Le projet sera rendu sous la forme d'un fichier `Rmd` compilable en pdf. Vous détaillerez les calculs et pourrez utiliser le fichier `projet.rmd` (disponible sur le site) comme base du rapport.

Mélange de Bernoulli

Considérons un vecteur aléatoire binaire $\boldsymbol{x} \in [0,1]^p$ de p variables x_j suivant chacune une distribution de Bernoulli $\mathcal{B}(\mu_j)$. La distribution du vecteur s'exprime comme:

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{j=1}^{p} \mu_j^{x_j}(1-\mu_j)^{1-x_j},$$

avec $\boldsymbol{x} = (x_1, \cdots, x_p)^T$ et $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)^T$.
Soit une distribution mélange à K composantes de Bernoulli

$$p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{M}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k)$$

où les π_k sont les proportions du mélange et les $p(\boldsymbol{x}|\boldsymbol{\mu}_k)$ sont des distributions de Bernoulli multivariées de paramètres $\boldsymbol{\mu}_k = (\mu_{k1}, \cdots, \mu_{kp})^T$, et $M = \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}^T$ la matrice des paramètres des densités de classes.
Dans la suite nous considérerons

- un échantillon observé $X = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ issu de cette distribution mélange,
- des variables latentes $Z = \{z_1, \cdots, z_n\}$ indiquant la composante d'origine de chaque $\boldsymbol{x}_i$.

Exercise 1 Simulation d'un mélange à trois composantes

Considérons un mélange à 3 composantes de Bernoulli.
Pour simuler un vecteur un valeur $\mu_{kj} \in]0,1[$ vous pouvez utiliser la fonction R `runif()`.

1. Simuler une matrice M de proportions dont les 3 lignes et 50 colonnes décrivent 3 vecteurs des proportions d'un mélange de Bernoulli dans un espace de dimension 50.
2. Considérons un mélange à 3 composantes mélangées en proportions égales $\pi_1 = \pi_2 = \pi_3$.
3. Simuler $Z = \{z_1, \cdots, z_n\}$ pour $n = 200$.
4. Simuler $X|Z$.
5. Permuter les lignes de votre matrice X à 200 lignes et 50 colonnes.
6. Visualiser la matrice.
7. Appliquer les `kmeans` (3 classes) sur X et commenter le résultat.
8. Visualiser la matrice classée.

Exercise 2 Équations de l'algorithme EM

1. Écrire la log-vraisemblance complète $\ln p(X,Z|\boldsymbol{\theta} = \{\boldsymbol{\pi},\boldsymbol{M}\})$.
2. Exprimer $t_{ik}^q = \mathbb{E}[Z_{ik}]$ par rapport à la loi $p_{\theta^q}(Z|X)$ où $Z_{ik} = \mathbb{I}_{(Z_i=k)}$.
3. Écrire $Q(\theta^q|\theta)$ l'espérance de cette log-vraisemblance par rapport à la loi $p_{\theta^q}(Z|X)$.
4. Donner la forme de $\theta^{q+1} = argmax_\theta\ Q(\theta^q|\theta)$.
5. Détailler les étapes de l'algorithme EM qui permet d'estimer $\boldsymbol{\theta}$.
6. Donner la forme $-\mathbb{E}[\ln p_{\theta_{q+1}}(Z|X)]$ par rapport à la loi $p_{\theta^{q+1}}(Z|X)$.
7. Donner la forme de $\ln p_{\hat{\theta}}(X|\boldsymbol{\theta} = \{\boldsymbol{\pi},\boldsymbol{M}\})$.
8. Écrire le critère BIC associé à un modèle à K classes.
9. Écrire le critère ICL associé à un modèle à K classes.
10. Détailler l'algorithme EM d'estimation des paramètres du modèle de mélange de Bernoulli.
11. Détailler l'algorithme CEM associé au même mélange.

Exercise 3 Programmation de l'algorithme EM

1. Écrire un fonction `E-step` qui produit les t_{ik} (voir Exercice 1) à partir de $\boldsymbol{\Theta}$. Vérifier les résultats en injectant les vrais paramètres de votre simulation et en comparant les t_{ik} estimé par rapport aux variables latentes Z de votre simulation.
2. Écrire la fonction `M-step` qui produit les estimateurs des $\boldsymbol{\Theta}$ à partir des données observées et des t_{ik}. Encore une fois pour vérifier cette étape vous pouvez utiliser votre simulation dont vous connaissez tous les paramètres.
3. Écrire l'algorithme EM qui estime les paramètres d'un mélange de Bernoulli en K classes.
4. Tracer l'évolution de la vraisemblance à chaque demi-étape (E et M) lorsque vous appliquez l'algorithme aux données simulées.
5. Programmer la fonction BIC qui prend la sortie de votre algorithme EM et rend le critère BIC.
6. Programmer la fonction ICL qui prend la sortie de votre algorithme EM et rend le critère ICL.

Exercise 4 Données `state-firearms`

1. Appliquer votre algorithme au jeu de données `state-firearms` sur les lignes et les colonnes et commentez.