Desigualdade de Hoeffding

$\xi_i \in \{0,1\}$ $\xi_1, \xi_2, \dots$ iid

$p = \mathbb{P}(\xi_1 = 1)$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i - p > \varepsilon\right) \leq e^{-2n\varepsilon^2}$$

$\Downarrow$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - p\right| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

$\downarrow$

$$\mathbb{P}\left(\left\{\frac{1}{n}\sum_{i=1}^{n}\xi_i > p+\varepsilon\right\} \cup \left\{\frac{1}{n}\sum_{i=1}^{n}\xi_i < p-\varepsilon\right\}\right)$$

$$\leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i > p+\varepsilon\right) + \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i < p-\varepsilon\right)$$

$\eta_1, \eta_2, \dots$ v.a iid

$\eta_i \in \{a,b\}$ onde $a<b$ em $\mathbb{R}$.

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\eta_i - \mathbb{E}(\eta_1)\right| > \varepsilon\right) \leq 2\exp\left\{-2n\frac{\varepsilon^2}{(b-a)^2}\right\}$$

Exercício

Solução:

Defino $\xi_i = \dfrac{\eta_i - a}{b-a}$ v.a iid

$$\mathbb{E}(\xi_i) = \mathbb{E}\left(\frac{\eta_i - a}{b-a}\right) = \frac{\mathbb{E}(\eta_i) - a}{b-a}$$

$$p = \mathbb{P}(\xi_i = 1) = \mathbb{P}(\eta_i = b)$$

$$\frac{1}{n}\sum_{i=1}^{n}\eta_i - \mathbb{E}(\eta_1) > \varepsilon$$

$$\frac{\sum_{i=1}^{n}(\eta_i - a) \overset{+na}{-} n\mathbb{E}(\eta_1)}{b-a} > \frac{n\varepsilon}{b-a}$$

$$\sum_{i=1}^{n} \left(\frac{x_i - a}{b-a} \right) + \frac{n\left[a - \mathbb{E}(x_1)\right]}{b-a} > \frac{n\varepsilon}{b-a}$$

$$\sum_{i=1}^{n} \xi_i - np > \frac{n\varepsilon}{b-a}$$

$$\mathbb{P}\left(\sum_{i=1}^{n} \xi_i - np > n\frac{\varepsilon}{b-a} \right) \leq e^{-2n\left(\frac{\varepsilon}{b-a}\right)^2}$$

Hoeffding vale para

$\xi_1, \xi_2, \ldots$ v.a iid com $\xi \in [a,b]$ então

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \xi_i - \mathbb{E}(\xi_1) \right| > \varepsilon \right) \leq 2e^{-n\left(\frac{\varepsilon}{b-a}\right)^2}$$

Teoria de Vapnik - Chervonenkis

$\mathcal{F}$ conjunto de classificadores

(X,Y) $\quad X \in \mathcal{X}$, $Y \in \{0,1\}$

queremos encontrar a relaçao entre X e Y a partir de uma amostra de treino

$D_n = (X_1,Y_1), \ldots, (X_n,Y_n)$ iid e tem a mesma lei que (X,Y)

$f \in \mathcal{F}$, risco de f $\quad R(f) = \mathbb{P}(f(X) \neq Y)$.

Risco empírico

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{f(X_i) \neq Y_i\}}$$

Usando Hoeffding

$$\mathbb{P}\left(|R(f) - R_n(f)| > \varepsilon \right) \leq 2e^{-2n\varepsilon^2}$$

$$\mathbb{P}\left(\bigcup_{f \in \mathcal{F}} |R_n(f) - R(f)| > \varepsilon \right)$$

Se $\mathcal{F}$ for finito ou numerável

$$\leq \sum_{f \in \mathcal{F}} \mathbb{P}\left(|R_n(f) - R(f)| > \varepsilon \right) = (*)$$

Se $\mathcal{F}$ for finito $(*) \leq 2|\mathcal{F}| e^{-2n\varepsilon^2}$

$\mathcal{F}$ enumerável

definir $\varepsilon(f)$ e ter

$$(*) \leq \sum_{f \in \mathcal{F}} \varepsilon(f) \leq \varepsilon \sum_{f \in \mathcal{F}} p(f) = \varepsilon$$

Se soubemos algo mais sobre (X, Y) talvez definir

$\varepsilon(f) = \varepsilon\, p(f)$ onde $0 \leq p(f) \leq 1$ e $\sum_{f \in \mathcal{F}} p(f) = 1$

Exemplo: VC para el exemplo seguinte.

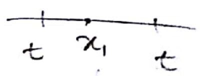
$\mathcal{X} = [0, 1]$

1) $\mathcal{F}_0 = \{ f : \mathcal{X} \to \{0,1\} : f(x) = 1_{\{x < t\}}$ para algún $t \in [0,1] \}$

2) $\mathcal{F}_1 = \{ f : \mathcal{X} \to \{0,1\} : f(x) = 1_{\{x < t\}}$ ou $f(x) = 1_{\{x \geq t\}}$, $t \in [0,1] \}$

1) Classifico bem todos os conjuntos de tamanho 1

$n = 1$ $\quad S(\mathcal{F}_0, 1) = 2^1$



$\Rightarrow VC(\mathcal{F}_0) = 1$

$n = 2$

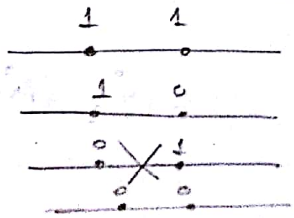
$S(\mathcal{F}_0, 2) = 3 < 2^2$

t maior que x_1 e x_2 $\longrightarrow$

t entre x_1 e x_2 $\longrightarrow$

? $\longrightarrow$

$t < \min\{x_1, x_2\}$ $\longrightarrow$



2) O exemplo que não dava para resolver antes com esta clase se pode

$f(x) = 1_{\{x \geq t\}}$ e tomo t entre x_1 e x_2, $x_1 < t < x_2$

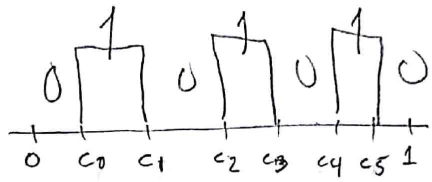
$S(\mathcal{F}_1, 1) = 1$

$S(\mathcal{F}_1, 2) = 4 \Rightarrow VC(\mathcal{F}_1) = 2$.

$S(\mathcal{F}_1, 3) < 2^3$

<u>Exercício</u>

Fixo $k \geq 1$, $\mathcal{X} = [0,1]$



$$\mathcal{F}_K = \left\{ f(x) = \sum_{i=0}^{K} 1_{\{c_{2i} \leq x \leq c_{2i+1}\}} : \text{onde } 0 \leq c_0 < c_1 < \dots < c_{2k+1} \leq 1 \right\}.$$

$S(\mathcal{F}_K, n) = ?$ $\qquad VC(\mathcal{F}_K) = ?$

Seleção estatística de uma árvore de contexto probabilística de altura $\leq k$. dada uma amostra. $\underbrace{X_{-k}, X_{-k+1}, \dots, X_n}$

amostra gerada por uma cadeia de alcance variável desconhecida.

$k \geq 1$ fixado

$k < \log n$

$$\tau \in \bigcup_{j=0}^{k} \mathcal{T}_j$$

$$\hat{\mathbb{P}}_{\tau}^{[j]}\left(X_0^n \mid \underbrace{X_{-k}^{-1}}_{\text{fixado}}\right) = \prod_{w \in \tau} \prod_{a \in A} \hat{P}_n^{[j]}(a|w)^{N_{0:n}(wa)} \qquad j \leq k$$

Teorema: Sobe a hipótese que amostra foi gerada por (τ, p)

$$\mathbb{P}\left(\bigcup_{a \in A} \bigcup_{w \in \tau} |\hat{P}_n^{\tau}(a|w) - p(a|w)| > \varepsilon \right) \leq |A|\,|\tau| \cdot \delta(n) \searrow_0$$

onde $\hat{p}_n^{\tau}(a|w) = \dfrac{N_{0:n}(wa)}{N_{0:n-1}(w)}$

$$\mathbb{P}\big(\underbrace{|\hat{p}_n(a|w) - p(a|w)| > \varepsilon}\big)^{\delta(n)}$$

Hoeffding
+
estimativa do valor do denominador $N_{0:n-1}(w)$.

Dado τ, sabemos estimar p

Problema: Cómo identificar τ ?

método: se a dependencia do pasado tem alcance k

[Hipotese nula]

então:

$$\mathbb{P}(X_n = b \mid X_{n-k}^{n-1} = a_{-k}^{-1}) =$$

$$= \mathbb{P}(X_n = b \mid X_{n-k}^{n-1} = a_{-k}^{-1}, \underbrace{X_{n-(k+1)} = z}_{\text{informação inútil}}) \quad \forall z$$

Por tanto:

$$\mathbb{P}\Big(\max_{a \in A} \max_{z, z' \in A} |\hat{p}_n(b|a_{-k}^{-1}z) - \hat{p}_n(b|a_{-k}^{-1}z')| > \varepsilon\Big) = \delta(n) \downarrow 0$$

$\xrightarrow{n \to \infty}$ $p(b|a_{-k}^{-1})$

$$\mathbb{P}\Big(|\underbrace{\hat{p}_n(b|a_{-k}^{-1}z) - p(b|a_{-k}^{-1}) + p(b|a_{-k}^{-1})}_{\text{somei e subtrai}} - \hat{p}_n(b|a_{-k}^{-1}z')| > \varepsilon\Big)$$

$$\leq \mathbb{P}\Big(|\hat{p}_n(b|a_{-k}^{-1}z) - p(b|a_{-k}^{-1})| + |p(b|a_{-k}^{-1}) - \hat{p}_n(b|a_{-k}^{-1}z')| > \varepsilon\Big)$$

$$\leq \mathbb{P}\left(\left| \hat{p}_n(b|a_{-k}^{-1} z) - p(b|a_{-k}^{-1}) \right| > \varepsilon/2 \right) +$$

$$\mathbb{P}\left(\left| p(b|a_{-k}^{-1}) - \hat{p}_n(b|a_{-k}^{-1} z') \right| > \varepsilon/2 \right)$$

Algoritmo contexto:

Entrada: Amostra

Saída: $(\hat{\tau}, \hat{p})$

Dada uma amostra $(X_{-k}, \dots, X_n)$ de símbolos no alfabeto A, para cada $w = (w_{-k}, \dots, w_{-k}) \in A^k$ defino

$$\Delta_n(w) = \max_{a,b \in A} \left| \hat{p}_n(a|w_{-k}^{-1}) - \hat{p}_n(a|w_{-k}^{-1} b) \right|$$

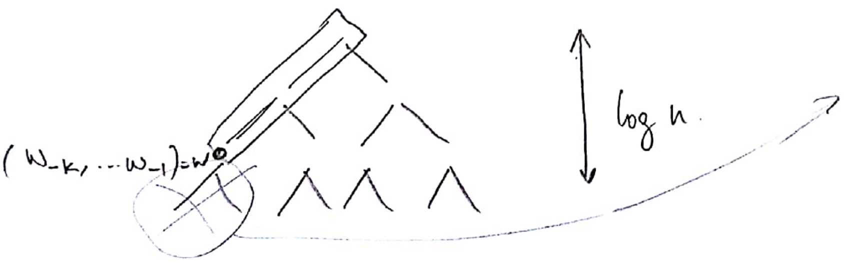
Fixo $\delta > 0$,

Para $k \leq \log n$

Se $\Delta(w) < \delta$, podamos os "filhos" de w (representados pela letra b)

Se $\Delta(w) > \delta$ mantemos as sequencias $b w_{-k}, \dots, w_{-1}$

Exemplo: $A = \{0,1\}$



Relevante? Sim: mantenho
Não: Podo.

pergunta: A informação referente ao paso $-(k+1)$ no pasado é pertimente?

isto é $\mathbb{P}(X_n = z \mid X_{n-k}^{n-1} = w_{-k}^{-1}, X_{n-(k+1)} = b)$

$$\overset{?}{=} \mathbb{P}(X_n = z \mid X_{n-k}^{n-1} = w_{-k}^{-1}).\ ?$$

1) Primeira questão da lista 3.

$$\mathbb{E}\left(2^{\sum_{m=1}^{n} \xi_m}\right) = \mathbb{E}\left(\prod_{m=1}^{n} 2^{\xi_m}\right) \overset{\text{indep}}{=} \prod_{m=1}^{n} \mathbb{E}\left(2^{\xi_m}\right)$$

$$\overset{\text{i.d}}{=} \underbrace{\mathbb{E}\left(2^{\xi_1}\right)}^{n} = (p+1)^n$$

$$\mathbb{E}\left(2^{\xi_1}\right) = 2^1\, \mathbb{P}(\xi_1 = 1) + 2^0\, \mathbb{P}(\xi_1 = 0) = 2 \cdot p + (1-p) = p+1$$