Modèles à variables latentes discrètes

Examen

Un formulaire de rappels est disponible à la fin du sujet

Exercice 1 : mélange de loi exponentielles

On dispose d'observations continues, univariées, et positives $\boldsymbol{X} = (X_i)_{i=1}^n$ indépendantes et que l'on souhaite modéliser à l'aide d'un mélange de loi exponentielles à K composantes

Pour $i = 1, \dots, n$,

1. $Z_i \sim \mathcal{M}_K(1, \pi)$,
2. $X_i \mid \{Z_{ik} = 1\} \sim \mathcal{E}(\gamma_k)$.

Les paramètres du modèles sont notés $\theta = \{\pi_k, \gamma_k\}_{k=1}^K$, $\sum_{k=1}^K \pi_k = 1$ et $\gamma_k > 0$.

1. Pour un modèle exponentiel simple ($K = 1$, pas de mélange), donner l'expression de la log-vraisemblance des données observées et calculer le maximum de vraisemblance $\hat{\gamma}$.
2. Écrire la log-vraisemblance des données observées $\log p_\theta(\boldsymbol{X})$ pour le modèle à K composantes. Existe-t-il une forme close pour l'estimateur du maximum de vraisemblance ?
3. Écrire la log-vraisemblance des données **complétées** $\log p_\theta(\boldsymbol{X}, \boldsymbol{Z})$ pour le modèle à K composantes. Supposons que l'on connaît $\boldsymbol{Z} = (Z_1, \dots, Z_n)$, donner l'estimateur du maximum de la vraisemblance complétée $\hat{\theta}_C$.
4. (**Question de cours**) Rappeler le principe général de l'algorithme EM : borne inférieur de la vraisemblance, E-step et M-step et pseudo-code de l'algorithme.
5. Pour le modèle de loi exponentiel à K composantes.
   a) Donner l'expression de la borne inférieure calculée lors de la E-step.
   b) Donner l'expression de $\hat{\theta}$ solution de la M-step.
6. Proposer une manière pertinente de trouver une partition (un clustering) des données $\boldsymbol{X}$.
7. Proposer une initialisation intéressante pour votre algorithme EM

Exercice 2 : chaîne de Markov cachée

Soit $\boldsymbol{Z} = Z_{1:n} \sim \mathcal{MC}(\nu, A)$ une chaîne de Markov à K états avec pour loi initiale ν et matrice de transition A. On considère le modèle de Markov caché

$$\begin{cases} Z_{1:n} & \sim \mathcal{MC}(\nu, A), \\ X_i \mid \{Z_{ik} = 1\} & \sim p_{\gamma_k} \end{cases}$$

avec lois d'émissions quelconques p_{γ_k}.

1. (**Question de cours**) Soit ν une loi stationnaire pour la chaîne, quelle équation ν doit-elle vérifier ? En 10 lignes maximum, décrire une stratégie possible pour trouver ν à l'aide d'une librairie de calcul scientifique (en français ou en pseudo-code, au choix).

2. (**Question de cours**) Écrire la log-vraisemblance des données complétées $(\boldsymbol{X}, \boldsymbol{Z})$ pour ce modèle. Comment appelle-t-on l'algorithme permettant le calcul efficace des quantités nécessaires pour l'étape E ?

3. On considère des émissions exponentielles $p_{\gamma_k}(y) = \gamma_k e^{-\gamma_k y} \mathbb{1}_{y>0}$. Donner l'expression des estimateur $\hat{\gamma}_k$ dans la M-step.

Exercice 3 : modèle à bloc stochastique

On considère le modèle à bloc stochastique (SBM) binaire à K classes pour la matrice d'adjacence $\boldsymbol{X} = (X_{ij})_{i,j=1}^n$

$$\begin{cases} Z_i & \sim \mathcal{M}_K(1, \pi), \\ X_{ij} \mid \{Z_{ik} Z_{jl} = 1\} & \sim \mathcal{B}(\gamma_{kl}). \end{cases}$$

Les paramètres du modèle sont $\theta = \{\pi, \gamma\}$ où $\gamma = (\gamma_{kl})_{k,l=1}^K$ sont dans $[0, 1]$.

1. (**Question de cours**) Peut-on facilement écrire la log-vraisemblance des **données complétées** $(\boldsymbol{X}, \boldsymbol{Z})$? Si oui, l'écrire, sinon expliquer pourquoi.

On propose d'approcher la distribution conditionnelle $p(\boldsymbol{Z} \mid \boldsymbol{X})$ par une loi dans la famille

$$\mathcal{Q} := \left\{ q(\boldsymbol{Z}) = \prod_{i=1}^n \mathcal{M}_K(Z_i \mid 1, \tau_i), \text{ t.q. } \sum_{k=1}^K \tau_{ik} = 1, \forall i = 1, \dots, n \right\}$$

2. (**Question de cours**) Quel nom donne-t-on à la forme factorisée de la famille variationnelle $\mathcal{Q}$?

3. Montrer que l'entropie s'écrit $\mathcal{H}(q) = -\sum_{i=1}^n \tau_{ik} \log \tau_{ik}$.

4. Quel problème d'optimisation doit être résolu dans l'étape VE de l'algorithme VEM ?

5. Supposons que le résultat de l'étape VE soit donné et connu. Donner l'estimateur $\hat{\theta} = (\hat{\pi}, \hat{\gamma})$ de la M-step.

Bonus : à faire uniquement si sujet terminé

On considère le HMM de l'exercice 2. Posons $r_i(k) = p(z_i = k \mid x_{i+1:n})$ et notons la matrice de "transition inversée" A^- telle que $A^-_{lk} = p(z_i = k \mid z_{i+1} = l)$.

1. Montrer que $r_i(k) = \sum_l p(z_{i+1} = l \mid x_{i+1:n}) A^-_{lk}$
2. Montrer que le terme $p(z_{i+1} = l \mid x_{i+1:n})$ peut s'exprimer en fonction de $r_{i+1}(l)$ et $p_{\gamma_l}(x_i)$. En déduire une formule récursive de r_i en fonction de r_{i+1}.

Quelques formules utiles

- Entropie d'une loi de probabilité q : $\mathcal{H}(q) = -\mathbb{E}_q[\log q(Z)]$
- La variable binaire $X \in \{0, 1\}$ suit une loi de Bernoulli de paramètre γ, notée $X \sim \mathcal{B}(\gamma)$, si sa densité s'écrit $p(x) = \gamma^x (1 - \gamma)^{1-x}$.
- Le vecteur aléatoire *discret* $Z \in \{0, 1\}^K$ suit une loi multinomiale de paramètre π, noté $Z \sim \mathcal{M}_K(1, \pi)$, si sa densité de probabilité s'écrit $p(z) = \prod_{k=1}^K \pi_k^{z_k}$, *i.e.* $p(Z_k = 1) = \pi_k$.
- La variable aléatoire *continue* X suit une loi exponentielle de paramètre γ, noté $X \sim \mathcal{E}(\gamma)$, si sa densité s'écrit
$$p_\gamma(x) = \gamma e^{-\gamma x} \, \mathbb{1}_{x>0}.$$