TP: Introduction aux mélanges gaussiens

INSA-ENSEEIHT ModIA – 2022 – 1 heure

1. Implémenter la fonction `generate_data` pour générer des échantillons d'un mélange de K gaussiennes 2D.
   Paramètres :
   — N : taille de l'échantillon
   — π : coefficients de mixage (taille (K,))
   — μ : moyennes des K gaussiennes (taille (K,2))
   — Σ : matrices de covariance des K gaussiennes (taille (K,2,2))
   Sortie :
   — X : échantillons (taille (N,2))
   — z : variable latente (taille (N,))
   Tester la fonction avec les valeurs suivantes (on affichera les données générées avec et sans la variable latente en couleurs RGB) :
   — $N = 3000$
   — $\pi = [\frac{1}{3}, \frac{1}{4}, \frac{5}{12}]$
   — $\mu = \begin{bmatrix} -1.5 & 0 \\ 0.5 & 0.5 \\ 2 & 1 \end{bmatrix}$
   — $\Sigma = \left[\begin{bmatrix} 1 & 0.01 \\ 0.01 & 0.01 \end{bmatrix}, \begin{bmatrix} 0.1 & 0.01 \\ 0.01 & 0.3 \end{bmatrix}, \begin{bmatrix} 1 & 0.01 \\ 0.01 & 0.01 \end{bmatrix} \right]$
2. Mettre en évidence les limites de l'algorithme K-means sur cet échantillon. On appliquera l'algorithme K-means sur les échantillons générés et on affichera les données avec les classes obtenues avec l'algorithme en couleurs RGB.
3. Implémenter la fonction `compute_posteriors` pour calculer les probabilités postérieures $\gamma_{nk} = p(z = k|x_n)$, sans connaître la variable latente z.
   Paramètres :
   — X : échantillon de données
   — π : coefficients de mixage (taille (K,))
   — μ : moyennes des K gaussiennes (taille (K,2))
   — Σ : matrices de covariance des K gaussiennes (taille (K,2,2))
   Sortie :
   — $(\gamma_{nk})_{n,k}$: probabilités postérieures (taille (N, K))
   Tester la fonction avec les échantillons générés précédemment et les valeurs théoriques (utilisées pour générer les échantillons) des paramètres π, μ et Σ. On affichera les données avec les probabilités postérieures en couleur).
4. **Bonus 1 :** Sur le graphe obtenu à la question 2., afficher les cercles correspondant aux différents clusters. Sur celui obtenu à la question 3., afficher les ellipses correspondant aux gaussiennes utilisées pour générer les données.

5. **Bonus 2 :** Utiliser un autoencodeur simple pour projeter les données dans $\mathbb{R}$ et interpréter le résultat comme une classification non supervisée.