Tópicos selecionados na aprendizagem de máquina supervisionada

Vladimir Pestov

twitter: @docente_errante

[1]Universidade Federal da Bahia
Salvador, BA, Brasil
(Professor Visitante)

[2]University of Ottawa / Université d'Ottawa
Ottawa, Ontario, Canadá
(Professor Emérito)

Departamento de Estatística, IME-USP, 18–29.11.2019

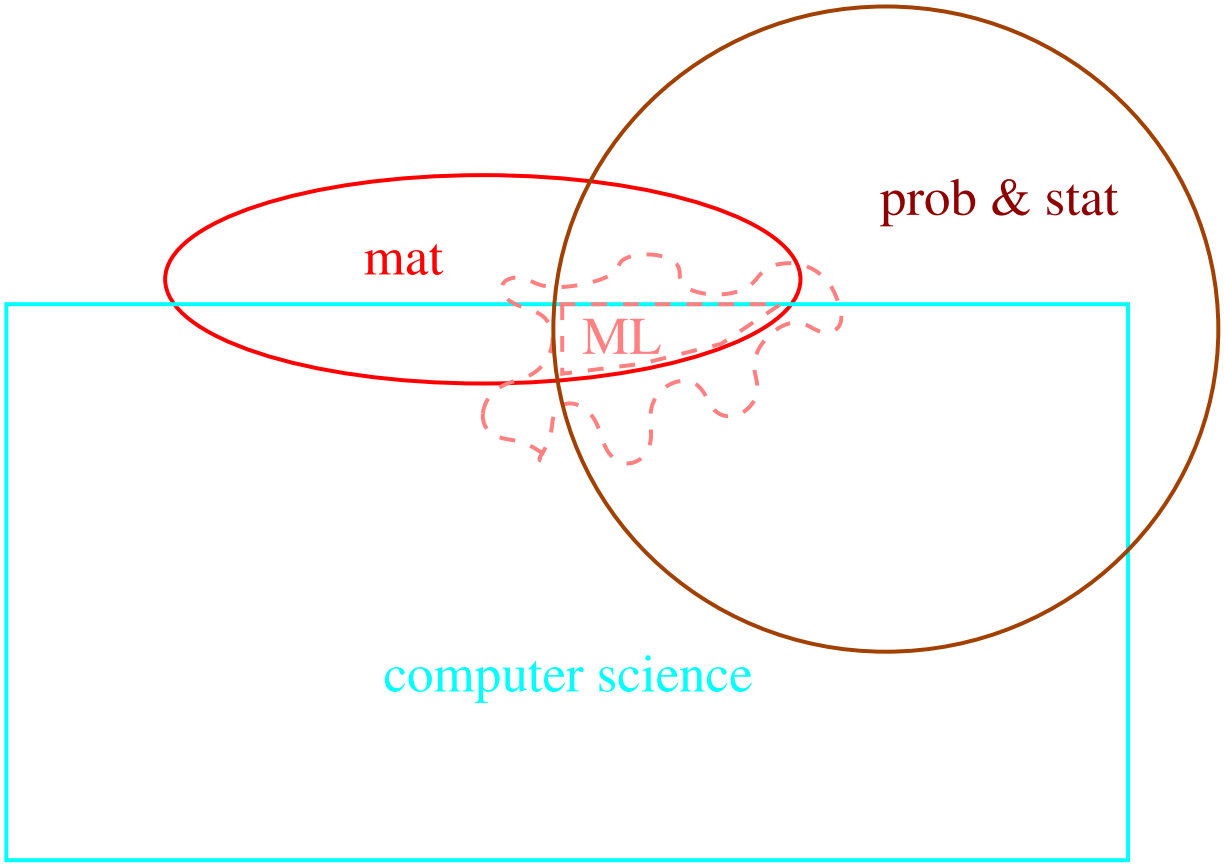
Aula 1. Fragmentação

Uma citação motivadora

Vladimir Vapnik

“Statistical learning theory does not belong to any specific branch of science: It has its own goals, its own paradigm, and its own techniques. Statisticians (who have their own paradigm) never considered this theory as part of statistics”.

Localização e tamanho do assunto



2017:

65,000 artigos publicados em ML + NN,

120,000 em toda matemática (MathSciNet).

O que este mini-curso é (e não é)

Modelo matemático de aprendizagem.

Noções matemáticas fundamentais neste contexto:

- regra de aprendizagem,
- aprendizagem provavelmente aproximadamente correta (PAC),
- dimensão de Vapnik–Chervonenkis (VC),
- classe de Glivenko–Cantelli, ...

Paradigmas de aprendizagem:

- dentro da classe,
- consistência universal

notas de curso: https://arxiv.org/abs/1910.06820

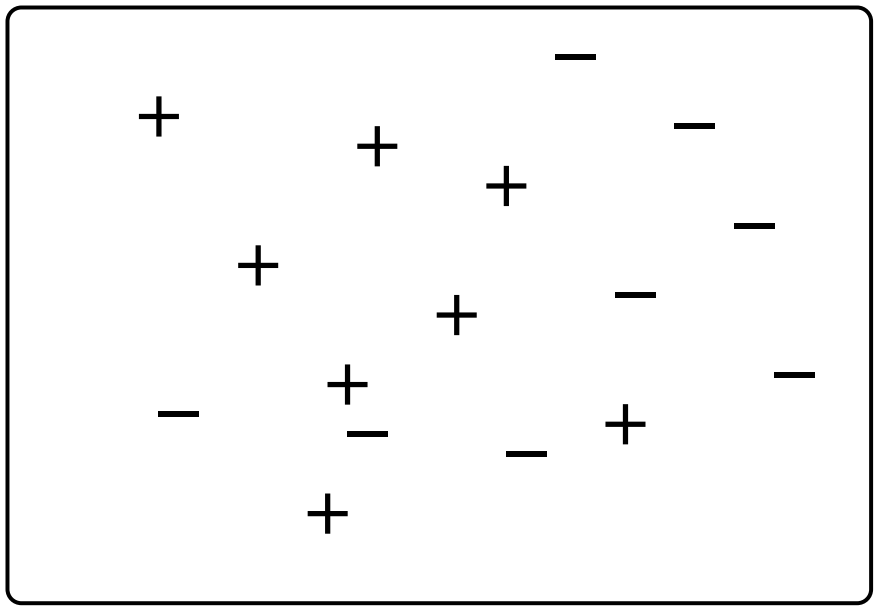
Cenário para criar e analizar novos algoritmos.

Implementações concretas: não.

Plano de curso (quão realista?)

- Fragmentação (shattering)
- Concentração de medida (? - talvez, já a Lei dos Grandes Números bastaria)
- Teorema de Benedek-Itai
- Classes de Glivenko-Cantelli
- Classificador k-NN, consistência universal
- Aproximação universal
- Compressão amostral

Problema de classificação binária



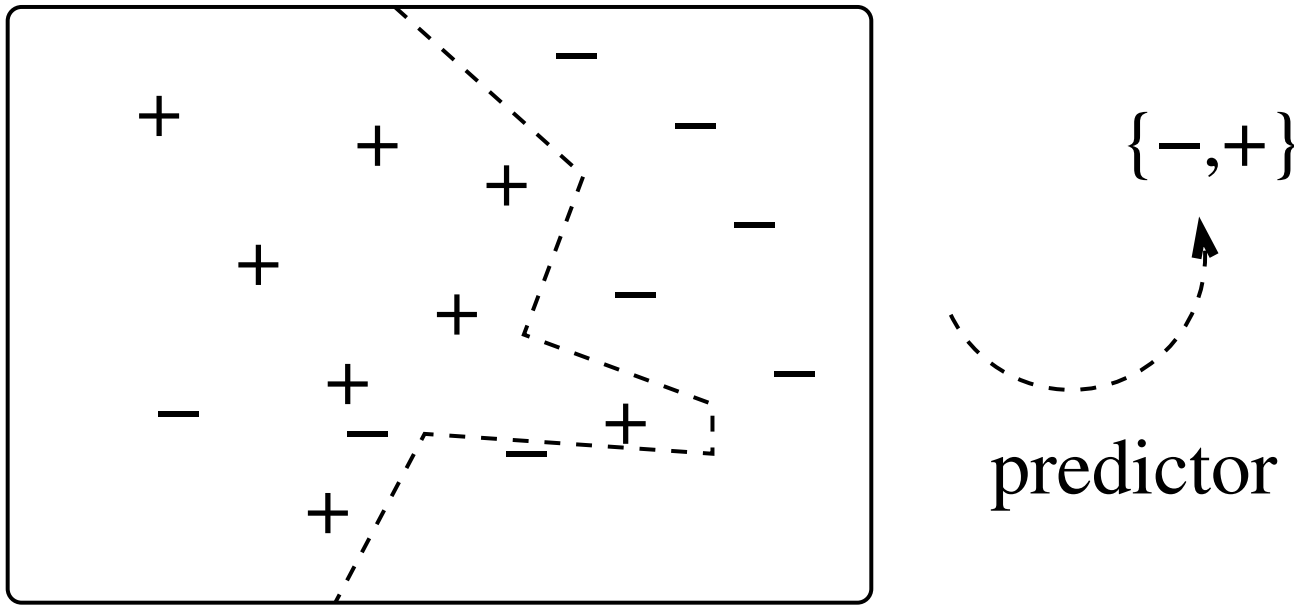
Os dados (pontos de *domínio*) divididos em duas classes (*amostra rotulada*)

Fragmente de um conjunto de dados da CSDM'2013 para deteção de intrusos na rede

39672	1.09	-0.03	-0.09	-0.49	-0.05	-0.15	1.11	1
39673	1.09	-0.03	-0.09	-0.49	-0.05	-0.15	1.11	1
39674	1.09	-0.03	-0.08	-0.49	-0.05	-0.15	-1.08	-1
39675	1.09	-0.03	-0.09	-0.49	-0.05	-0.15	1.11	1
39676	1.09	-0.03	-0.09	-0.49	-0.05	-0.15	1.11	1
39677	1.09	-0.03	-0.09	-0.49	-0.05	-0.15	1.11	1
39678	-1.00	-0.03	-0.09	-0.49	-0.05	-0.15	-1.08	1
39679	1.09	-0.03	-0.09	-0.49	-0.05	-0.15	1.11	1

$\sigma \in \mathbb{R}^7 \times \{0, 1\}$,
$n = |\sigma| = 77,959$,
incluindo $71,758$ sessões normais $(+1)$ e $6,201$ sessões ataque (-1)

Problema de classificação binária

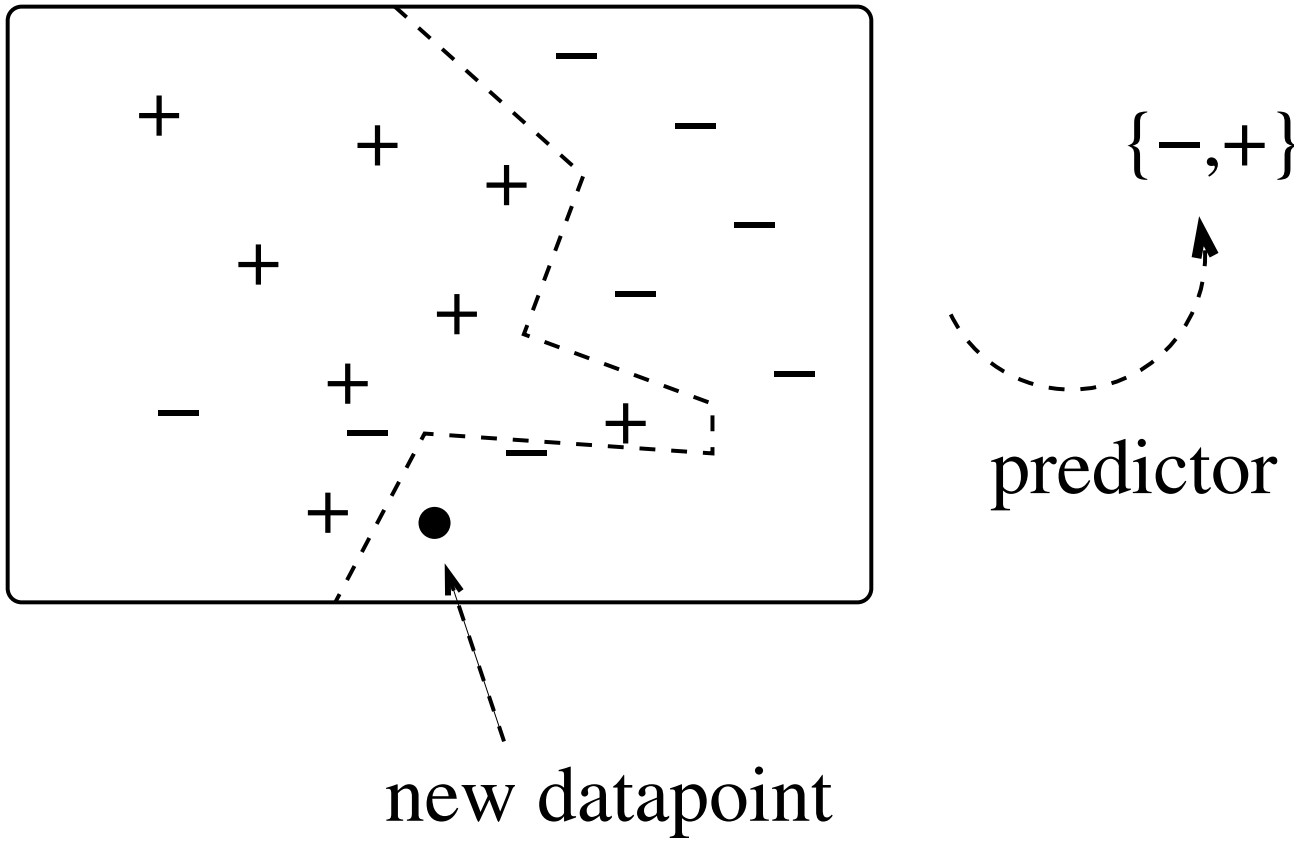


construir uma função binária

$$T : [0,1]^2 \to \{0,1\}$$

(*classificador* / *preditor* / *função de transferência*)

Problema de classificação binária



capaz de predizer com alta confiança o rótulo de novos pontos

(Aprendizagem automática estatística *supervisionada*)

Rotulagens, conceitos, hipóteses

Ω, um domínio (conjunto)

$\sigma = (x_1, x_2, \ldots, x_n) \in \Omega^n$ uma amostra (não rotulada)

Rotulagem de σ: uma sequência de rótulos,

$$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \in \{0, 1\}$$

Conjunto $\{0, 1\}^n$ de todas as rotulagens possíveis de σ é o *cubo de Hamming de posto n*

$C \subseteq \Omega$ um *conceito* (desconhecido, a ser aprendido)

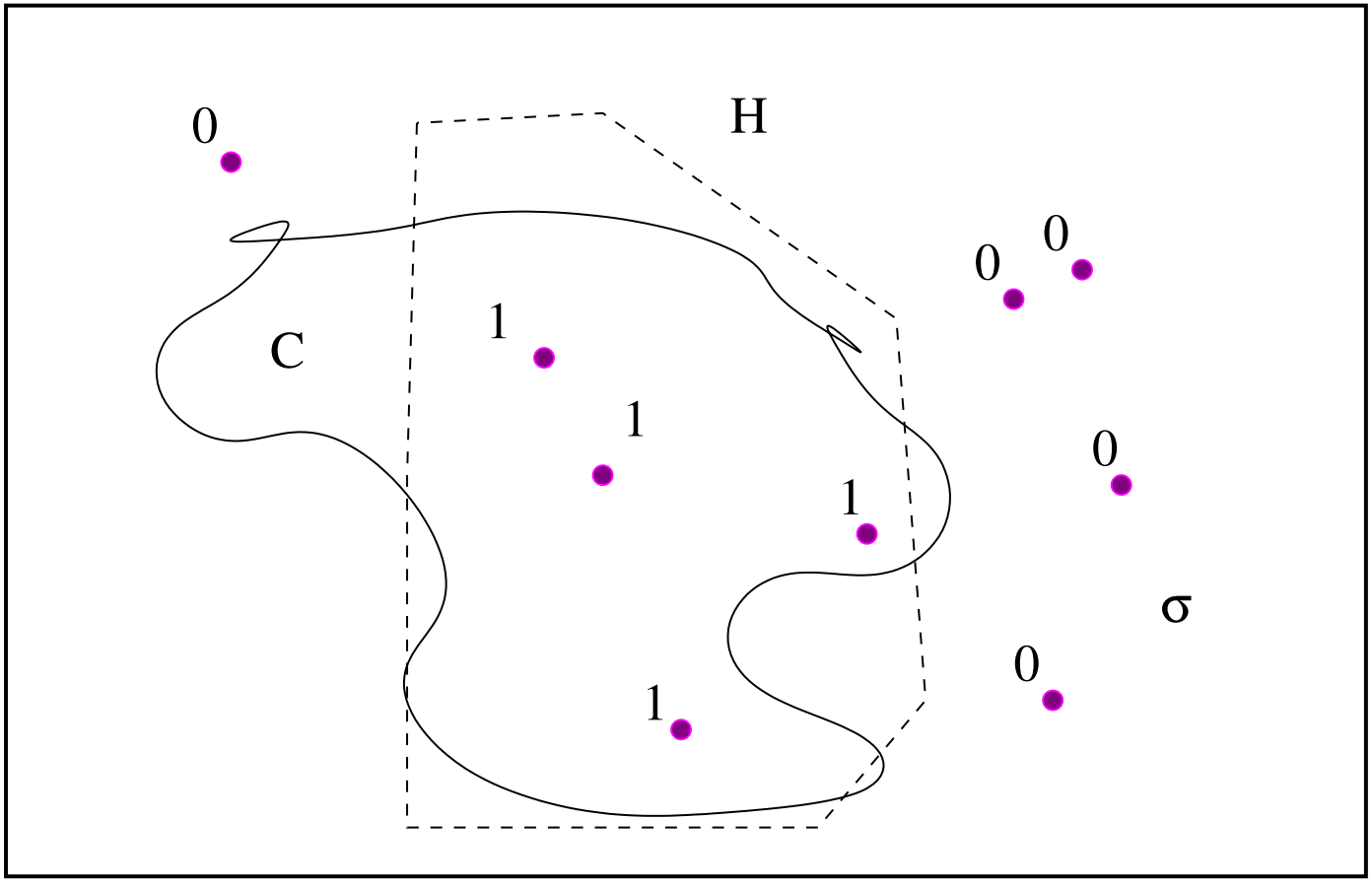
O conceito C gera uma rotulagem, $C \restriction \sigma$ (ou $C \cap \sigma$):

$$\varepsilon_1 = \chi_C(x_1), \varepsilon_2 = \chi_C(x_2), \ldots, \varepsilon_n = \chi_C(x_n)$$

Amostra rotulada: $(x_1, x_2, \ldots, x_n, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$.

Tarefa: adivinhar C a partir da amostra rotulada, gerando uma *hipótese*, $H \subseteq \Omega$.

Rotulagens, conceitos, hipóteses



Classes de conceitos

Ω, um domínio (conjunto)

$C \subseteq \Omega$ é um *conceito*

Conceitos, $C \leftrightarrow$ funções binárias, $T = \chi_C$

Há conceito desconhecido, C, a ser aprendido

Dada uma amostra, $x_1, x_2, \ldots, x_n$, o conceito C gera uma rotulagem, $\mathscr{C} \upharpoonright \sigma$:

$$\varepsilon_1 = \chi_C(x_1), \varepsilon_2 = \chi_C(x_2), \ldots, \varepsilon_n = \chi_C(x_n)$$

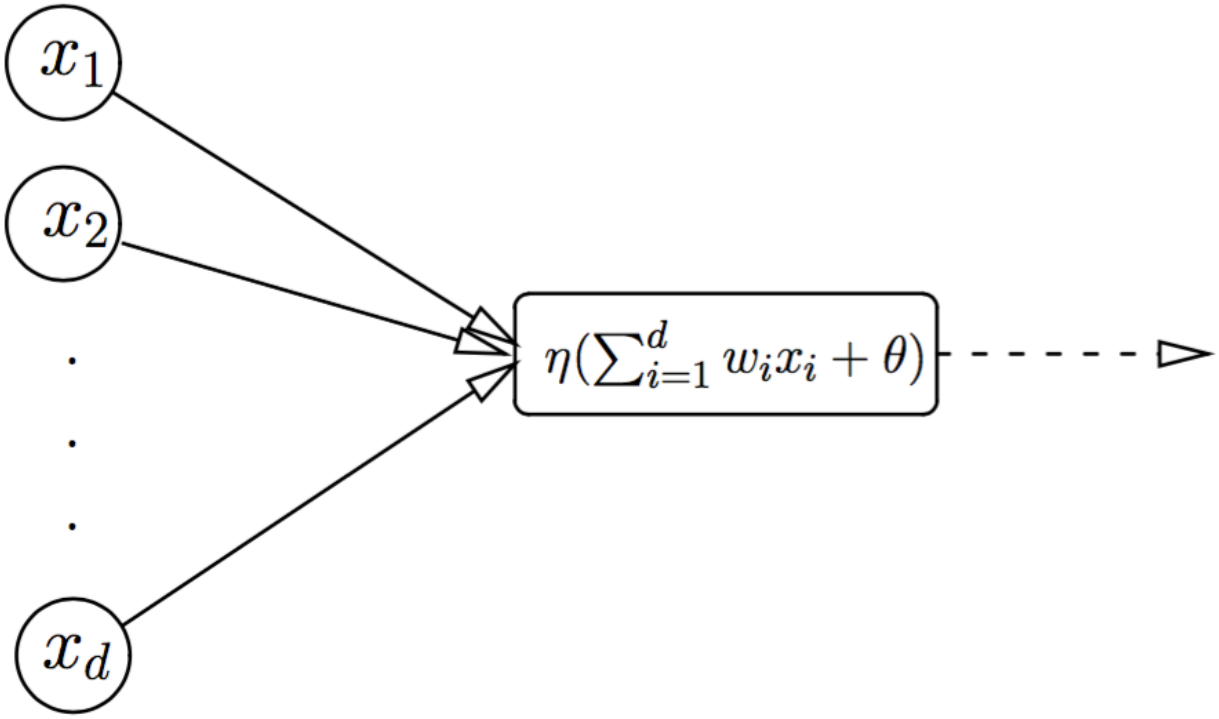
Um algoritmo de aprendizagem produz uma família de classificadores / conceitos, que formam uma *classe de conceitos,* $\mathscr{C}$ (*concept class*).

O cenário para hoje: uma classe de conceitos, $\mathscr{C} \subseteq 2^\Omega$, uma amostra $\sigma = (x_1, x_2, \ldots, x_n)$, e as rotulagens geradas por $\mathscr{C}$ sobre σ:

$$\mathscr{C} \upharpoonright \sigma \subseteq \{0,1\}^n$$

Perceptron

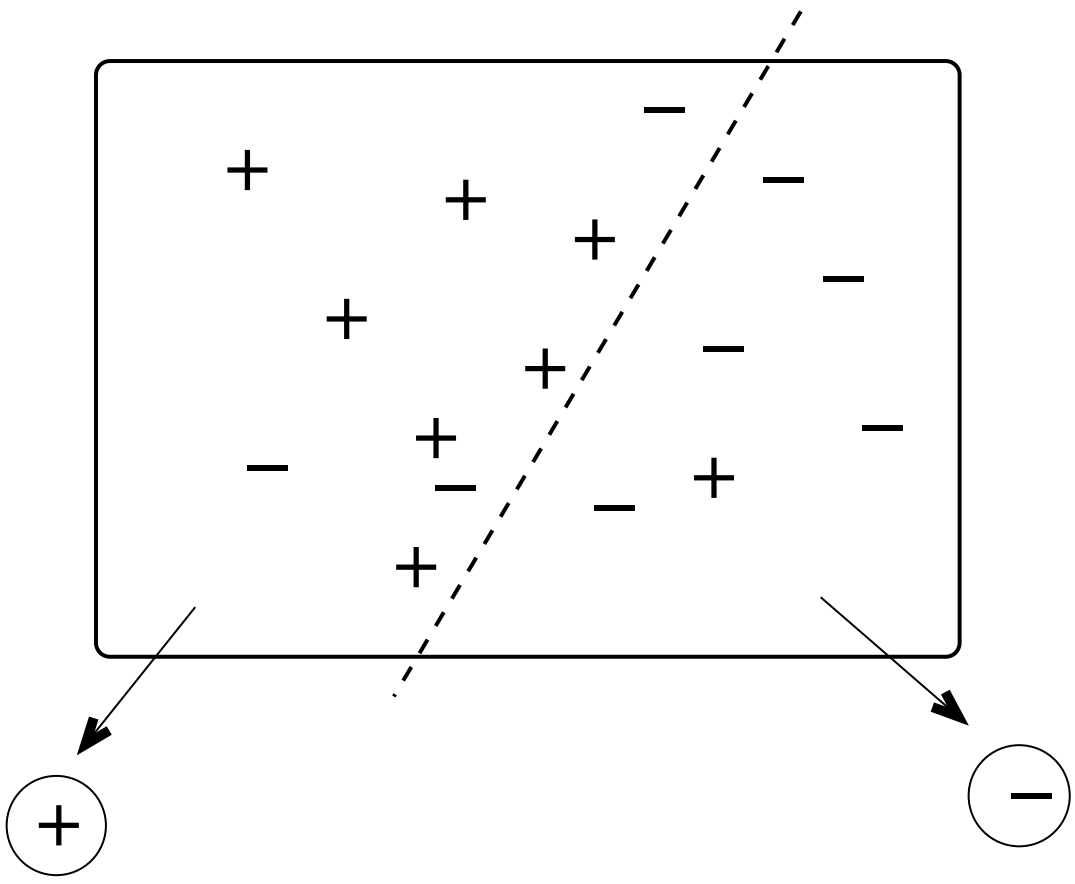
Família de classificadores sobre $\mathbb{R}^d$



w_i: pesos,
θ: parâmetro limiar;
η: função de Heaviside:

$$\eta(x) = \begin{cases} 1, & \text{se } x \geq 0, \\ 0, & \text{se } x < 0. \end{cases}$$

Perceptron

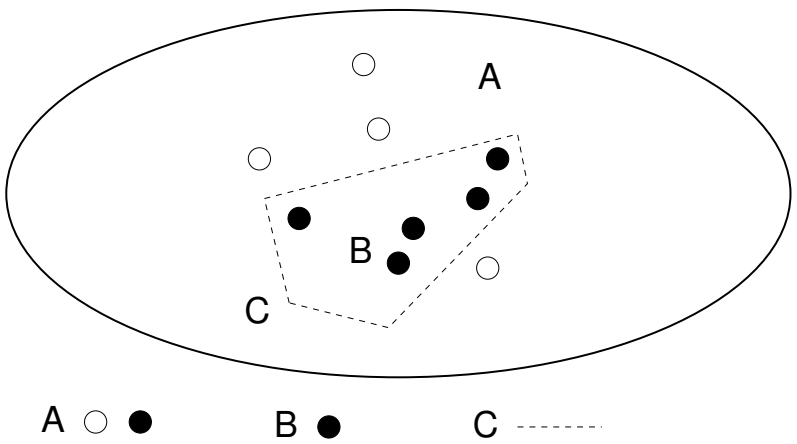


Perceptron realiza uma separação linear

Fragmentação (shattering)

Um subconjunto finito $A \subseteq \Omega$ é *fragmentado* (shattered) por uma classe de conceitos $\mathscr{C}$, se

$$\mathscr{C} \restriction A = \{0,1\}^A.$$



$$\forall B \subseteq A \; \exists C \in \mathscr{C} \;\; C \cap A = B$$

Dimensão de Vapnik–Chervonenkis de $\mathscr{C}$:

VC-dim($\mathscr{C}$), o supremo de cardinalidades de subconjuntos finitos, A, fragmentados por $\mathscr{C}$.

Dimensão de Vapnik–Chervonenkis

Classe de um conceito só

$\Omega \neq \emptyset$

$\mathscr{C} = \{C\}$.

O conjunto vazio é fragmentado por $\mathscr{C}$:

$$\mathscr{C} \upharpoonright \emptyset = \{\emptyset\} = 2^{\emptyset}$$

(Todas rotulagens possíveis sobre $\emptyset$ — ou seja, a única rotulagem, vazia – podem ser geradas por $\mathscr{C}$...)

Logo, VC-dim$(\mathscr{C}) \geq 0$.

Nenhum conjunto unitário é fragmentado:

$$\mathscr{C} \upharpoonright \{x\} = \{\emptyset\} \text{ ou } \{\{x\}\},$$

nunca $\{\emptyset, \{x\}\}$.

$\therefore$ VC-dim$(\mathscr{C}) = 0$.

Dimensão de Vapnik–Chervonenkis

Classe de dois conceitos

$\Omega \neq \emptyset$

$\mathscr{C} = \{\Omega, \emptyset\}$

- qualquer conjunto unitário $\{x\}$ é fragmentado por $\mathscr{C}$:

$$\mathscr{C} \restriction \{x\} = \{\emptyset, \{x\}\}$$

$\therefore$ VC-dim$(\mathscr{C}) \geq 1$

- ao mesmo tempo, nenhum conjunto com dois pontos $\{x, y\}$, $x \neq y$, é fragmentado: e.g. $\{x\} \notin \mathscr{C} \restriction \{x, y\}$

$\therefore$ VC-dim$(\mathscr{C}) \leq 1$

$$* * *$$

Mais geralmente, uma classe finita satisfaz
VC-dim$(\mathscr{C}) \leq \log_2 \sharp\mathscr{C}$

Dimensão de Vapnik–Chervonenkis

Classe de intervalos finitos em $\mathbb{R}$

$\Omega = \mathbb{R}$

$\mathscr{C} = \{[a, b] : a, b \in \mathbb{R},\ a \leq b\}$

E.g., $\{0, 1\}$ é fragmentado por $\mathscr{C}$:

$$\emptyset = \{0, 1\} \cap [3, 4],\ \{0\} = \{0, 1\} \cap [-1, 0], ...$$

$\therefore$ VC-dim$(\mathscr{C}) \geq 2$

Ao mesmo tempo, *nenhum* conjunto com três pontos é fragmentado: se $a < b < c$, então $\{a, c\} \neq \{a, b, c\} \cap [x, y]$, quaisquer que sejam x, y.

$\therefore$ VC-dim$(\mathscr{C}) \leq 2$

Dimensão de Vapnik–Chervonenkis

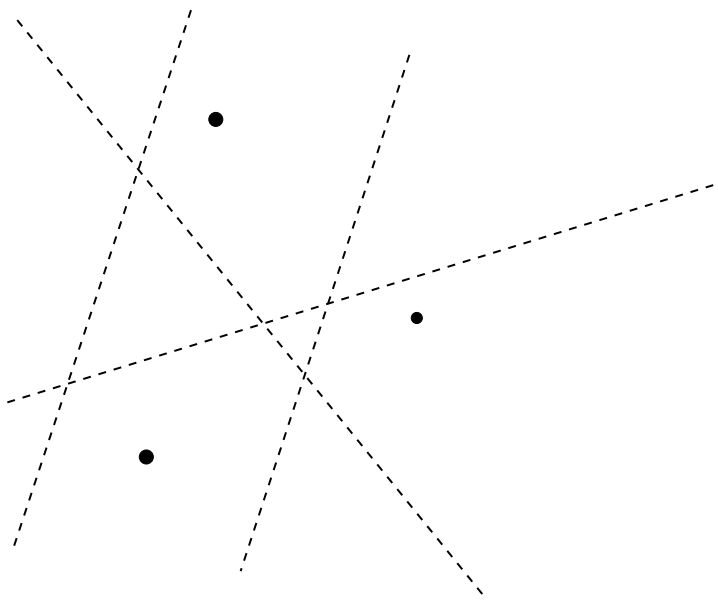
Semi-planos fechados em $\mathbb{R}^2$

$\Omega = \mathbb{R}^2$

$\mathscr{C}$ consiste de todos os semi-planos fechados:

$$H \equiv H_{\vec{v},b} = \{\vec{x} \in \mathbb{R}^2 : \langle \vec{x}, \vec{v} \rangle \geq b\}, \quad \vec{v} \in \mathbb{R}^2, \quad b \in \mathbb{R}$$

Sejam $\{a, b, c\}$ quaisquer, não colineares:



$\therefore$ VC-dim$(\mathscr{C}) \geq 3$

Dimensão de Vapnik–Chervonenkis

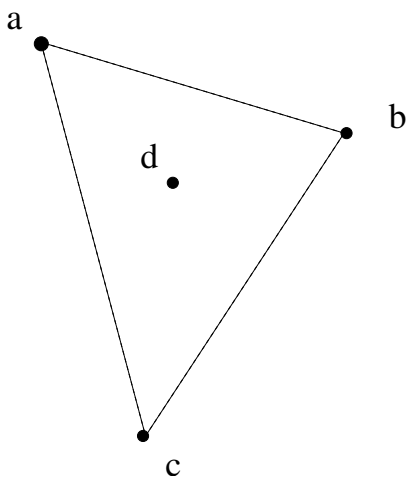
Semi-planos fechados em $\mathbb{R}^2$

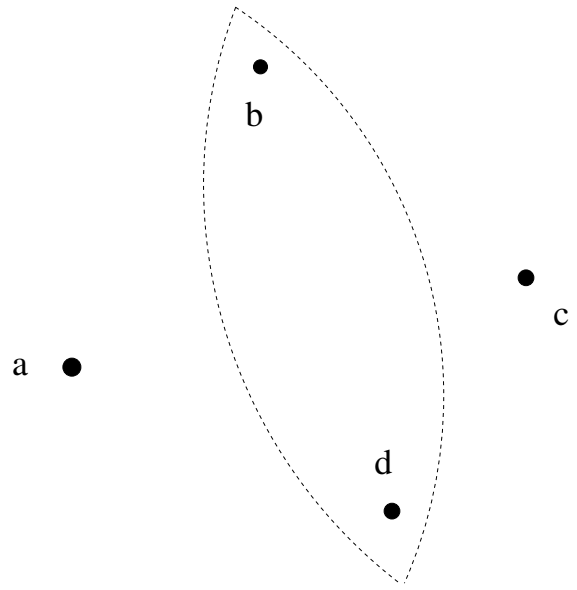
$\Omega = \mathbb{R}^2$

$\mathscr{C}$ consiste de todos os semi-planos fechados:

$$H \equiv H_{\vec{v},b} = \{\vec{x} \in \mathbb{R}^2 : \langle \vec{x}, \vec{v} \rangle \geq b\}, \quad \vec{v} \in \mathbb{R}^2, \quad b \in \mathbb{R}$$

Nenhum conjunto com 4 pontos é fragmentado. Dois casos:





$\therefore$ VC-dim$(\mathscr{C}) \leq 3$

Dimensão de Vapnik–Chervonenkis

Semi-espaços fechados em $\mathbb{R}^d$ (perceptron com d inputs)

$\Omega = \mathbb{R}^d$

$\mathscr{C}$ consiste de todos os semi-espaços fechados:

$$H \equiv H_{\vec{w},b} = \{\vec{x} \in \mathbb{R}^d : \langle \vec{x}, \vec{w} \rangle \geq b\}, \quad \vec{v} \in \mathbb{R}^d, \quad b \in \mathbb{R}$$

A classe gerada pelo perceptron com d inputs:

$$x \mapsto \eta(\langle x, w \rangle + b) \in \{0, 1\}.$$

VC-dim$(\mathscr{C}) = d + 1$. Várias provas, a mais simples segue-se de um argumento algébrico:

Dimensão VC e dimensão vetorial

teorema. Para uma função $f\colon \Omega \to \mathbb{R}$, denotemos

$$P_f = \{x \in \Omega\colon f(x) \geq 0\}.$$

Seja V um sub-espaço vetorial de $\mathbb{R}^\Omega$. Então,

$$\text{VC-dim}\{P_f\colon f \in V\} \leq d = \dim_{\mathbb{R}} V.$$

$\triangleleft$ Dado $x \in \Omega$,

$$\hat{x}(f) = f(x)$$

é um funcional linear sobre V. Sejam $x_1, x_2, \ldots, x_d, x_{d+1} \in \Omega$ distintos, fragmentados por P_f, $f \in V$. Pode supor que, no espaço V^*,

$$\hat{x}_{d+1} = \sum_{i=1}^{d} \lambda_i \hat{x}_i.$$

Seja $f \in V$ t.q. $f(x_i) \geq 0 \iff \lambda_i \geq 0$. Então,
$f(x_{d+1}) = \hat{x}_{d+1}(f) \geq 0$, e $x_1, \ldots, x_{d+1}$ não é fragmentado. $\triangleright$

Dimensão de Vapnik–Chervonenkis

Semi-espaços fechados em $\mathbb{R}^d$ (perceptron com d inputs)

$\Omega = \mathbb{R}^d$

$\mathscr{C}$ consiste de todos os semi-espaços fechados:

$$H \equiv H_{\vec{w},b} = \{\vec{x} \in \mathbb{R}^d : \langle \vec{x}, \vec{w} \rangle \geq b\}, \quad \vec{v} \in \mathbb{R}^d, \quad b \in \mathbb{R}$$

A classe gerada pelo perceptron com d inputs:

$$x \mapsto \eta(\langle x, w \rangle + b) \in \{0, 1\}.$$

Temos:

$$H_{\vec{w},b} = P_{\langle \vec{x}, \vec{w} \rangle - b}$$

O espaço de funções afins sobre $\mathbb{R}^d$ tem dimensão $d + 1$, concluimos: VC-dim$(\mathscr{C}) \leq d + 1$.

É fácil verificar que $0, e_1, e_2, \ldots, e_d$ é fragmentado pelos semi-espaços.

Teorema de Pajor

teorema: *Seja $\mathscr{C}$ uma classe de conceitos com m elementos, $m \geq 1$. Então $\mathscr{C}$ fragmenta pelo menos m subconjuntos de Ω dois a dois diferentes.*

Prova: indução em m.

$m = 1$: a classe contém um conceito só, e fragmenta o conjunto vazio.

Suponha que a afirmação seja válida para $1 \leq i \leq m$. Seja $\sharp\mathscr{C} = m + 1$. Então, existe $x_0 \in \cup\mathscr{C} \setminus \cap\mathscr{C}$.

$$\mathscr{C}_0 = \{A \in \mathscr{C} : A \ni x_0\}, \quad \sharp\mathscr{C}_0 = k \geq 1,$$

$$\mathscr{C}_1 = \{B \in \mathscr{C} : B \not\ni x_0\}, \quad \sharp\mathscr{C}_1 = \ell \geq 1, \quad k + \ell = m + 1.$$

Segundo a hipótese, existem

$$\underbrace{A_1, A_2, \ldots, A_k}_{\text{distintos, fragmentados por } \mathscr{C}_0} \quad , \quad \underbrace{B_1, B_2, \ldots, B_l}_{\text{distintos, fragmentados por } \mathscr{C}_1}$$

Teorema de Pajor /2

teorema: *Seja $\mathscr{C}$ uma classe de conceitos com m elementos, $m \geq 1$. Então $\mathscr{C}$ fragmenta pelo menos m subconjuntos de Ω dois a dois diferentes.*

$$\mathscr{C}_0 = \{A \in \mathscr{C} : A \ni x_0\}, \quad \sharp\mathscr{C}_0 = k,$$

$$\mathscr{C}_1 = \{B \in \mathscr{C} : B \not\ni x_0\}, \quad \sharp\mathscr{C}_1 = \ell.$$

Segundo a hipótese, existem

$$\underbrace{A_1, A_2, \ldots, A_k}_{\text{distintos, fragmentados por } \mathscr{C}_0} \quad , \quad \underbrace{B_1, B_2, \ldots, B_l}_{\text{distintos, fragmentados por } \mathscr{C}_1}$$

Suponha $A_i = B_j$, ou seja, fragmentado por $\mathscr{C}_0$ e por $\mathscr{C}_1$.

Logo, $A_i \cup \{x_0\} = B_j \cup \{x_0\}$ é fragmentado por $\mathscr{C} = \mathscr{C}_0 \cup \mathscr{C}_1$ (mas não por $\mathscr{C}_0$ nem por $\mathscr{C}_1$)

Substituímos B_j por $B_j \cup \{x_0\}$. Etc. □

Lema de Sauer–Shelah

teorema: Suponha VC-dim($\mathscr{C}$) $\leq d$, e seja σ uma amostra com n elementos. Então, o número de rotulagens diferentes induzidas sobre σ por $\mathscr{C}$ satisfaz

$$\sharp(\mathscr{C} \restriction \sigma) \leq \sum_{i=0}^{d} \binom{n}{i} \quad \left(= \sharp[\sigma]^{\leq d}\right)$$
$$< \left(\frac{en}{d}\right)^{d}.$$

◁ Caso contrário, segundo t. de Pajor, o número de subconjuntos de σ fragmentados por $\mathscr{C}$ é maior que a cardinalidade da família de todos os conjuntos com $\leq d$ elementos, logo existe um conjunto com $d+1$ elementos fragmentado por $\mathscr{C}$, logo VC-dim($\mathscr{C}$) $> d$. ▷

Segunda estimativa: usa-se a desigualdade de Euler, $\left(1+\frac{a}{x}\right)^{x} < e^{a}$ $(x > 0)$.

Lema de Sauer–Shelah /2

Para todo $0 \leq i \leq d$, temos

$$\left(\frac{n}{d}\right)^d \left(\frac{d}{n}\right)^i = \left(\frac{n}{d}\right)^{d-i} \geq 1,$$

e por conseguinte,

$$\begin{aligned}
\sum_{i=0}^{d} \binom{n}{i} &\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \\
&\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^i \\
&= \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \\
&< \left(\frac{n}{d}\right)^d e^d \\
&= \left(\frac{en}{d}\right)^d .
\end{aligned}$$

Coeficientes de fragmentação

n-ésimo coeficiente de fragmentação (*n-th shattering coefficient*) de uma classe $\mathscr{C}$ é o maior número de rotulagens induzidas por $\mathscr{C}$ sobre n-amostras:

$$s(n, \mathscr{C}) = \sup\{\sharp \mathscr{C}|_\sigma : \sigma \subseteq \Omega,\ \sharp\sigma \leq n\}.$$

Por exemplo, VC-dim$(\mathscr{C}) = \sup\{n\colon\ s(n, \mathscr{C}) = 2^n\}$.

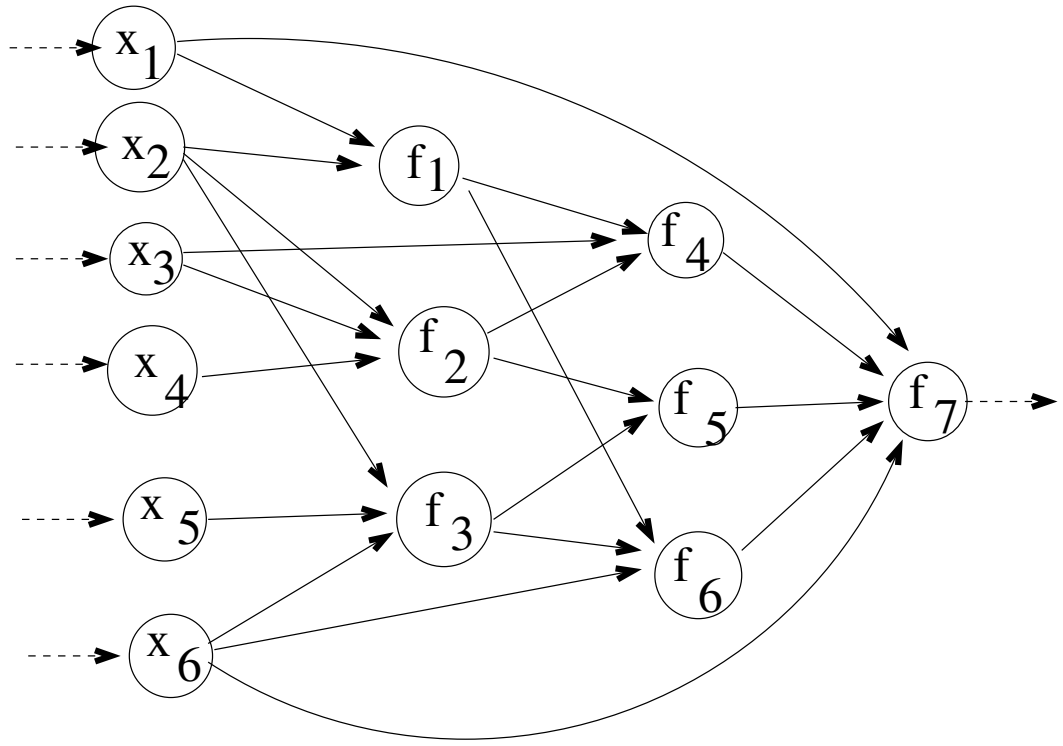
Lema de Sauer–Shelah:

$$s(n, \mathscr{C}) \leq \sum_{i=0}^{d} \binom{n}{i} < \left(\frac{en}{d}\right)^d.$$

A primeira desigualdade é exata (exercício)

Redes de unidades computacionais

Estrutura mais geral do que ANNs



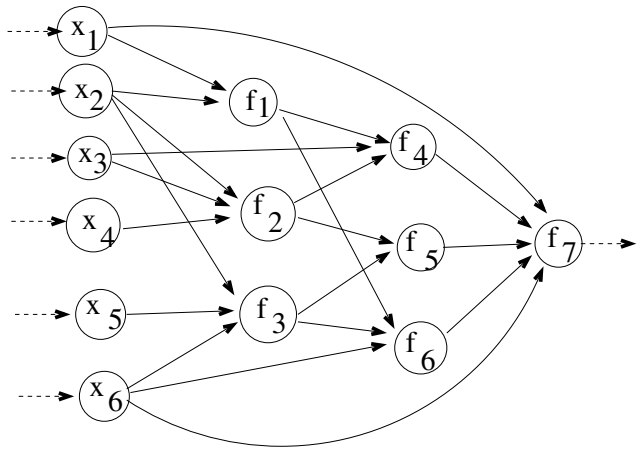
Um grafo dirigido, sem ciclos.

1a camada: entrada, inputs $=$ elementos de $\mathbb{R}^d = \Omega$.

Outras camadas: unidades computacionais (funções binárias dependendo de parâmetros, por exemplo, perceptrons).

Última camada: única unidade, a de saída (0 ou 1).

Redes de unidades computacionais



teorema. Se a rede $\mathscr{N}$ tem k unidades computacionais, $W = \sum_u \text{VC-dim}(u)$, então para cada n,

$$s(\mathscr{N}, n) \leq \left(\frac{enk}{W}\right)^W,$$

e

$$\text{VC-dim}(\mathscr{N}) \leq 2W \log_2\left(\frac{2k}{\log 2}\right) = O(W \log k)$$

Coeficientes de fragmentação de $\mathscr{N}$ /1

Escolhemos uma ordem total entre unidades,

$$u_1, u_2, \ldots, u_k,$$

de modo que se existe conexão $u_i \to u_j$, então $i < j$.

Estado ω da rede: totalidade de parâmetros.

Fixemos uma amostra, $\sigma = (x_1, x_2, \ldots, x_n)$, $x_j \in \Omega = \mathbb{R}^d$.

Relação de equivalência $\omega \overset{i}{\sim} \omega'$: para cada input x_j, $j = 1, \ldots, n$, unidades $u_1, \ldots, u_i$ produzem mesmos valores.

$\sharp$ (classes mod $\overset{1}{\sim}$) $\leq s(u_1, n) \leq (en/d_1)^{d_1}$

Coeficientes de fragmentação de $\mathscr{N}$ /2

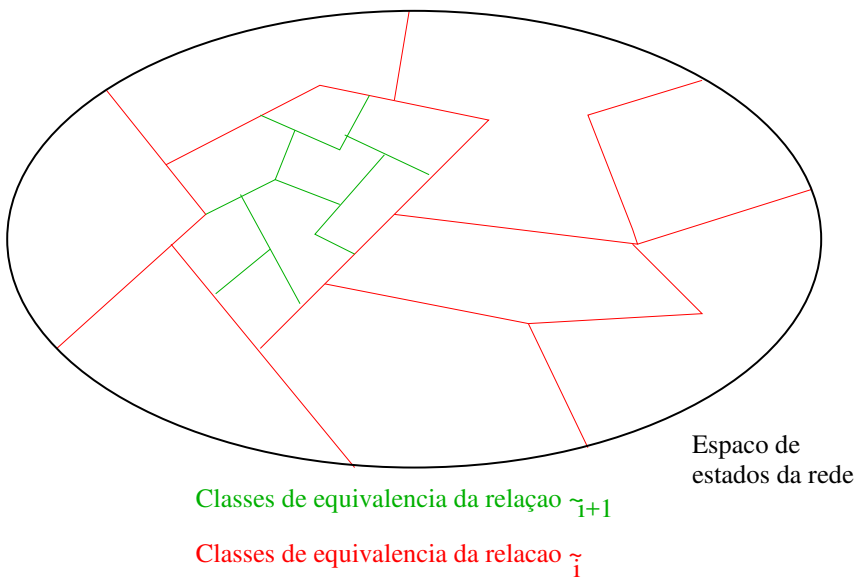
Unidades, $u_1, u_2, \ldots, u_k$, se $u_i \to u_j$, então $i < j$.

Estado ω da rede: totalidade de parâmetros.

Fixemos uma amostra, $\sigma = (x_1, x_2, \ldots, x_n)$, $x_j \in \Omega = \mathbb{R}^d$.

Relação de equivalência $\omega \overset{j}{\sim} \omega'$: para cada input x_j, $j = 1, \ldots, n$, unidades $u_1, \ldots, u_i$ produzem mesmos valores.

$\sharp\,(\text{classes} \mod \overset{1}{\sim}) \leq s(u_1, n) \leq (en/d_1)^{d_1}$



$\sharp\,(\text{classes} \mod \overset{i+1}{\sim})\ \sharp\,(\text{classes} \mod \overset{i}{\sim}) \times (en/d_{i+1})^{d_{i+1}}$

Coeficientes de fragmentação de $\mathscr{N}$ /3

Escolhemos uma ordem total entre unidades,

$$u_1, u_2, \ldots, u_k,$$

de modo que se existe conexão $u_i \to u_j$, então $i < j$.

Fixemos uma amostra, $\sigma = (x_1, x_2, \ldots, x_n)$, $x_j \in \Omega = \mathbb{R}^d$.

Relação de equivalência $\omega \overset{i}{\sim} \omega'$: para cada input x_j, $j = 1, \ldots, n$, unidades $u_1, \ldots, u_i$ produzem mesmos valores.

$\sharp$ (classes mod $\overset{1}{\sim}$) $\leq s(u_1, n) \leq (en/d_1)^{d_1}$

$\sharp$ (classes mod $\overset{i+1}{\sim}$) $\sharp$ (classes mod $\overset{i}{\sim}$) $\times (en/d_{i+1})^{d_{i+1}}$

n-ésimo coeficiente de fragmentação da rede $\leq \sharp$ classes de equivalência mod $\overset{k}{\sim}$, $s(\mathscr{N}, n) \leq \prod_{i=1}^{k} (en/d_i)^{d_i}$

$$\therefore \log s(\mathscr{N}, n) \leq \sum_{i=1}^{k} d_i \log\left(\frac{en}{d_i}\right).$$

Entropia de Claude Shannon

Seja X uma variável aleatória, com valores $x_1, x_2, \ldots, x_n$ e probabilidades p_i. A *entropia* de X é a quantidade

$$H(X) = \sum_{i=1}^{n} -p_i \log p_i.$$

lema. $H(X) \leq \log n$, atingido sobre a distribuição uniforme:

$$p_i = \frac{1}{n},\ i = 1, 2, \ldots, n.$$

◁ O logaritmo é uma função côncava. Logo, para qualquer que seja a coleção $\lambda_i > 0$, $i = 1, 2, \ldots, n$,

$$\log\left(\sum_{i=1}^{n} p_i \lambda_i\right) \geq \sum_{i=1}^{n} p_i \log(\lambda_i).$$

No caso $\lambda_i = 1/p_i$,

$$\log n \geq \sum_{i=1}^{n} p_i \log\left(\frac{1}{p_i}\right) = H(X) \quad \triangleright.$$

Coeficientes de fragmentação de $\mathscr{N}$ /4

$$
\begin{aligned}
\log s(\mathscr{N}, n) &\leq \sum_{i=1}^{k} d_i \log\left(\frac{en}{d_i}\right) \\
&= W \sum_{i=1}^{k} \frac{d_i}{W}\left[\log\frac{W}{d_i} + \log(en) - \log W\right] \\
&= W \cdot H(X) + W \log\frac{en}{W} \\
&\leq W \log k + W \log\frac{en}{W} \\
&= W \log\frac{enk}{W}.
\end{aligned}
$$

Dimensão de Vapnik–Chervonenkis de $\mathscr{N}$

VC-dim$(\mathscr{N}) \leq n \iff s(\mathscr{N}, n) \leq 2^n$, em particular, quando

$$\left(\frac{enk}{W}\right)^W \leq 2^n, \text{ ou seja, } n \geq W \log_2 \left(\frac{enk}{W}\right).$$

lema. Para todos $\alpha, x > 0$,

$$\log x \leq \alpha x - \log \alpha - 1,$$

com a igualdade se e apenas se $\alpha x = 1$. □

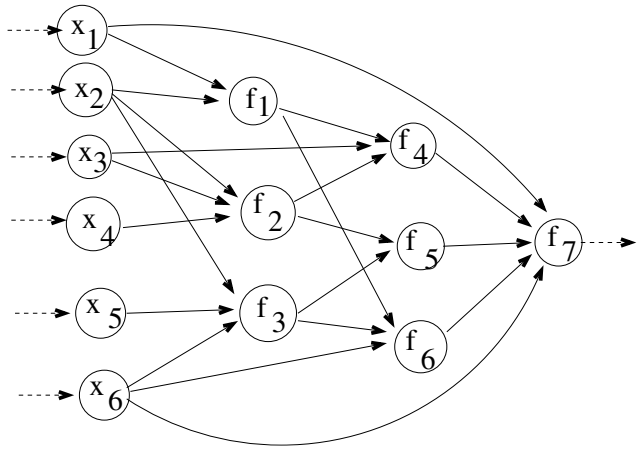
Apliquemos o lema com $x = \frac{enk}{W}$ e $\alpha = \frac{\log 2}{2ek}$:

$$\log \left(\frac{enk}{W}\right) \leq \frac{n \log 2}{2W} - \log \left(\frac{\log 2}{2ek}\right) - 1,$$

$$W \log_2 \left(\frac{enk}{W}\right) \leq \frac{n}{2} + W \log_2 \left(\frac{2k}{\log 2}\right). \quad \triangleright$$

Redes de unidades computacionais

Problema de “sobreajuste benigno” de DNNs



teorema. Se a rede $\mathscr{N}$ tem k unidades computacionais, $W = \sum_u \text{VC-dim}(u)$, então para cada n,

$$s(\mathscr{N}, n) \leq \left(\frac{enk}{W}\right)^W,$$

e

$$\text{VC-dim}(\mathscr{N}) \leq 2W \log_2\left(\frac{2k}{\log 2}\right) = O(W \log k)$$

Como a taxa de crescimento depende da geometria da DNN?

Precisa de uma análise mais fina.