UNIVERSITÀ PARTHENOPE
DiST
Dipartimento di Scienze e Tecnologie

Machine Learning (part II)

PAC learning model

Angelo Ciaramella

Introduction

- The probably approximately correct (PAC) learning model
  - A formal, mathematical model of learnability
  - It origins from a paper by Valiant ("A theory of the learnable", 1984)
  - It is very theoretical
  - Has only very few results that are usable in practice

Formalization

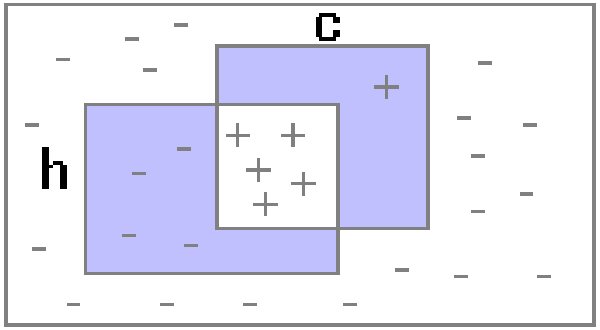
- It formalizes the concept learning task as follows
  - X
    - the space from which of the objects (*examples*) come
    - e.g., if the objects are described by two continuous features, then $X=\mathbf{R}^2$
  - c is a concept $c \subseteq X$
    - but c can also be interpreted as an $X \rightarrow \{0, 1\}$ mapping (each point of X either belongs to c or not)
  - C concept class
    - a set of concepts
    - In the following we always assume that the c concept we have to learn is a member of C

PAC learnability

- L denotes the learning algorithm
  - goal
    - to learn a given c
  - output
    - hypothesis *h* from concept class C
  - help
    - it has access to a function Ex(c,D)
      - training examples in the form of <x,c(x)> pairs
      - The examples are random, independent, and follow a fixed (but arbitrary) probability distribution D

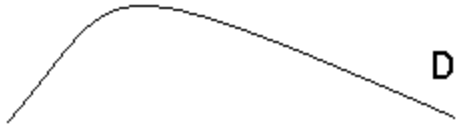
PAC learnability

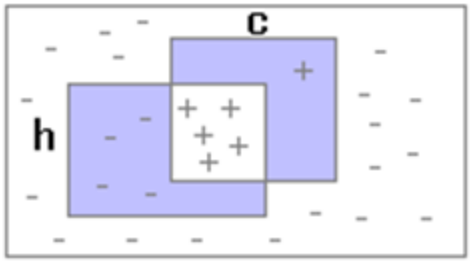
- Example
  - $X = R^2$
    - the 2D plane
  - C
    - all such rectangles of the plain that are parallel with the axes
  - c
    - one fixed rectangle
  - h
    - another given rectangle
  - D
    - a probability distribution defined over the plain
  - Ex(c, D)
    - It gives positive and negative examples from c



Error

- After learning, how can we measure the final error of h?
  - error(h)=cΔh=(c\h)$\cup$(h\c)
  - the symmetric difference of c and h – the size of the blue area in the figure





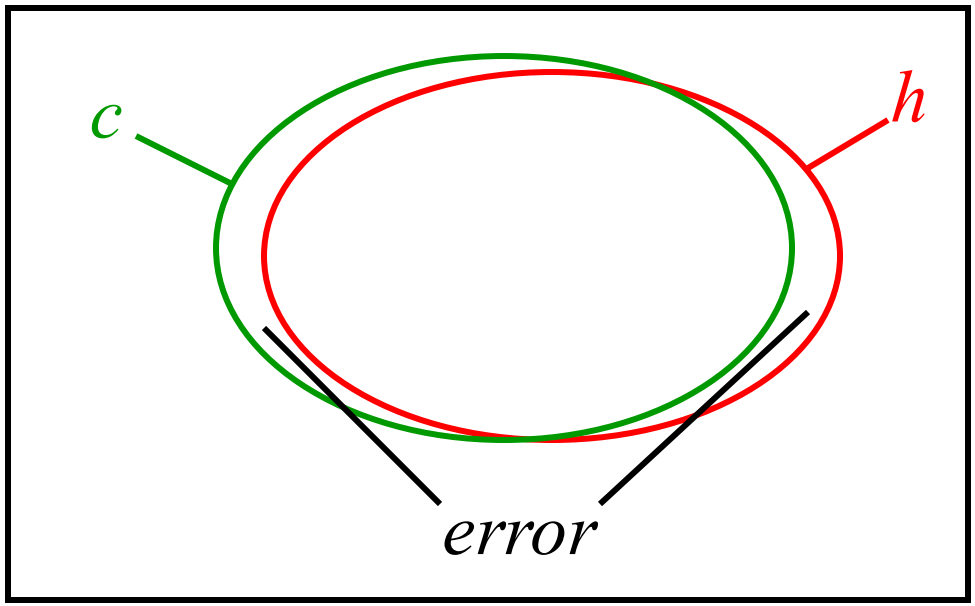
- It is not good - why?
  - D is the probability of picking a given point of the plain randomly
  - we would like our method to work for any possible distribution D
  - If D is 0 in a given area, we won't get samples from there
    - we cannot learn in this area
  - So we cannot guarantee error(h)=cΔh to become 0 for any arbitrary D

- D is the same during testing
  - we won't get samples from that area during testing either
  - no problem if we couldn't learn there!

True error of a Hypothesis

- *True error* of a hypothesis *h* with respect to target concept *c* and distribution *D* is the probability that *h* will misclassify an instance drawn at random according to *D*

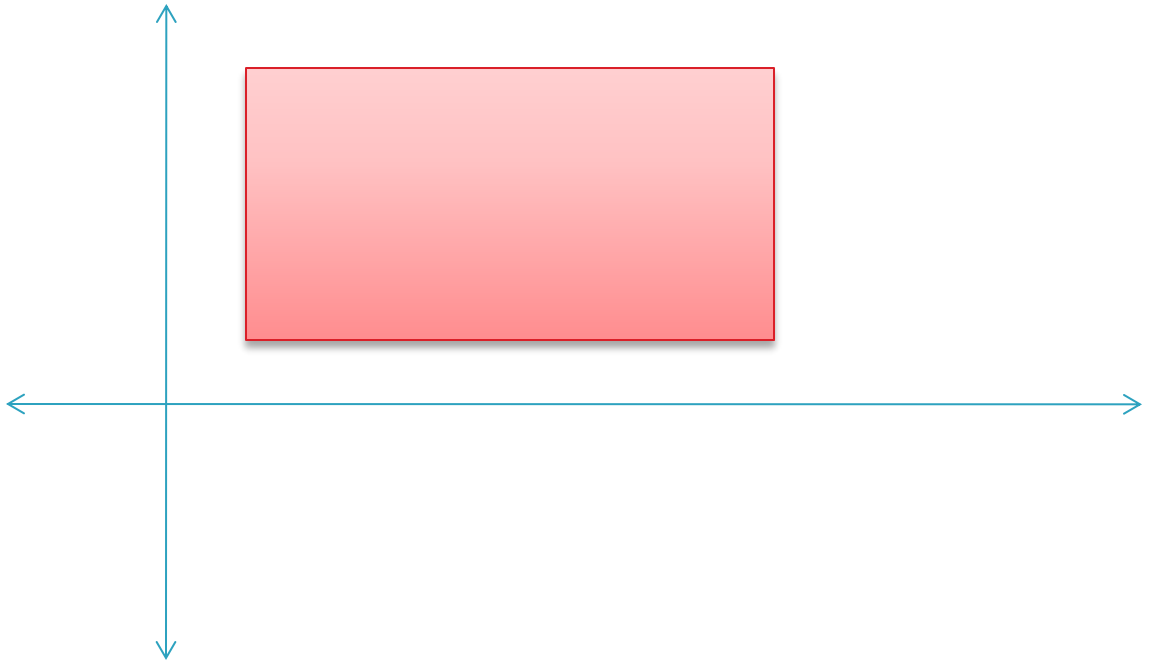
$$error_D(h) \equiv \Pr_{x \in D}[c(x) \neq h(x)]$$



Instance Space X

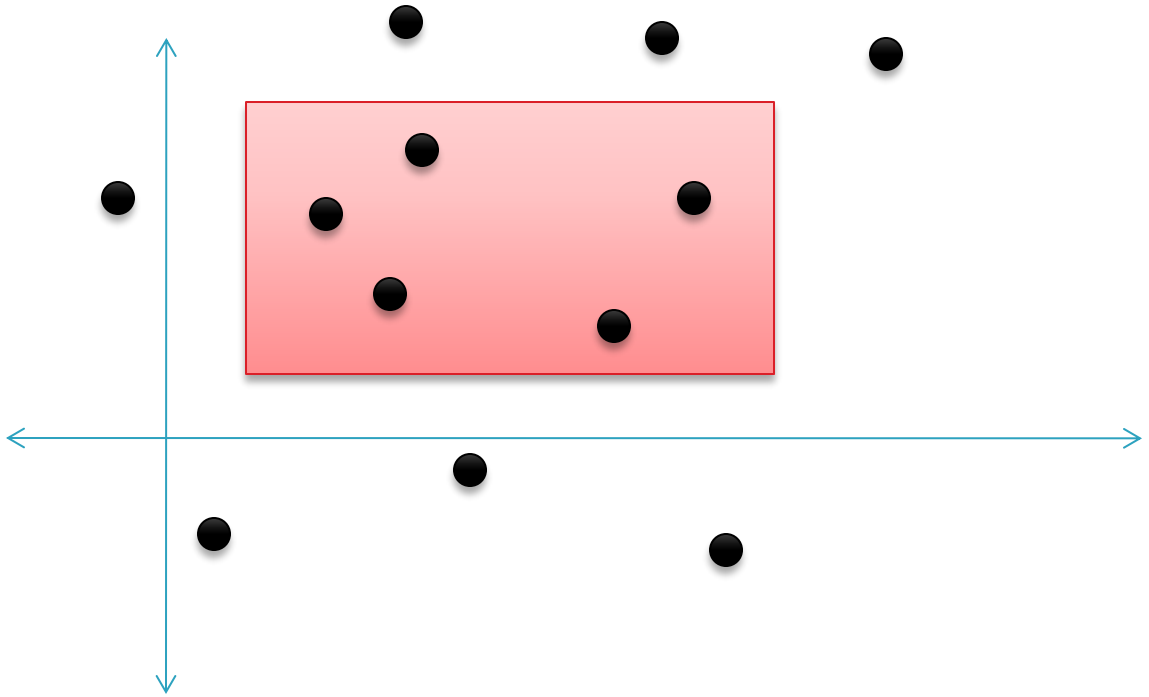
Rectangle game

- Fix a rectangle (unknown to you)



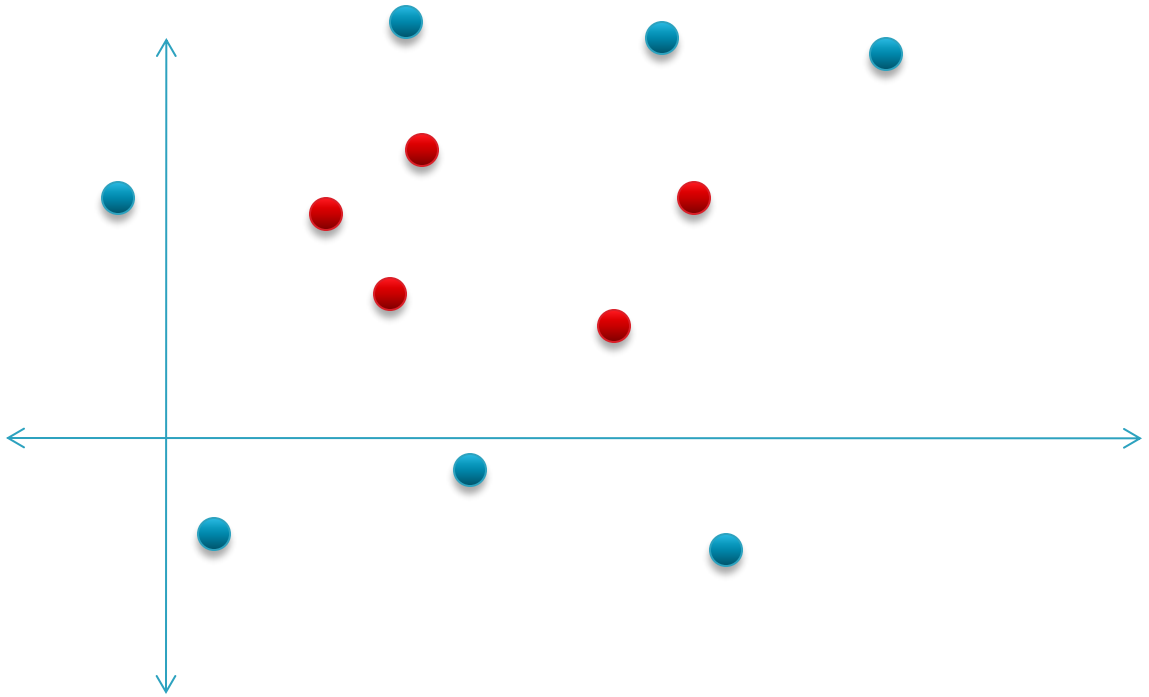
Rectangle game

- Draw points from some fixed unknown distribution



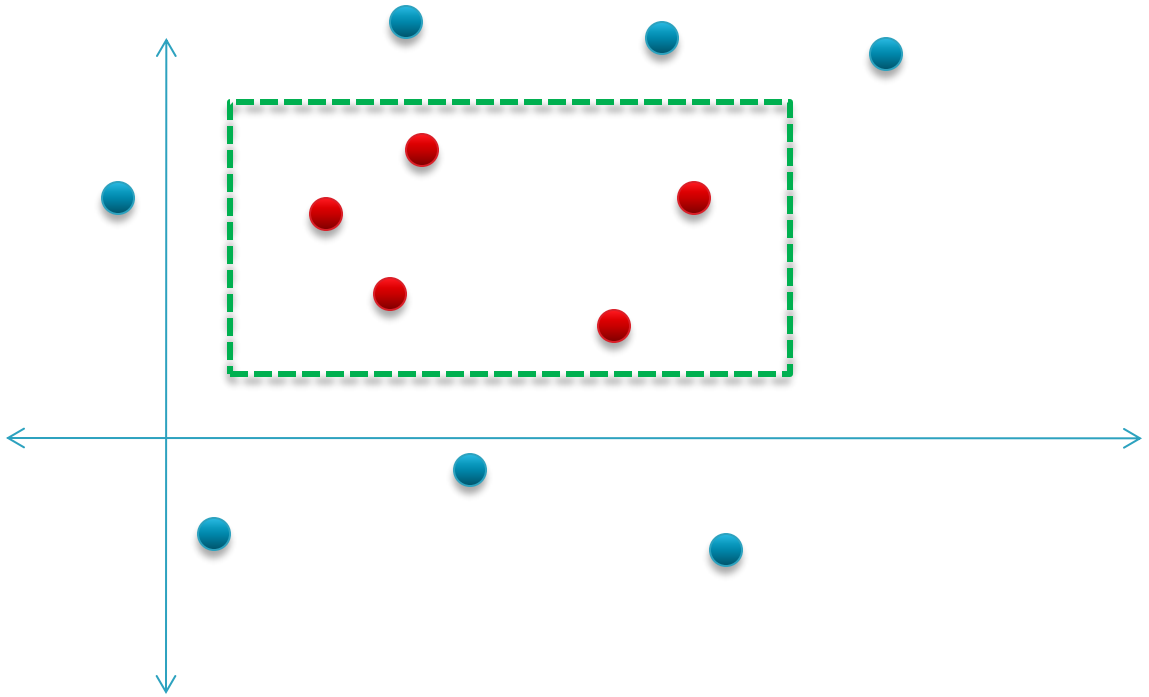
Rectangle game

- You are told the points and whether they are in or out



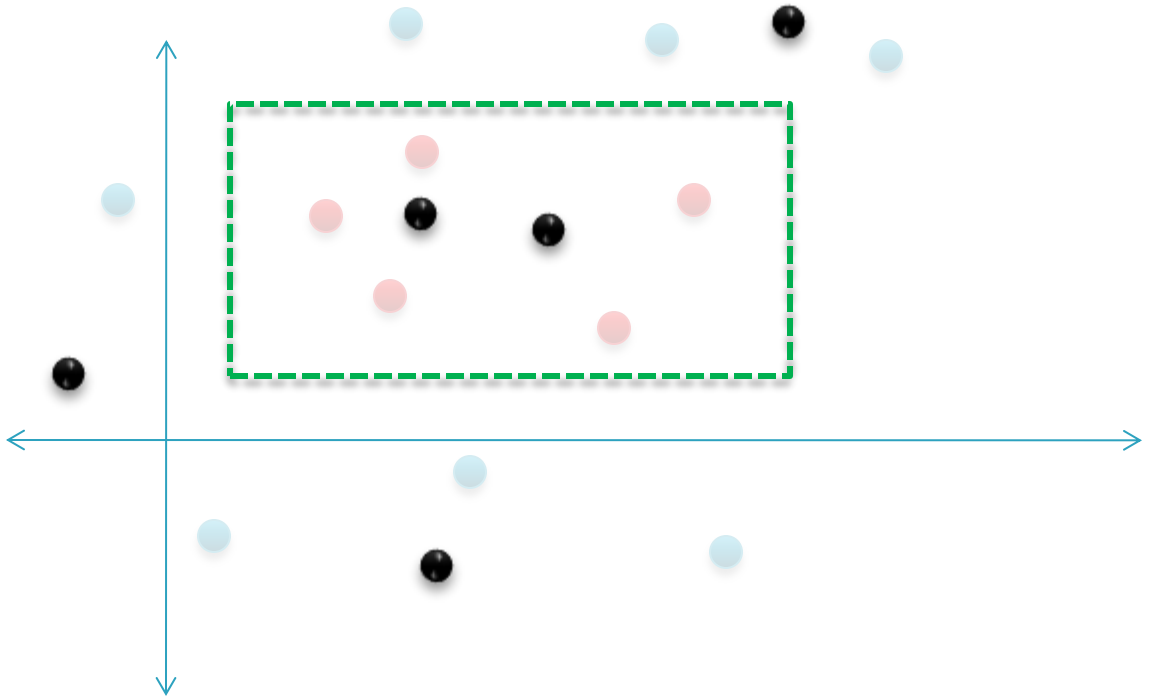
Rectangle game

- You propose a hypothesis



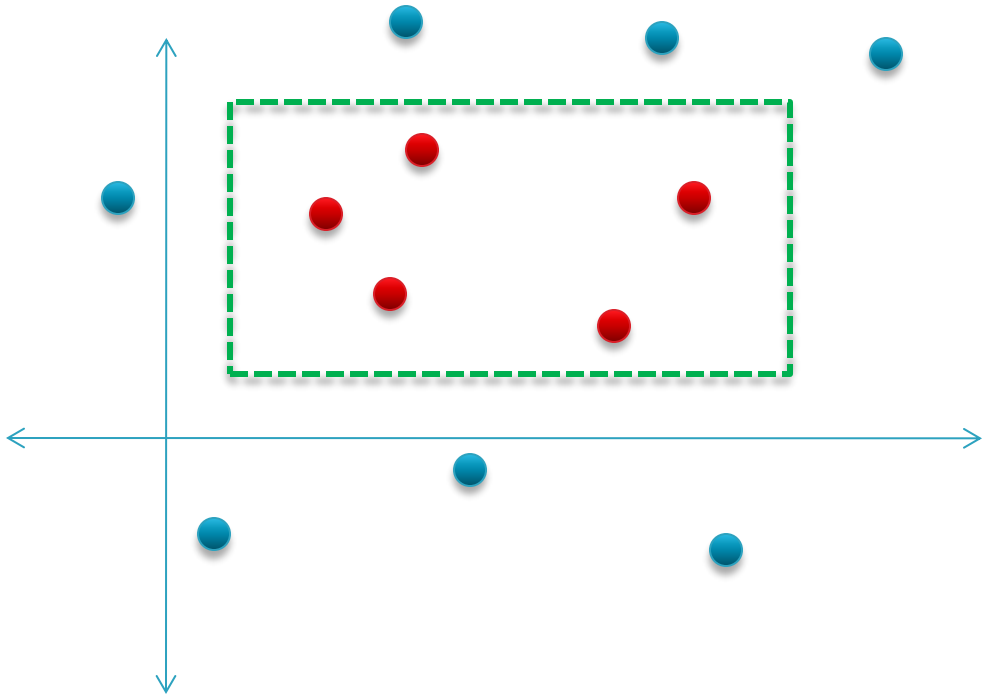
Rectangle game

- Your hypothesis is tested on points drawn from the same distribution



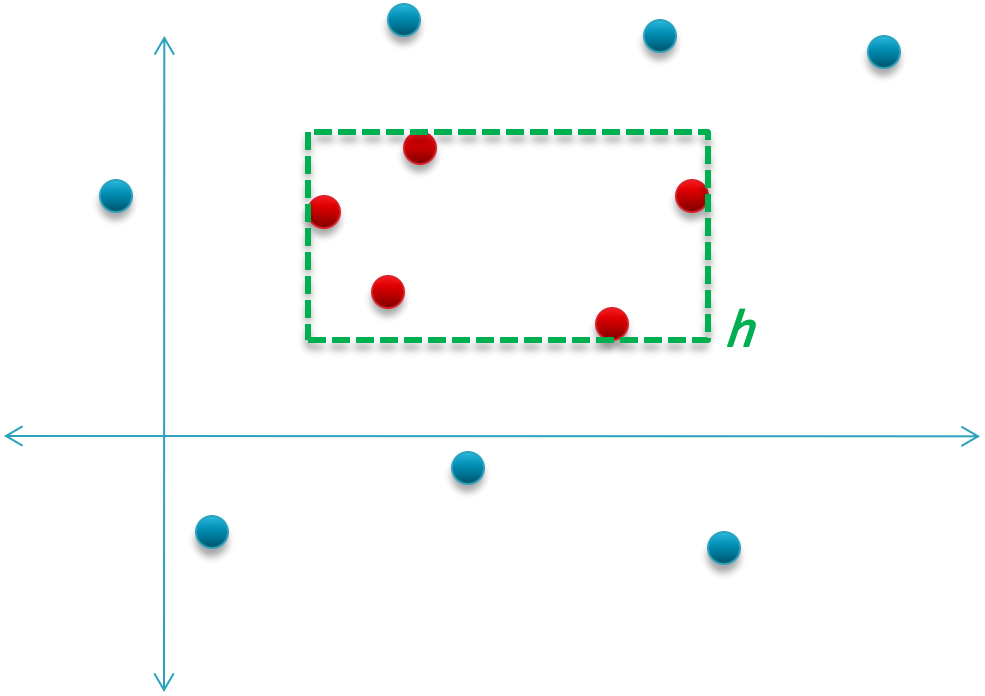
Rectangle game

- We want an algorithm that
  - with high probability will choose a hypothesis that is approximately correct



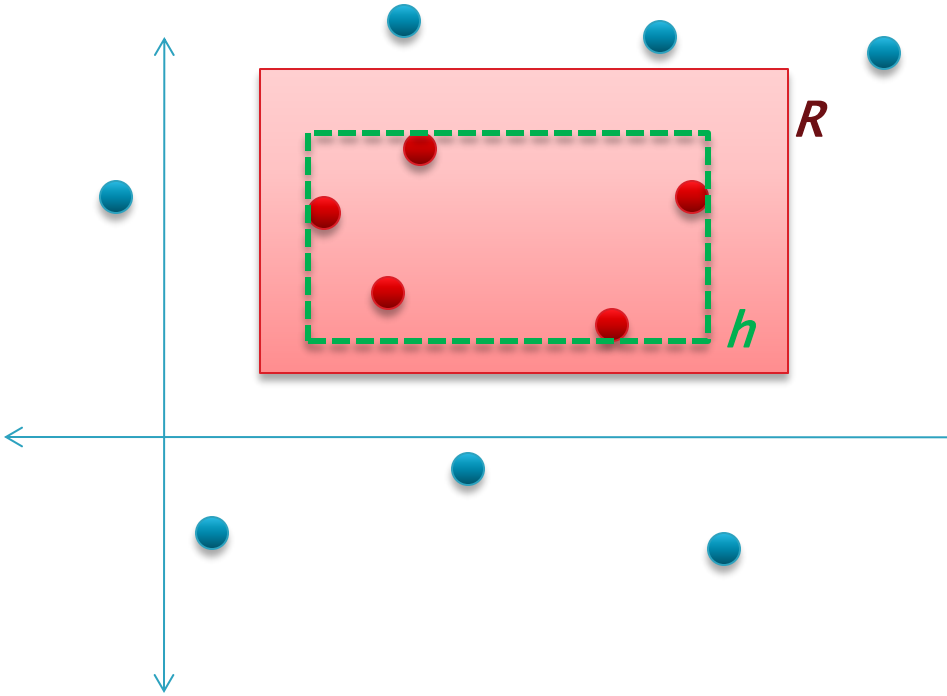
Rectangle game

- Choose the minimum area rectangle containing all the positive points



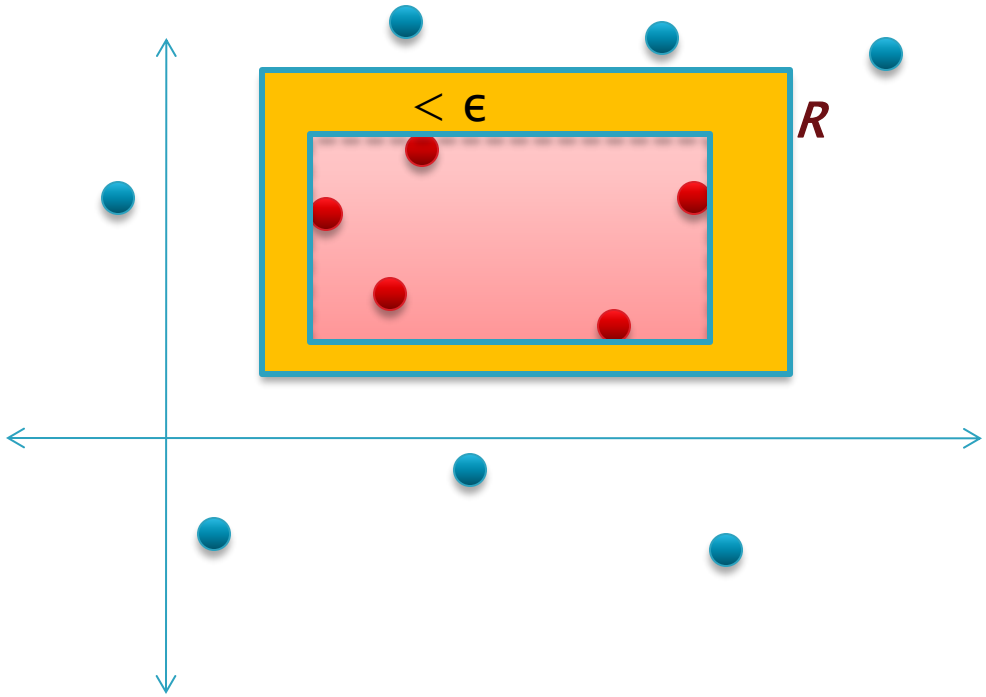
How good is this?

- Derive a PAC bound
- For fixed
  - R : Rectangle
  - D : Data Distribution
  - ε : Test Error
  - δ : Probability of failing
  - m : Number of Samples



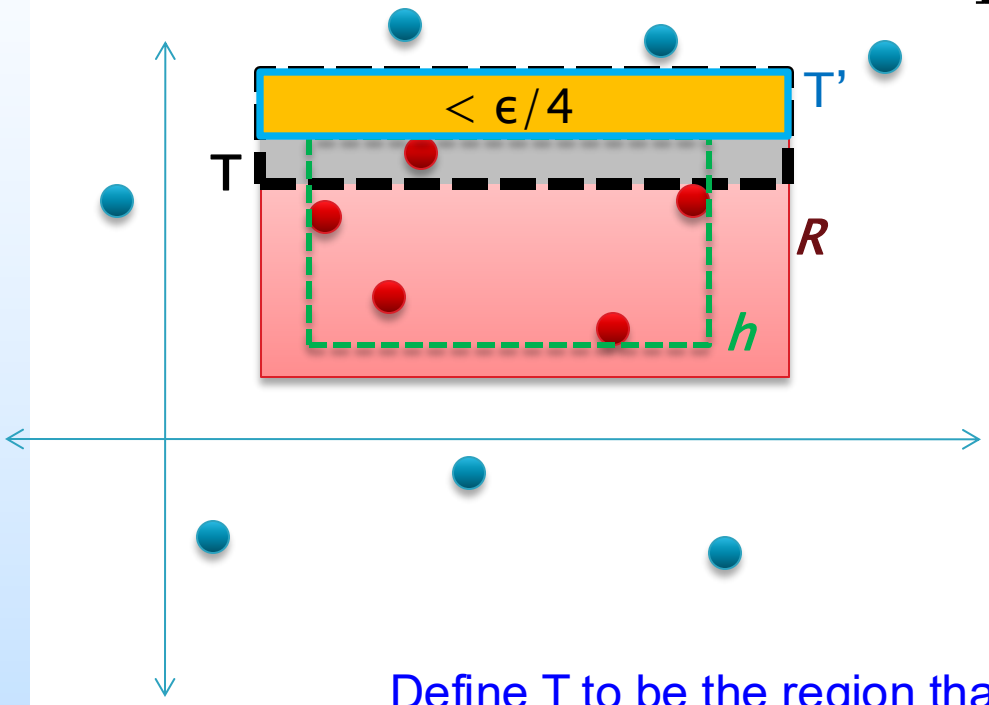
$$\mathbf{P}\left(\text{error}_{\text{test}}(h) \leq \epsilon\right) \leq 1 - \delta$$

Proof



We want to show that with high probability the area below measured with respect to *D* is bounded by ε

Proof



$$\mathbf{P}\left(m \text{ samples miss } T\right) = \left(1 - \frac{\epsilon}{4}\right)^m$$

probability that all *m* samples miss T

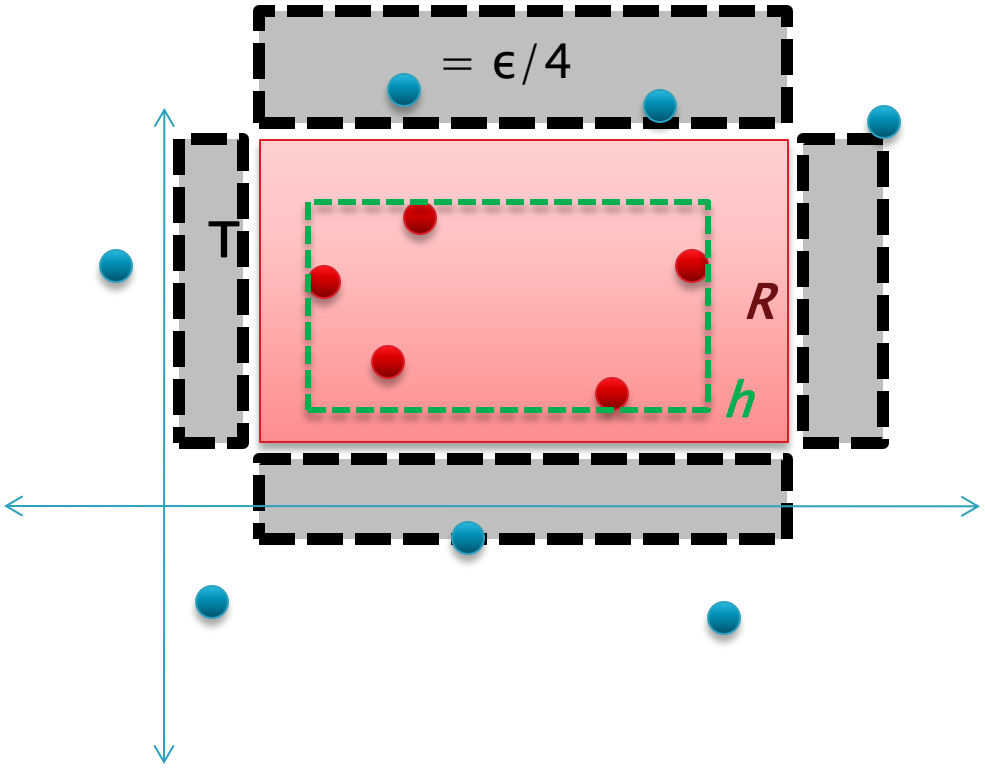
Define T to be the region that contains exactly ε/4 of the mass in D sweeping down from the top of R

p(T') > ε/4 = p(T) IFF T' contains T

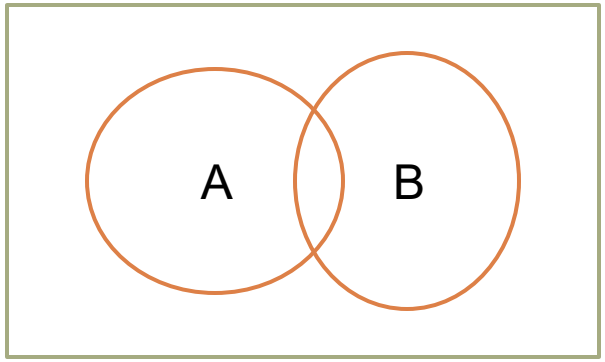
T' contains T IFF none of our *m* samples are from T

What is the probability that all samples miss T

Proof



$$\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$$



What is the probability that we miss any of the rectangles:

Union Bound

$$\mathbf{P}(m \text{ samples miss any } T) \leq 4\left(1 - \frac{\epsilon}{4}\right)^m$$

Proof

- Probability that any region has weight greater than $\varepsilon/4$ after m samples is at most

$$4\left(1-\frac{\epsilon}{4}\right)^m$$

- If we fix *m* such that

$$4\left(1-\frac{\epsilon}{4}\right)^m \leq \delta$$

- Than with probability 1- δ we achieve an error rate of at most ε

Inequality

- Common Inequality

$$1 - x \leq e^{-x}$$

- We can show

$$4\left(1 - \frac{\epsilon}{4}\right)^m \quad \leq \quad 4e^{-m\epsilon/4}$$

- Obtain a lower bound on the samples

$$m \quad \geq \quad \frac{4}{\epsilon} \ln\left(\frac{4}{\delta}\right)$$

PAC learning

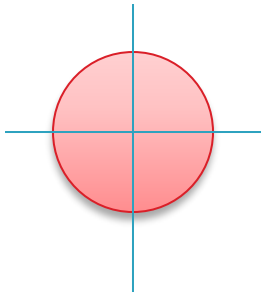
- Consider a class C of possible target concepts defined over a set of instances *X* of length *n*, and a learner *L* using hypothesis space *H*

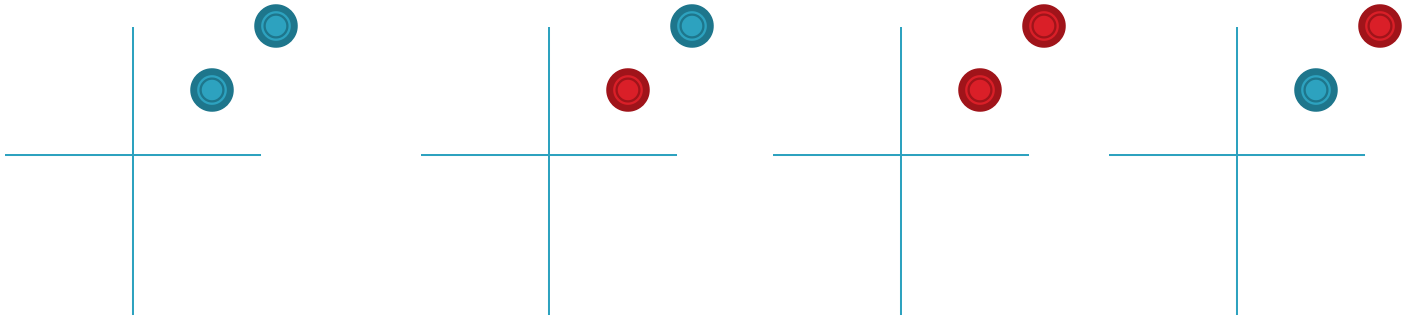
- Definition
  - C is PAC-learnable by *L* using *H* if for all $c \in C$, distributions *D* over *X*, ε such that $0 < \varepsilon < 1/2$, and δ such that $0 < \delta < 1/2$, learner L will with prob. at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $error_D(h) \leq \varepsilon$, in time that is polynomial in $1/\varepsilon$, $1/\delta$, *n* and *size*(c).

Vapnik-Chervonenkis dimenison

- Provides a measure of the complexity of a "hypothesis space" or the "power" of "learning machine"
- Higher VC dimension implies the ability to represent more complex functions
- The VC dimension is the maximum number of points that can be arranged so that **f** shatters them
- What does it mean to shatter?

VC dimenison





A classifier **f** can shatter a set of points if and only if for all truth assignments to those points **f** gets zero training error

example: f(x,b) = sign(x·x - b)

H not finite

What if |H| can not be determined?

- It is still possible to come up with estimates based not on counting how many hypotheses, but based on how many instances can be completely discriminated by H

- Use the notion of a shattering of a set of instances to measure the complexity of a hypothesis space

- VC Dimension measures this notion and can be used as a stand in for |H|

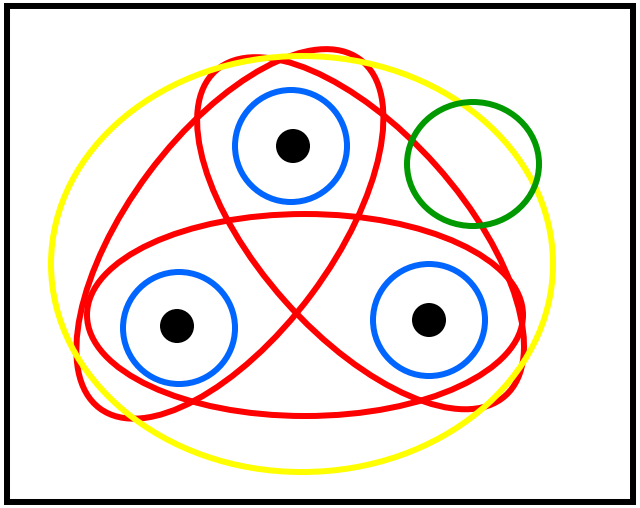
H not finite

- Definition
  - a dichotomy of a set S is a partition of S into two disjoint subsets

- Definition
  - a set of instances S is shattered by hypothesis space *H* iff for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy

H not finite

Example:
3 instances
shattered

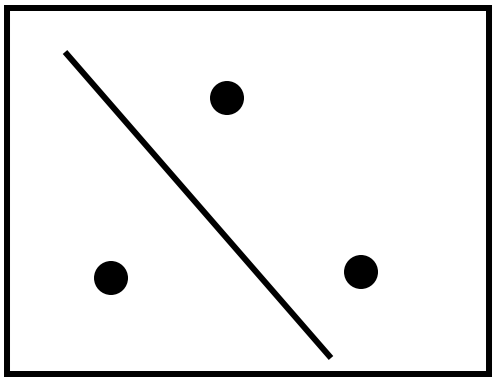


Instance space X

VC dimension

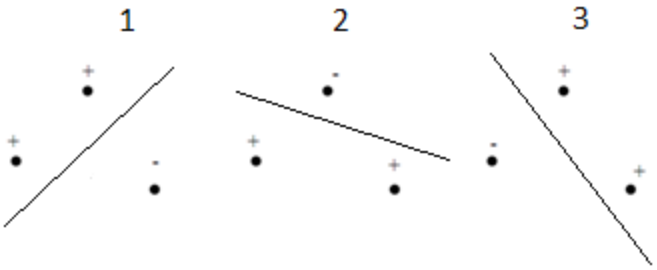
- Definition
  - the Vapnik-Chervonenkis (VC) dimension, VC(H), of hypothesis space H defined over *instance space X* is the size of the largest finite subset of X shattered by *H*. If arbitrarily large finite sets of X can be shattered by H, then VC(H) = ∞

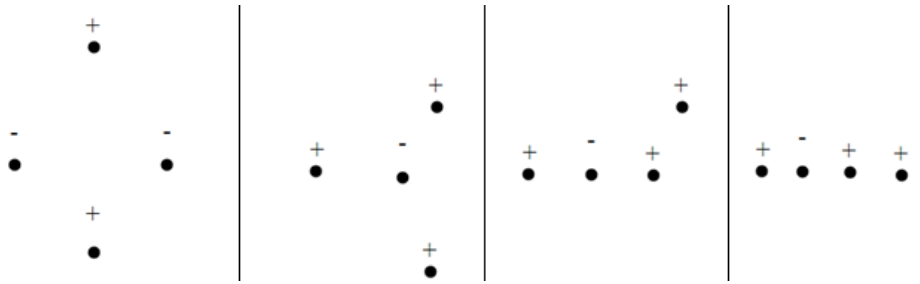
- Example
  - VC dimension of linear decision surfaces is 3



VC dimension

- Separating the two classes by lines on the plane VCD=3
  - in d-dimensional space VCD=d+1
    - VCD ≥3, as these 3 points can be shattered
      (all labeling configurations should be tried!)
    - VCD<4, as no 4 points can be shattered:
      (all point arrangements should be tried!)

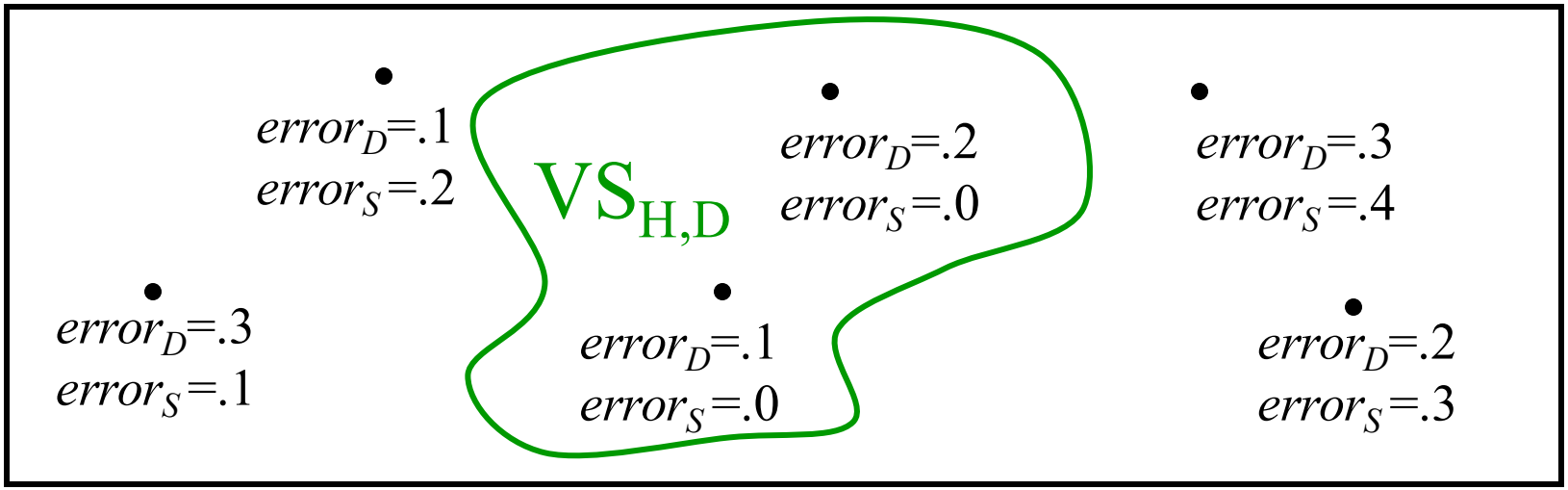




Exhausting the Version Space

- Definition
  - the version space $VS_{H,D}$ is said to be *ε-exhausted* with respect to c and D, if every hypothesis h in $VS_{H,D}$ has error less than ε with respect to c and D

$$(\forall h \in VS_{H,D})\, error_D(h) < \varepsilon$$



Sample Complexity with VC Dimension

- How many randomly drawn examples suffice to ε - exhaust $VS_{H,D}$ with probability at least $(1 - \delta)$?

$$m \geq \frac{1}{\varepsilon}\left(4\log_2\left(\frac{2}{\delta}\right) + 8\,VC(H)\log_2\left(\frac{13}{\varepsilon}\right)\right)$$

VC dimension

- Goal
  - Worst-case performance for a particular trained network
  - Binary ouputs
- input vectors
  - generated from some probability distribution $P(\mathbf{x})$
- target data
  - generated by a noisless function $h(\mathbf{x})$
- Model
  - $y(\mathbf{x})$
  - average generalization ability $g(y)$ to be the probability that $y(\mathbf{x}) = h(\mathbf{x})$

VC dimension

- Problem
  - we cannot calculate g(y) directly because we do not know P(x) and h(x)
- Finite data set
  - N samples
  - $g_N(y)$ is the measure of the fraction of the training set which the network y(**x** ; **w**) correctly classifies (estimation)
  - $g_N(y) \rightarrow g(y)$ for $N \rightarrow \infty$
- Maximum of discrepancy
  - Set of all function $\{y\}$

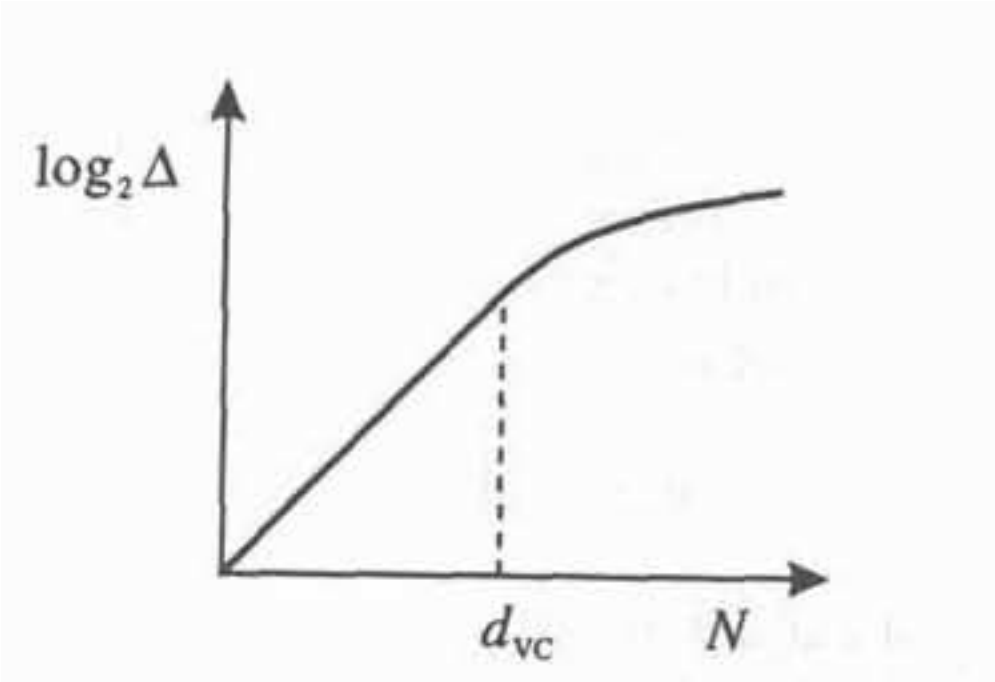
$$\max_{\{y\}} |g_N(y) - g(y)|$$

Vapnik-Chervonenkis dimension

- Goal
  - Worst-case performance for a particular trained network
  - Binary ouputs
- Theorem

$$\Pr\left(\max_{\{y\}} |g_N(y) - g(y)| > \epsilon\right) \leq 4\Delta(2N)\exp(-\epsilon^2 N/8)$$

Vapnik-Chervonenkis dimension



$$\Delta(N) \leq N^{dvc} + 1$$

Vapnik-Chervonenkis dimension

- NN
  - M units, W weights

$$d_{\mathrm{VC}} \leq 2W \log_2(eM)$$

$$N \geq \frac{W}{\epsilon} \log_2 \left(\frac{M}{\epsilon} \right)$$

  - Two layers and threshold units

$$d_{\mathrm{VC}} \geq 2\lfloor M/2 \rfloor d$$ d inputs

$$Md \simeq W$$

For large networks

$$N_{\min} \simeq W/\epsilon.$$