Exercice 1 On considère un échantillon indépendant $((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ d'individus $\boldsymbol{X}_i$ décrits par d variables réelles ($\boldsymbol{X}_i \in \mathbb{R}^d$) d'une population de K classes telle que $Y_i \in \mathcal{Y}$ est la classe de l'individu $\boldsymbol{X}_i$. On dispose d'un échantillon d'apprentissage $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ et d'un échantillon de test $((\boldsymbol{x}_{n+1}, y_{n+1}), \ldots, (\boldsymbol{x}_m, y_m))$ issus de cette même population. On s'intéresse à prédire les classes des données des test sur la base d'un modèle probabiliste appris sur les données d'apprentissage. On considère le cas de l'analyse discriminante gaussienne.

La prédiction s'effectue par la règle du maximum a posteriori (MAP) qui consiste à affecter l'individu $\boldsymbol{x}_i$ à la classe y_i maximisant la probabilité a posteriori :

$$y_i = \arg\max_{k \in \mathcal{Y}} \mathbb{P}(Y_i = k | \boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta}) \tag{1}$$

où $\pi_k = \mathbb{P}(Y_i = k)$ est la probabilité a priori de la classe k, $f_{\boldsymbol{X}|Y}(\boldsymbol{x}_i | Y_i = k; \boldsymbol{\theta}_k)$ est sa densité définie par On utilise une modélisation gaussienne pour chacune des classes, i.e.,

$$f_{\boldsymbol{X}|Y}(\boldsymbol{x}_i | Y_i = k; \boldsymbol{\theta}_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top {\boldsymbol{\Sigma}_k}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \tag{2}$$

pour la classe k avec $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_K^\top, \mathrm{vech}(\boldsymbol{\Sigma}_1), \ldots, \mathrm{vech}(\boldsymbol{\Sigma}_K))^T$ étant le vecteur paramètre du modèle.

Problème à deux classes : On considère le cas à deux classes (binaire) où $Y_i \in \mathcal{Y} = [\![0, 1]\!]$. Dans ce cas le vecteur paramètre est donné par $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\mu}_0^\top, \boldsymbol{\mu}_1^\top, \mathrm{vech}(\boldsymbol{\Sigma}_0), \mathrm{vech}(\boldsymbol{\Sigma}_1))^T$

On suppose que $\pi_0 = \pi_1 = 0.5$ avec $\pi_0 = \mathbb{P}(Y_1 = 0)$ et $\pi_1 = \mathbb{P}(Y_i = 1)$, $\boldsymbol{\mu}_0 = (1, 1)^t$, $\boldsymbol{\mu}_1 = (-1, -1)^t$, et $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \mathbf{I}_2$, $\mathbf{I}_d$ étant la matrice identité de dimension d.

1. Générer un échantillon $((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ de taille $n = 200$. On utilisera le processus génératif suivant : pour i allant de 1 à n,
$$Y_i \sim \mathcal{B}(1/2)$$
ensuite
$$\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_{Y_i}, \boldsymbol{\Sigma})$$
2. Afficher sur le même graphique les données de chacune des deux classes (on obtiendra un graphique similaire à celui vu en TD)
3. Déterminer et implémenter l'expression de $\mathbb{P}(Y_i = 0 | \boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta})$ et $\mathbb{P}(Y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta})$
4. Calculer et tracer la frontière de décision entre les deux classes
5. Prédire les classes des observations suivantes :
$$\boldsymbol{x}_1 = (-2, 3)^t, \quad \boldsymbol{x}_2 = (1, -1)^t, \quad \boldsymbol{x}_3 = (-2, 2)^t, \quad \boldsymbol{x}_4 = (1, -2)^t$$
6. Calculer l'erreur de prédiction sur l'ensemble des données générées

Exercice 2 On considère le cadre de l'exercice précédent et on suppose maintenant que $\boldsymbol{\Sigma}_0 = \mathbf{I}_2$ et $\boldsymbol{\Sigma}_1 = 2\,\mathbf{I}_2$.

1. $((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ de taille $n = 200$. On utilisera le processus génératif suivant :
   Pour i allant de 1 à n,
$$Y_i \sim \mathcal{B}(1/2)$$
   ensuite
$$\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_{Y_i}, \boldsymbol{\Sigma}_{Y_i})$$
2. Afficher sur le même graphique les données de chacune des deux classes
3. Déterminer et implémenter l'expression de $\mathbb{P}(Y_i = 0|\boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta})$ et $\mathbb{P}(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta})$
4. Calculer et tracer la frontière de décision entre les deux classes
5. Prédire les classes des observations suivantes :
$$\boldsymbol{x}_1 = (-2, 3)^t, \quad \boldsymbol{x}_2 = (1, -1)^t, \quad \boldsymbol{x}_3 = (-2, 2)^t, \quad \boldsymbol{x}_4 = (1, -2)^t$$
6. Calculer l'erreur sur l'ensemble des données générées

Exercice 3 On considère un échantillon $((X_1, Y_1), \ldots, (X_n, Y_n))$ d'un couple (X, Y) issue d'une population hétérogène à deux classes tel que au sein de la première classe la distribution du couple est définie par la densité
$$f_{(X,Y)}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) \tag{3}$$
et pour les données de la deuxième classe la densité est
$$f_{(X,Y)}(x, y) = \frac{1}{2\pi} \exp\left(-\frac{(x-2)^2 + y^2}{2}\right). \tag{4}$$
ρ étant le coefficient de corrélation linéaire ($|\rho| < 1$)). On note par $(Z_1, \ldots, Z_n)$ les classes des données de l'échantillon observées avec $Z_i \in \{0, 1\}$ la classe du couple (X_i, Y_i) $(i = 1, \ldots, n)$ et on cherche à prédire la classe Z_i pour un nouveau couple (X_i, Y_i) $(i = n+1, \ldots, m)$. On suppose que les classes sont à proportions identiques.

1. Générer un échantillon $((x_1, y_1, z_1), \ldots, (x_n, y_n, z_n))$ du triplet (X, Y, Z) de taille $n = 200$.
2. Afficher sur le même graphique les données de chacune des deux classes
3. Déterminer et implémenter la règle de décision de Bayes entre les deux classes
4. Calculer et tracer la forntière de décision pour $\rho = 0$. Commenter
5. Calculer et tracer la forntière de décision pour différentes valeurs de ρ de votre choix. Commenter