INFÉRENCE VARIATIONNELLE

ELBO

Dans ce chapitre, on considère un modèle à données latentes (non observées). Soit (Z, X) un couple de variables aléatoires dans $\mathbb{R}^d \times \mathbb{R}^m$. On suppose que la loi du couple a une densité notée $(z, x) \mapsto p(z, x)$ par rapport à une mesure de référence (cette dernirère peut dépendre d'un paramètre θ, ce que nous ignorons dans ces notes). Dans ce contexte, on écrit en général,

$$(z, x) \mapsto p(z, x) = p(z)p(x|z)\,,$$

où $z \mapsto p(z)$ est une densité a priori pour Z et $x \mapsto p(x|z)$ est une densité pour la loi de X sachant Z. Dans ce cadre, nous n'avons en général pas accès à la densité de la loi de Z sachant X, car celle-ci s'écrit :

$$z \mapsto p(z|x) = \frac{p(z)p(x|z)}{p(x)}\,,$$

où $p(x) = \int p(z)p(x|z)\mathrm{d}z$ est une intégrale que l'on ne sait en général pas calculer.

Dans le cadre de l'inférence variationnelle, on propose une famille de densités candidates pour approcher $z \mapsto p(z|x)$. On introduit alors $\mathcal{D} = \{q_\phi\}_{\phi\in\Phi}$ où Φ est un espace de paramètres où les densités q_ϕ sont choisies telles que :

- pour tout ϕ, q_ϕ est simple à évaluer ;
- pour tout ϕ, q_ϕ est simple à simuler.

On remarque alors que pour tout x et tout ϕ,

$$\begin{aligned}
\mathrm{KL}\left(q_\phi \| p(\cdot|x)\right) = \int q_\phi(z) \log \frac{q_\phi(z)}{p(z|x)}\mathrm{d}z &= \mathbb{E}_{q_\phi}[\log q_\phi(Z)] - \mathbb{E}_{q_\phi}[\log p(Z|x)]\,,\\
&= \mathbb{E}_{q_\phi}[\log q_\phi(Z)] - \mathbb{E}_{q_\phi}[\log p(Z, x)] + \log p(x)\,,\\
&= -\mathrm{ELBO}(q_\phi) + \log p(x)\,.
\end{aligned}$$

En appliquant l'inégalité de Jensen on remarque que $\mathrm{KL}\left(q_\phi \| p(\cdot|x)\right) \geq 0$ et donc que

$$\mathrm{ELBO}(q_\phi) \leq \log p(x)\,.$$

Cette inégalité justifie le nom Evidence Lower Bound de la quantité ELBO. Dans le cadre de l'inférence variationnelle, on souhaite alors approcher $p(\cdot|x)$ par q_{ϕ_*} où :

$$\phi_* \in \mathrm{argmax}_{\phi\in\Phi}\, \mathrm{ELBO}(q_\phi)\,.$$

Coordinate ascent variational inference

L'approche la plus répandue pour résoudre le problème précédent est tout d'abord de choisir une approche "mean field" i.e. de choisir $\mathcal{D}$ de la forme

$$\mathcal{D} = \left\{ z \mapsto q(z) = \prod_{j=1}^{d} q_j(z_j)\ ;\ q_j \text{ est une densité} \right\}.$$

Dans ce cas, on peut écrire, pour tout $q \in \mathcal{D}$, et pour tout $j \in \{1, \dots, d\}$,

$$
\begin{aligned}
\mathrm{ELBO}(q) &= \mathbb{E}_q\left[\log \frac{p(Z_1, \dots, Z_d, x)}{\prod_{j=1}^d q_j(Z_j)}\right] , \\
&= \mathbb{E}_{q_j}\left[\mathbb{E}_{q_{-j}}\left[\log p(Z_j|Z_{-j}, x)\right]\right] - \mathbb{E}_{q_j}\left[\log q_j(Z_j)\right] + \mathrm{cste}
\end{aligned}
$$

où cste est un terme ne dépendant pas de q_j, $z_{-j} = (z_\ell)_{1\le \ell\le d; \ell\ne j}$ et $q_{-j}(z_{-j}) = \prod_{1\le \ell\le d;\ell\neq j} q_\ell(z_\ell)$. Si l'on fixe les densités q_ℓ, $\ell \ne j$ et que l'on maximise la quantité précédente sur q_j, on obtient comme solution la densité proportionnelle à

$$
z_j \mapsto \exp\left(\mathbb{E}_{q_{-j}}\left[\log p(z_j|Z_{-j}, x)\right]\right) .
$$

L'algorithme CAVI propose donc de mettre à jour les q_j, $1 \le j \le d$ jusqu'à atteindre un critère d'arrêt.

Exemple : mélange de lois gaussiennes

On considère un mélange de K gaussiennes de moyennes $\mu = (\mu_k)_{1\leqslant k\leqslant K}$ et de variance 1. Les variables $\mu = (\mu_k)_{1\leqslant k\leqslant K}$ sont (i.i.d.) de loi gaussienne de moyenne 0 et de variance σ^2. Le poids de la composante k est noté ω_k. Conditionnellement à μ, les observations $(X_i)_{1\leqslant i\leqslant n}$ sont i.i.d. et la densité de probabilité de X_1 est:

$$
x \mapsto p(x|\mu) = \sum_{k=1}^K \omega_k \varphi_{\mu_k,1}(x) ,
$$

où φ_{μ_k,η^2} est la densité gaussienne de moyenne μ_k et de variance η^2. La vraisemblance jointe est alors :

$$
p(x_1, \cdots, x_n) = \int p(x_1, \cdots, x_n|\mu)p(\mu)\mathrm{d}\mu = \int \prod_{i=1}^n p(x_i|\mu)p(\mu)\mathrm{d}\mu = \int \prod_{i=1}^n \left(\sum_{k=1}^K \omega_k\varphi_{\mu_k,1}(x_i)\right) p(\mu)\mathrm{d}\mu
$$

Notre objectif est d'approcher $p(\mu, c|x)$ où $c = (c_1, \cdots, c_n)$ sont les composantes des observations. L'approximation 'mean-field' considérée s'écrit:

$$
q(\mu, c) = \prod_{k=1}^K \varphi_{m_k,s_k}(\mu_k) \prod_{i=1}^n \mathrm{Cat}_{\phi_i}(c_i) ,
$$

ce qui signifie que sous q:

- μ et c sont indépendantes.
- $(\mu_k)_{1\leqslant k\leqslant K}$ sont des gaussiennes indépendantes de moyennes $(m_k)_{1\leqslant k\leqslant K}$ et variances $(s_k)_{1\leqslant k\leqslant K}$.
- $(c_i)_{1\leqslant i\leqslant n}$ sont indépendantes de distribution multinomiales de paramètres $(\phi_i)_{1\leqslant i\leqslant n}$: $q(c_i = k) = \phi_i(k)$ pour $1 \leqslant k \leqslant K$.

Notons $\mathcal{D}$ cette famille de distributions où les moyennes $(m_k)_{1\leqslant k\leqslant K} \in \mathbb{R}^K$, les variances $(s_k)_{1\leqslant k\leqslant K} \in (\mathbb{R}_+^*)^K$ et les $(\phi_i)_{1\leqslant i\leqslant n} \in \mathcal{S}_K^n$ où $\mathcal{S}_K$ est le simplexe de dimension K.

L'objectif est de trouver le "meilleur candidat" dans $\mathcal{D}$ pour approcher $p(\mu, c|x)$, i.e. celui qui minimise la distance de Kullback suivante :

$$
q^* = \mathrm{Argmin}_{q\in\mathcal{D}}\, \mathrm{KL}\left(q(\mu, c)\|p(\mu, c|x)\right) .
$$

Notons que :

$$\begin{aligned}
\mathrm{KL}\left(q(\mu,c)\|p(\mu,c|x)\right) &= \mathbb{E}_q[\log q(\mu,c)] - \mathbb{E}_q[\log p(\mu,c|x)]\,,\\
&= \mathbb{E}_q[\log q(\mu,c)] - \mathbb{E}_q[\log p(\mu,c,x)] + \log p(x)\,,\\
&= -\mathrm{ELBO}(q) + \log p(x)\,,
\end{aligned}$$

où l'Evidence Lower Bound est

$$\mathrm{ELBO}(q) = -\mathbb{E}_q[\log q(\mu,c)] + \mathbb{E}_q[\log p(\mu,c,x)]\,.$$

Ainsi, minimiser la divergence de Kullback revient à maximiser la ELBO, avec $\log p(x) \geqslant \mathrm{ELBO}(q)$. La complexité de $\mathcal{D}$ détermine la complexité du problème d'optimisation. L'algorithme CAVI calcule itérativement pour $1 \leqslant k \leqslant K$,

$$q(\mu_k) \propto \exp\left(\mathbb{E}_{\tilde{q}_{\mu_k}}[\log \tilde{p}_k(\mu_k|x)]\right)$$

et pour tout $1 \leqslant i \leqslant n$,

$$q(c_i) \propto \exp\left(\mathbb{E}_{\tilde{q}_{c_i}}[\log \tilde{p}_i(c_i|x)]\right)\,,$$

où

- $\tilde{p}_i(c_i|x)$ est la distribution conditionnelle de c_i sachant les observations et les autres paramètres et $\tilde{p}_k(\mu_k|x)$ est la loi conditionnelle de μ_k sachant les observations et les autres paramètres.
- $\mathbb{E}_{\tilde{q}_z}$ est l'espérance sous la loi variationnelle de toutes les variables sauf z.

Soit $\tilde{p}_i(c_i|x)$ la distribution conditionnelle de c_i sachant les observations et les autres paramètres.

$$\tilde{p}_i(c_i|x) \propto p(c_i)p(x_i|c_i,\mu) \propto p(c_i)\prod_{k=1}^{K}\left(\varphi_{\mu_k,1}(x_i)\right)^{1_{c_i=k}}\,.$$

Ainsi,

$$\mathbb{E}_{\tilde{q}_{c_i}}[\log \tilde{p}_i(c_i|x)] = \log p(c_i) + \sum_{k=1}^{K} 1_{c_i=k}\mathbb{E}_{\tilde{q}_{c_i}}[\log \varphi_{\mu_k,1}(x_i)]$$

et

$$\begin{aligned}
\exp\left(\mathbb{E}_{\tilde{q}_{c_i}}[\log \tilde{p}_i(c_i|x)]\right) &\propto p(c_i)\exp\left(\sum_{k=1}^{K} 1_{c_i=k}\mathbb{E}_{\tilde{q}_{c_i}}[\log \varphi_{\mu_k,1}(x_i)]\right)\\
&\propto p(c_i)\exp\left(\sum_{k=1}^{K} 1_{c_i=k}\mathbb{E}_{\tilde{q}_{c_i}}[-(x_i-\mu_k)^2/2]\right)\\
&\propto p(c_i)\exp\left(\sum_{k=1}^{K} 1_{c_i=k}\mathbb{E}_{\tilde{q}_{c_i}}[-(x_i-\mu_k)^2/2]\right)\,.
\end{aligned}$$

La mise à jour s'écrit alors :

$$\varphi_i(k) \propto p(c_i=k)\exp\left(m_k x_i - \frac{m_k^2+s_k}{2}\right)\,.$$

Puisque $\tilde{p}_k(\mu_k|x)$ est la loi conditionnelle de μ_k sachant les observations et les autres paramètres,

$$\tilde{p}_k(\mu_k|x) \propto p(\mu_k)\prod_{i=1}^{n} p(x_i|c_i,\mu)\,.$$

Ainsi,

$$\mathbb{E}_{\tilde{q}_{\mu_k}}[\log \tilde{p}_k(\mu_k|x)] = \log p(\mu_k) + \sum_{i=1}^{n} \mathbb{E}_{\tilde{q}_{\mu_k}}[\log p(x_i|\mu, c_i)]$$

et

$$\begin{aligned}
\exp\left(\mathbb{E}_{\tilde{q}_{\mu_k}}[\log \tilde{p}_i(c_i|x)]\right) &\propto p(\mu_k)\exp\left(\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{\tilde{q}_{\mu_k}}[1_{c_i=k} \log \varphi_{\mu_k,1}(x_i)]\right) \\
&\propto p(\mu_k)\exp\left(\sum_{i=1}^{n} \phi_i(k)\mathbb{E}_{\tilde{q}_{\mu_k}}[\log \varphi_{\mu_k,1}(x_i)]\right) \\
&\propto \exp\left(-\frac{\mu_k^2}{2\sigma^2} - \frac{1}{2}\sum_{i=1}^{n} \phi_i(k)(x_i - \mu_k)^2\right) , \\
&\propto \exp\left(-\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^{n} \phi_i(k)x_i\mu_k - \frac{1}{2}\sum_{i=1}^{n} \phi_i(k)\mu_k^2\right) .
\end{aligned}$$

La mise à jour s'écrit :

$$\mu_k = \frac{\sum_{i=1}^{n} \phi_i(k)x_i}{1/\sigma^2 + \sum_{i=1}^{n} \phi_i(k)} \quad \text{and} \quad s_k = \frac{1}{1/\sigma^2 + \sum_{i=1}^{n} \phi_i(k)} .$$