Teoría de generalización → ¿Qué tan buena es la red al predecir?

¿Cuál es la influencia de la cantidad de ejemplos de entrenamiento?

$\bar{w}$: conjunto de parámetros de la red

$\tilde{f}(\vec{x})$: función blanco

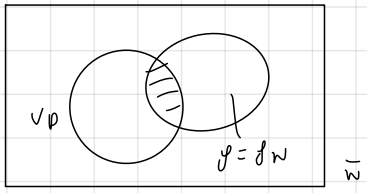
$\vec{x}$: vectores de entrada

$(x^\mu, y^\mu = \tilde{f}(\vec{x}^\mu))$ $\mu = 1, \dots, P$ pares a aprender → conjunto de entrenamiento

$\vec{x}^\mu$ son extraídos de $P(\vec{x})$ (independientes)

distribución de prob de los parámetros

$f_{\bar{w}}(\vec{x})$ función implementada por $\bar{w}$



$V_0 = \int dW\, P(W)$

volumen total en el espacio de acoplamientos

$\Theta(\bar{w}) = \begin{cases} 1 & \text{si } f(\vec{x}) = f_w(\vec{x}) \\ 0 & \text{c.c} \end{cases}$

$V(f) = \int d\bar{w}\, P(\bar{w})\, \Theta_f(\bar{w})$

$I(f_w, \vec{x}^\mu) = \begin{cases} 1 & \text{si } f_w(\vec{x}^\mu) = \tilde{f}(\vec{x}^\mu) \\ 0 & \end{cases}$

$V_P = \int d\bar{w}\, P(\bar{w}) \left[\prod_{\mu=1}^{P} I(f_{\bar{w}}, \vec{x}^\mu) \right]$

$V(P(f)) = \int dW\, P(w)\, \Theta_f(\bar{w}) \left[\prod_{\mu=1}^{P} I(f_{\bar{w}}, \vec{x}^\mu) \right]$

$V_P(f) = \left[\int d\bar{w}\, P(\bar{w})\, \Theta_f(\bar{w}) \right] \left[\prod_{\mu=1}^{P} I(f, \vec{x}^\mu) \right]$

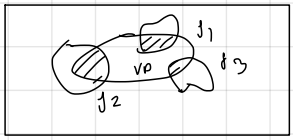
$\langle \dots \rangle$ sobre $P(\vec{x})$

$\langle V_P(f) \rangle = \left[\int d\bar{w}\, P(\bar{w})\, \Theta_f(\bar{w}) \right] \langle \prod_{\mu=1}^{P} I(f, \vec{x}^\mu) \rangle$

$= \left[\int d\bar{w}\, P(\bar{w})\, \Theta_f(\bar{w}) \right] \underbrace{\langle I(f, \vec{x}) \rangle^P}_{dy\,:\, g(f)}$

$I(f, \vec{x}) = \begin{cases} 1 & \text{si } f(\vec{x}) = \tilde{f}(\vec{x}) \\ 0 & \text{c.c} \end{cases}$

$\langle V(P, f) \rangle = V_0(f)\, g(f)^P$



$\bar{w}$

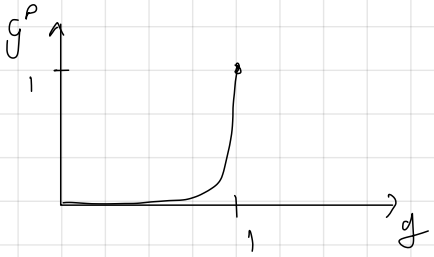
$$P_p(y) \propto V_p(y)\, g(y)^p$$

$$P_p(g) = \sum_y P_p(y)\, \delta(g - g(y)) \propto \sum_y v_0(y)\, g(y)^p\, \delta(g - g(y))$$

$$\propto g^p \underbrace{\sum_y v_0(y)\, \delta(g - g(y))}_{g_0(g)}$$

$$P_p(g) = \frac{g^p P_0(g)}{\int_0^1 g'^p P_0(g')\, dg'}$$

$$\langle g \rangle = \int_0^1 g\, P_p(g)\, dg = \frac{\int_0^1 g^{p+1} P_0(g)\, dg}{\int_0^1 g^p P_0(g)\, dg}$$



$$\simeq \frac{\int_0^1 g^{p+1} dg\, \cancel{P_0(1)}}{\int_0^1 g^p dg\, \cancel{P_0(1)}} = \frac{\frac{1}{p+2}}{\frac{1}{p+1}} = \frac{p+1}{p+2} = \frac{1+p}{1+2p} = \left(1 + \frac{1}{p}\right)\left(1 - \frac{2}{p}\right)$$

$$\simeq 1 - \frac{1}{p}$$

escoger de como se tiene que incrementar el nº de ejemplos para mejorar la generalización

Interesa cual es la probabilidad del peor caso

COTA DEL PEOR CASO

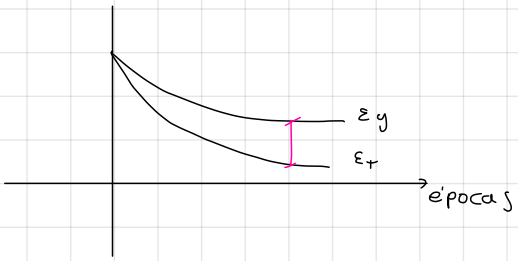
Por simplicidad tomemos funciones booleanas

$\tilde{f}(\bar{x})$: función booleana

$(\bar{x}^\mu, \bar{\nu}^\mu \equiv \tilde{f}(\bar{x}^\mu)) \quad \mu = 1, \ldots, p$

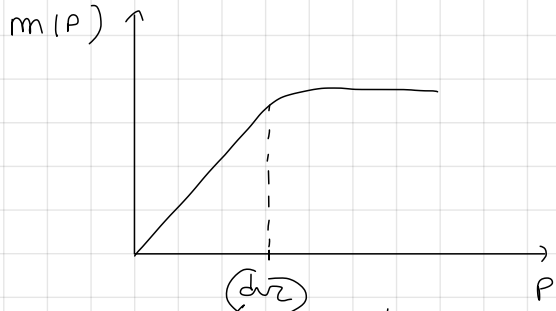
$g(y) = \langle I(\tilde{f}, \tilde{x}) \rangle$

$$g_P(f) = \left\langle \prod_{k=1}^{P} I(f, \bar{x}^k) \right\rangle \rightarrow$$ performance sobre el conjunto de entrenamiento



típicamente $g_P(f) > g(f)$

$$\text{PROB}\left(\max_f |g_P(f) - g(f)| \right) \leq 4\, m(2,0)\, e^{-\varepsilon^2 P/8} \rightarrow 0 \quad P \to \infty \ \forall \varepsilon$$



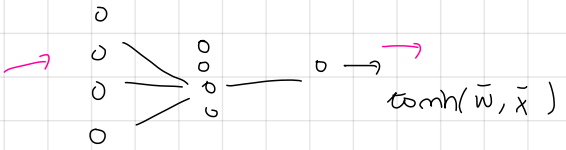
$m(P) \leq P^{d_{VC}} + 1$

potencia mínima de P para que sature → dimensión Vapnik-Chervonenkis

resume la complejidad de la red

perceptrón 2 capas $2N = d_{VC}$

$d_{VC} = N m \lg(m)$

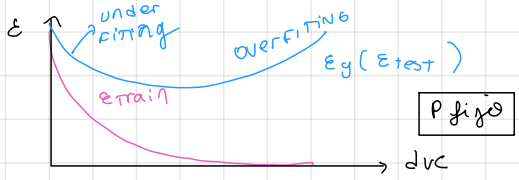


escriben cartas
ven películas
cocinan
piden verduras
eligen películas

- en psicología
- reunión mates

d_{VC} debe depender del tamaño de los pesos

1 capa continua $d_{VC} \leq \min\{D^2|W|^2, 2N+1\}$



Cambiamos la complejidad de la red

ε para tener una cota que deje crecer