Dimensión VC

Se puede considerar como una métrica de la complejidad de un algoritmo de ML.

Consideremos un algoritmo de clasificación binaria.

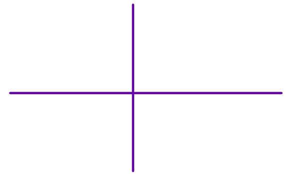
$$\begin{array}{c} \vec{x} \longrightarrow \\ y \longrightarrow \end{array} \boxed{\text{Alg. de ML}} \longrightarrow h \rightarrow \text{Modelo "hipótesis"}$$

ej. h es una recta que separa el espacio $\mathbb{R}^2$ en 2 subconjuntos

Llamemos $\mathcal{H}$ "espacio de hipótesis" al conjunto de todos los posibles modelos arrojados por el algoritmo. ej. todas las rectas en $\mathbb{R}^2$

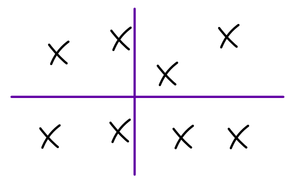
Sea X el espacio de instancias a clasificar.

Ej $\mathbb{R}^2$



Consideremos un subconjunto de instancias $S \subseteq X$.

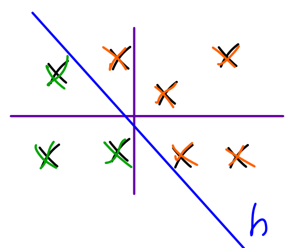
Ej



Cada modelo h separa el conjunto S en dos subconjuntos

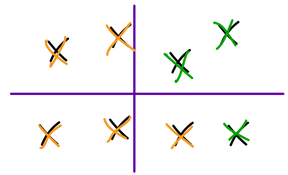
En otras palabras, cada modelo genera una dicotomía en el conjunto S

Ej

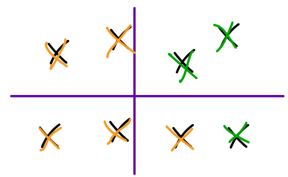


Diremos entonces que h "representa" dicha dicotomía

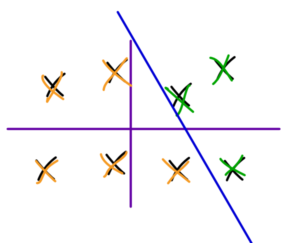
Ahora, dada cualquier dicotomía en S



podemos preguntarnos si hay un modelo h que la representa



?

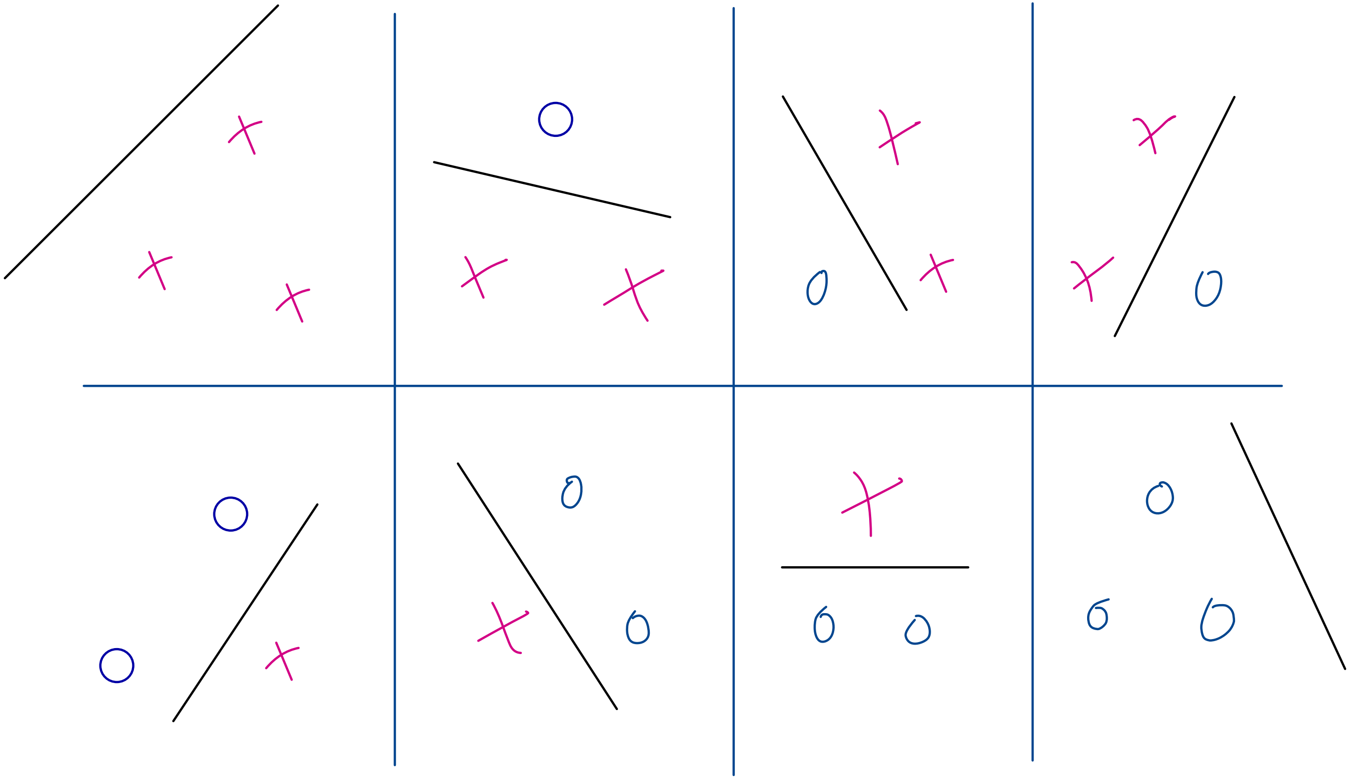


para ésta sí!

Siempre la hay?

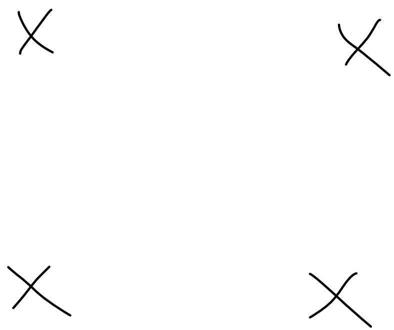
NO!

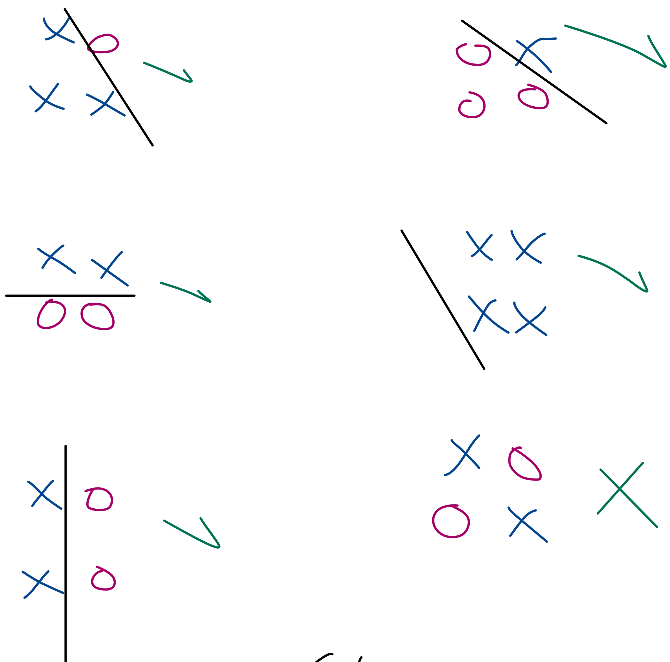
hay algún conjunto S para el cual toda dicotomía tenga un modelo que la represente?



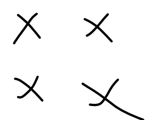
En este caso, decimos que $\mathcal{H}$ separa al conjunto de tres puntos de arriba

Se puede para 4? H separa al conjunto?





¿Algún conjunto de 4 puntos se puede separar por $\mathcal{H}$?

El conjunto 

No se puede "separar" por $\mathcal{H}$

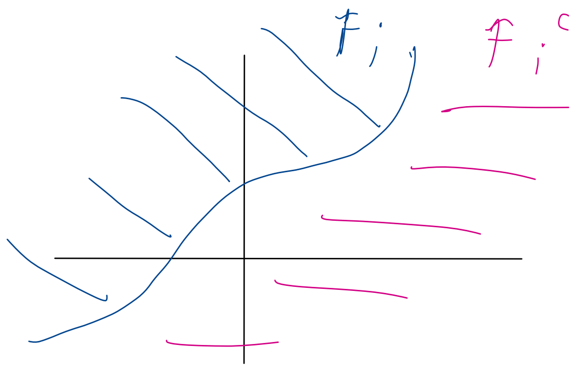
Ningún conjunto de 4 puntos se puede separar por $\mathcal{H}$!

Decimos entonces que la Dimensión VC de $\mathcal{H}$ es $\boxed{3}$

Definicion formal

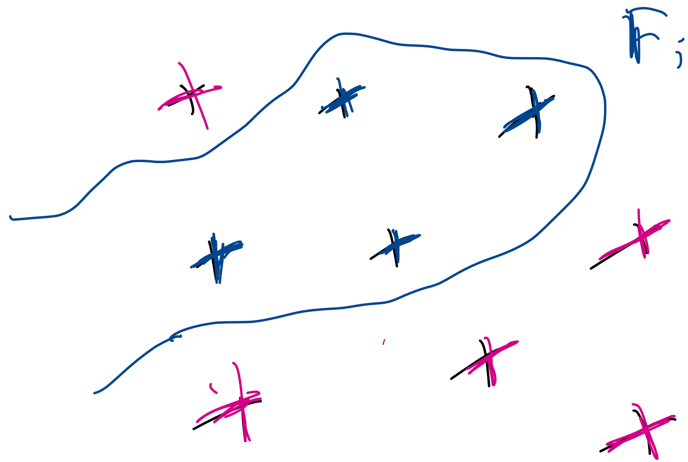
Un algoritmo de clasificación binaria genera una familia $\mathcal{F}$ de subconjuntos del espacio X de instancias.

$$\mathcal{F} = \{ F_1, F_1^c, F_2, F_2^c, \ldots\ldots \}$$



<u>Definición</u> Una familia $\mathcal{F}$ de subconjuntos separa (atomiza) un conjunto $S = \{s_1, \dots, s_n\}$ si

$$\forall X \subseteq S, \ \exists F_i \in \mathcal{F} \quad X = F_i \cap S$$



Definición La dimensión VC de una familia $\mathcal{F}$ es el máximo número D tal que existe un conjunto S de tamaño D tal que $\mathcal{F}$ separa a S. Si no existe un máximo, entonces la dimension es ∞

La dimension VC de un algoritmo de clasificación es la dimensión VC de su familia $\mathcal{F}$ asociada.

Observaciones

- La dimensión VC en general es difícil de calcular!

- Si la dimensión VC es grande, el algoritmo es más complejo (y más propenso al sobreajuste)

Ejercicio

¿Cuál es la dimensión VC de la familia de planos en $\mathbb{R}^3$?

Ejercicio

Generalización a hiperplanos: demostrar que la dimención VC de los hiperplanos en $\mathbb{R}^n$ es $n+1$

Ejercicio

La dimensión VC de una familia finita $\mathcal{F}$ es a lo más $\log_2 |\mathcal{F}|$ ¿Por qué?

Dimensión VC en Máquinas de soporte vectorial

Con Kernel lineal $\equiv$ hiperplanos ($VC\text{-}dim = n+1$)

Con Kernel RBF : $VC\text{-}dim = \infty$

Dimensión VC en redes neuronales

Si E es el conjunto de aristas de una red neuronal y V es el conjunto de vértices (Nodos), entonces:

Si la función de activación es la función de salto (step function)

$$VC\text{-dim} \leq O(|E| \cdot \log |E|)$$

Si la activación es la sigmoide $VC\text{-dim} \leq O(|E|^2, |V|^2)$

Si los pesos vienen de una familia finita, $VC\text{-dim} \leq O(E)$ (para ambas activaciones)

Una cota para el error de testeo

$$P\left(\text{test error} \leq \text{training error} + \sqrt{\frac{1}{N}\left[D\left(\log\left(\frac{2N}{D}\right)+1\right) - \log\left(\frac{\alpha}{4}\right)\right]}\right) = 1-\alpha$$

VC-dim

tamaño del conjunto de entrenamiento