<u>Amostras</u>: de cadeias assumindo valores em A (A alfabeto finito)

$X_1^1 \; X_2^1 \; \ldots \; X_n^1 \; — \; \hat{\tau}_n^1$

$X_1^2 \; X_2^2 \; \ldots \; X_n^2 \; — \; \hat{\tau}_n^2$

$X_1^M, \; X_2^M, \; \ldots, \; X_n^M \; — \; \hat{\tau}_n^M$

Modelos que o algoritmo atribui a cada amostra

índice da amostra

tamanho da amostra

$\mathcal{M}$: clase de modelos

$\mathcal{A}$: algoritmo: amostra $\rightarrow$ modelo

<u>Perguntas</u>:

Há modelos <u>idênticos</u> em $\hat{\tau}_n^1, \ldots, \hat{\tau}_n^M$?
"próximos" $\longrightarrow$ Isto depende de uma distancia definida em $\mathcal{M}$

Cadeias estocástica com memória de alcance variavel
1983 - J Rissanen
"A universal system for data compression"
Minimal Desciption Length

$k \geq 1$

$$\mathcal{M}_k(A) = \left\{ p : A^k \times A \rightarrow [0,1] : \forall a_{-1}^{-k} \in A^k, \text{ vale } \sum_{b \in A} p(b|a_{-1}^{-k}) = 1 \right\}$$

$p \in \mathcal{M}_k(A)$ e $(X_n)_n$ foi gerada por p

cadeia de markov de alcance k

τ árvore de contextos de altura k.

$$\mathcal{M}_\tau(A) = \left\{ p : \tau \times A \rightarrow [0,1], \; \forall w \in \tau, \; \sum_{b \in A} p(b|w) = 1 \right\}$$

Se altura de $\tau = k$

$$M_\tau(A) \subset M_k(A)$$

$\tau \to$ "captura" informações sobre a dependência temporais da cadeia

$\bigcup_{k=1}^{\infty} M_k(A)$ $\qquad$ $M_\infty(A)$ = clase de cadeias de alcance finito

$\bigcup$

$\bigcup_{\tau \in \mathcal{T}} M_\tau(A)$

$\uparrow$

arvores finitas

Seja $p \in M_k(A)$, gero amostra $X_1, X_2, \ldots, X_n$ usando p

Perguntas: 1) Conhecendo k, estimar p

2) Se k e p for desconhecidos, seleccionar $\hat{k}_n$ y $\hat{p}_n$ que "melhor" descrebam a amostra.

Seja $p \in M_k(A)$ com τ finito, gero a amostra $X_1, \ldots, X_n$ usando (τ, p).

3) Se τ e p foram desconhecidos, selecionar $\hat{\tau}_n$ e $\hat{p}_n$ que "melhor" descrevam a amostra.

Perguntas:

1) Dada a amostra $X_1, \ldots, X_n$ qual é o máximo alcance k que podemos identificar?

Resposta: Só conseguimos identificar alcances k tais que

$$|A|^k << n$$

$$e^{k \ln |A|} < e^{\ln n} \Rightarrow \boxed{k < \frac{\ln n}{\ln |A|}} \Leftarrow$$

T. Shannon Mc. Millan Breine

Lema de Kac

2. Como seleccionar k (ou τ) com alcance $< \dfrac{\ln n}{\ln |A|}$?

Estimador de máxima verosimilhança em $M_k(A)$ ou $M_\tau(A)$

amostra: $X_{-k}^{n} = X_{-k}, X_{-k+1}, \ldots, X_n$

$$k = \frac{\ln n}{\ln |A|} \qquad N_{0:n}(a_{-k}^{-1} b) = \sum_{t=0}^{n} \mathbb{1}\{X_{t-k}^{t-1} = a_{-k}^{-1}, X_t = b\}$$

$$\hat{p}_n^{[k]}(b|a_{-k}^{-1}) = \frac{N_{0:n}(a_{-k}^{-1} b)}{\sum_{z \in A} N_{0:n}(a_{-k}^{-1} z)} = \boxed{\frac{N_{0:n}(a_{-k}^{-1} b)}{N_{0:n-1}(a_{-k}^{-1})}}$$

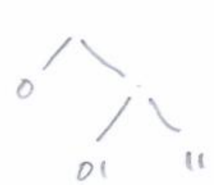
Para caso : τ árvore finita

$$\hat{p}_n^{\tau}(b|w) = \frac{N_{0:n}(wb)}{N_{0:n-1}(w)}, \quad \forall w \in \tau.$$

onde: $N_{0:n}(wb) = \sum_{t=0}^{n} \mathbb{1}\{X_{t-\ell(w)}^{t-1} = w, X_t = b\}$, onde

$\ell(w)$ = comprimento de w.

<u>Exemplo</u>: $A = \{0,1\}$ $\quad \tau =$



$\tau = \{0, 01, 11\}$

$$N_{0:n}((01)0) = \sum_{t=0}^{n} \mathbb{1}\{X_{t-2}^{t-1} = (0,1), X_t = 0\}$$

Vimos que se a amostra foi gerada por $p \in M_k(A)$ (ou por $M_\tau(A)$)

então $\hat{p}_n^{[k]}(b|a_{-k}^{-1}) \xrightarrow[n\to\infty]{} p(b|a_{-k}^{-1})$

$\left(\hat{p}_n^{\tau}(b|w) \xrightarrow[n\to\infty]{} p(b|w)\right)$ Lei dos grandes números.

Se $p \in M_k(A)$ e (X_n) foi gerada por p.

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

$$\parallel$$

$p(a \mid a_{-k}^{-1})$ qualquer seja z

$$\hat{p}_n^{[k+1]}(b \mid \underbrace{a_{-k}^{-1}}_{\substack{\text{últimos} \\ k \text{ valors}}} \; z) = \frac{N_{0:n}(z a_{-k} \dots a_{-1} b)}{N_{0:n-1}(z a_{-k} \dots a_{-1})} \longrightarrow p(b \mid a_{-k}^{-1})$$

valor $k+1$ passos atrás

isto converge a porque não depende de z

Então $\hat{p}_n^{[k+1]}(b \mid a_{-k}^{-1} z) \cong \hat{p}_n^{[k]}(b \mid a_{-k}^{-1})$ se n for grande

?!?! | o que quer dizer estatisticamente igual

Dado $\varepsilon > 0$, ε pequenho

como calcular $\delta(\varepsilon, n)$?

$$\mathbb{P}\left(\left| \hat{p}_n^{[k]}(b \mid a_{-k}^{-1}) - p(b \mid a_{-k}^{-1}) \right| > \varepsilon \right) \leq \delta(\varepsilon, n)$$

pequenho sob a hipotese nula: A cadeia foi gerada por p.

Teorema Limite Central

$$\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^{n} \xi_i - p \right] \xrightarrow{\text{Distribuição}} N(0, p(1-p))$$

com qué velocidade?

Ley dos grandes números

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

$$\frac{1}{n} \sum_{i=1}^{n} \xi_i \xrightarrow[n \to \infty]{} \mathbb{P}(\xi_1 = 1) = p$$

com qué velocidade?

$$\mathbb{P}\left(\hat{p}(b|w) - p(b|w) > \varepsilon \right)$$

$$= \mathbb{P}\left(\frac{N_{0:n}(wb)}{N_{0:n-1}(w)} - p(b|w) > \varepsilon \right)$$

$$\left[\begin{array}{c} N_{0:n}(wb) = N(wb) \\ p(b|w) = p \end{array} \right]$$

$$= \mathbb{P}\left(N(wb) - p\,N(w) > N(w)\varepsilon \right)$$

$$M_n = N_{0:n}(wb) - N_{0:n-1}(w)\, p(b|w)$$

Isto é um MARTINGALE!! Doob (anos 50)

$$\mathbb{E}\left(M_n \mid X_{-k}^{n-1} \right) = M_{n-1}$$

esperança condicional

$$\mathbb{P}\left(M_n > \underbrace{N(w)\varepsilon} \right)$$

se fose um número e não uma v.a podeíamos usar a Des. de Hoeffding para martingais.

$$\mathbb{P}\left(M_n > N(w)\varepsilon \,;\, N(w) < m \right) \leq \mathbb{P}\left(N(w) < m \right)$$

caso pésimo

$$+ \underbrace{\mathbb{P}\left(M_n > N(w)\varepsilon \,;\, N(w) \geq m \right)}$$

ótimo, w apareça mais de m vezes

$$\leq \mathbb{P}\left(M_n > m\varepsilon \,;\, \underbrace{N(w) \geq m} \right)$$

evento altamente provavel

Juntando temos:

$$\mathbb{P}\left(\hat{p}(b|w) - p(b|w) > \varepsilon \right) \leq \mathbb{P}\left(N(w) < m \right) + \underbrace{\mathbb{P}\left(M_n > m\varepsilon \right)}$$

Aqui posso usar Hoeffding

$(X_1, Y_1), \dots, (X_n, Y_n)$ iid

$X_n \in \mathcal{X}$

$Y_n \in \{0,1\}$

$\mathcal{G} : \mathcal{X} \to \{0,1\}$

Clase de funçoes ("modelos") candidatas a classificador

Objetivo. Encontrar o classificador que a cada amostra X associa Y a partir de amostra $(X_1, Y_1), \dots, (X_n, Y_n)$

Risco do classificador $g \in \mathcal{G}$

$$R(g) = \mathbb{P}(g(X) \neq Y)$$

Risco mínimo: $R^* = \inf \{ R(f) : f : \mathcal{G} \to \{0,1\} \}$

f mensurável | $\{x : f(x) = 1\}$ é um evento conjunto messurável

Classificador de Bayes:

$$f^*(x) = 1\{\eta(x) \geq 1/2\} \quad \text{onde} \quad \eta(x) = \mathbb{P}(Y=1 \mid X=x)$$

função de regresão

Teorema:

$$\boxed{R(f^*) = R^*}$$

Demostração ver página

Risco empírico: podemos calcular a partir da amostra

$$R_n(g) = \frac{1}{n} \sum_{n=1}^{n} 1\{g(X_i \neq Y_i)\}$$

variavel aleatoria

Pregunta central: Quão longe $R_n(g)$ está de $R(g)$?

$R(g)$ se não conhecemos $\mathbb{P}$ não podemos calcular.

$$\mathbb{P}\left(R_n(g) - R(g) > \varepsilon \right) \leq \underbrace{e^{-2n\varepsilon^2}}_{\delta}$$

$$\Downarrow$$

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

$$\delta = e^{-2n\varepsilon^2} \Rightarrow \ln \delta = -2n\varepsilon^2 \Rightarrow \varepsilon^2 = -\frac{\ln \delta}{2n} \Rightarrow \underline{\varepsilon = \sqrt{\frac{\ln 1/\delta}{2n}}}$$

$$\mathbb{P}\left(R_n(g) - R(g) > \underbrace{\sqrt{\frac{\ln 1/\delta}{2n}}}_{} \right) \leq \delta$$

precisão de estimação $\Rightarrow$ cresce quando $n \uparrow \infty$
decresce quando $\delta \searrow 0$

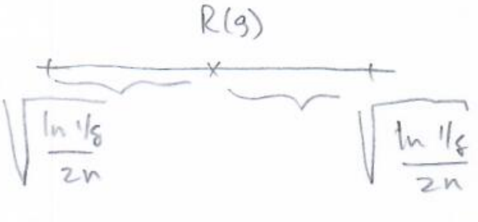
$1/\delta \nearrow +\infty$ quando $\delta \searrow 0$

<u>Idem</u>

$$\mathbb{P}\left(|R_n(g) - R(g)| > \sqrt{\frac{\ln 1/\delta}{2n}} \right) \leq \delta$$

Ótimo, porque essa majoração não depende da função g
(só usamos os fatos: g assume os valores 0 e 1 e $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid)

$R_n(g)$



$$R(g) \leq R_n(g) + \sqrt{\frac{\ln 1/\delta}{2n}} \quad \text{com probabilidade} \geq 1-\delta$$

C_g = conjunto ruim

$$C_g = \left\{ |R_n(g) - R(g)| > \sqrt{\frac{\ln 1/\delta}{2n}} \right\}$$

$$\mathbb{P}(C_g) \leq \delta$$

<u>Problema</u>: Se usamos uma outra função $\tilde{g}$, teremos un outro conjunto ruim $C_{\tilde{g}}$

Vamos supor que $\mathcal{G} = \{g_1, g_2\}$. Quero encontrar uma stimativa que seja boa para ambas (8)

$$\mathbb{P}\left(\exists g : \mathcal{G} : |R_n(g) - R(g)| > \sqrt{\frac{\ln 1/\delta}{2n}}\right)$$

$$= \mathbb{P}\left(\bigcup_{g\in\mathcal{G}} C_g\right) \leq \sum_{g\in\mathcal{G}} \mathbb{P}(C_g) = \mathbb{P}(C_{g_1}) + \mathbb{P}(C_{g_2}) = 2\delta$$

Se $\mathcal{G} = \{g_1, \dots, g_N\}$

$$\mathbb{P}\left(\bigcup_{g\in\mathcal{G}} \left\{|R_n(g) - R(g)| > \sqrt{\frac{\ln 1/\delta}{2n}}\right\}\right) \leq \sum_{i=1}^{N} \mathbb{P}(C_{g_i}) = N\cdot\delta$$

Vamos refazer desde o començo:

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

$$\{R_n(f) - R(f) > \varepsilon\} = D_f$$

$$\mathbb{P}(\exists f \in \mathcal{G} : R_n(f) - R(f) > \varepsilon) \leq \sum_{f\in\mathcal{G}} \mathbb{P}(R_n(f) - R(f) > \varepsilon)$$

$$\boxed{\text{Se está usando } P\left(\bigcup_{i=1}^{N} D_i\right) \leq \sum_{i=1}^{N} P(D_i)}$$

$$= |\mathcal{G}|\, e^{-2n\varepsilon^2} = N e^{-2n\varepsilon^2}$$

si $\mathcal{G} = \{g_1, \dots, g_N\}$

tamanho da amostra

\# de funçoes na clase

$$\delta = N\cdot e^{-2n\varepsilon^2}$$

$$\Rightarrow \varepsilon = \sqrt{\frac{\ln N + \ln 1/\delta}{2n}}$$

$$\mathbb{P}\left(\bigcup_{f\in\mathcal{G}} R_n(f) - R(f) > \sqrt{\frac{\ln N + \ln 1/\delta}{2n}}\right) \leq \delta$$

Ou seja

$$\mathbb{P}\left(R_n(f) < R(f) + \sqrt{\frac{\ln N + \ln 1/\delta}{2n}} \ , \ \forall f \in \mathcal{G} \right) \geq 1-\delta$$

O numerador $\nearrow +\infty$, para compensarlo o tamanho da amostra $n \gg \ln N + \ln 1/\delta$

Quero que:

$$\sup_{g \in \mathcal{G}} |R_n(g) - R(g)| \leq \varepsilon \quad \text{e típicamente } |\mathcal{G}| \text{ é infinito!!!}$$

Cómo enfrentar isso? Teoria de VC.

$\boxed{\text{Demostração que classificador de Bayes realiza o risco de Bayes}}$

$$\eta(x) = \mathbb{P}(Y=1 \mid X=x)$$

$$f^*(x) = \mathbb{1}\{\eta(x) \geq 1/2\}$$

<u>Teorema</u>: $R(f^*) = R^*$

onde $R^* = \inf\{R(g) : g : \mathcal{X} \to \{0,1\}\}$ g meesuravel

Temos que demostrar que:

$$R(g) - R(f^*) \geq 0 \quad \text{para toda função } g \text{ meesurável}$$

Ora:

$$R(g) - R(f^*) = \mathbb{P}(g(X) \neq Y) - \mathbb{P}(f^*(X) \neq Y)$$

$$\mathbb{P}(g(X) \neq Y) = \int_{\mathcal{X}} \mathbb{P}_X(dx)\, \mathbb{P}(g(X) \neq Y \mid X=x)$$

$$\mathbb{P}(f^*(X) \neq Y) = \int_{\mathcal{X}} \mathbb{P}_X(dx)\, \mathbb{P}(f^*(X) \neq Y \mid X=x)$$

ou seja

$$R(g) - R(f^*) = \int \mathbb{P}_X(dx) \left[\mathbb{P}(g(X) \neq Y \mid X=x) - \mathbb{P}(f^*(X) \neq Y \mid X=x) \right]$$

Basta demostrarnos que esta diferença é sempre ≥ 0

$$\mathbb{P}(g(X) \neq Y \mid X = x) = 1 - \mathbb{P}(g(X) = Y \mid X = x)$$

$$= 1 - [\mathbb{P}(g(x) = 1, Y = 1 \mid X = x) + \mathbb{P}(g(x) = 0, Y = 0 \mid X = x)]$$

$$= 1 - [\mathbb{1}_{\{g(x)=1\}} \underbrace{\mathbb{P}(Y = 1 \mid X = x)}_{\eta(x)} + \mathbb{1}_{\{g(x)=0\}} \underbrace{\mathbb{P}(Y = 0 \mid X = x)}_{1-\eta(x)}]$$

$$= 1 - [\mathbb{1}_{\{g(x)=1\}} \eta(x) + \mathbb{1}_{\{g(x)=0\}} (1 - \eta(x))]$$

Idem:

$$\mathbb{P}(f^*(X) \neq Y \mid X = x) = 1 - [\mathbb{1}_{\{f^*(x)=1\}} \eta(x) + \mathbb{1}_{\{f^*(x)=0\}} (1 - \eta(x))$$

$$\mathbb{P}(g(X) \neq Y \mid X = x) - \mathbb{P}(f^*(X) \neq Y \mid X = x)$$

$$= (2\eta(x) - 1)[\mathbb{1}_{\{f^*(x)=1\}} - \mathbb{1}_{\{g(x)=1\}}]$$

Analisar casos: $\eta(x) \geq 1/2$ e $\eta(x) < 1/2$, para ambos casos da ≥ 0.