Clasificação não supervisionada

A alfabeto finito.

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

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

τ árvore em A

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

$$\mathcal{M}_\infty(A) = \bigcup_{\tau \text{ árvores}} \mathcal{M}_\tau(A)$$

Note que: se altura$(\tau) = k$, então $\mathcal{M}_\tau(A) \subset \mathcal{M}_k(A)$

Altura(τ) = maior comprimento de ramos em τ

$$\mathcal{M}(A) = \bigcup_{k=0}^{\infty} \mathcal{M}_k \cup \mathcal{M}_{\tau \text{ finito}}(A)$$

Problema: Dada uma amostra $(X_0, X_1, \dots, X_n) = X_0^n$ de símbolos em A, queremos encontrar $\hat{\tau}(X_0^n) \in \mathcal{M}_{\text{finito}}(A)$

$$\hat{\tau}(X_0^n) = \hat{\tau}_n$$

Pergunta: como fazer uma selecção de $\hat{\tau}_n$?

Vimos que:

Seja $p \in \mathcal{M}_k(A)$, geramos $X_{-k}, X_{-k+1}, X_0, \dots, X_n = X_{-k}^n$

Problema statistico: Conhecendo k, "stimar" p

Estimador de máxima verossimilhança:

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

Notação:

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

Observação:

$$\sum_{z\in A} N_{0:n}(a_{-k}^{-1}z) = N_{0:n-1}(a_{-k}^{-1})$$

Sabemos que:

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

Sei que $p \in \bigcup_{j=0}^{\infty} \mathcal{M}_j(A)$ mas não sei o alcance da memória de p.

1. Começo com o maior valor possível para k dada miha amostra. (se a amostra tem n símbolos, $k \leq \log n$)

2. Seja K o maior valor possível dado o tamanho da amostra, estimo $\hat{p}_n^{[K]}(\cdot|\ )$ e $\hat{p}_n^{[K-1]}(\cdot|\ )$ y pergunto: são iguais?

   $=$ igualdade no sentido estatístico.

3. Se $\hat{p}_n^{[K]} \not\cong \hat{p}_n^{[K-1]}$ então concluo que o alcance de p que gerou a amostra é K.

4. Se $\hat{p}_n^{[K]} \cong \hat{p}_n^{[K-1]}$ recomeço e comparo. $\hat{p}_n^{[K-1]}$ e $\hat{p}_n^{[K-2]}$

Para $j = 1, \dots, K-1$ faça:

comparo $\hat{p}_n^{[K-j]} \overset{?}{\neq} \hat{p}_n^{[K-(j+1)]}$

se são diferentes

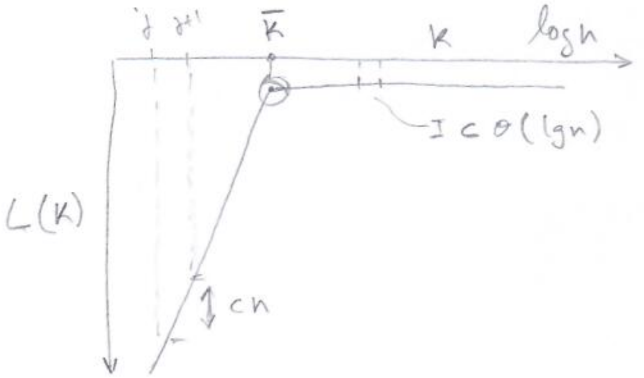
$\hat{K}_n \leftarrow K - j$

se são iguais continuo.

PARO quando der iguais pela 1ra vez.

Amostra dada X_0^n e estimamos $\hat{p}_n^{[k]}$ para $k = 0, \dots, \log n$ e calculamos $\log \hat{\mathbb{P}}_n^{[k]}(X_0^n) =$

$$L(k) = \sum_{b \in A} \sum_{a_{-k}^{-1} \in A^k} N(a_{-k}^{-1} b) \log \hat{p}_n^{[k]}(b \mid a_{-k}^{-1})$$



- L(k) é uma função crescente de k, por isso não é o bom critério pegar o k que maximize verosimilhança.
- Mas esse gráfico amostra um ponto de inflexão interessante

Dúvidas

1. Qual é o maior valor que podemos <u>testar</u> para o alcance da cadeia?
2. O que significa $\hat{p}_n^{[j]} \cong \hat{p}_n^{[j+1]}$?

Respostas:

1. Sai do famoso teorema de Shannon-McMillan-Breiman que diz que: $\forall \varepsilon > 0$ (ε pequeno) existe $\bar{n} = \bar{n}(\varepsilon)$ tal que $\forall n \geq \bar{n}$ vale o seguinte:

Existe um conjunto $B(\varepsilon,n) \subset A^n$ tal que:

- $\mathbb{P}(X_1^n \in B(\varepsilon,n)) \geq 1-\varepsilon$

entropia

- para $b_1^n \in B(\varepsilon,n)$, $\mathbb{P}(X_1^n = b_1^n) \sim e^{-nh}$ onde $h \geq 0$ é uma constante associado á cadeia $(X_n)_n$

$p \in M_1(A)$

(X_n) gerada por p

$$- \log \mathbb{P}(X_0^n = a_0^n)$$

$$- \log \left[\mathbb{P}(X_0 = a_0) \prod_{t=1}^{n} P(X_t = a_t \mid X_{t-1} = a_{t-1}) \right]$$

$$= - \log \mathbb{P}(X_0 = a_0) - \sum_{(x,y)\in A^2} N_{1:n}(xy) \log p(y|x)$$

Quero que $n \to +\infty$

$$\frac{1}{n} \log \mathbb{P}(X_0 = a_0) + \sum_{(x,y)} \frac{N_{1:n}(x,y)}{n} \log p(y|x)$$

$$\frac{N_{1:n}(x,y)}{n} \xrightarrow[n\to\infty]{\text{ley dos grandes números}} \mu(x)\, p(y|x)$$

entropia da cadeia

$$- \frac{1}{n} \log \mathbb{P}(X_0^n = a_0^n) \longrightarrow \sum_{(x,y)\in A^2} \mu(x) p(y|x) \log p(y|x) \stackrel{\text{def}}{=} h(p)$$

$p \in M_1(A)$

Então $- \frac{1}{n} \log \mathbb{P}(X_0^n = a_0^n) \longrightarrow h$ | Equivalente a $\mathbb{P}(X_0^n = a_0^n) \to e^{-nh}$.

Compresão de arquivos usando o algoritmo Zip (Ziv-Lipman) 5

Vamos descompor em subsequencias:

cada subsequencia é a menor subsequência de símbolos futuros que ainda não apareceu.

$\boxed{0}\boxed{0\,1}\boxed{0\,1\,1}\boxed{1}\boxed{0\,0}\boxed{\underbrace{0\,1}_{\text{já visto}}\,\underline{0}}\,1\,1\,0\,0\,0\,0$

símbolo nuevo.

R_N: instante de o primeiro reaparecimento da sequencia de N símbolos iniciais.

$$\frac{1}{N}\log R_N \xrightarrow[N\to\infty]{} h$$

"moralmente" $R_N \sim e^{Nh}$

Para poder calcular $\hat{p}^{(k)}(b|a_{-k}^{-1})$ é preciso que a_{-k}^{-1} apareza muitas veces.

Ora: Típicamente o tempo que una sequencia de tamaño k leva para reaparecer é e^{hk}

Para aparecer MUITAS veces na amostra X_1^n é preciso n seja $\gg e^{hk}$

No pior caso $h = \ln(A)$, queremos

$$\underset{\substack{\uparrow\\ e^{\ln n}}}{n} \gg e^{k\ln(A)}$$

Então é preciso que $\ln(n) \gg k\ln|A|$

$$k \ll \frac{\ln n}{\ln|A|} \qquad k \leq (1-\varepsilon)\frac{\ln n}{\ln|A|}$$