Astuces d'échantillonnage discret

Maximilien

July 2020

Notions MIASHS probabilité, règle de Baye

1 Introduction

Beaucoup de problèmes en *machine learning*, en statistiques, etc., se ramènent à de l'estimation de probabilités. Ces dernières peuvent engendrer diverses problèmes dans certains cas : *overflow* numérique, problèmes discrets, etc.

Ces notes abordent quelques astuces à ce propos.

2 Astuce de normalisation du *softmax*

Un ordinateur représente chaque nombre au travers d'un séquence binaire finie. Lorsque le nombre devient trop grand pour être représenté, il se produit une erreur qu'on appelle un *overflow*.

Il est fréquents de travailler avec ce qu'on appelle les *logits* $z \in \mathbb{R}^n$. En *deep learning* ou dans diverses stratégies d'inférences statistiques. Cependant, dès que nous sommes confrontés à du *big data*, l'amplitude de ces *logit* s'accroît. En pratique, leur valeur ne nous intéresse pas et nous cherchons plutôt la probabilité qu'ils représentent :

$$\pi_k = \frac{e^{z_k}}{\sum_j e^{z_j}}$$

On imagine sans mal que la quantité e^{z_k} peut rapidement produire un *overflow* si z_k est trop grand alors qu'en pratique, seule la probabilité (entre 0 et 1) nous intéresse et ne devrait pas produire cet *overflow*. L'astuce de normalisation du *softmax* résout ce problème.

Soit $c = \max_k z_k$. On observe sans difficulté :

$$\pi_k = \frac{e^{z_k}}{\sum_j e^{z_j}} \frac{e^{-c}}{e^{-c}} = \frac{e^{z_k - c}}{\sum_j e^{z_j - c}}.$$

Ainsi, l'exponentielle la plus grande prendra la valeur 0 et le problème est résolu.

3 Astuce de *max-Gumbel*

3.1 La loi de Gumbel

La loi de Gumbel est définie sur $]-\infty, +\infty[$ et est continue. Elle est paramétré par $\mu \in \mathbb{R}$ et $\beta > 0$. Son espérance est donnée par $\mu + \beta\gamma$ où γ est la constante d'Euler-Mascheroni [1]. Sa médiane est donnée par $\mu - \beta\ln(\ln(2))$ et sa variance par $\frac{\pi^2}{6}\beta^2$. Sa densité de probabilité est :

$$f(x; \mu, \beta) = \frac{\exp(-z)z}{\beta}, \text{ où } z = \exp\Big(-\frac{x-\mu}{\beta}\Big),$$

et sa fonction de répartition :

$$F(x; \mu, \beta) = \exp(-\exp(\frac{\mu - x}{\beta})).$$

Lorsque $\mu = 0$ et $\beta = 1$, on parle de loi Gumbel standard.

Tout cela semble très compliqué, mais il est en réalité très facile d'échantillonner selon une loi Gumbel. En effet, soit $X \sim \mathcal{U}(0, 1)$, une variable aléatoire uniforme sur $[0, 1]$. On obtient une variable Gumbel à partir de X de la manière suivante. Soit $Y = \mu - \beta\ln(-\ln(X))$, alors Y est une Gumbel de paramètres μ et β.

3.2 L'astuce

Soit un vecteur $x \in \mathbb{R}^n$ qu'on associera à des log-probabilités non normalisées. Soit le paramètre suivant :

$$\pi_k = \frac{\exp(x_k)}{\sum_j \exp(x_j)}$$

L'objectif est de simuler $k \sim \text{Bernoulli}(\pi)$. Cela est finalement assez simple. Il suffit de simuler $t \sim \mathcal{U}(0, 1)$ puis de parcourir les π_k successivement en cumulant les probabilités jusqu'à ce qu'on dépasse t. L'indice k ainsi construit correspond à la variable simulée.

Cette stratégie implique cependant de devoir calculer π, ce qui peut provoquer des *overflow*. La méthode précédente permet de résoudre ce problème. De plus, il est courant que ce genre de calculs doivent être faits plusieurs fois au cours d'une procédure d'inférence. Malheureusement, le coût de calcul de l'exponentiation commencera à se faire sentir.

Une autre stratégie, appelée astuce de max-Gumbel est possible et est illustrée par la proposition suivante.

Proposition 1. *Soit $g_1, ..., g_n$, i.i.d. selon une loi de Gumbel de paramètres $\mu = 0$ et $\beta = 1$. Soit :*

$$y = \underset{k}{\text{argmax}}\, x_k + g_k$$

La variable y est alors distribuée selon une Bernoulli de paramètre π.

1. $\gamma = \lim_{n\to\infty}(1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} - \ln(n))$

Démonstration. Soit $g_1, ..., g_n$, i.i.d. selon une loi de Gumbel de paramètres $\mu = 0$ et $\beta = 1$. Notons $z_k = x_k + g_k$. Évidemment, z_k suit une loi de Gumbel de paramètre $\mu = x_k$ et $\beta = 1$. La probabilité que la k^{eme} dimension soit supérieure à toutes les autres est définies de la manière suivante (via la fonction de répartition d'une Gumbel) :

$$\mathbb{P}(k = \underset{j}{\operatorname{argmax}}\, x_j + g_j | z_k, x_1, ..., x_n) = \prod_{j \neq k} \exp(-\exp(x_j - z_k)).$$

(La fonction de répartition nous permet de calculer la probabilité qu'un $x_j + z_j$ aille au plus jusqu'à z_k) Désintégrons le long de z_k, nous obtenons :

$$\begin{aligned}
\mathbb{P}(k = \operatorname{argmax} \dots | x_1, \dots) &= \int \left[\exp\Big(x_k - z_k - \exp(x_k - z_k)\Big) \right. \\
&\qquad \left. \prod_{j \neq k} \exp(-\exp(x_j - z_k)) \right] dz_k \\
&= \int \left[\exp(x_k - z_k - \sum_j \exp(x_j - z_k)\Big) \right]] dz_k \\
&= \int \left[\exp(x_k - z_k - \exp(-z_k) \sum_j \exp(x_j)\Big) \right] dz_k
\end{aligned}$$

Cette intégral est tractable et nous trouvons :

$$\mathbb{P}(k = \operatorname{argmax} \dots | x_1, \dots) = \left[\frac{\exp(x_k - e^{-z_k} \sum_j e^{x_k})}{\sum_j e^{x_j}} \right]_{-\infty}^{+\infty} = \frac{e^{x_k}}{\sum_j e^{x_j}} = \pi_k$$

On retombe bien sur notre Bernoulli. □

Nous avons vu comment simuler selon une loi de Gumbel plus tôt. Cependant, il faut toujours calculer le logarithme, opération coûteuse. L'avantage vient de la répétition de l'expérience. Ici, le logarithme n'est pas calculé sur les données mais pourrait être précalculé de manière vectorielle sur un grand nombre de tirages uniformes.