Modèles à variables latentes discrètes

Devoir maison facultatif

Exercice 1 : Équivalence entre modèles de mélanges Gaussiens contraints, K-Means et EM Classification

Le Modèle de Mélange Gaussien (GMM) et l'algorithme de K-means visent tous deux à regrouper les données en clusters. Dans cet exercice, vous explorerez l'équivalence entre leurs objectifs et algorithmes sous certaines hypothèses.

Rappel de cours – L'algorithme Classification-EM (CEM), est un algorithme EM standard où l'étape E est modifiée. Dans CEM, l'étape E est remplacée par une affectation stricte (étape CE) basée sur les valeurs maximales des probabilités a posteriori $\tau_{ik}^{(t)}$. On attribue chaque point x_i à un seul cluster, via l'estimateur

$$\hat{z}_{ik}^{(t)} = \begin{cases} 1 & \text{if } k = \arg\max_l \tau_{il}^{(t)}, \\ 0 & \text{otherwise} \end{cases}$$

L'étape CE utilise ensuite le critère de la vraisemblance complète $Q(\theta \mid \theta^{(t)}) = \log p_\theta(X, \hat{Z}^{(t)})$. Cela contraste avec l'EM standard où chaque x_i contribue à chaque cluster à hauteur des probabilités τ_{ik}.

(a) **Log-vraisemblance d'un mélange Gaussien:** Supposons que les données soient $\{x_i\}_{i=1}^N$ et que les paramètres du modèle soient $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$. On note $z_i \in \{0,1\}^K$ les variables latentes. Écrivez les fonctions de log-vraisemblance et de log-vraisemblance complète pour un Modèle de Mélange Gaussien avec K composantes.

(b) **Objectif de K-Means :** Écrivez la fonction objectif minimisée par l'algorithme de K-means en fonctions des z_{ik}.

(c) **GMM contraint :** Considérez le GMM auquel on impose les contraintes suivantes sur les paramètres

- Les matrices de covariance sont isotropes et identiques : $\Sigma_k = \sigma^2 I$.
- Les poids des mélanges sont uniformes : $\pi_k = \frac{1}{K}$.

Montrez que le critères K-means et la log-vraisemblances **complète** sont proportionels à une constante négative près (dépendant de σ^2).

(d) **Équivalence entre algroithme du K-Means et CEM :**

(i) Écrivez la règle de mise à jour pour l'étape CE dans l'algorithme CEM dans le cadre du GMM contraint. Expliquez comment cette étape est équivalente à l'étape d'affectation des clusters dans K-means.

(ii) Écrivez la règle de mise à jour pour l'étape M dans l'algorithme CEM dans le cadre du GMM contraint. Montrez que cette étape correspond à la mise à jour des centroïdes des clusters dans K-means.

(iii) Qu'en déduisez-vous ?

(e) **Bilan:** Selons vous quels sont les avantages et inconvénients d'une affectation stricte (comme dans K-means et CEM) par rapport aux affectations souples (comme dans EM standard).

Exercice 2 : Algorithme EM pour le Mélange de Distributions Gamma

Dans cet exercice, vous dériverez et implémenterez l'algorithme EM pour le modèle de mélange de distribution Gamma. La loi Gamma est une loi continue à 2 paramètres défini par la densité de probabilité suivante :

$$\text{Gamma}(x \mid \alpha_k, \beta_k) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} x^{\alpha_k - 1} e^{-\beta_k x} \mathbb{1}_{x>0},$$

où α_k est le paramètre de forme et β_k le paramètre d'échelle.

(a) **Configuration du Modèle :** Considérez un ensemble de données positives $\{x_i\}_{i=1}^N$ tirées d'un mélange de K distributions Gamma. Écrivez la densité de probabilité pour le modèle de mélange et définissez les paramètres γ_k de chaque composante.

(b) **Étape E :** Dérivez l'expression pour les probabilités a posteriori τ_{ik} indiquant qu'un point de données x_i appartient à la composante k.

(c) **Étape M :** Dérivez les équations de mise à jour pour les paramètres des distributions Gamma (α_k et β_k) et les poids du mélange (π_k). Expliquez comment ces mises à jour sont calculées de manière itérative.

(d) **Implémentation de l'Algorithme :** Décrivez les étapes de l'algorithme EM pour le mélange de distributions Gamma sous forme de pseudocode.

(e) **Considérations Pratiques :** Discutez des défis liés à l'implémentation de l'algorithme EM pour ce modèle, en particulier en ce qui concerne l'initialisation, les critères de convergence et la stabilité numérique.

Exercicee 3: VEM for binary Stochastic Block Model (SBM)

The stochastic block model (SBM) assumes a graph with n nodes divided into K groups. The presence or absence of an edge between any two nodes is modeled as a Bernoulli random variable, where the probability of an edge depends on the groups to which the nodes belong.

Let:

- Let $q(Z) = \prod_{i=1}^n \prod_{k=1}^K \tau_{ik}^{z_{ik}}$ denote a probability distribution over all latent variable,
- γ_{kl} denote the probability of an edge between nodes in groups k and l,
- π_k denote the prior group proportion for group k, where $\sum_{k=1}^K \pi_k = 1$.

Questions

(a) How do we call a fully factorized distribution like q ?

(b) Write the Evidence lower bound (ELBO) for the binary SBM with the variational distribution q as above.

(c) **Variational E-step (VE-step)** What optimization has to be solved in the VE-step ? Does it have an exact solution ? If not, write down an iterative algorithm to solve it and explicit its stopping criterion(s).

(d) **M-step updates** Write down the updates for the parameters γ_{kl} and π_k using the current posterior membership probabilities.

(e) How would you cluster the nodes of the network after the VEM ?