Aprenditagem { supervisionado / não supervisionado

$(X_i; Y_i)_{i=1,\dots,n}$ v.a iid
↑ classificador

Problema: Encontrar $\hat{f}_n \in \mathcal{F}$ tal que $\mathbb{P}(\hat{f}_n(X) = Y) \leq \epsilon(n)$

Desigualdades | Markov, Cramer - Chernoff, Hoeffding → desenvolver Teoria VC Vapnik-Chervonenkis

- Classificação não supervisionado de sequências simbólicas | Kolmogorov, Vitany, Rissanen
- Predição

Amostra: $X^{(N)} = \left(X_0^{(N)}, \dots, X_n^{(N)}\right) \in A^n$

Clase de modelos $\mathcal{M}$

Algoritmo $\mathcal{A} : \left(X^{(1)}, \dots, X^{(N)}\right) \longrightarrow \hat{\tau} \in \mathcal{M}$

modelo de geração de sequências simbólicas indo de passado para futuro
geração ↓ sequencial.

A = alfabeto finito

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

Notação

$a_{-k}, \dots, a_{-1} = a_{-k}^{-1}$

$X_m, X_{m+1}, \dots, X_n = X_m^n$

$(X_n)_n$ é uma cadeia de markov de alcance k assumindo valores em A se ela puder ser gerada por um algoritmo do tipo:

Algoritmo:

- Sejam $U_0, U_1, \dots$ v.a iid uniformemente distribuidas no intervalo $[0,1]$
- seja $k \geq 1$ fixado, A alfabeto finito fixado, seja $p \in \mathcal{M}_k(A)$ fixado e seja $f : A^k \times [0,1] \to A$ tal que $\forall a_{-k}^{-1} \in A^k, b \in A$ tenhamos $\mathbb{P}(f(a_{-k}^{-1}, U) = b) = p(b \mid a_{-k}^{-1})$, $U \sim \text{Unif}_{[0,1]}$

1. Escolho de alguma forma fixada ...

2. Para $n \geq 0$, faça

$$X_n = f\left(X_{n-k}^{n-1}, u_n\right)$$

Exemplo

$A = \{1, 2, 3\}$, $k = 1$

Dado $p = \{p(b|a) : a \in A, b \in A\}$ con $p \in \mathcal{M}_1(A)$

$$f(a,u) = \begin{cases} 1, & \text{se } u \in [0, p(1|a)[\\ 2 & \text{se } u \in [p(1|a), p(1|a) + p(2|a)[\\ 3 & \text{se } u \in [p(1|a) + p(2|a), 1] \end{cases}$$

Por abuso de notaçao as vezes diremos que $(X_n)_{n \geq -k} \in \mathcal{M}_k(A)$ si $(X_n)_n$ for gerado por um algoritmo associado a $p \in \mathcal{M}_k(A)$.

Propocisao: Se $(X_n)_{n \geq -k}$ for gerado por um algoritmo associado a $p \in \mathcal{M}_k(A)$, então para todo $n \geq 0$

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

$$= p\left(b \mid a_{n-k}^{n-1}\right)$$

$\mathcal{M}_k(A)$, $k \geq 1$

$$\mathcal{M}_0(A) = \left\{ p : A \to [0,1] : \sum_{b \in A} p(b) = 1 \right\}$$

Dado $p \in \mathcal{M}_0(A)$, (X_n) é uma sequencia de variaveis aleatorias assumindo valors em A independentes identicamente distribuidas com distribuçao p, se $(X_n)_{n \geq 0}$ puede ser gerado por um algoritmo do tipo:

Algoritmo

0. Seja A alfabedo finito
Seja $p \in \mathcal{M}_0(A)$
Seja $U_0, U_1, \ldots$ v.a iid com distribuçao uniforme em $[0,1]$
Seja $f : [0,1] \to A$ tal que $\mathbb{P}(f(U) \ldots$

1. Para $n \geq 0$ faça

$$X_n = f(U_n)$$

$$\boxed{\mathcal{M}(A) = \bigcup_{k \geq 0} \mathcal{M}_k(A)}$$

Clase dos modelos de memória finita
ó
Clase das cadeias de Markov

Dada uma amostra $\overbrace{X_{-k}, X_{-k+1}, \ldots, X_0, X_1, \ldots X_n}^{X_{-k}^{n}}$ de símbolos em A, queremos encontrar $\hat{k}_n$ e $\hat{p}_n$ em $\mathcal{M}_{\hat{k}}(A)$ que "melhor se ajuste" á amostra.

Que algoritmo para isso? | Selecçao estatística de modelos

Nao se vai falar hoje sobre $\hat{k}_n$ | Algoritmo contexto
MDL (Minimum description length)

navalha de Occam

Dado A:
- Fixo $k \geq 1$
- Fixo $p \in \mathcal{M}_k(A)$

Quero calcular: $\mathbb{P}\left(X_0^n = a_0^n \mid X_{-k}^{-1} = a_{-k}^{-1}\right)$ | Verossimilhança da Amostra

$$= \mathbb{P}\left(X_0 = a_0 \mid X_{-k}^{-1} = a_{-k}^{-1}\right) \mathbb{P}\left(X_1 = a_1 \mid X_0 = a_0, X_{-k}^{-1} = a_{-k}^{-1}\right) \ldots \mathbb{P}\left(X_n = a_n \mid X_{-k}^{n-1} = a_{-k}^{n-1}\right)$$

$$\mathbb{P}\left(X_1 = a_1 \mid X^0_{-(k-1)} = a^0_{(-k-1)}\right) \ldots \mathbb{P}\left(X_n = a_n \mid X_{n-k}^{n-1} = a_{n-k}^{n-1}\right)$$

$$= \prod_{m=0}^{n} p(a_n \mid a_{m-k}^{m-1})$$

$$= \prod_{u_{-k}^{-1} \in A^k} \prod_{v \in A} p(v \mid u_{-k}^{-1})^{N(u_{-k}^{-1} v)}$$

$N(u_{-k}^{-1} v)$ = número de veces em que a sequencia u_{-k}^{-1} seguida do
símbolo v aparece na amostra

$$N_n(u_{-k}^{-1} v) = N_{n-1}(u_{-k}^{-1} v) + \mathbb{1}_{\{X_{n-k}^{n-1} = u_{-k}^{-1},\, X_n = v\}}$$

$$N_0(u_{-k}^{-1} v) = \mathbb{1}_{\{X_{-k}^{-1} = u_{-k}^{-1},\, X_0 = v\}}$$

$$\sum_{v \in A} \sum_{u_{-k}^{-1} \in A^k} N_n(u_{-k}^{-1} v) = n+1$$

$$\log \mathbb{P}(X_0^n = a_0^n \mid X_{-k}^{-1} = a_{-k}^{-1}) = \sum_{u_{-k}^{-1} \in A^k} \sum_{v \in A} N_n(u_{-k}^{-1} v) \log p(v \mid u_{-k}^{-1})$$

$N_n(u_{-k}^{-1} v) \sim \text{cte}(u_{-k}^{-1} v) \cdot n$

número aleatorio, depende da amostra $X_{-k}, \dots X_n$

Lei dos grandes números garante:

$$\frac{N_n(u_{-k}^{-1} v)}{n} \xrightarrow[n \to \infty]{} \mu(u_{-k}^{-1})\, \underbrace{p(v \mid u_{-k}^{-1})}$$

$p \in \mathcal{M}_k(A)$ que gerou a cadeia

$\mu : A^k \to [0,1]$ e uma medida de probabilidade

isto e' $\displaystyle\sum_{u_{-k}^{-1} \in A^k} \mu(u_{-k}^{-1}) = 1 \qquad (\mu \in \mathcal{M}_0(A^k))$

μ satisface o sistema de equaçoes

$$\forall\, u_{-k}^{-1} \in A^k, \quad \mu(u_{-k}^{-1}) = \sum_{\substack{z_{-k}^{-1} :\, z_{-1} = u_{-2} \\ \vdots \\ z_{-(k-1)} = u_{-k}}} \mu(z_{-k}^{-1})\, p(u_{-1} \mid z_{-k}^{-1})$$

$$-\frac{1}{n}\log \mathbb{P}\left(X_0^n \mid X_{-k}^{-1}\right)$$

$$= -\sum_{u_{-k}^{-1}\in A^k}\ \sum_{v\in A} \frac{N_n(u_{-k}^{-1}v)}{n}\log p(v\mid u_{-k}^{-1})$$

$$\longrightarrow -\sum_{u_{-k}^{-1}\in A}\ \sum_{v\in A} \mu(u_{-k}^{-1})\,p(v\mid u_{-k}^{-1})\log p(v\mid u_{-k}^{-1}) = h(p)$$

$p\in \mathcal{M}_k(A)$.

$$\frac{N_n(u_{-k}^{-1}v)}{\sum_{z\in A} N_n(u_{-v}^{-1}z)} = \hat{p}_n(v\mid u_{-k}^{-1})$$

Estimador de maxima verosimilanza.

$$\hat{h}_n = -\sum_{u_{-k}^{-1}\in A^k}\ \sum_{v\in A}\frac{N_n(u_{-k}^{-1}v)}{n}\log \hat{p}_n(v\mid u_{-k}^{-1})$$

$\hat{h}_n \xrightarrow{n\to\infty} h(p)$, $\quad \frac{N_n(u_{-k}^{-1}v)}{n} \to \mu(u_{-k}^{-1})p(v\mid u_{-k}^{-1})$, $\quad \hat{p}_n(v\mid u_{-k}^{-1}) \to p(v\mid u_{-k}^{-1})$

<u>Problema</u> Dado A, dado $k\geq 1$, dada uma amostra X_{-k}^{-n} gerada por $\bar{p}\in \mathcal{M}_k(A)$, com $\bar{p}$ desconhecido, quero "estimar" $\bar{p}$.

$$X_{-k}^{-n} \longrightarrow \hat{p} = \hat{p}(X_{-k}^{-n}) \in \mathcal{M}_k(A) \quad . \quad \text{Qué algoritmo?}$$

Criterio: Seleccionо $\hat{p}_n$ que <u>maximiza</u> a verosimilhança da amostra.

$$q\in \mathcal{M}_k(A) \longmapsto \mathbb{P}_q\left(X_0^n \mid X_{-k}^{-1}\right) \in [0,1]$$

$$L(q) = \log \prod_{v\in A}\ \prod_{u_{-k}^{-1}\in A^k} q(v\mid u_{-k}^{-1})^{N_n(u_{-k}^{-1}v)}$$

$$L(q) = \sum_{u_{-k}^{-1} \in A^k} \sum_{v \in A} N_n(u_{-k}^{-1} v) \log q(v \mid \underbrace{u_{-k}^{-1}}_{u})$$

$$\hat{p} = \arg\max \{ L(q) : q \in \mathcal{M}_k(A) \}$$

$$\hat{p} = \hat{p}(k, X_{-k}^{n}) \leftarrow \text{matrix aleatoria}$$

Notaçao: $u = u_{-k}^{-1} \in A^k$

Fixados $N_n(uv)$ quero encontrar $\hat{p}$ que maximiza $q \mapsto L(q)$

Máximo com vinculo

Vinculo: $p \in \mathcal{M}_k(A)$, ou seja $\forall u = u_{-k}^{-1} \in A^k$ devemos ter $\sum_{u \in A^k} q(v|u) = 1$

solução: usando multiplicadores de Lagrange

$$\lambda = (\lambda(u) : u \in A_{-k}^{-1})$$

$$F(q, \lambda) = L(q) + \sum_{u \in A^k} \lambda(u) \left[1 - \sum_{v \in A} q(v|u) \right]$$

Para cada $u = u_{-k}^{-1} \in A$, $v \in A$

derivo $\left. \dfrac{\partial F}{\partial q(v|u)} \right|_{q = \hat{p}} = 0$, $\left. \dfrac{\partial F}{\partial \lambda(u)} \right|_{\hat{\lambda}} = 0$

$$\frac{\partial F}{\partial q(v|u)} = \frac{\partial}{\partial q(v|u)} N(uv) \log q(v|u) + \frac{\partial}{\partial q(v|u)} \left[-\lambda(u) q(v|u) \right]$$

$$= N(uv) \frac{\partial}{\partial q(v|u)} \log q(v|u) + \frac{\partial}{\partial q(v|u)} \left[-\lambda(u) q(v|u) \right]$$

$$= \frac{N(uv)}{q(v|u)} - \lambda(u) \frac{\partial}{\partial q(v|u)} q(v|u)$$

$$= \frac{N(uv)}{q(v|u)} - \lambda(u) \Big|_{(\hat\lambda, \hat q)} = 0 \implies \boxed{\hat q(v|u) = \frac{N(uv)}{\hat\lambda(u)}}$$

$$\frac{\partial F}{\partial \lambda(u)} = \frac{\partial}{\partial \lambda(u)} \lambda(u) \left[1 - \sum_v q(v|u)\right]$$

$$= 1 - \sum_{v \in A} q(v|u) \Big|_{\hat q = 0} \implies \boxed{1 = \sum_{v \in A} \hat q(v|u)}$$

$$\hat\lambda(u) = \sum_{z \in A} N(uz)$$

$$\hat q = \hat p_n(v|u_{-k}^{-1}) = \frac{N_n(u_{-k}^{-1} v)}{\sum_{z \in A} N_n(u_{-k}^{-1} z)}$$

Estimador de máxima verosimilhança em $M_k(A)$

<u>Fato</u>: lei dos grandes números garante que

$$\hat p_n^{(k)}(v|u_{-k}^{-1}) \xrightarrow[n \to \infty]{} p(v|u_{-k}^{-1})$$

se p for a matriz que gerou a amostra.

Se $(x_n)_n$ for gerado por $p \in M_k(A)$ e se

$$\hat p_m^{[k]}(v|u_{-k}^{-1}) \longrightarrow p(v|u_{-k}^{-1})$$

o que acontece se estimamos

8

$$\hat{p}_n^{[k+1]}\left(v \mid u_{-(k+1)}^{-1}\right) = \frac{N_n\left(u_{-(k+1)}^{-1} v\right)}{\sum_{z \in A} N\left(u_{-(k+1)}^{-1} z\right)} \xrightarrow[n \to \infty]{\text{lei dos grandes números}} \mathbb{P}\left(X_0 = v \mid X_{-(k+1)}^{-1} = u_{-(k+1)}^{-1}\right)$$

$$\parallel$$

$$p\left(v \mid u_{-k}^{-1}\right)$$

k basta!

τ árvore finita de contextos

$A = \{0,1\}$

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

(1) pseudo-código

0. Defina $u_0, u_1, \dots$ sequencia de v.a iid com distribuição uniforme em $[0,1]$,

$r(X_{n-1}) = p(1 \mid X_{n-1})$

1. $X_{-1} = 0$

2. Para $n \geq 0$, faça

$$X_n = \mathbb{1}\{u_n < r(X_{n-1})\}$$

(2) 0. Defino para todo $w \in \tau$, o parâmetro $r(w) = p(1 \mid w)$

Tomo $u_0, u_1, \dots$ v.a iid uniformes em $[0,1]$

1. $X_{-2} = X_{-1} = 1$

2. Para $n \geq 0$ faço

Se $X_{n-1} \in \tau$, então $X_n = \mathbb{1}\{u_n < r(X_{n-1})\}$

Se $X_{n-1} \notin \tau$, então $X_n = \mathbb{1}\{u_n < r(X_{n-2}, X_{n-1})\}$

(3) $X_{-1}^{-20} = (0,1,0,1,0,1,1,1,0,1,0,1,0,1,0,0,0,1,0,1)$

$\hat{p}(0 \mid 1) = \frac{7}{9}$ $\quad \hat{p}(1 \mid 1) = \frac{2}{9}$

$\hat{p}(1 \mid 0) = \frac{8}{10}$ $\quad \hat{p}(0 \mid 0) = \frac{2}{10}$

$$\hat{P}_n(u \mid v) = \frac{N_n(vu)}{\sum_{u'} N_n(vu')}$$

(4) $X_{-1}^{-12} = (0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1)$

$p(0|0)$ $\quad p(1|0)$ $\quad p(1|10)$ $\quad p(1|11)$ $\quad p(0|11)$ $\quad p(0|0)$ $\quad p(0|0)$ $\quad p(1|0)$ $\quad p(0|10)$ $\quad p(1|0)$ $\quad p(1|10)$ $\quad p(1|11)$

$$\log \mathbb{P}(X_{-1}^{-12}) = 3 \cdot \log p(0|0) + 3 \log p(1|0) + 2 \log p(1|10) + \log p(0|10) + 2 \log p(1|11) + \log p(0|11) = \sum \sum N_n(wa) \log p(a|w)$$