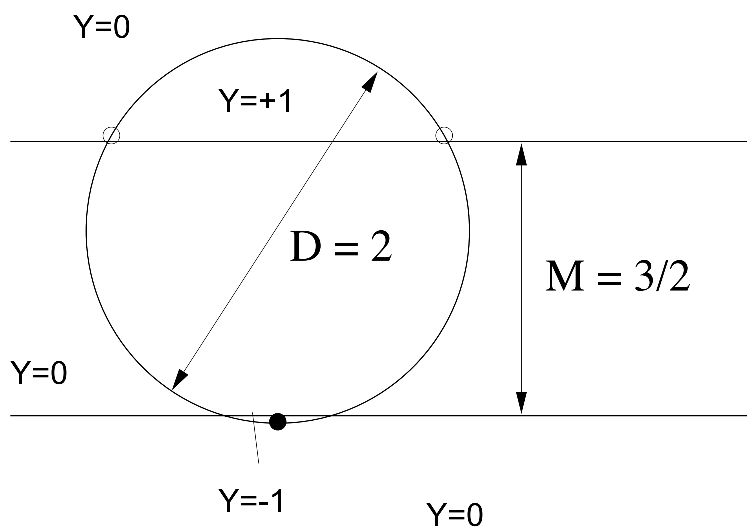
VC dimension = $\min\left(d, \frac{D^2}{M^2}\right)$

$M = 2\gamma = 2\frac{1}{||w||}$ → **SVM attempts to *minimize* $||w||^2$, which *minimizes* VC-dimension!!!**

[Figure from Chris Burges]

Gap tolerant classifiers

- Suppose data lies in R^d in a ball of diameter **D**
- Consider a hypothesis class H of linear classifiers that can only classify point sets with margin at least **M**
- What is the largest set of points that H can shatter?



VC dimension = $\min\left(d, \frac{D^2}{M^2}\right)$

[Figure from Chris Burges]

$$K(\vec{u},\vec{v}) = \exp\left(-\frac{||\vec{u}-\vec{v}||_2^2}{2\sigma^2}\right)$$

What is R=D/2 for the Gaussian kernel?

$$\begin{aligned} R &= \max_x ||\phi(x)|| \\ &= \max_x \sqrt{\phi(x)\cdot\phi(x)} \\ &= \max_x \sqrt{K(x,x)} \\ &= 1 \quad \text{!!!} \end{aligned}$$

What is $||w||^2$?

$$||w||^2 = \left(\frac{2}{M}\right)^2$$

$$\begin{aligned} ||w||^2 &= ||\sum_i \alpha_i y_i \phi(x_i)||_2^2 \\ &= \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \end{aligned}$$

What you need to know

- Finite hypothesis space
  - Derive results
  - Counting number of hypothesis
- Complexity of the classifier depends on number of points that can be classified exactly
  - Finite case – number of hypotheses considered
  - Infinite case – VC dimension
  - VC dimension of gap tolerant classifiers to justify SVM
- Bias-Variance tradeoff in learning theory