Rapport de stage

Alphonse Paix

Année 2022–2023

Table des matières

1 Les processus ponctuels temporels

Définition 1.1 (Processus ponctuel)**.** Un processus ponctuel temporel est une distribution de probabilité sur des séquences de longueur variable.

Une réalisation d'un processus ponctuel est un objet de la forme $\{x_1, \dots, x_N\}$ avec $x_i \in \mathcal{X}$ et N une variable aléatoire à valeurs dans $\mathbb{N}$.

Le choix de l'espace $\mathcal{X}$ va distinguer les différents types de processus ponctuels. Par exemple pour $\mathcal{X} = \mathbb{R}^2$, chaque réalisation pourrait représenter des coordonnées spatiales.

Pour $\mathcal{X} \subseteq [0, +\infty[$, on parle de processus ponctuels temporels. Les événements, notés t_i, apparaissent au cours du temps sur la demi-droite. Une propriété importante est que ces événements sont ordonnés, c'est-à-dire que la séquence $t = (t_1, \dots, t_N)$ vérifie $0 < t_1 < t_2 < \cdots < t_N$. Une autre propriété est que l'événement t_i est habituellement seulement influencé par les événements qui le précèdent, ce qui n'est pas le cas pour des processus spatio-temporels.

1.1 Le modèle auto-régressif

Comment peut-on générer une séquence $\{t_1, \dots, t_N\}$ dans l'intervalle $[0, T]$? On commence par tirer t_1 de sa distribution que l'on note $P(t_1)$. Si $t_1 > T$ on s'arrête et on obtient la séquence vide t. Sinon on continue et on tire t_2 de sa distribution $P(t_2 \mid t_1)$. Si $t_2 > T$ on s'arrête, sinon on continue

et on tire t_3 de sa distribution $P(t_3 \,|\, t_1, t_2)$ et ainsi de suite jusqu'à tirer $t_{N+1} > T$. On obtient alors la séquence t constituée de N éléments vérifiant $t_1 < t_2 < \cdots < t_N < T$.

À chaque étape la distribution du temps d'arrivée est conditionnée par l'*historique* des éléments qui précèdent. On note $\mathcal{H}(t_i) = \{t_j, j < i\}$ et

$$P(t_i \,|\, \mathcal{H}(t_i))$$

la distribution du temps d'arrivée t_i. On note également cette distribution $P^*(t_i)$ où l'astérisque dénote le conditionnement par rapport au passé.

Comment peut-on représenter cette distribution ?

1.2 La fonction d'intensité conditionnelle

En apprentissage statistique, une distribution $P^*(t_i)$ est souvent caractérisée par sa fonction de densité notée $t \mapsto f_i^*(t)$. La valeur $f_i^*(t_i)\,dt$ représente la probabilité que l'événement t_i se produise dans l'intervalle $[t, t + dt]$ où dt est infinitésimal.

On peut également utiliser la fonction de répartition $t \mapsto F_i^*(t) = \int_0^t f_i^*(s)\,ds$ qui donne la probabilité que l'événement se produise avant t et la fonction de survie $t \mapsto S_i^*(t) = 1 - F_i^*(t)$ qui est la probabilité que l'événement se produise après t.

Définition 1.2 (Fonction de danger)**.** Une autre possibilité pour caractériser une distribution est la fonction de danger h_i^*, définie par

$$t \mapsto h_i^*(t) = \frac{f_i^*(t)}{S_i^*(t)}$$

où la quantité $h_i^*(t)\,dt$ représente la probabilité que l'événement t_i se produise dans l'intervalle $[t, t + dt]$ *sachant* qu'il ne s'est pas produit avant t.

On considère le scénario où l'événement t_{i-1} vient de se produire et notre horloge est au temps $t = t_{i-1}$. Le temps passe et on se situe au nouveau temps t et on considère la probabilité que l'événement se produise dans l'intervalle $[t, t + dt]$. Cette probabilité, notée $\mathbb{P}(t_i \in [t, t + dt] \,|\, \mathcal{H}_t)$ n'est plus égale à $f_i^*(t)\,dt$. Il faut conditionner par rapport au fait que l'événement t_i ne s'est pas produit avant t. Pour cela on renormalise la densité de sorte qu'elle vaille 1 sur l'intervalle $[t, +\infty]$:

$$f_i^*(t \,|\, t_i > t) = \frac{f_i^*(t)}{\int_t^{+\infty} f_i^*(t)\,dt} = \frac{f_i^*(t)}{S_i^*(t)} = h_i^*(t).$$

On peut également retrouver la densité à partir de h_i^* :

$$h_i^*(t) = \frac{f_i^*(t)}{S_i^*(t)} = \frac{-\frac{d}{dt} S_i^*(t)}{S_i^*(t)} = -\frac{d}{dt} \log S_i^*(t),$$

soit

$$S_i^*(t) = \exp\left(- \int_{t_i - 1}^t h_i^*(s)\,ds \right).$$

On dérive pour obtenir la densité :

$$f_i^*(t) = -\frac{d}{dt} S_i^*(t) = h_i^*(t) \exp\left(- \int_{t_i - 1}^t h_i^*(s)\,ds \right).$$

La fonction de danger est souvent utilisée en statistiques lorsqu'on s'intéresse par exemple à la probabilité d'apparition d'une panne dans un système lorsqu'on sait qu'elle ne s'est pas produite avant un certain instant.

Pour résumé, on peut caractériser le processus ponctuel de plusieurs manières, par exemple en spécifiant les fonctions de densité de chaque événément $\{f_1^*, f_2^*, \dots, f_N^*\}$ ou les fonctions de danger $\{h_1^*, h_2^*, \dots, h_N^*\}$. Mais cela introduit de la lourdeur dans la notation. À la place, on peut spécifier la fonction d'intensité conditionnelle :

Définition 1.3 (Fonction d'intensité conditionnelle)**.** La fonction d'intensité conditionnelle, notée λ^* est définie par :

$$\lambda^*(t) = \begin{cases} h_1^*(t) & \text{si } t \leq t_1, \\ h_2^*(t) & \text{si } t_1 \leq t < t_2, \\ \vdots & \\ h_N^*(t) & \text{si } t_{N-1} < t \leq t_N \\ h_{N+1}^*(t) & \text{si } t_N < t \leq T. \end{cases}$$

L'astérisque de λ^* dénote le conditionnement par rapport au passé :

$$\lambda^*(t) = \lambda^*(t \,|\, \mathcal{H}(t))$$

où $\mathcal{H}(t)$ est l'historique des événements antérieurs à t.

Ainsi, pour spécifier la distribution d'un processus ponctuel, il nous suffit juste de définir une fonction à valeurs positives qui prend un premier paramètre, le temps $t \in [0, T]$ et un deuxième paramètre qui est une séquence de longueur variable $\{t_1, \dots, t_{i-1}\}$ représentant l'historique des événements antérieurs. Cette fonction définit complètement la distribution.

1.3 Définir un processus avec la fonction d'intensité

La fonction d'intensité comme vu précédemment permet de définir complètement la distribution d'un processus ponctuel. On pourrait choisir une fonction d'intensité indépendante du temps :

$$\lambda^*(t) = g(t)$$

avec g une fonction positive.

Remarque. Cette définition correspond aux processus de Poisson.

Des valeurs de g élevées correspondent à un taux d'occurence plus important. On pourrait imaginer un intervalle $[0, T]$ qui représenterait le temps d'une journée et des valeurs de g élevées le matin et le soir pour caractériser le traffic internet d'un site web par exemple.

Un autre exemple d'intensité conditionnelle est le suivant :

$$\lambda^*(t) = \mu + \sum_{t_j \in \mathcal{H}(t)} \alpha \exp(-(t - t_j))$$

avec μ et α deux constantes positives qui représentent le taux d'occurence de fond et le saut effectué par la fonction à chaque occurence. La fonction augmente de α à chaque apparition d'un nouvel événement et décroît exponentiellement vers le taux de fond μ jusqu'à l'apparition d'un nouvel événement. La somme caractérise le phénomène de cascade engendré par un événement, qui va temporairement augmenter la probabilité d'en faire apparaître un autre.

1.4 Le processus de comptage

Il est également possible de définir le processus ponctuel comme un processus de comptage. Chaque réalisation d'un processus est une fonction $t \mapsto N(t) \in \mathbb{N}$ où N est une fonction croissante du temps. Bien sûr la fonction N est définie par

$$N(t) = \sum_{i=1}^{N} \mathbb{1}_{t_j < t}(t)$$

et représente le nombre d'événements qui se sont produits avant l'instant t.

Comment caractériser la distribution du processus ? On va spécifier la fonction d'intensité conditionnelle. On considère une quantité infinitésimale dt et on s'intéresse à la variation de N dans l'intervalle $[t, t + dt]$:

$$\begin{aligned}
\mathbb{E}\left[N(t+dt) - N(t) \,|\, \mathcal{H}(t)\right] &= \mathbb{P}(\text{prochain événement } t_i \text{ dans } [t, t+dt] \,|\, \mathcal{H}(t)) \\
&\quad + 0 \times \mathbb{P}(\text{pas d'événement dans } [t, t+dt] \,|\, \mathcal{H}(t)) \\
&= \mathbb{P}(\text{prochain événement } t_i \text{ dans } [t, t+dt] \,|\, \mathcal{H}(t)) \\
&= h_i^*(t)\, dt \\
&= \lambda^*(t)\, dt.
\end{aligned}$$

On obtient alors

$$\lambda^*(t) = \lim_{dt \to 0} \frac{\mathbb{E}\left[N(t+dt) - N(t) \,|\, \mathcal{H}(t)\right]}{dt}$$

qui est la fonction d'intensité du processus. Elle représente alors le nombre d'occurences attendues dans un petit intervalle par unité de temps.

2 Un modèle neuronal génératif

On va définir le modèle de manière auto-regréssive, c'est-à-dire qu'à chaque étape $i = 1, 2, \ldots$ nous allons définir la distribution $P^*(t_i)$ du prochain temps d'arrivée t_i en conditionnant par rapport à l'historique des événements antérieurs $\mathcal{H}(t_i) = \{t_1, \ldots, t_{i-1}\}$.

2.1 Représentation des données

Nous allons travailler avec les temps d'attente. Pour une séquence $t = (t_1, t_2, \ldots, t_N)$ on définit les temps d'attente comme la différence entre chaque temps d'arrivée consécutif, en prenant $t_0 = 0$ et $t_{N+1} = T$ on a alors :

$$\tau_i = t_i - t_{i-1}$$

et on obtient la séquence complémentaire $\tau = \{\tau_1, \ldots, \tau_N, \tau_{N+1}\}$ avec le dernier temps d'attente défini comme la différence entre l'horizon T et le temps d'apparation du dernier événement.

Pour la construction du modèle génératif, on dispose de N séquences de longueurs variables $l_i, i = 1, \ldots, N$ où l_i est la longueur de la i-ième séquence. On commence alors par calculer les temps d'attente et obtenir les séquences correspondantes pour chaque élément, puis on rajoute le nombre de 0 nécessaire à chaque séquence pour qu'elles aient toutes la même taille. Si on dispose

en entrée de trois séquences de tailles 3, 4 et 2 respectivement, on obtient après transformation une matrice de trois lignes et cinq colonnes :

$$\begin{pmatrix} t_1^{(1)} & t_2^{(1)} & t_3^{(1)} & \\ t_1^{(2)} & t_2^{(2)} & t_3^{(2)} & t_4^{(2)} \\ t_1^{(3)} & t_2^{(3)} & & \end{pmatrix} \quad \Longrightarrow \quad \begin{pmatrix} \tau_1^{(1)} & \tau_2^{(1)} & \tau_3^{(1)} & \tau_4^{(1)} & 0 \\ \tau_1^{(2)} & \tau_2^{(2)} & \tau_3^{(2)} & \tau_4^{(2)} & \tau_5^{(2)} \\ \tau_1^{(3)} & \tau_2^{(3)} & \tau_3^{(3)} & 0 & 0 \end{pmatrix}$$

2.2 Paramétrisation de la distribution

Le but du modèle est de paramétriser la distribution du prochain temps d'attente. Ceci s'effectue en plusieurs étapes :

1. on encode l'historique $\mathcal{H}(t_i) = \{t_1, \dots, t_{i-1}\}$ dans un vecteur c_i appelé vecteur contexte de taille fixe à l'aide d'un réseau de neurones ;
2. on choisit une distribution paramétrique définie par sa fonction de densité $f(\cdot \mid \theta)$ qui définit une variable aléatoire positive (notre prochain temps d'attente) ;
3. on utilise le vecteur contexte c_i pour obtenir θ_i qui caractérise la distribution du prochain temps d'attente $P_i^*(\tau_i)$.

2.2.1 Encodage de l'historique dans un vecteur

Chaque temps t_i est encodé sous la forme $y_i = (\tau_i, \log \tau_i)$ et on initialise le vecteur contexte c_1 de manière aléatoire. Le vecteur c_1 est le premier état du réseau en plus d'être utilisé pour la paramétrisation du premier temps d'attente. À chaque nouvel événement, on calcule l'état suivant du vecteur contexte en utilisant l'équation du réseau de neurones :

$$c_{i+1} = \mathrm{rnn}(c_i, y_i)$$

où rnn est une fonction qui implémente une transformation régie par l'architecture du réseau.

2.3 Choix de la distribution paramétrique

Le choix de la distribution doit respecter quelques contraintes :

1. il faut pouvoir calculer les fonctions de densité et de survie analytiquement pour le calcul de la log-vraisemblance (qui sera l'objectif d'entrainement du modèle) ;
2. il faut pouvoir effectuer des tirages à partir de la distribution choisie ;
3. la distribution doit être assez complexe pour pouvoir représenter de manière fidèle la variable qu'elle modélise.

Un premier choix est la distribution de Weibull dont on donne les fonctions de densité et de survie :

$$f(\tau \mid b, k) = bk\tau^{k-1} \exp(-b\tau^k) \qquad S(\tau \mid b, k) = \exp(-b\tau^k)$$

avec $\tau > 0$ et $(b, k) \in \mathbb{R}_+^2$ les deux paramètres de la distribution.

2.3.1 Paramétrisation de la distribution

Les paramètres b_i et k_i sont obtenues comme résultat de l'application d'une composition d'une transformation linéaire et d'une fonction d'activation non linéaire sur le vecteur contexte c_i :

$$b_i = \sigma(v_d^T c_i + d_b) \qquad k_i = \sigma(v_k^T c_i + d_k)$$

avec σ qui ici vient appliquer les containtes sur les paramètres et où les matrices $v_d, v_k \in \mathbb{R}^C$ et $b_d, b_k \in \mathbb{R}$ sont les paramètres entraînables du modèle (C est la taille du vecteur contexte c_i).

Remarque. On peut choisir $\sigma = \text{softplus}$ où softplus est définie par $\text{softplus}(x) = \frac{1}{\beta}\log(1+\exp(\beta \cdot x))$ avec $\beta \cdot x$ le produit élément par élément des termes de la matrice x par le scalaire β.

2.4 Calcul de la log-vraisemblance

La log-vraisemblance est l'objectif pour l'entraînement du modèle génératif. Comment obtient-on la log-vraisemblance pour un processus ponctuel ? On suppose qu'on a observé qu'un seul événement dans l'intervalle $[0, T]$. Cela se traduit par

$$\begin{aligned}\mathbb{P}(\{t_1\}) &= \mathbb{P}(\text{premier événement dans } [t_1, t_1 + dt]) \\ &\quad + \mathbb{P}(\text{second événement après } T \mid t_1) \\ &= f_1^*(t_1)\, dt \times S_2^*(T).\end{aligned}$$

Ce résultat se généralise lorsqu'on observe pas un seul mais plusieurs événements $t = \{t_1, \ldots, t_N\}$ dans l'intervalle :

$$\mathbb{P}(t) = (dt)^N \left(\prod_{i=1}^{N} f_i^*(t_i) \right) S_{N+1}^*(T)$$

avec $(dt)^N$ un terme que l'on ignore durant l'entraînement.

On obtient la log-vraisemblance :

$$\log \mathbb{P}(t) = \sum_{i=1}^{N} \log f_i^*(t_i) + \log S_{N+1}^*(T),$$

soit en repassant aux temps d'attente :

$$\log \mathbb{P}(t) = \sum_{i=1}^{N} \log f_i^*(\tau_i) + \log S_{N+1}^*(\tau_{N+1}).$$

2.5 Entraînement

L'objectif de l'entraînement est la minimisation de la log-vraisemblance négative, qui s'effectue sur plusieurs itérations. On utilise un optimiseur tel que `rmsprop` ou `adam` qui implémente un algorithme de rétropropagation (par descente de gradient stochastique) dont le but est de minimiser la perte par rapport aux paramètres entraînables du modèle.

Pour rappel les paramètres entraînables du modèle sont les paramètres du réseau de neurones récurrents et les paramètres de la transformation linéaire activée qui achève de paramétriser la distribution recherchée.

Plusieurs hyperparamètres tels que le nombre d'itérations complètes sur les données, le choix de l'optimiseur, la vitesse à laquelle s'effectue la descente de gradient et la taille C du vecteur contexte produit par le réseau de neurones viennent influer sur les performances du modèle.

Références

[1] Kyunghyun Cho et al. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. arXiv : 1406.1078 [cs.CL].

[2] Kelian Dascher-Cousineau, Oleksandr Shchur et Emily Brodsky. « Flexible and Scalable Earthquake Forecasting ». In : *AGU Fall Meeting Abstracts*. T. 2021. 2021, S33B-02.

[3] Oleksandr Shchur. « Modeling Continous-time Event Data with Neural Temporal Point Processes ». Thèse de doct. Technische Universität München, 2022.

[4] Oleksandr Shchur. *Temporal Point Processes 1: The Conditional Intensity Function*. 2020. url : https://shchur.github.io/blog/2020/tpp1-conditional-intensity (visité le 24/05/2023).

[5] Oleksandr Shchur. *Temporal Point Processes 2: Neural TPP Models*. 2021. url : https://shchur.github.io/blog/2021/tpp2-neural-tpps (visité le 24/05/2023).

[6] Oleksandr Shchur et al. *Neural Temporal Point Processes: A Review*. 2021. arXiv : 2104.03528 [cs.LG].