MACHINE LEARNING

Vapnik-Chervonenkis (VC) Dimension

Alessandro Moschitti

Department of Information Engineering and Computer Science
University of Trento
Email: moschitti@disi.unitn.it

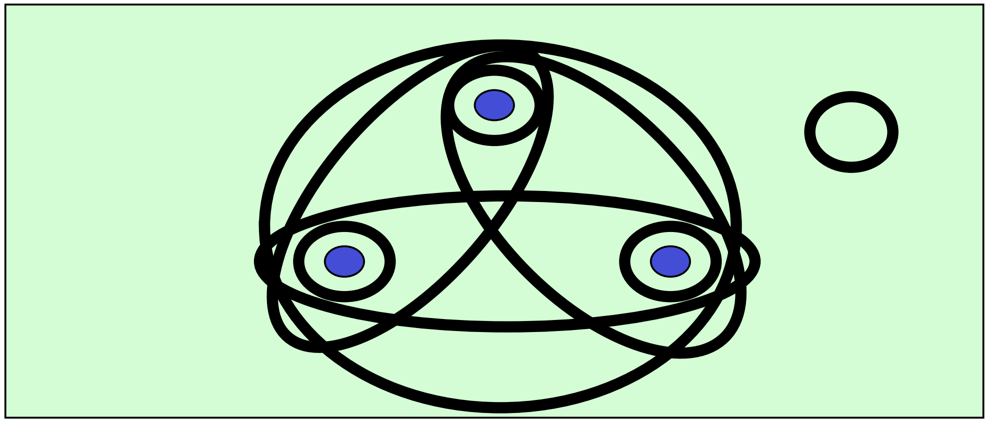
Computational Learning Theory

- The approach used in rectangular hypotheses is just one case:
  - Medium-built people
  - No general rule has been derived
- Is there any means to determine if a function is PAC learnable and derive the right bound?
- The answer is yes and it is based on theVapnik-Chervonenkis dimension (VC-dimension, [Vapnik 95])

VC-Dimension definition (1)

- Def.1: (***set shattering***): a subset S of instances of a set X is shattered by a collection of function F if $\forall$ S'$\subseteq$ S there is a function $f \in F$ such data:

$$f(x) = \begin{cases} 1 & x \in S' \\ 0 & x \in S - S' \end{cases}$$



VC-Dimension definition (2)

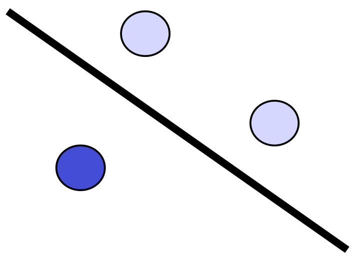
- Def. 2: the VC-dimension of a function set F (VC-dim(F)) is the cardinality of the largest dataset that can be shattered by F.
- Observation: the type of the functions used for shattering data determines the VC-dim

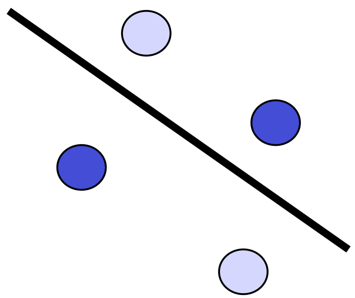
VC-Dim of linear functions (hyperplane)

- In the plane (hyperplane = line):
  - VC(Hiperlpanes) is at least 3
  - VC(Hiperlpanes)< 4 since there is no set of 4 points, which can be shattered by a line.

$\Rightarrow$ VC(H)=3. In general, for a k-dimension space VC(H)=k+1

- NB: It is useless selecting a set of linealy independent points





Upper Bound on Sample Complexity

Theorem 2.9 *(upper bound on sample complexity, [Blumer* et al.*, 1989])*
Let H and F be two function classes such that $F \subseteq H$ and let A an algorithm that derives a function $h \in H$ consistent with m training examples. Then, $\exists c_0$ such that $\forall f \in F$, $\forall D$ distribution, $\forall \epsilon > 0$ and $\delta < 1$ if

$$m > \frac{c_0}{\epsilon}\left(VC(H) \times ln\frac{1}{\epsilon} + \frac{1}{\delta}\right)$$

then with a probability $1 - \delta$,

$$error_D(h) \leq \epsilon,$$

where VC(H) is the VC dimension of H and $error_D(h)$ is the error of h according to the data distribution D.

Lower Bound on Sample Complexity

Theorem 2.10 *(lower bound on sample complexity, [Blumer* et al.*, 1989])*
To learn a concept class F whose VC-dimension is d, any PAC algorithm requires $m = O(\frac{1}{\epsilon}(\frac{1}{\delta} + d))$ examples.

Bound on the Classification error using VC-dimension

Theorem 2.11 *(Vapnik and Chervonenkis, [Vapnik, 1995])*
Let H be a hypothesis space having VC dimension d. For any probability distribution D on $X \times \{-1, 1\}$, with probability $1-\delta$ over m random examples S, any hypothesis $h \in H$ that is consistent with S has error no more than

$$error(h) \leq \epsilon(m, H, \delta) = \frac{2}{m}\left(d \times ln\frac{2e \times m}{d} + ln\frac{2}{\delta}\right),$$

provided that $d \leq m$ and $m \geq 2/\epsilon$.

Example: Rectangles have VC-dim > 4

- We must choose 4-point set, which can be shattered in all possible ways
- Given such 4 points, we assign them the {+,-} labels, in all possible ways.
- For each labeling it must exist a rectangle which produces such assignment, i.e. such classification

Example (cont'd)

- Our classifier: inside the recatagle positive and outside negative examples, respectively
- Given 4 points (linearly independent), we have the following assignments:

a) All points are "+" $\Rightarrow$ use a rectangle that includes them

b) All points are "-" $\Rightarrow$ use a empty rectangle

c) 3 points "-" and 1 "+" $\Rightarrow$ use a rectangle centered on the "+" points

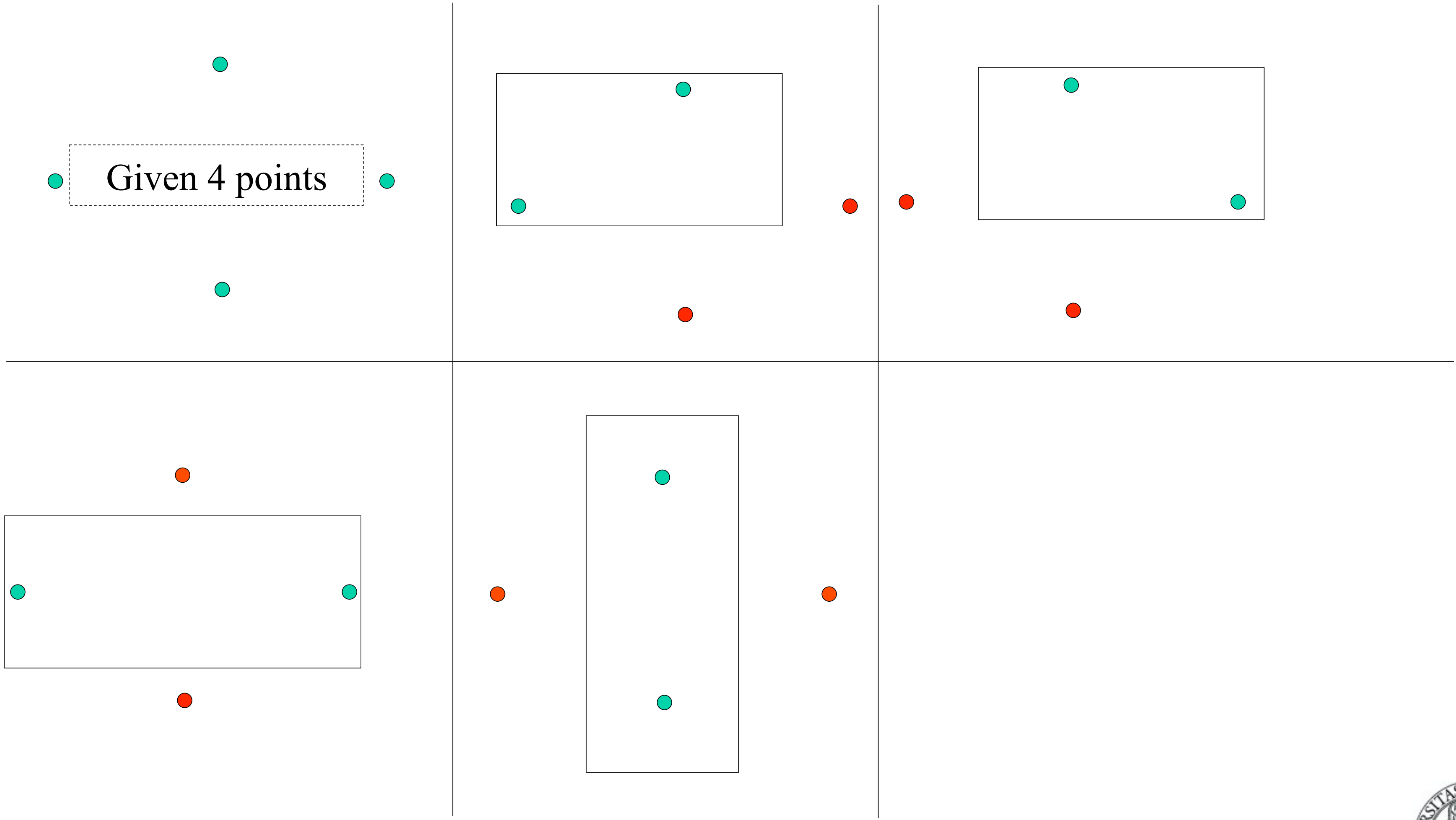
Example (cont'd)

d) 3 points "+" and one "-" $\Rightarrow$ we can always find a rectangle which exclude the "-" points

e) 2 points "+" and 2 points "-" $\Rightarrow$ we can define a rectangle which includes the 2 "+" and excludes the 2 "-".

- To show d) and e) we should check all possibilities

For example, to prove e)



VC-dim cannot be 5

- For any 5-point set, we can define a rectangle which has the most extern points as vertices
- If we assign to such vertices the "+" label and to the internal point the "-" label, there will not be any rectangle which reproduces such assigment

Bound Comparison

- $m > (4/\varepsilon) \cdot \ln(4/\delta)$ (ad hoc bound)
- $m > (1/\varepsilon) \cdot \ln(1/\delta) + 4/\varepsilon =$ (lower bound based on VC-dim)
- $(4/\varepsilon) \cdot \ln(4/\delta) > (1/\varepsilon) \cdot \ln(1/\delta) + 4/\varepsilon$
- $4 \cdot \ln(4/\delta) > \ln(1/\delta) + 4$
- $\ln(4/\delta) > \ln((1/\delta)1/4) + 1$
- $4/\delta > (1/\delta)1/4 \cdot e$
- $4 > \delta\ 3/4 \cdot e$
- $4 > (<1) \cdot (<3)$ verified

References

- VC-dimension:
  - **MY SLIDES: http://disi.unitn.it/moschitti/teaching.html**
  - **MY BOOK:**
    - **Automatic text categorization: from information retrieval to support vector learning**
    - **Roberto Basili and Alessandro Moschitti**

References

- *A tutorial on Support Vector Machines for Pattern Recognition*
  - **Downloadable from the web**
- *The Vapnik-Chervonenkis Dimension and the Learning Capability of Neural Nets*
  - **Downloadable from the web**
- Computational Learning Theory
  (Sally A Goldman Washington University St. Louis Missouri)
  - **Downloadable from the web**
- *AN INTRODUCTION TO SUPPORT VECTOR MACHINES*
  (and other kernel-based learning methods)
  N. Cristianini and J. Shawe-Taylor Cambridge University Press
  - **You can buy it also on line**