Nicolas Sendrier

Majeure d’informatique

Introduction la théorie de l’information

Cours n$^\circ$4

Les séquences typiques et l’AEP

Définitions

Soit une source (un processus stochastique) constituée de la suite de variables aléatoires $X_1, X_2, \ldots, X_n, \ldots$ à valeur dans un alphabet $\mathcal{X}$. On suppose que l'entropie par lettre de cette source est définie

$$\mathcal{H} = H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} \log_2 H(X_1, \ldots, X_n)$$

Définition Ensemble de séquences typiques (de longueur n)

$$A_\varepsilon^{(n)} = \left\{ (x_1, \ldots, x_n) \in \mathcal{X}^n, \left| \frac{1}{n} \log_2 \frac{1}{P(x_1, \ldots, x_n)} - \mathcal{H} \right| \leq \varepsilon \right\}$$

Définition (*Asymptotic Equipartition Property*)
Un processus stochastique (une source) vérifie l'AEP si :

$$\forall \varepsilon > 0, \lim_{n\to\infty} Pr(A_\varepsilon^{(n)}) = 1.$$

Propriétés des séquences typiques

Proposition Pour toute source vérifiant l'AEP

(i) $\frac{1}{n} \log_2 \frac{1}{P(x_1, \ldots, x_n)} \underset{n \to \infty}{\longrightarrow} \mathcal{H}$ presque sûrement

(ii) $Pr(A_\varepsilon^{(n)}) 2^{n(\mathcal{H}-\varepsilon)} \leq |A_\varepsilon^{(n)}| \leq 2^{n(\mathcal{H}+\varepsilon)}$

Autrement dit :

Il y a $2^{n\mathcal{H}}$ séquences typiques, ces séquences sont équiprobables de probabilité $2^{-n\mathcal{H}}$.

Théorème de Shannon

Définition Soit φ un codage de $\mathcal{X}$, sa longueur moyenne par lettre est définie par

$$\bar{N}(\varphi) = \lim_{n \to \infty} \frac{1}{n} \sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n) |\varphi(x_1, \ldots, x_n)|,$$

lorsque cette limite existe.

Théorème (Shannon) Pour toute source discrète $\mathcal{X}$.

1. Tout codage régulier φ de $\mathcal{X}$ vérifie $\bar{N}(\varphi) \geq H(\mathcal{X})$, si $H(\mathcal{X})$ existe.
2. Si la source vérifie l'AEP, alors pour tout $\varepsilon > 0$ il existe un codage régulier φ de $\mathcal{X}$ tel que $\bar{N}(\varphi) \leq H(\mathcal{X}) + \varepsilon$.

Processus stochastiques

Définitions Un processus stochastique est une suite $X_1, X_2, \ldots$ de variables aléatoires à valeur dans un alphabet $\mathcal{X}$.

- Il est dit *stationnaire* si pour tous entiers n et l,

$$P_{X_1 \ldots X_n}(x_1, \ldots, x_n) = P_{X_{1+l} \ldots X_{n+l}}(x_1, \ldots, x_n).$$

- Il est dit *markovien* d'ordre s si pour tout entier positif $n \geq s$ et tout $(x_1, \ldots, x_n) \in \mathcal{X}^n$

$$P(x_n \mid x_{n-1}, \ldots, x_1) = P(x_n \mid x_{n-1}, \ldots, x_{n-s}).$$

  Nous parlerons aussi de source markovienne.

- Un processus markovien est dit *irréductible* si tout état peut être atteint en un nombre fini d'étapes à partir de tout autre avec une probabilité > 0.

AEP en pratique

Proposition Une source sans mémoire vérifie l'AEP.

Proposition Une source markovienne stationnaire irréductible vérifie l'AEP.

Proposition Une source stationnaire ergodique vérifie l'AEP.