UE Modélisation :

Topic 2 :

7

Prédiction de mots par chaînes de Markov.

ou « comment écrire comme votre auteur préféré »

Dans cette leçon qui peut être vue comme une introduction au NLP (Natural Language Processing), nous allons faire de la prédiction de mots.

<u>Idée</u> : Voir un texte comme une réalisation de chaîne de Markov sur $E = \{\text{Mots}\}$

En fait Andrei Andreevich Markov (1856 - 1922) a conçu les chaînes de Markov avec l'exemple du TEXTE - juste avec deux états $E = \{ V = \text{"voyelle"}, C = \text{"consonne"} \}$.

[I] Rappel :

Pour ce cours, nous avons uniquement besoin de 2 notions de la séance précédente :

(1) Définition des chaînes de Markov. (CM)

Si $E = \{1, 2, \dots, n\}$ espace d'états fini.

$\mathbb{P} = (P_{ij})_{1 \leq i, j \leq n}$ matrice stochastique $\left(P_{ij} \geq 0 \quad \sum_{j=1}^{n} P_{ij} = 1 \ \forall i \right)$

Alors la CM avec $\begin{cases} \text{espace d'états } E \\ \text{matrice de transit}^{\circ}\ \mathbb{P} \end{cases}$

est le processus $X = (X_n ;\ n \in \mathbb{N})$ donné par

$$\begin{cases} X_0 \in E \quad \text{(éventuellement aléatoire)} \\ \mathbb{P}_{x,y} = \mathbb{P}(X_{n+1} = y \mid X_n = x) = \mathbb{P}(X_{n+1} = y \mid X_n = x, X_{n-1}, \dots, X_0) \end{cases}$$

Moralement c'est l'exemple le plus simple après une suite i.i.d. On parle souvent de dependance faible car la dependance est uniquement en l'étape précédente.

(≠ AR(p) ou MA(q) si p, q >> 1, **qui seront vues en séries chronologiques**)

(2) Thm ergodique de Birkhoff :

ou "loi forte des grands nombres"

Si X irred, ~~apériodique~~

Alors $\frac{1}{T} \sum_{t=0}^{T} f(X_t) \xrightarrow[T \to +\infty]{p.s} \mathbb{E}_{\mathcal{L}(X)=\pi}(f(X))$

$= \sum_{x \in E} \pi_x f(x).$

où π unique mesure invariante

$(\pi P = \pi)$

? Lien ?

Problème : $E = \{$ Mots dans un texte $\}$

$P =$ Matrice de transitions pour la chaîne de Markov dont UNE réalisation est notre texte.

Comment estimer ce P ?

→ **Statistique**

II L'exemple historique de Markov (Andrei)

Andrei Markov était en désaccord avec Pavel Nekrasov sur le fait que l'indépendance est nécessaire pour la loi (faible) des grands nombres.

A cet effet, dans son papier de 1913, Markov a choisi une phrase de longueur $T = 20\,000$ lettres du livre "Eugene Onegin" de Pushkin.

Il a estimé une chaîne de Markov sur $E = \{V, C\}$

avec $$\hat{P}_T = \begin{array}{c} \\ V \\ C \end{array} \begin{array}{c} \begin{array}{cc} V & C \end{array} \\ \begin{pmatrix} 0{,}128 & 0{,}872 \\ 0{,}663 & 0{,}337 \end{pmatrix} \end{array}$$

Puis il a calculé $\pi = (0{,}432 \; ; \; 0{,}568)$ où $\pi P = \pi$.

et prouvé la loi (faible) des grands nombres / le thm de CV à l'équilibre (version faible de Birkhoff).

Question : Comment a-t-il trouvé ce $\hat{P}_T$?

↳ Il a pris ce texte de longueur $T = 20\,000$.

↳ Il a soigneusement noté la réalisat° correspondte de la CM X de matrice P.

Ex: Eugene Onegin.

VVCVCV VCVCVC

↳ Calcul : $$\hat{P}_T = \begin{pmatrix} \hat{\alpha}_T & 1-\hat{\alpha}_T \\ 1-\hat{\beta}_T & \hat{\beta}_T \end{pmatrix} \xrightarrow[T \to +\infty]{?} P = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}$$

Converge (en Proba ou p.s)

où $$\hat{\alpha}_T = \frac{\#\{0 \leq t \leq T-1 \mid X_t = X_{t+1} = V\}}{\#\{0 \leq t \leq T \mid X_t = V\}} \xrightarrow{?} \alpha$$

$$\hat{\beta}_T = \frac{\#\{0 \leq t \leq T-1 \mid X_t = X_{t+1} = C\}}{\#\{0 \leq t \leq T \mid X_t = C\}} \xrightarrow{?} \beta$$

↳ Pour quoi cela fonctionne-t-il ?

Théorème : Soit $(X_t ; t \in \mathbb{N})$ une chaîne de Markov sur E.

Alors $[\hat{P}_T]_{ij} := \dfrac{n_{ij}}{\sum_{k=1}^{n} n_{ik}}$ où $n_{ij} = \# \{ 0 \leq t \leq T-1 \mid (X_t, X_{t+1}) = (i,j) \}$

$\left(\sum_k n_{ik} = \# \{ 0 \leq t \leq T-1 \mid X_t = i \} \right)$

donne
- une matrice stochastique $\hat{P}_T$.
- $\hat{P}_T$ est un estimateur consistant de P

$$\left(\hat{P}_T \xrightarrow[T \to +\infty]{\mathbb{P}/ps} P \right)$$

III La preuve probabiliste par le thm ergodique de Birkhoff.

Exercice guidé dans le cas Markov

$$P = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix} \quad \text{avec} \quad \begin{cases} \alpha = \mathbb{P}(X_{t+1} = V \mid X_t = V) \\ \beta = \mathbb{P}(X_{t+1} = C \mid X_t = C) \end{cases}$$

(1) Par Birkhoff :

$$\frac{1}{T} \# \{ 0 \leq t \leq T \mid X_t = V \} = \frac{1}{T} \sum_{t=0}^{T} \mathbb{1}_V(X_t) \xrightarrow[T \to +\infty]{} ? \quad \mathbb{E}_\pi \mathbb{1}_V(X) = \mathbb{P}_\pi(X = V) = \pi_V$$

$$\frac{1}{T} \# \{ 0 \leq t \leq T \mid X_t = C \} \xrightarrow[T \to +\infty]{} ? \quad \pi_C$$

(2) Que reste-t-il à montrer pour que

$$\hat{\alpha}_T = \frac{\frac{1}{T} \# \{ 0 \leq t \leq T-1 \mid X_t = X_{t+1} = V \}}{\frac{1}{T} \# \{ 0 \leq t \leq T-1 \mid X_t = V \}} \xrightarrow[T \to +\infty]{} \alpha = P_{VV}$$

Il reste

$$\frac{1}{T} \# \{ 0 \leq t \leq T-1 \mid X_t = X_{t+1} = V \} \xrightarrow{T \to +\infty} \pi_V P_{VV}$$

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{1}_{\{V,V\}}(X_t, X_{t+1})$$ A besoin d'instants successifs ! Pas Birkhoff

Invitation à se demander si (X_t, X_{t+1}) chaîne de Markov.

Sondage : OUI ① ou NON ② ?

Lemme : $((X_t, X_{t+1})\ ;\ t \in \mathbb{N})$ est une chaîne de Markov de matrice de transition $Q = \left(Q_{(i,j),(k,\ell)}\right)_{\substack{1\le i,j\le n \\ 1\le k,\ell\le n}}$

$$Q_{(i,j),(k,\ell)} = \mathbb{P}\left((X_{t+1}, X_{t+2}) = (k,\ell) \mid (X_t, X_{t+1}) = (i,j), \dots \right)$$

$$= \mathbb{1}_{\{j=k\}}\, P_{k,\ell}$$

et de mesure invariante π_Q où

$$\pi_Q(k,\ell) = \pi_P(k)\, P_{k,\ell}.$$

Preuve : $$Q_{(i,j),(k,\ell)} = \mathbb{P}\left((X_{t+1}, X_{t+2}) = (k,\ell) \mid (X_t, X_{t+1}) = (i,j), \dots \right)$$

$$= \mathbb{1}_{\{j=k\}}\, \mathbb{P}\left(\cancel{X_{t+1} = k},\ X_{t+2} = \ell \mid X_{t+1} = k,\ X_t = i, \dots \right)$$

$$\overset{\text{Markov}}{=} \mathbb{1}_{\{j=k\}}\, \mathbb{P}\left(X_{t+2} = \ell \mid X_{t+1} = k \right)$$

$$= \mathbb{1}_{\{j=k\}}\, P_{k,\ell}.$$

$$\pi_Q(k,\ell) \overset{\text{def}}{=} \pi_P(k)\, P_{k,\ell} \overset{?}{=} (\pi_Q\, Q)(k,\ell)$$

$$= \sum_{i,j} \pi_Q(i,j)\, Q\big((i,j),(k,\ell)\big)$$

$$\overset{\text{juste avant}}{=} \sum_{i,j} \pi_Q(i,j)\, \mathbb{1}_{\{j=k\}}\, P_{k,\ell}$$

$$= \sum_i \pi_Q(i,k)\, P_{k,\ell} = \left[\sum_i \pi_Q(i,k)\right] P_{k,\ell}$$

$$\overset{\text{def } \pi_Q}{=} \underbrace{\left(\sum_i \pi_P(i)\, P_{i,k}\right)}_{\pi_P(k) \text{ def de } \pi_P \text{ mesure invariante}} P_{k,\ell}$$

$$= \pi_P(k)\, P_{k,\ell}$$

$$= \pi_Q(k,\ell)$$

$\boxed{///}$

Ainsi en appliquant Birkhoff à (X_t, X_{t+1}) (✱)

$$\frac{1}{T} \# \{ 0 \leq t \leq T-1 \mid X_t = X_{t+1} = V \}$$

$$= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{1}_{\{V,V\}}(X_t, X_{t+1})$$

$$\xrightarrow[T \to +\infty]{(✱)} \mathbb{E}_{\pi_Q} \mathbb{1}_{\{V,V\}}(X,Y) = \mathbb{P}_{\pi_Q}\big((X,Y) = (V,V)\big)$$

$$= \pi_Q(V,V)$$

$$\overset{\text{def}}{=} \pi_V\, P_{VV}$$

CQFD.

Sondage : Qui avait repéré où était utilisée la formule $[\hat{P}_T]_{ij} = \frac{n_{ij}}{\sum_{k=1}^{m} n_{ik}}$ dans le TP ?

① J'avais repéré.

② Non je n'avais repéré.

③ Je ne comprends pas la question.

Au programme aujourd'hui : Faire les preuves du thm

i.e $\hat{P}_T \xrightarrow[T\to+\infty]{ps/\mathbb{P}} P$ ← la vraie matrice de transition.

Rappel : En statistique, on dit que $\hat{P}_T$ est un estimateur consistant.

III La preuve probabiliste par le thm. ergodique de Birkhoff (suite et fin).

$$[\hat{P}_T]_{ij} = \frac{n_{ij}}{\sum_{k=1}^{m} n_{ik}} = \frac{\frac{1}{T}\#\{0\leq t\leq T-1 \mid (X_t, X_{t+1}) = (i,j)\}}{\frac{1}{T}\#\{0\leq t\leq T-1 \mid X_t = i\}}$$

$$\overset{?}{\underset{T\to+\infty}{\longrightarrow}} P_{ij}$$

A montrer pour (i,j) fixé

Comme pour le cas 2×2 avec les consonnes-voyelles de "Eugene Onegin", le dénominateur est traité par le thm de Birkhoff (en supposant irréductible + apériodique).

$$\frac{1}{T} \# \{ 0 \leq t \leq T-1 \mid X_t = i \} \qquad \mathbb{E}_\pi [\mathbb{1}_{\{i\}}(X)]$$

$$= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{1}_{\{i\}}(X_t) \xrightarrow[T \to +\infty]{\text{Birkhoff (p.s)}} \pi(i) \quad \text{où } \pi \text{ mesure invariante}$$

Pour le numérateur, on écrit :

$$\frac{1}{T} \# \{ 0 \leq t \leq T-1 \mid (X_t, X_{t+1}) = (i,j) \}$$

$$= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{1}_{\{(i,j)\}}(X_t, X_{t+1})$$

$$\xrightarrow[T \to +\infty]{p.s} \pi_Q(i,j)$$

par Birkhoff appliqué à

⚠ Par le lemme, $((X_t, X_{t+1})\,;\, t \in \mathbb{N})$ est une chaîne de Markov

où π_Q probabilité invariante de $Y = ((X_t, X_{t+1})\,;\, t \in \mathbb{N})$

Rappel (Cours 2) : $\pi_Q(k,\ell) = \pi_P(k) P_{k,\ell}$.

Donc $\frac{1}{T} \# \{ 0 \leq t \leq T-1 \mid (X_t, X_{t+1}) = (i,j) \} \xrightarrow[T \to +\infty]{p.s.} \pi(i) P_{ij}$

Par quotient :

$$[\hat{P}_T]_{ij} = \frac{\frac{1}{T} \# \{ 0 \leq t \leq T-1 \mid (X_t, X_{t+1}) = (i,j) \}}{\frac{1}{T} \# \{ 0 \leq t \leq T-1 \mid X_t = i \}} \xrightarrow[T \to +\infty]{p.s} \frac{\cancel{\pi(i)} P_{ij}}{\cancel{\pi(i)}} = P_{ij}$$

CQFD. ▨

Exo: Irréd. $\Rightarrow \forall i \in E, \pi(i) > 0$.

Commentaire final : Sondage. Maintenant que vous avez vu la preuve et les arguments qu'elle utilise (CM, Birkhoff, Birkhoff pour les paires, Ergodicité) ;

ne pensez-vous pas que la formule était plutôt intuitive et n'avait pas besoin de tout cela ?

(1) Oui cette preuve est un overkill. 16

(2) Non la preuve est nécessaire et la formule n'est pas intuitive. 12

Commentaire final (bis).

AR(1) : $X_{t+1} = a X_t + \varepsilon_t$ $\quad X = (X_t \,;\, t \in \mathbb{Z})$

X a la propriété de Markov :

$$\mathcal{L}(X_{t+1} \mid \underset{x}{\overset{\shortparallel}{X_t}}, X_{t-1}, \dots) = \mathcal{L}(ax + \varepsilon).$$

AR(2) : $X_{t+1} = a_0 X_t + a_1 X_{t-1} + \varepsilon_t$

X n'est pas Markov.

Mais $Y = ((X_t, X_{t+1}) \,;\, t \in \mathbb{Z})$ est Markov.

En effet : $Y_t = \begin{bmatrix} X_t \\ X_{t+1} \end{bmatrix}$

$$Y_{t+1} = \begin{bmatrix} X_{t+1} \\ X_{t+2} \end{bmatrix} = \begin{bmatrix} X_{t+1} \\ a_0 X_{t+1} + a_1 X_t + \varepsilon_t \end{bmatrix} = \begin{bmatrix} 0 \\ \varepsilon_t \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ a_1 & a_0 \end{bmatrix} \overset{Y_t}{\overset{\shortparallel}{\begin{bmatrix} X_t \\ X_{t+1} \end{bmatrix}}}$$

$$= \begin{bmatrix} 0 \\ \varepsilon_t \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ a_1 & a_0 \end{bmatrix} Y_t$$

Plus généralement, dans le AR(p), une "fenêtre" de longueur p pour avoir la propriété de Markov.

$$Y^{(p)} = ((X_t, X_{t+1}, \dots, X_{t+p}) \,;\, t \in \mathbb{Z}) \text{ Markov}$$

⇝ Moralement, notre TP 2 était une histoire de processus autoregressif (AR) sur l'espace des mots.

IV La preuve du statisticien (avec une bonne dose d'optimisat°)

« Méta-thm de la statistique » :

Il faut considérer l'estimateur du maximum de vraisemblance (MLE : Maximum Likelihood Estimator).

Exercice pour se rafraichir la mémoire.

$(X_1, X_2, \dots, X_n)$ un échantillon iid de temps d'attente à l'arrêt d'un bus.

On suppose une loi exponentielle de densité

$$f_\theta(x) = \theta e^{-\theta x} \mathbb{1}_{\{x>0\}} \qquad \theta \text{ param inconnu.}$$

Rappel : $\mathbb{E} X = \frac{1}{\theta} \leadsto \hat{\theta}_n = \frac{1}{\hat{m}_n} = \frac{n}{\sum_{i=1}^n X_i}$

Calculer le MLE : θ_n^* estimateur consistant (loi des gds nombres).

$$\hat{\theta}_n^* = \operatorname{Argmax}_\theta L_n(\theta)$$

L_n ↳ vraisemblance de l'échantillon $(X_1, \dots, X_n) = (x_1, \dots, x_n)$

$$= \operatorname{Argmax}_\theta \log L_n(\theta)$$

"Vraisemblance" = Proba / densité de voir ce que je vois.

$\leadsto L_n(\theta) =$ Densité de $(X_1, X_2, \dots, X_n)$ en $(x_1, x_2, \dots, x_n)$

$$\overset{\text{indépendance}}{=} f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n)$$

$$\rightsquigarrow \quad \log L_n(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i)$$

$$= \sum_{i=1}^{n} \log\left(\theta e^{-\theta x_i} \mathbb{1}_{\{x_i > 0\}}\right)$$

$$= \sum_{i=1}^{n} \Big[\log\theta + \underbrace{\log e^{-\theta x}}_{-\theta x_i} + \underbrace{\log \mathbb{1}_{\{x_i > 0\}}}_{\substack{-\infty \\ \mathbb{1}_{\{x_i \leq 0\}}}}\Big]$$

$$= n\log\theta - \theta\sum_{i=1}^{n} x_i$$

$$-\infty\, \mathbb{1}_{\{x_1 \leq 0 \text{ ou } x_2 \leq 0 \text{ ou } \dots \text{ ou } x_n \leq 0\}}$$

Donc $\hat{\theta}_n^* = \underset{\theta}{\text{Argmax}} \underbrace{n\log\theta - \theta\sum_{i=1}^{n} x_i}_{\varphi(\theta)}$

$$\varphi'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i \overset{?}{=} 0$$

$$\Longleftrightarrow \frac{n}{\theta} = \sum_{i=1}^{n} x_i \Longleftrightarrow \theta = \frac{n}{\sum_{i=1}^{n} x_i}$$

Donc $\theta_n^* = \text{MLE} = \dfrac{n}{\sum_{i=1}^{n} x_i} \overset{!}{=} \hat{\theta}_n$. ∎

Revenons à notre chaîne de Markov.

On observe un échantillon $(X_0, X_1, X_2, \dots, X_T)$

↳ Paramètre : $\theta = P$. $= (x_0, x_1, x_2, \dots, x_T)$

↳ Calcul de la vraisemblance

$$L = L(P) = \mathbb{P}(X_T = x_T, X_{T-1} = x_{T-1}, \dots, X_0 = x_0)$$

$$= \mathbb{P}(X_T = x_T \mid X_{T-1} = x_{T-1}, \dots, X_0 = x_0) \overset{\text{Markov}}{=} \mathbb{P}(X_T = x_T \mid X_{T-1} = x_{T-1})$$

$$\times\, \mathbb{P}(X_{T-1} = x_{T-1}, \dots, X_0 = x_0)$$

$$= P_{x_{T-1}, x_T}$$

$$= P_{x_{T-1}, x_T} \times \mathbb{P}(X_{T-1} = x_{T-1}, \dots, X_0 = x_0)$$

$= \dots$ Récurrence immediate.

$$= P_{x_{T-1}, x_T} \times P_{x_{T-2}, x_{T-1}} \times \dots \times P_{x_1, x_0}$$

$$= \left(\prod_{i=0}^{T-1} P_{x_i, x_{i+1}} \right) \times \mathbb{P}(X_0 = x_0)$$

$\mathbb{P}(X_0 = x_0)$

↳ Ré-écriture :

$$L(P) = \mathbb{P}(X_0 = x_0) \times \prod_{1 \leq i,j \leq m} P_{ij}^{\#\{(i,j) \text{ apparait dans } (x_0, x_1, \dots, x_T)\}}$$

$$= \mathbb{P}(X_0 = x_0) \times \prod_{1 \leq i,j \leq m} P_{ij}^{n_{ij}}$$

Donc MLE est donné par

$$\hat{P}_T = \underset{P}{\text{Argmax}} \log L(P)$$

$$= \underset{P}{\text{Argmax}} \cancel{\log \mathbb{P}(X_0 = x_0)} + \sum_{1 \leq i,j \leq m} \log P_{ij}^{n_{ij}}$$

$$\Rightarrow \boxed{\hat{P}_T = \underset{P}{\text{Argmax}} \underbrace{\sum_{1 \leq i,j \leq m} n_{ij} \log P_{ij}}_{\mathcal{L}(P)}}$$

⚠ Il faut optimiser sur un espace de matrices P de taille $m \times n$ sous contrainte $\sum_{j=1}^{n} P_{ij} = 1 \; \forall i$

matrice stochastique

↳ Calcul par la méthode des multiplicateurs de Lagrange :

Poser $\mathcal{L}_c(P, \lambda) = \mathcal{L}(P) - \sum_{i=1}^{m} \lambda_i \Big(\underbrace{\sum_{j=1}^{n} P_{ij} - 1}_{\text{contrainte}} \Big)$

Loss précédente

multiplicateur de Lagrange pour la contrainte i.

PUIS optimiser comme si P et $\lambda = (\lambda_1, \dots, \lambda_m)$ sont libres / sans contraintes.

(P, λ) point critique

$$\iff \begin{cases} \forall i, & \dfrac{\partial \mathcal{L}_c}{\partial \lambda_i} = 0 \\ \forall i,j, & \dfrac{\partial \mathcal{L}_c}{\partial P_{ij}} = 0 \end{cases} \iff \begin{cases} \sum_{j=1}^{n} P_{ij} - 1 = 0 \;\; \forall i \\ \forall i,j, \quad \dfrac{n_{ij}}{P_{ij}} - \lambda_i = 0 \end{cases}$$

$$\iff \begin{cases} \forall i,j, & P_{ij} = \dfrac{n_{ij}}{\lambda_i} \\ \forall i, & 1 = \sum_{j=1}^{n} P_{ij} = \Big(\sum_{j=1}^{n} n_{ij}\Big) / \lambda_i \rightsquigarrow \lambda_i = \sum_{j=1}^{n} n_{ij} \end{cases}$$

Donc $\boxed{\hat{P}_{ij} = \dfrac{n_{ij}}{\sum_{k=1}^{n} n_{ik}}} = [\hat{P}_T]_{ij}$

C'est bien la formule de Markov

(Markov a donné le MLE

et $\hat{P}_T \xrightarrow[T \to +\infty]{\mathbb{P}} P$ en vertu

des thms généraux de CV du MLE)