TD 10

Exercice 1 – Inférence Bayésienne

Soit un ensemble de données d'apprentissage i.i.d. $\mathcal{X} = \{\mathbf{x^i}\}_{i=1,...,N}, \mathbf{x^i} \in \mathbb{R}$, avec $\forall i, x_i \sim \mathcal{N}(\mu, \sigma^2)$. Dans cet exercice nous considérons σ connu pour simplifier.

Q 1.1 Selon le prior $p(\mu) = \mathcal{N}(\mu_0, \lambda^2)$, donner le paramètre μ^* maximisant la probabilité a posteriori de l'échantillon d'apprentissage.

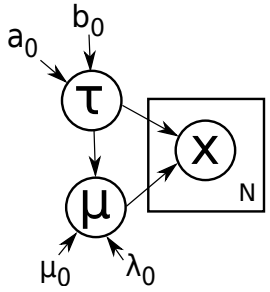
Q 1.2 Cette recherche de MAP nous donne un "point-estimate" de la distribution postérieure $p(\mu|\mathcal{X})$. Dans de nombreuses applications (gestion de l'incertitude, active sampling, génération, etc.), nous sommes intéressés par l'estimation complète de cette distribution postérieure. En donner la formulation complète et indiquer la difficulté de cette formulation.

Q 1.3 Heureusement dans cet exemple, nous avons utilisé des distibutions conjuguées (un prior gaussien est le conjugué d'une vraissemblance gaussienne). Cela signifie que la distribution postérieure est de la même famille que ce prior (la postérieure peut alors servir comme nouveau prior si on observe de nouvelles données d'apprentissage). Pour l'observer, nous proposons de développer l'expression précédente (en notant que $p(\mathcal{X})$ n'est qu'une constante de normalisation) pour faire apparaître cette formulation gaussienne de la probabilité postérieure du modèle.

Q 1.4 Donner alors la distribution "posterieure prédictive" $p(\tilde{x}|\mathcal{X})$ d'un nouvel exemple $\tilde{x}$ suivant la même loi que les données d'apprentissage. Indication : utiliser la relation $\int_{-\infty}^{\infty} e^{-ax^2+bx}\, dx = \sqrt{\frac{\pi}{a}} e^{\frac{b^2}{4a}}, \forall a > 0$.

Exercice 2 – Inférence Variationnelle

On considère maintenant le modèle suivant :



Dans ce diagramme "plate", les cercles représentent les variables aléatoires, les flèches les dépendences. x correspond à une observation, τ correspond à la précision du modèle (i.e., l'inverse de la variance), μ l'espérance. On a :

$$\begin{aligned} x &\sim \mathcal{N}(\mu, \tau^{-1}) \\ \mu &\sim \mathcal{N}(\mu_0, (\lambda_0\tau)^{-1}) \\ \tau &\sim \text{Gamma}(a_0, b_0) \end{aligned} \tag{24}$$

Q 2.1 Donner l'expression de la loi jointe $p(\mathcal{X}, \mu, \tau)$ pour un ensemble d'observations $\mathcal{X} = \{\mathbf{x^i}\}_{i=1,...,N}$.

Q 2.2 Du fait des dépendences entre variables, la loi de la postérieure $p(\tau, \mu|\mathcal{X})$ est difficile à déterminer directement [1]. Une possibilité pour l'approximer est d'employer l'inférence variationnelle qui consiste à utiliser une distribution approchée $q(\tau, \mu)$ plus simple à manipuler. Dans la suite, on suppose $q(\tau, \mu) = q_\tau(\tau)q_\mu(\mu)$ (Mean Field Approximation). On cherche à faire tendre $q(\tau, \mu)$ vers $p(\tau, \mu|\mathcal{X})$ en minimisant la divergence de Kullback-Leibler définie par :

$$D_{\mathrm{KL}}(Q||P) \triangleq \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log \frac{Q(\mathbf{Z})}{P(\mathbf{Z} \mid \mathbf{X})}$$

avec $\mathbf{X}$ les variables observées et $\mathbf{Z}$ les variables cachées du modèle, $Q(\mathbf{Z})$ la distribution variationnelle et $P(\mathbf{Z} \mid \mathbf{X})$ la distribution postérieure à approcher. Donner la formulation de la KL pour notre modèle.

Q 2.3 À partir de la formulation de la KL précédente, montrer que l'on peut écrire :

$$\log P(\mathcal{X}) = D_{\mathrm{KL}}(Q||P) + \mathcal{L}(Q)$$

avec $\mathcal{L}(Q)$ une quantité à définir.

Q 2.4 Justifier l'appellation ELBO (pour Evidence Lower Bound) de $\mathcal{L}(Q)$.

Q 2.5 Dire pourquoi la maximisation de l'ELBO selon les paramètres de la distribution variationnelle permet de minimiser la divergence $D_{\mathrm{KL}}(Q||P)$.

Q 2.6 Il peut être montré que pour chaque facteur indépendant Z_j de la distribution variationnelle, la distribution $q_j^*(Z_j)$ maximisant l'ELBO est donnée par :

$$q_j^*(\mathbf{Z}_j \mid \mathbf{X}) = \frac{e^{\mathrm{E}_{i\neq j}[\ln p(\mathbf{Z},\mathbf{X})]}}{\int e^{\mathrm{E}_{i\neq j}[\ln p(\mathbf{Z},\mathbf{X})]}\, d\mathbf{Z}_j}$$

Avec $\mathrm{E}_{i\neq j}$ l'espérance selon tous les facteurs sauf j. En pratique, on s'intéresse au logarithme :

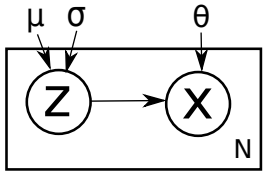
$$\ln q_j^*(\mathbf{Z}_j \mid \mathbf{X}) = \mathrm{E}_{i\neq j}[\ln p(\mathbf{Z}, \mathbf{X})] + \text{constant}$$

Où la constante correspond au terme de normalisation de la distribution, qui ne dépend pas de Z_j et peut être déterminée selon la loi de probabilité du facteur Z_j. Donner les formulations correspondantes pour les facteurs de la distribution variationnelle de notre modèle.

Q 2.7 En déduire un algorithme itératif pour l'approximaxtion des lois postérieures de μ et τ.

Exercice 3 – Variationnal Auto-Encoder

Soit le modèle suivant :



où $\forall i \in \{1, ..., N\}, z_i \in \mathbb{R}^{dz}$ est une représentation latente d'une observation $x_i \in \mathbb{R}^{dx}$, avec $x_i \sim \mathcal{N}(f_\theta^{dec}(z_i), g_\theta^{dec}(z_i)Id)$. Id est la matrice identité $dx \times dx$, f_θ^{dec} et g_θ^{dec} sont deux fonctions de décodage $\mathbb{R}^{dz} \to \mathbb{R}^{dx}$ retournant respectivement l'espérance et la variance diagonale de x à partir de z. $\forall i \in \{1, ..., N\}, z_i \sim \mathcal{N}(\mu, \sigma^2 Id)$, avec μ le vecteur moyen de z et $\sigma^2 Id$ sa matrice de co-variance diagonale. θ est un ensemble de paramètres à apprendre, μ et σ sont des hyper-paramètres du modèle.

1. En fait c'est une normal-gamma, qui est le prior conjugué d'une vraisemblance normale pour laquelle on ne connaît ni la moyenne ni la variance, et pour laquelle on dispose de l'estimateur exact. Mais nous n'allons pas utiliser ce résultat ici, afin d'illustrer l'inférence variationnelle sur un exemple simple.

Q 3.1 Donner l'expression de $p(X|\theta, \mu, \sigma)$

Q 3.2 Passer cette expression au log. Quel est le problème d'une recherche de MAP selon les paramètres θ ?

Q 3.3 Comment une distribution variationnelle pourrait-elle aider ? En donner la forme, ainsi que l'ELBO correspondante.