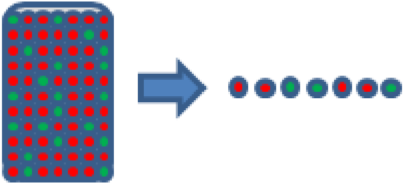
PAC, Generalizzation and SRM

Corso di AA, anno 2017/18, Padova

Fabio Aiolli

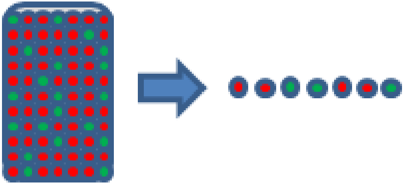
11 Ottobre 2017

A simple experiment



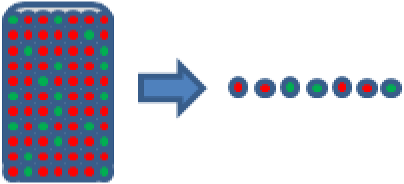
- $P(\text{red}) = \pi$
- $P(\text{green}) = 1 - \pi$
- π is unknown
- Pick N marbles (the *sample*) independently from the bin
- $\sigma =$ fraction of red marbles in the sample

A simple experiment



- Does σ say anything about π?
- Short answer... NO
- Ans: Sample can be mostly green while bin is mostly red
- Long answer... YES
- Ans: Sample frequency σ is likely close to bin frequency π

What does σ say about π

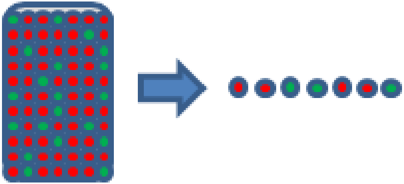


In a big sample (large N), the value σ is likely close to π (within ϵ)
More formally (Hoeffding's Inequality),

$$P(|\sigma - \pi| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

That is, $\sigma = \pi$ is P.A.C. (Probably Approximately Correct)

What does σ say about π



$$P(|\sigma - \pi| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

- Valid for all N and ϵ
- Bound does not depend on π
- Tradeoff: N, ϵ, and the bound
- $\sigma \approx \pi \Rightarrow \pi \approx \sigma$, that is "$\mu$ tends to be close to σ"

Connection to Learning

- In the Bin example, the unknown is π
- In the Learning example the unknown is $f : \mathcal{X} \to \mathcal{Y}$

- The bin is the input space $\mathcal{X}$
- Given an hypothesis h, green marbles correspond to examples where the hypothesis is right, i.e. $h(\mathbf{x}) = f(\mathbf{x})$
- Given an hypothesis h, red marbles correspond to examples where the hypothesis is wrong, i.e. $h(\mathbf{x}) \neq f(\mathbf{x})$

So, for *this* h, σ (empirical error) actually generalizes to π (ideal error) but... this is verification, not learning!

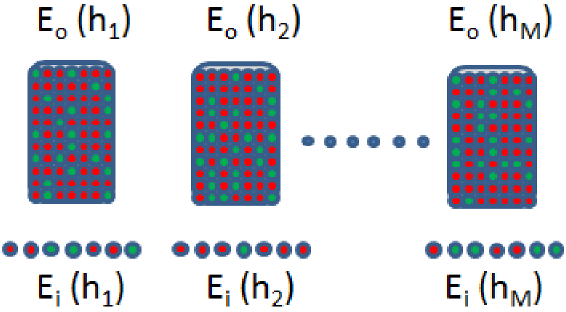
Connection to Learning

We need to choose from multiple hypotheses!
π and σ depend on which h

Change of notation

- $\sigma \rightarrow E_i(h)$
- $\pi \rightarrow E_o(h)$
- then, $P(|E_i(h) - E_o(h)| > \epsilon) \leq 2e^{-2\epsilon^2 N}$

Multiple Bins



Hoeffding's inequality does not apply here!

Analogy: Head and Cross

- If you toss a (fair) coin 10 times, which is the probability that you will get 10 heads?
- $(0.5)^{10} = 0.0009765625 \approx 0.1\%$
- If you toss 1000 (fair) coins 10 times each, which is the probability that *some coin* will get 10 heads?
- $(1 - (1 - 0.001)^{1000}) = 0.6323045752290363 \approx 63\%$

Going back to the learning problem

Is the learning feasible?

$$
\begin{aligned}
P(|E_i(g) - E_o(g)| > \epsilon) \quad \leq \quad & P(|E_i(h_1) - E_o(h_1)| > \epsilon \\
& \textbf{or} |E_i(h_2) - E_o(h_2)| > \epsilon \\
& \ldots \\
& \textbf{or} |E_i(h_M) - E_o(h_M)| > \epsilon) \\
\leq \quad & \sum_{m=1}^{M} P(|E_i(h_m) - E_o(h_m)| > \epsilon) \leq 2Me^{-2\epsilon^2 N}
\end{aligned}
$$

Going back to the learning problem

- Testing: $P(|E_i(g) - E_o(g)| > \epsilon) \leq 2e^{-2\epsilon^2 N}$
- Training: $P(|E_i(g) - E_o(g)| > \epsilon) \leq 2Me^{-2\epsilon^2 N}$

In fact M can be substituted by $m(\mathcal{H}) \leq 2^N$ which is related to the *complexity* of the hypothesis space!

Remember that $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

So, when the bad events are very overlapping (low complexity of the hypothesis space), then the value $m(\mathcal{H}) \ll 2^N$

Shattering: Given $S \subset X$, S is shattered by the hypothesis space $\mathcal{H}$ iff

$$\forall S' \subseteq S,\ \exists h \in \mathcal{H},\ \text{such that } \forall x \in S,\ h(x) = 1 \Leftrightarrow x \in S'$$

($\mathcal{H}$ is able to implement all possible dichotomies of S)

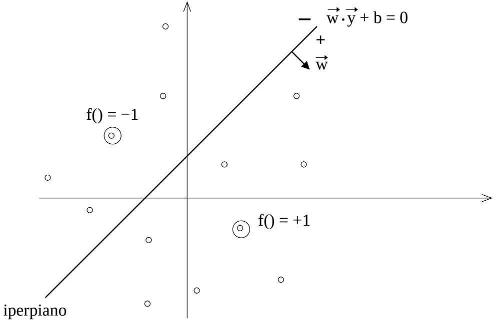
VC-dimension: The VC-dimension of a hypothesis space $\mathcal{H}$ defined over an instance space X is the size of the largest finite subset of X shattered by $\mathcal{H}$:

$$VC(\mathcal{H}) = \max_{S \subseteq X} |S| : \; S \text{ is shattered by } \mathcal{H}$$

If arbitrarily large finite sets of X can be shattered by $\mathcal{H}$, then $VC(\mathcal{H}) = \infty$.

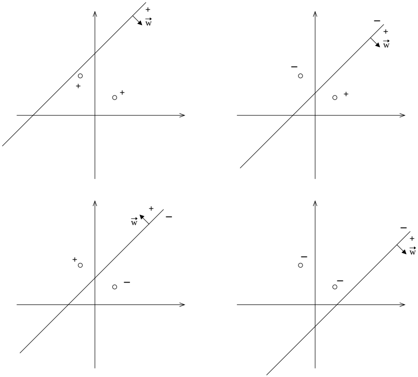
VC-dimension: Example

What is the VC-dimension of $\mathcal{H}_1$?
$\mathcal{H}_1 = \{f_{(\vec{w},b)}(\vec{y}) | f_{(\vec{w},b)}(\vec{y}) = sign(\vec{w} \cdot \vec{y} + b), \vec{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$



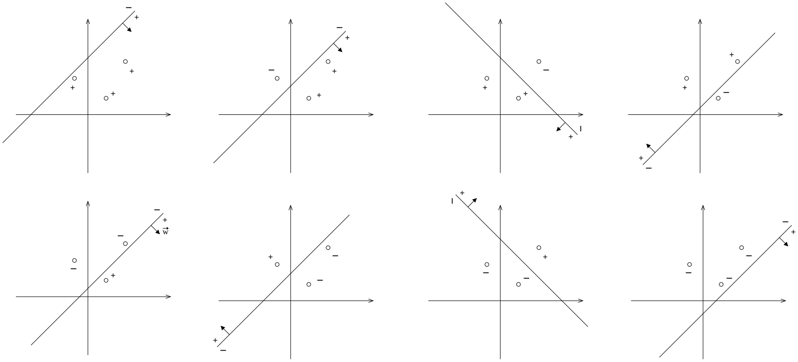
VC-dimension: Example

What is the VC-dimension of $\mathcal{H}_1$?
$VC(\mathcal{H}) \geq 1$ trivial. Let consider 2 points:



VC-dimension: Example

What is the VC-dimension of $\mathcal{H}_1$?
Thus $VC(\mathcal{H}) \geq 2$. Let consider 3 points:

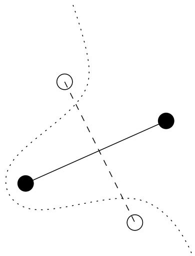


VC-dimension: Example

What is the VC-dimension of $\mathcal{H}_1$?
Thus $VC(\mathcal{H}) \geq 3$. What happens with 4 points ?

VC-dimension: Example

What is the VC-dimension of $\mathcal{H}_1$?
Thus $VC(\mathcal{H}) \geq 3$. What happens with 4 points ? It is impossible to shatter 4 points!!
In fact there always exist two pairs of points such that if we connect the two members by a segment, the two resulting segments will intersect. So, if we label the points of each pair with a different class, a curve is necessary to separate them! Thus $VC(\mathcal{H}) = 3$



What if $n > 2$?

Generalization Error

Consider a binary classification learning problem with:

- Training set $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- Hypothesis space $\mathcal{H} = \{h_\theta(\mathbf{x})\}$
- Learning algorithm $\mathcal{L}$, returning the hypothesis $g = h_\theta^*$ minimizing the empirical error on $\mathcal{S}$, that is $g = \arg\min_{h \in \mathcal{H}} \text{error}_S(h)$.

It is possible to derive an upper bound of the ideal error which is valid with probability $(1 - \delta)$, δ being arbitrarily small, of the form:

$$\text{error}(g) \leq \text{error}_S(g) + F\left(\frac{\text{VC}(\mathcal{H})}{n}, \delta\right)$$

Analysis of the bound

Let's take the two terms of the bound

- $A = \text{error}_S(g)$
- $B = F(VC(\mathcal{H})/n, \delta)$

- The term A depends on the hypothesis returned by the learning algorithm $\mathcal{L}$.
- The term B (often called VC-confidence) does not depend on $\mathcal{L}$. It only depends on:
  - the training size n (inversely),
  - the VC dimension of the hypothesis space VC($\mathcal{H}$) (proportionally)
  - the confidence δ (inversely).

Structural Risk Minimization

Problem: as the VC-dimension grows, the empirical risk (A) decreases, however the VC confidence (B) increases !

Because of that, Vapnik and Chervonenkis proposed a new inductive principle, i.e. Structural Risk Minimization (SRM), which aims to minimizing the right hand of the confidence bound, so to get a tradeoff between A and B:

Consider $\mathcal{H}_i$ such that

- $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{H}_n$
- $VC(\mathcal{H}_1) \leq \cdots \leq VC(\mathcal{H}_n)$
- select the hypothesis with the smallest bound on the true risk

Example: Neural networks with an increasing number of hidden units