PAC, Generalizzation and SRM

Corso di AA, anno 2016/17, Padova

Fabio Aiolli

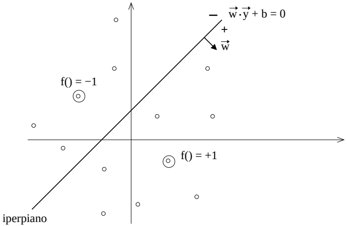
12 Ottobre 2016

Università degli Studi di Padova

Hypothesis Space: Example 1

Hyperplanes in $\mathbb{R}^2$

- Instance space: points in the plane $\mathcal{X} = \{y | y \in \mathbb{R}^2\}$
- Hypothesis space: dichotomies induced by hyperplanes in $\mathbb{R}^2$, that is $\mathcal{H} = \{f_{\mathbf{w},b}(y) = \text{sign}(\mathbf{w} \cdot y + b), \mathbf{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$



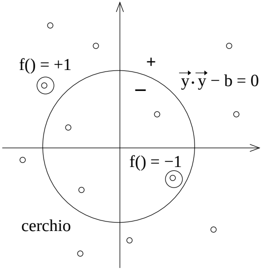
What changes in $\mathbb{R}^n$?

Università degli Studi di Padova

Hypothesis Space: Example 2

Circles in $\mathbb{R}^2$

- Instance space: points in the plane $\mathcal{X} = \{y | y \in \mathbb{R}^2\}$
- Hypothesis space: dichotomies induced by circles centered in the origin in $\mathbb{R}^2$, that is $\mathcal{H} = \{f_b(y) = \text{sign}(||y||^2 - b), b \in \mathbb{R}\}$

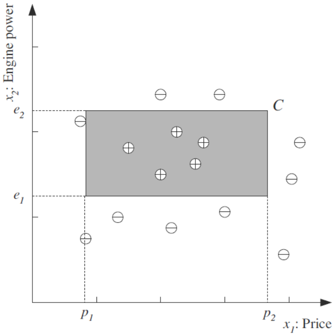


What changes in $\mathbb{R}^n$?

Hypothesis Space: Example 3

Rectangles in $\mathbb{R}^2$

- Instance space: points in the plane $\mathcal{X} = \{(p, e) | (p, e) \in \mathbb{R}^2\}$
- Hypothesis space: dichotomies induced by rectangles in $\mathbb{R}^2$, that is $\mathcal{H} = \{f_\theta(y) = [p_1 \leq p \leq p_2 \cap e_1 \leq e \leq e_2], \theta = \{p_1, p_2, e_1, e_2\}\}$ where $[z] = +1$ if $z =$ True, -1 otherwise.



What changes in $\mathbb{R}^n$

Hypothesis Space: Example 4

Conjunction of m positive literals

- Instance space: strings of m bits, $\mathcal{X} = \{s | s \in \{0,1\}^m\}$
- Hypothesis space: all the logic sentences involving positive literals $l_1, \ldots, l_m$ (l_1 is true if the first bit is 1, l_2 is true if the second bit is 1, etc.) and just containing the operator $\wedge$ (**and**)

$$\mathcal{H} = \{f_{\{i_1,\ldots,i_j\}}(s) | f_{\{i_1,\ldots,i_j\}}(s) \equiv l_{i_1} \wedge l_{i_2} \wedge \cdots \wedge l_{i_j}, \{i_1, \ldots, i_j\} \subseteq \{1, \ldots, m\}\}$$

E.g. $m = 3$, $X = \{0,1\}^3$
Examples of instances: $s_1 = 101$, $s_2 = 001$, $s_3 = 100$, $s_4 = 111$
Examples of hypotheses: $h_1 \equiv l_2$, $h_2 \equiv l_1 \wedge l_2$, $h_3 \equiv true$, $h_4 \equiv l_1 \wedge l_3$, $h_5 \equiv l_1 \wedge l_2 \wedge l_3$
h_1, h_2, and h_5 are false for s_1, s_2 and s_3 and true for s_4; h_3 is true for any instance; h_4 is true for s_1 and s_4 but false for s_2 and s_3

UNIVERSITÀ DEGLI STUDI DI PADOVA

Hypothesis Space: Example 4

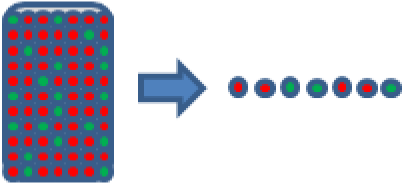
Conjunction of m positive literals

- Question 1: how many and which are the distinct hypotheses for $m = 3$?
  - Ans.(which): *true*, l_1, l_2, l_3, $l_1 \wedge l_2$, $l_1 \wedge l_3$, $l_2 \wedge l_3$, $l_1 \wedge l_2 \wedge l_3$
  - Ans.(how many): 8
- Question 2: how many distinct hypotheses there are as a function of m?
  - Ans.: 2^m, in fact for each possible bit of the input string the corresponding literal may occur or not in the logic formula, so:

$$\underbrace{2 \cdot 2 \cdot 2 \cdots 2}_{m \text{ times}} = 2^m$$

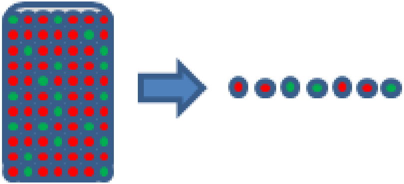
UNIVERSITÀ DEGLI STUDI DI PADOVA

A simple experiment



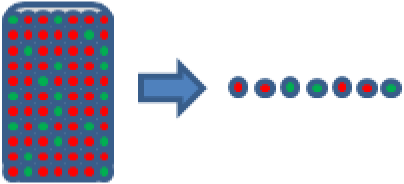
- $P(\text{red}) = \pi$
- $P(\text{green}) = 1 - \pi$
- π is unknown
- Pick N marbles (the *sample*) independently from the bin
- $\sigma =$ fraction of red marbles in the sample

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)

Università degli Studi di Padova

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 $\mathcal{X}$
- Green marbles correspond to examples where the hypothesis is right $h(\mathbf{x}) = f(\mathbf{x})$
- Red marbles correspond to examples where the hypothesis is right $h(\mathbf{x}) \neq f(\mathbf{x})$

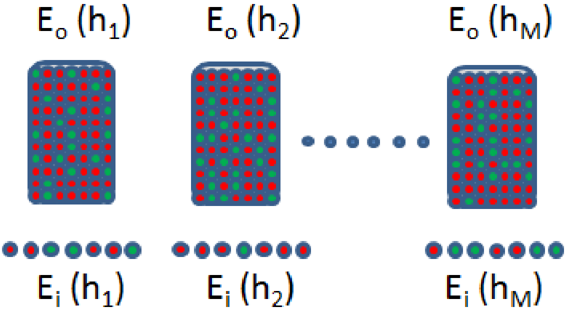
So, for *this* h, σ (empirical error) generalizes to π (ideal error)
but... this is verification, not learning!

Connection to Learning

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?
- $(1 - (1 - 0.001)^{1000}) = 0.6323045752290363 \approx 63\%$

UNIVERSITÀ DEGLI STUDI DI PADOVA

Going back to the learning problem

Is the learning feasible?

$$
\begin{aligned}
P(|E_i(g) - E_o(g)| > \epsilon) \quad \leq \quad & P(|E_i(h_1) - E_o(h_1)| > \epsilon \\
& \textbf{or} |E_i(h_2) - E_o(h_2)| > \epsilon \\
& \dots \\
& \textbf{or} |E_i(h_M) - E_o(h_M)| > \epsilon) \\
\leq \quad & \sum_m P(|E_i(h_m) - E_o(h_m)| > \epsilon) \leq 2Me^{-2\epsilon^2 N}
\end{aligned}
$$

Università degli Studi di Padova

Going back to the learning problem

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

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

Università degli Studi di Padova

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)

Measuring the complexity of the hypothesis space

VC-dimension

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 $\mathcal{H}$:

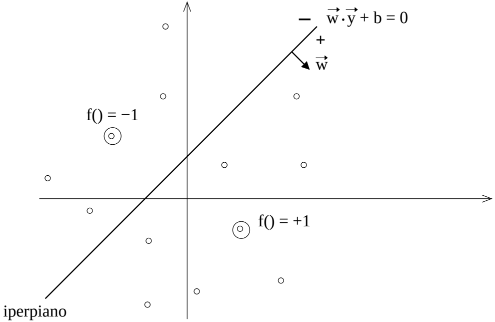
$$VC(\mathcal{H}) = \max_{S \subseteq X} |S| : \; S \text{ is shattered by } \mathcal{H}$$

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

Università degli Studi di Padova

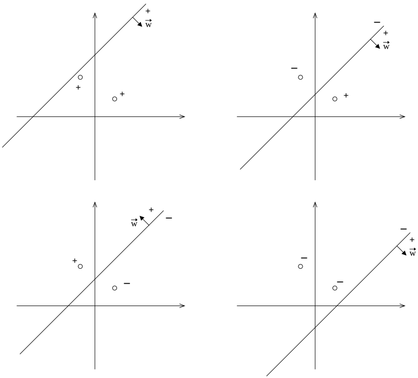
VC-dimension: Example

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

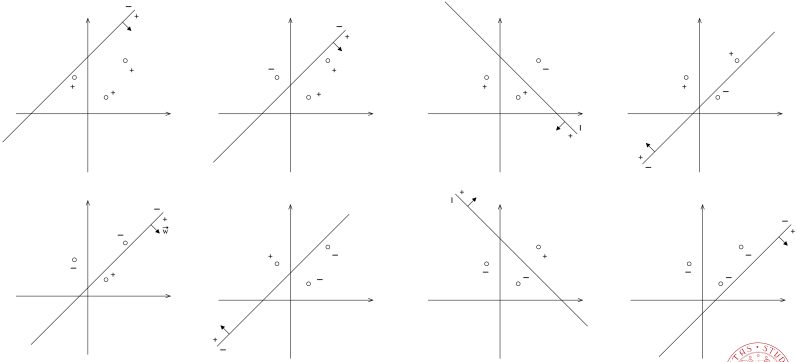
What is the VC-dimension of $\mathcal{H}_1$?
$VC(\mathcal{H}) \geq 1$ trivial. Let consider 2 points:



Università degli Studi di Padova

VC-dimension: Example

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

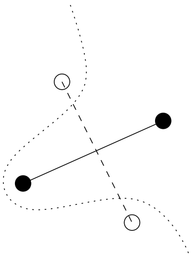


VC-dimension: Example

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 $\mathcal{S} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$
- 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:

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

UNIVERSITÀ DEGLI STUDI DI PADOVA

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