Note de lecture sur le papier sur l'Entropie croisée

29 mars 2020

CE = Cross-entropy
La CE est un modèle utilisé pour estimer des probabilités d'événements rares (permet d'avoir moins d'itérations que pour les méthodes classiques d'estimation). Ce modèle peut aussi être utilisé pour résoudre des problèmes d'optimisation combinatoire (COPs). C'est fait en transformant l'optimisation « déterministe » en une optimisation « stochastique » pour ensuite utiliser les méthodes de modélisation d'événements rares.
La méthode de CE est une méthode itérative dont chaque itération peut être décomposée en deux étapes :

1. Génération d'un échantillon aléatoire selon un mécanisme spécifique
2. Mise à jour du paramètre de ce mécanisme en se basant sur les données qui donnent les meilleurs résultats.

Pour estimer des probabilités d'événements rares, on pourrait penser à la méthode d'*Important sampling*. Un des désavantages de cette méthode est que les paramètres optimaux (*tilting*) sont difficiles à obtenir. Au contraire, l'avantage de la méthode de CE est d'avoir une procédure simple pour estimer les paramètres optimaux.

Première partie

Exemple d'un problème d'optimisation combinatoire

On suppose que l'on a un vecteur binaire $y = (y_1, \dots, y_n)$ qu'on cherche à deviner. On ne connait pas y mais on a un « oracle » qui pour chaque *input* nous donne une *réponse* ou *performance*. Par exemple : $S(x) = n - \sum_{j=1}^{n} |x_j - y_j|$ qui donne le nombre d'éléments de $x = (x_1, \dots, x_n)$ égaux à y.
Une méthode naïve est de générer des vecteur $X = (X_1, \dots, X_n)$ de façon à ce que $X_1, \dots, X_n$ soient des Bernoulli indépendantes de paramètre $p_1, \dots, p_n$. Donc $X \sim \mathcal{B}(p)$ et si $p = y$ (loi dégénérée) on a $S(X) = n$ et $X = y$.

FIGURE 1 – Boîte noire utilisée pour décoder le vecteur y

L'algorithme de CE consiste à transformer le problème en un événement rares, créer une séquence de paramètres $\hat{p_0}, \hat{p_1}, \ldots$, et des niveaux (*levels*) $\hat{\gamma_1}, \hat{\gamma_2}, \ldots$ de telles sorte que $\hat{\gamma_1}, \hat{\gamma_2}, \ldots$ converge vers la valeur optimale (ici n) et que la suite $\hat{p_0}, \hat{p_1}, \ldots$ converge vers le paramètre optimal (ici y).

Description de l'algorithme :

1. On commence avec un certain $\hat{p}_0$, par exemple $\hat{p_0} = (1/2, \ldots, 1/2)$ et $t := 1$
2. On génére des échantillons $X_1, \ldots, X_N$ selon une loi de bernoulli de paramètre $\hat{p}_{t-1}$. On calcule le score $S(X_i)$ et on trie ces données du plus petit au plus grand : $S_{(1)} \leq \cdots \leq S_{(N)}$. On note $\hat{\gamma}_{t-1}$ le quantile $1-\rho$ du score : $\hat{\gamma}_{t-1} = S_{\lceil (1-\rho)N \rceil}$.
3. On utilise le même échantillon pour calculer $\hat{p}_t = (\hat{p}_{t,1}, \ldots \hat{p}_{t,n})$ avec la formule :
$$\hat{p}_{t,j} = \frac{\sum_{i=1}^{N} \mathbb{1}_{S(X_i) \geq \hat{\gamma}_t} \mathbb{1}_{X_{i,j}=1}}{\sum_{i=1}^{N} \mathbb{1}_{S(X_i) \geq \hat{\gamma}_t}}$$
Cette condition s'interprète de la façon suivante : pour mettre à jour la jème probabilité, on compte le nombre de vecteurs de X qui ont un score plus grand que $\hat{\gamma}_t$ et dont la jème composante est égale à 1 et on divise ce nombre par le nombre de vecteurs de X qui ont un score plus grand que $\hat{\gamma}_t$.
4. On arrête si on rencontre le critère d'arrêt ($\hat{\gamma}_t$ constant ou $\hat{p}_t$ loi dégénérée), sinon $t := t + 1$.

Deuxième partie

La méthode CE pour les événements rares

Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire qui prend ses valeurs dans $\mathcal{X}$, $\{f(\cdot; v), v \in \mathcal{V}\}$ une famille de densités de probabilités – *probability density functions* (pdfs) – sur $\mathcal{X}$ associés à une mesure ν(mesure de Lebesgue ou mesure de comptage). Pour toute fonction mesurable H on a :

$$\mathbb{E}H(X) = \int_{\mathcal{X}} H(x) f(x; v) \nu(\mathrm{d}x)$$

Soit S une fonction réelle sur $\mathcal{X}$, on suppose que l'on s'intéresse à la probabilité que $S(x)$ soit supérieure à un certain nombre γ sous la densité $f(\cdot, u)$. Cette probabilité s'écrit :

$$l = \mathbb{P}_u(S(X) \geq \gamma) = \mathbb{E}_u \mathbb{1}_{\{S(x) \geq \gamma\}}$$

Si cette probabilité est petite (plus petite que 10^{-5}), $\{S(X) \geq \gamma\}$ est appelé événement rare (*rare event*). Une façon simple d'estimer l est d'utiliser brutalement une méthode de simulation de Monte-Carlo : on échantillonne $X_1, \ldots, X_N$ de loi $f(\cdot, u)$ et ensuite

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{S(X_i) \geq \gamma\}}$$

nous donne un estimateur sans biais de l. Le problème est que $\{S(X) \geq \gamma\}$ est un événement rare : il faut beaucoup de simulations pour estimater avec précisions l (i.e. : avec une faible erreur ou un petit intervalle de confiance).
Une façon alternative est d'utiliser l'important sampling : on tire un échantillon $X_1, \ldots, X_N$ à partir d'une densité g sur $\mathcal{X}$ d'important sampling et on calcule l en utilisation l'estimateur du rapport de vraisemblance – *likelihood ratio* (LR) :

$$\hat{l} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{S(X_i) \geq \gamma\}} \frac{f(X_i; u)}{g(X_i)}$$

La meilleure façon d'estimer l est d'utiliser la densité

$$g^*(x) := \frac{\mathbb{1}_{\{S(x) \geq \gamma\}} f(x; u)}{l} \implies \mathbb{1}_{\{S(X_i) \geq \gamma\}} \frac{f(X_i; u)}{g(X_i)} = l$$

Et donc l'estimateur précédent est de variance nulle. Le problème évident est que g^* dépend de l. De plus il est plus pratique de choisir g dans $\{f(\cdot; v),\ v \in \mathcal{V}\}$. L'idée est de choisir un paramètre de référence – *reference parameter* ou *tilting parameter* – v de façon à ce que la distance entre g^* et $f(\cdot, v)$ soit minimale. La distance entre deux densités g et h qui est utilisée est la distance de *Kullback-Leibler*, aussi appelée *cross-entropy* entre g et h :

$$\mathfrak{D}(g, h) = \mathbb{E}_g \ln \frac{g(X)}{h(X)} = \int g(x) \ln g(x) \mathrm{d}x - \int g(x) \ln h(x) \mathrm{d}x$$

Minimiser la distance de Kullback-Leibler entre g^* et $f(\cdot, v)$ est équivalent à choisir v qui minimise $-\int g^*(x) \ln f(x; v) \mathrm{d}x$, soit à résoudre le problème de maximusation :

$$\max_v \int g^*(x) \ln f(x; v) \mathrm{d}x \iff \max_v \int \frac{\mathbb{1}_{\{S(x) \geq \gamma\}} f(x; u)}{l} \ln f(x; v) \mathrm{d}x$$

On obtient le programme :

$$\max_v D(v) = \max_v \mathbb{E}_u \mathbb{1}_{\{S(X) \geq \gamma\}} \ln f(X; v)$$

On peut encore une fois utiliser l'important sampling avec une nouvelle mesure $f(\cdot, w)$ et le problème précédent devient :

$$\max_v D(v) = \max_v \mathbb{E}_w \mathbb{1}_{\{S(X) \geq \gamma\}} W(X; u, w) \ln f(X; v)$$

Avec $W(x;u,w) = \dfrac{f(x;u)}{f(x;w)}$ le rapport de vraisemblance, en x, entre $f(\cdot,u)$ et $f(\cdot,w)$. La solution optimale s'écrit :

$$v^* = \underset{v}{\operatorname{argmax}}\ \mathbb{E}_w \mathbb{1}_{\{S(X)\geq\gamma\}} W(X;u,w) \ln f(X;v) \tag{1}$$

On estime v^* à partir du programme stochastique – *stochastic counterpart* – suivant :

$$\max_v \hat{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{S(X_i)\geq\gamma\}} W(X_i;u,w) \ln f(X_i;v) \tag{2}$$

Avec $X_1,\ldots,X_N$ tirés aléatoirement selon $f(\cdot,w)$. Dans les applications classiques, $\hat{D}$ est différentable et convexe en v, la solution de 2 peut alors être obtenue en résolvant les équations :

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{S(X_i)\geq\gamma\}} W(X_i;u,w) \nabla_v \ln f(X_i;v) = 0 \tag{3}$$

Remarque : Pour que le programme de CE 2 soit utile, il faut que la probabilité de l'événement $\{S(X) \geq \gamma\}$ ne soit pas trop petite, sinon les $\mathbb{1}_{\{S(X_i)\geq\gamma\}}$ seront souvent égales à 0.
Pour surmonter cette difficulté, on utilise un algorithme multi-niveau – *multi-level algorithm.* L'idée est de construire une suite de paramètres de références $(v_t)_{t\geq 0}$ et une suite de niveaux $(\gamma_t)_{t\geq 1}$, l'algorithme consiste à construire itérativement cette suite. L'idée est de choisir $\hat{v}_0 = u$ et de prendre $\hat{\gamma}_1$ plus petit que γ : $\hat{v}_1$ rendra l'événement $\{S(X) \geq \gamma\}$ un peu moins rare, donc $\hat{\gamma}_2$ peut être plus proche de γ.

Description de l'algorithme de CE pour la simulation d'événements rares :

1. On pose $\hat{v}_0 = u$, $t = 1$.
2. On génère un échantillon $X_1, \ldots, X_N$ de loi $f(\cdot, v_{t-1})$, on calcule le quantile $(1-\rho)$ de la fonction score qui donne $\hat{\gamma}_t$:

$$\hat{\gamma}_t = S_{\lceil (1-\rho)N \rceil}$$

3. On utilise le **même** échantillon $X_1, \ldots, X_N$ pour résoudre le problème stochastique :

$$\hat{v}_t = \underset{v}{\operatorname{argmax}}\, \hat{D}(v) = \underset{v}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{S(X_i) \geq \hat{\gamma}_t\}} W(X_i; u, \hat{v}_{t-1}) \ln f(X_i; v)$$

La solution nous donne $\hat{v}_t$.

4. Si $\hat{\gamma}_t < \gamma$, $t = t + 1$ et on recommence à l'étape 2. Sinon on estime la probabilité l en utilisant le rapport de vraisemblance

$$\hat{l} = \frac{1}{N_1} \sum_{i=1}^{N_1} \mathbb{1}_{\{S(X_i) \geq \gamma_t\}} W(X_i; u, \hat{v}_T)$$

Avec T le nombre total d'itération.

Troisième partie

La méthode de CE pour l'optimisation combinatoire

1 Description générale

1.1 Le problème d'optimisation combinatoire

On considère le problème de maximisation suivant : soit $\mathcal{X}$ un ensemble fini d'états (*states*) et S une fonction de performance. On cherche à trouver le maximum de S sur $\mathcal{X}$ et les points pour lesquels ce maximum est atteint. Si on note γ^* ce maximum, on cherche donc :

$$S(x^*) = \gamma^* = \max_{x \in \mathcal{X}} S(x) \tag{4}$$

Le point de départ de la méthode de CE est d'associer un problème d'estimation avec un problème d'optimisation précédent. Pour cela on définit un ensemble d'indécatrices $\mathbb{1}_{\{S(x) \geq \gamma\}}$ sur $\mathcal{X}$ pour plusieurs niveaux $\gamma \in \mathbb{R}$.

On note $\{f(\cdot;v),\, v \in \mathcal{V}\}$ une famille discrète de probabilités sur $\mathcal{X}$, paramétrée par un paramètre vectoriel v.
Pour un certain $u \in \mathcal{V}$ on associer le problème de maximisation 4, le problème d'estimation du nombre

$$l(\gamma) = \mathbb{P}_u(S(X) \geq \gamma) = \sum_x \mathbb{1}_{S(x)\geq\gamma} f(x;u) = \mathbb{E}_u \mathbb{1}_{S(x)\geq\gamma} \tag{5}$$

C'est le problème stochastique associé – *associated stochastic problem* (ASP). Cela permet de voire l'analogie avec un problème d'estimation d'une probabilité d'événement rare. En effet, si $f(\cdot;u)$ est la densité uniforme sur $\mathcal{X}$ alors $l(\gamma^*) = f(x^*;u) = \frac{1}{|\mathcal{X}|}$ et donc si $\mathcal{X}$ contient beaucoup d'événements, $S(X) \geq \gamma^*$est un événement rare.

Description de l'algorithme de CE pour l'optimisation :
Il faut définir en avance $\hat{v}_0$, la taille de l'échantillon N et le nombre ρ (10% dans notre projet sur Master Mind)

1. On prend $\hat{v}_0$ fixé arbitrairement , $t = 1$.
2. On génère un échantillon $X_1, \ldots, X_N$ de loi $f(\cdot, v_{t-1})$, on calcule le quantile $(1-\rho)$ de la fonction score qui donne $\hat{\gamma}_t$:

   $$\hat{\gamma}_t = S_{\lceil (1-\rho)N \rceil}$$

   Si $\hat{\gamma}_t \geq \gamma$ on prend $\hat{\gamma}_t = \gamma$.
3. On utilise le **même** échantillon $X_1, \ldots, X_N$ pour résoudre le problème stochastique (avec $W = 1$) :

   $$\hat{v}_t = \underset{v}{\operatorname{argmax}}\, \hat{D}(v) = \underset{v}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{S(X_i)\geq\hat{\gamma}_t\}} \ln f(X_i;v) \tag{6}$$

4. Si pour un certain $t \geq d$, (par exemple $d = 5$), on a :

   $$\hat{\gamma}_t = \hat{\gamma}_{t-1} = \cdots = \hat{\gamma}_{t-d}$$

   alors on arrête l'algorithme.

1.2 Smoothed updating

Plutôt que de mettre à jour directement $\hat{v}_{t-1}$ à partir de la formule du problème 6, on réalise souvent une mise à jour lissée – *smoothed updating* :

$$\hat{v}_t = \alpha \hat{w}_t + (1-\alpha)\hat{v}_{t-1} \tag{7}$$

avec $\hat{w}_t$ la valeur obtenue en résolvant le problème 6 ($\hat{w}_t$ correspondait donc à la précédente valeur qu'on utilisait avant pour $\hat{v}_t$). C'est en particulier pertinent pour les problèmes d'optimisaiton avec des variables discrètes : cela permet d'éviter l'occurence de 0 ou de 1 dans les paramètres des vecteurs. En effet, le

problème est que dès lors qu'on a un 0 ou un 1, il le reste souvent pour toujours, ce qui a des effets indésirables. On trouve empiriquement que $0,4 \leq \alpha \leq 0,9$ donne des meilleurs résultats.
Si $\alpha = 1$ on retrouve l'algorithme précédent.
Dans beaucoup d'applications on observe une convergence numérique de $f(\cdot; \hat{v}_t)$ vers une mesure dégénérée (une mesure de Dirac), c'est-à-dire que l'on affecte toute la masse de probabilité à un seul état x_T pour lequel, par définition, on a $S(x_T) \geq \hat{\gamma}_T$.

1.3 Estimation du maximum de vraisemblance

Le problème 6 est étroitement lié au problème de maximisation de la vraisemblance. En effet, le problème de maximisasion de vraisemblance revient à a

$$\hat{v}_t = \underset{v}{\operatorname{argmax}} \sum_{i=1}^{N} \mathbb{1}_{\{S(X_i) \geq \hat{\gamma}_t\}} \ln f(X_i; v) \tag{8}$$

La seule différence avec le problème 6 est la présence de $\mathbb{1}_{\{S(X_i) \geq \hat{\gamma}_t\}}$. Le problème d'optimisation combinatoire peut être réécrit en :

$$\hat{v}_t = \underset{v}{\operatorname{argmax}} \sum_{X_i \,:\, S(X_i) \geq \hat{\gamma}_t} \mathbb{1}_{\{S(X_i) \geq \hat{\gamma}_t\}} \ln f(X_i; v) \tag{9}$$

En d'autres termes, $\hat{v}_t$ est égal à l'estimateur de maximum de vraisemblance de $\hat{v}_{t-1}$ calculé uniquement à partir des vecteurs X_i qui ont une performance supérieure ou égale à $\hat{\gamma}_t$.

1.4 Paramètres

Le choix de la taille de l'échantillon N et du paramètre ρ dépendent de la taille du problème et du nombre de paramètres dans le problème de maximisation.
En particulier dans les problème de type s*tochastic node networks* (SNN), on prend généralement $N = cn$ avec n le nombre de *nodes* et c une constante ($c > 1$), souvent $5 \leq c \leq 10$.
Dans les problème de type *stochastic edge networks* (SEN) on prend généralement $N = cn^2$ avec n^2 le nombre de *edges* dans le réseau. La taille de l'échantillon est donc liée au nombre de paramètres à estimer (n et n^2).
Pour estimer k paramètres, il faut prendre une taille d'échantillon **au minimum** égale à $N = ck$ avec $c > 1$.
Pour ρ il est souvent conseillé de prendre ρ autour de 0,01 si n est grand ($n \geq 100$) et de prendre un ρ plus grand, $\rho \simeq \frac{\ln(n)}{n}$ pour $n < 100$.
Les paramètres N et ρ peuvent aussi être déterminés de façon itérative : c'est ce qui est fait dans l'algorithme FACE ou dans Homem-de-Mello, T. and Rubinstein, R. (2002). Rare event estimation for static models via cross-entropy and importance sampling. Submitted for publication.

Dans notre cas, il faut donc prend $N = c \times m \times n$ avec $c > 1$.

2 Application au mastermind

Dans notre projet, on a n boules et m couleurs. On numérote les couleurs de 1 à m.
On a donc $\mathcal{X} = \{1, 2, \ldots, m\}^n$.
Les composantes du vecteur $X = (X_1, \ldots, X_n) \in \mathcal{X}$ sont tirées aléatoirement de façons à ce que leur distribution soit déterminée par une suite $p_1, \ldots, p_n$ de vecteurs de probabilités, la $j$ème composante de p_i étant égale à $p_{ij} = \mathbb{P}(X_i = j)$. On peut représenter cette loi par une matrice

$$P = (p_{i,j})_{i,j} = (\mathbb{P}(X_i = j))_{i,j} \in \{\text{matrices stochastiques de } \mathcal{M}_{n,m}(\mathbb{R})\}$$

La densité s'écrit :

$$f(X; p) = \prod_{i=1}^{n} \prod_{j=1}^{m} p_{i,j}^{\mathbb{1}_{\{X_i=j\}}}$$

Il vient :

$$\ln f(X; p) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}_{\{X_i=j\}} \ln p_{i,j}$$

On suppose dans un premier temps qu'il n'y a pas de lien entre les $p_{i,j}$ (comme si on n'avait pas $\sum_{j=1}^{m} p_{i,j} = 1$) :

$$\frac{\partial}{\partial p_{k,l}} \ln f(X; p) = \frac{\partial}{\partial p_{k,l}} \sum_{j=1}^{m} \mathbb{1}_{\{X_k=j\}} \ln p_{k,j} = \frac{\mathbb{1}_{\{X_k=l\}}}{p_{k,l}}$$

Le programme de maximisation peut s'écrire comme un problème de maximisation sous contrainte :

$$\begin{cases} \max\limits_{p_{i,j}} & \sum_{i=1}^{N} \mathbb{1}_{\{S(X_i) \geq \hat{\gamma}_t\}} \ln f(X_i; v) \\ s.c & \forall i \ : \ \sum_{j=1}^{m} p_{i,j} = 1 \end{cases}$$

Le Lagrangien s'écrit :

$$\mathcal{L} = \sum_{i=1}^{N} \mathbb{1}_{\{S(X_i) \geq \hat{\gamma}_t\}} \ln f(X_i; v) + \sum_{i=1}^{n} \lambda_i \left(\sum_{j=1}^{m} p_{i,j} - 1 \right)$$

La condition d'optimalité en $p_{k,l}$:

$$\frac{\partial}{\partial p_{k,l}} \mathcal{L} = 0 \implies \sum_{i=1}^{N} \mathbb{1}_{\{S(X_i) \geq \hat{\gamma}_t\}} \frac{\mathbb{1}_{\{X_{i,k}=l\}}}{p_{k,l}} - \lambda_k = 0$$

Soit :

$$\sum_{i=1}^{N} \mathbb{1}_{\{S(X_i) \geq \hat{\gamma}_t\}} \mathbb{1}_{\{X_{i,k}=l\}} = \lambda_k p_{k,l}$$

En sommant sur l et en utilisant la condition sur les $p_{i,j}$, il vient :

$$\lambda_k = \sum_{i=1}^{N} \mathbb{1}_{\{S(X_i)\geq\hat{\gamma}_t\}} \underbrace{\sum_{l=1}^{m} \mathbb{1}_{\{X_{i,k}=l\}}}_{=1}$$

D'où la formule de mise à jour :

$$p_{k,l} = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{S(X_i)\geq\hat{\gamma}_t\}} \mathbb{1}_{\{X_{i,k}=l\}}}{\sum_{i=1}^{N} \mathbb{1}_{\{S(X_i)\geq\hat{\gamma}_t\}}}$$