VC dimension : exercises

October 15, 2017

Notation :

We will consider functions $f : \chi \mapsto \{0,1\}$.

If F is a class of such functions and $x_1, \cdots x_n$ is a family of n points in χ, we define the set $N_F(x_1, \cdots x_n)$ as the set of all images of this family of points by the functions in F :

$$N_F(x_1, \cdots x_n) = \{(f(x_1), \cdots f(x_n)), f \in F\}$$

We define then the shattering coefficient of F with respect to n points sets in χ, denoted $S(F, n)$, as :

$$S(F, n) = \max |N_F(x_1, \cdots x_n)|$$

where the maximum is taken over all possible sets $(x_1, \cdots x_n) \in \chi^n$.

Finally, we define the VC dimension of F as :

$$\mathrm{VC}(F) = \max \{n \geq 1,\ S(F, n) = 2^n\}$$

Exercises :

Determine the VC dimension of the next sets of functions where $\chi = [0,1]$:

- $F = \{f : \chi \mapsto \{0,1\},\ f(x) = 1_{x<t},\ \text{t} \ \in [0,1]\}$
- $F' = \{f : \chi \mapsto \{0,1\},\ f(x) = 1_{x<t}$ or $f(x) = 1 - 1_{x<t},\ \text{t} \ \in [0,1]\}$
- $F = \{f : \chi \mapsto \{0,1\},\ f(x) = 1_{t_1 \leq x < t_2},\ t_1 < t_2 \in [0,1]\}$
- $F' = \{f : \chi \mapsto \{0,1\},\ f(x) = 1_{t_1 \leq x < t_2}$ or $f(x) = 1 - 1_{t_1 \leq x < t_2},\ t_1 < t_2 \in [0,1]\}$
- $F_k = \{f : \chi \mapsto \{0,1\},\ f(x) = \sum_{i=0}^{k} 1_{t_{2i} \leq x < t_{2i+1}}$, for $0 \leq t_0 < \cdots < t_{2k+1} \leq 1\}$ for any $k \geq 1$

Note here that for any F, F' is essentially the same set of functions, the only difference being that it allows to label the points indifferently 1 against 0, or 0 against 1. This apparently harmless technical enhancement is actually not totally insignificant as the VC dimension of F and F' are different.

Solutions

- Obviously any set of one point can be shattered, so VC(F) ≥ 1. Moreover, if you take two points x_1 and x_2 (assume $x_1 < x_2$ without loss of generality) then if x_1 is labeled 1 and x_2 labeled 0, the set cannot be shattered by any function in F. Therefore VC(F) $= 1$.

- Now, if you take two points x_1 and x_2 (assume $x_1 < x_2$ without loss of generality) all possible labeling of the points is reachable by putting $x_1 < t < x_2$, $t < x_1$ or $t > x_2$. So VC(F') ≥ 2. If you take three points x_1, x_2 and x_3 (assume $x_1 < x_2 < x_3$ without loss of generality), then for example there is no way that you can label x_1 and x_3 with the value 1, and x_2 with the value 0. So VC(F') $= 2$.

- If you take two points x_1 and x_2 (assume $x_1 < x_2$ without loss of generality) all possible labeling of the points is reachable by putting $t_1 < x_1 < t_2 < x_1$, $x_1 < t_1 < x_2 < t_2$, $t_1 < x_1 < x_2 < t_2$ or $x_1 < t_1 < t_2 < x_2$. So VC(F) ≥ 2. If you take three points x_1, x_2 and x_3 (assume $x_1 < x_2 < x_3$ without loss of generality) then there is no way that you can label x_1 and x_3 with the value 1, and x_2 with the value 0. Thus VC(F) $= 2$.

- With F' you can label x_1 and x_3 with the value 1, and x_2 with the value 0. This was the only labeling that was impossible with the previous F, therefore VC(F') ≥ 3. With four points x_1, x_2, x_3 and x_4 (assumed increasing as always), you cannot label x_1 and x_3 with the value 1 and x_2 and x_4 with the value 0 for example. So VC(F') $= 3$.

- It's clear that the "worst" labeling you can encounter is when the labels are alternating $(0, 1, 0, 1 \cdots)$. Here "worst" means that if you can do this one you can do any other labeling. Now in this kind of configuration, if you have $k+1$ labels 1 (and therefore $2(k+1)$ points in your set) it's clear that you can label all of them by putting one of the $k+1$ "doors" of your function over each one of the $k+1$ labels 1 of your set of points (the set F_k being the set of all functions with $k+1$ doors). So VC(F_k) $\geq 2(k+1)$. Moreover if you have $2(k+1)+1$ points, you can create a configuration of alternating labels with $k+2$ labels 1, by starting and ending by 1 $(1, 0, 1, \cdots, 0, 1)$. this last configuration is unreachable with $k+1$ doors. Therefore VC(F_k) $= 2(k+1)$.

Svolgimento Dettagliato, $\mathcal{X}=[0,1]$

Schema Generale procedura

$VCdim(F) = VC(F) = d$ se esiste un sottoinsieme di X, diciamo Y, tale che $|F_Y| = 2^d = 2^{|Y|}$

dove $F_Y = \{(f(x_1), \dots, f(x_d)) \, ; \, f \in F\}$

e $\forall Z \subset X$ con $|Z| = d+1$ si ha $|F_Z| < 2^{d+1} = 2^{|Z|}$

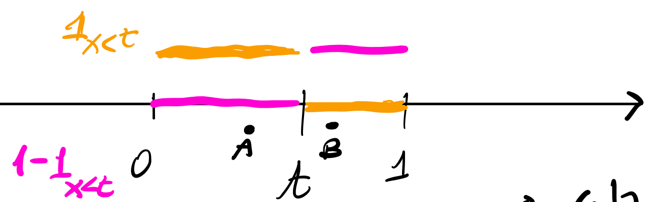
⊙ $F = \{f : \mathcal{X} \to \{0,1\} : f(x) = \mathbb{1}_{x<t} \, , \, t \in [0,1]\}$

Preso un qualsiasi $x \in [0,1)$ si ha che $f(x)=1$ per $t=1$. D'altro canto, prendendo $t \le x$ si ha $f(x)=0$ quindi $VCdim(F) \geq 1$.

D'altro canto $\forall$ coppia di punti distinti $x<y$ in $[0,1]$ si ha che la funzione $(x,y) \to (0,1)$ non è restrizione di f per alcuna $f \in F$.

Quindi $VCdim(F) = 1$

⊙ $F' = \{f : \mathcal{X} \to \{0,1\} : f(x) = \mathbb{1}_{x<t}$ o $f(x) = 1 - \mathbb{1}_{x<t} \, , \, t \in [0,1]\}$

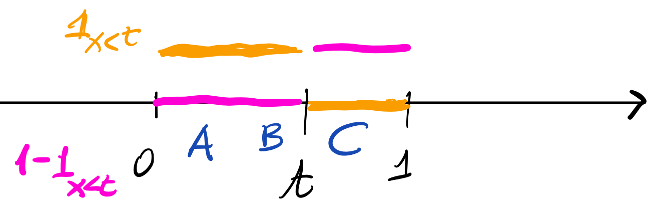


Chiaramente $VCdim(F') \geq 1$ (vedi sopra e considera che $F' \supset F \Rightarrow VCdim(F') \geq VCdim(F)$.

Consideriamo A, B come sopra, allora è chiaro che muovendo t possiamo ottenere $(0,0), (1,0), (0,1), (1,1)$ quindi $VCdim(F') \geq 2$

→ Segue

Segue →



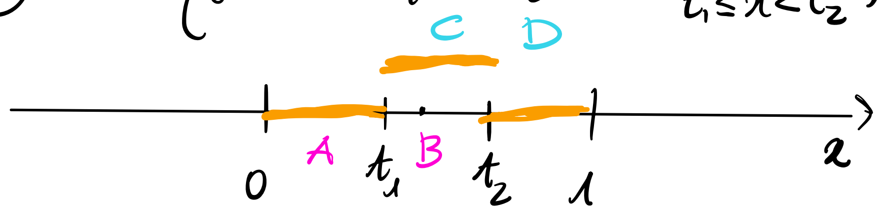
Quali terne di $\{0,1\}^3$ possiamo realizzare muovendo t? Ora le — uniche configurazioni possibili sono

$t<A<B<C \qquad A<t<B<C$

$A<B<t<C \qquad A<B<C<t$

con esse si possono produrre tutte le stringhe binarie, tranne $(0,1,0)$ e $(1,0,1)$ quindi la dimensione $VCdim(F')=2$.

$$F=\{f:\mathcal{X}\to\{0,1\};\ f(x)_{t_1\le x<t_2},\ t_1<t_2\in[0,1]\}$$



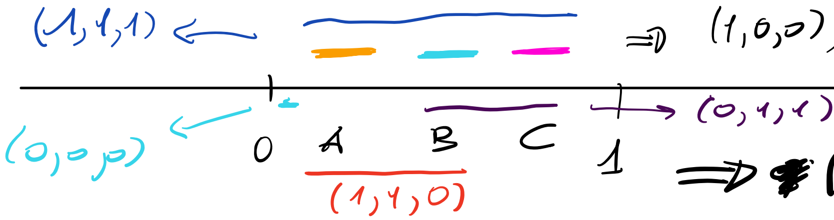
Chiaramente $VCdim(F)\geq 1$ e anche due punti sono shattered

$A<B\leq t_1<t_2 \to (1,1)$

$t_1<t_2<A<B \to (0,0)$

$A<t_1<B<t_2 \Rightarrow (0,1)$

$t_1<A<t_2<B \Rightarrow (1,0)$

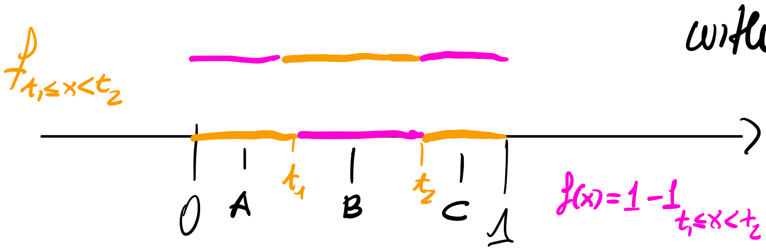


$\Rightarrow (1,0,0),(0,1,0),(0,0,1)$

$\Rightarrow (1,0,1)$ non è possibile

→ quindi $VCdim(F) = 2$ (5)

• $F' = \{ f : \mathcal{X} \to \{0,1\} ; f(x) = 1_{t_1 \le x < t_2}$ o $f(x) = 1 - 1_{t_1 \le x < t_2}$ with $t_1 < t_2 \in [0,1] \}$

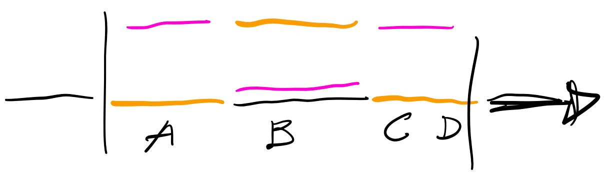


$F' \supset F \Rightarrow VCdim(F') \geq 2$

Controlliamo 3 punti $A < B < C$, considerando che $F' \supset F$ ~~con~~ guardiamo solo il caso $(1,0,1)$ che prima non era possibile: si vede che $f(x) = 1 - 1_{t_1 \le x < t_2}$ con $A < t_1 < B < t_2 < C$ produrrà $(1,0,1)$

quindi $VCdim(F') \geq 3$

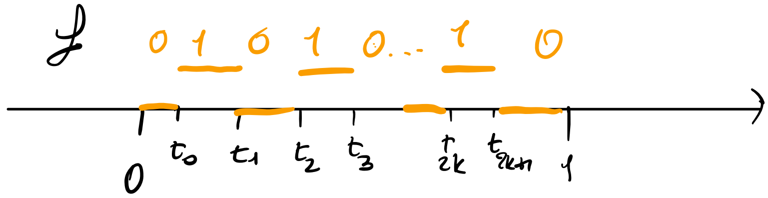
Consideriamo 4 punti e vediamo che non c'è modo di produrre ad esempio $(1,0,1,0)$



Quindi la $VCdim(F') = 3$

(a) $F_k = \left\{ f : \mathcal{X} \to \{0,1\} : f(x) = \sum_{i=0}^{k} 1_{t_{2i} \le x < t_{2i+1}}, 0 \le t_0 < \dots < t_{2k+1} \le 1 \right\}$

— con $k \geq 1$.



La sequenza alternate sono il problema.

....

$(1,0,1,0,1,0 \dots 1,0) \Rightarrow 2(k+1)$

(Stesso con $(0,1,0,1,\dots,0,1)$ sempre $2(k+1)$)

Se prendo $2(k+1)+1$ ottengo che $(1,0,1,0,\dots,1,0,1)$ non si può produrre, perchè ha $k+2$ Celle uguali a "1" il che non è possibile

Quindi $VCdim(F_k) = 2(k+1)$ _

Fine