Nicolas Sendrier

Majeure d'informatique

Introduction la théorie de l'information

Cours n°4

Les séquences typiques et l'AEP

Processus stochastiques

Une source produit une suite de lettres dans l'alphabet $\mathcal{X}$.

Pour décrire cette suite de lettres nous utiliserons une suite de variables aléatoires $X_1, X_2, \ldots$ à valeurs dans $\mathcal{X}$. Ces variables ne sont pas nécessairement indépendantes. On parlera de *processus stochastique* pour décrire cette suite de variable aléatoires.

Entropie par lettre

Définition Un processus stochastique est *stationnaire* si son comportement ne varie pas lorsque l'on décale l'observation dans le temps. Pour tous entiers positifs n et j, et tout $(x_1, \ldots, x_n) \in \mathcal{X}^n$

$$p_{X_1 \ldots X_n}(x_1, \ldots, x_n) = p_{X_{1+j} \ldots X_{n+j}}(x_1, \ldots, x_n)$$

Théorème Pour tout processus stochastique stationnaire, les limites ci-dessous existent et sont égales

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

La quantité $H(\mathcal{X})$ est appelée *entropie par lettre*.

Processus markovien invariant dans le temps

Définition Un processus stochastique est dit *markovien* si pour tout entier positif n et tout $(x_1, \ldots, x_n) \in \mathcal{X}^n$

$$\Pr[X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = X_{n-1}] = \Pr[X_n = x_n \mid X_{n-1} = x_{n-1}]$$

Le processus est dit *invariant dans le temps* si ces probabilités ne dépendent pas de n.

Nous noterons $p(x_2 \mid x_1) = \Pr[X_n = x_2 \mid X_{n-1} = x_1]$.

Théorème L'entropie par lettre d'un processus markovien invariant dans le temps (irréductible) est égal à

$$H(\mathcal{X}) = H(X_2|X_1) = \sum_{x_1,x_2} -\lambda(x_1)p(x_1 \mid x_2) \log_2 p(x_1 \mid x_2)$$

où $\lambda(x), x \in \mathcal{X}$ est la *distribution stationnaire*.

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 :

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

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) $\Pr\left[A_\varepsilon^{(n)}\right] 2^{n(\mathcal{H}-\varepsilon)} \leq |A_\varepsilon^{(n)}| \leq 2^{n(\mathcal{H}+\varepsilon)}$

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

$$\mathcal{L}(\varphi) = \lim_{n\to\infty} \frac{1}{n} \sum_{x_1,\ldots,x_n} p(x_1,\ldots,x_n)|\varphi(x_1,\ldots,x_n)|,$$

lorsque cette limite existe.

Théorème (Shannon) Soit une source discrète d'alphabet $\mathcal{X}$ et d'entropie par lettre $\mathcal{H}$ qui vérifie l'AEP.

1. Tout codage régulier φ de $\mathcal{X}$ vérifie $\mathcal{L}(\varphi) \geq \mathcal{H}$.
2. Il existe un codage régulier φ de $\mathcal{X}$ tel que $\mathcal{L}(\varphi) \leq \mathcal{H} + \varepsilon$, $\forall \varepsilon > 0$.

AEP en pratique

Proposition Une source sans mémoire vérifie l'AEP.

Proposition Une source markovienne irréductible vérifie l'AEP.

Proposition Une source stationnaire ergodique vérifie l'AEP.