TP14 – Classification bayésienne

L'objectif est de réaliser un classifieur bayésien permettant de classer les images de trois espèces de fleurs. Lancez le script `donnees.m`, qui affiche des images de pensées, d'œillets et de chrysanthèmes. Vous observez que ces images n'ont pas toutes la même taille.

Exercice 1 : calcul de la couleur moyenne d'une image

Dans un premier temps, vous allez classer les images selon la couleur moyenne de chaque espèce de fleurs. En chaque pixel de chaque image, les trois niveaux de couleur $(R, V, B) \in [0, 255]^3$, qui sont entiers, sont d'abord transformés en *niveaux de couleur normalisés* (r, v, b) définis de la manière suivante :

$$(r, v, b) = \frac{1}{\max\{1, R + V + B\}} (R, V, B)$$

L'intérêt des niveaux de couleur normalisés est que deux valeurs parmi (r, v, b) permettent de déduire la troisième, puisque $r + v + b = 1$, sauf dans le cas exceptionnel où $(r, v, b) = (0, 0, 0)$. On peut donc caractériser une image par les moyennes $(\bar{r}, \bar{v}, \bar{b})$, ou plus simplement par $(\bar{r}, \bar{v})$, puisque $\bar{r} + \bar{v} + \bar{b} = 1$, c'est-à-dire par un vecteur $\mathbf{x} = [\bar{r}, \bar{v}] \in \mathbb{R}^2$ qu'on appelle sa *couleur moyenne*. Cela revient à effectuer une *réduction de dimension* plus souple que l'ACP, dans la mesure où les images peuvent avoir des tailles différentes. Compte tenu des différences de couleurs moyennes entre les trois espèces de fleurs, on postule que ce vecteur suffira à les distinguer.

Écrivez la fonction `moyenne`, appelée par le script `exercice_1.m`, qui calcule la couleur moyenne d'une image. N'oubliez pas de convertir les niveaux de couleur (R, V, B) au format `double`. Le script `exercice_1.m` est censé afficher les couleurs moyennes de l'ensemble des images de fleurs sous la forme de trois nuages de points de $\mathbb{R}^2$. Au regard de cette figure, la couleur moyenne vous semble-t-elle une caractéristique suffisamment discriminante de ces trois espèces de fleurs ?

Exercice 2 : classification par l'algorithme des k-moyennes

En utilisant la fonction `kmeans` de Matlab, écrivez un script de nom `exercice_2.m` qui calcule le pourcentage de bonnes classifications des données d'apprentissage. Comme pour le TP13, utilisez l'option `'emptyaction'` avec la valeur `'error'`, et l'option `'start'` avec comme valeur une matrice contenant les moyennes des classes.

Exercice 3 : estimation de la vraisemblance de chaque espèce de fleurs

Les trois nuages de points de l'exercice 1 peuvent également être modélisés par des lois normales bidimensionnelles. Il est rappelé que la densité de probabilité d'une loi normale s'écrit, en dimension d :

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} (\det \Sigma)^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mu) \Sigma^{-1} (\mathbf{x} - \mu)^\top \right\} \tag{1}$$

où :

- μ désigne l'espérance (la moyenne) des vecteurs $\mathbf{x} \in \mathbb{R}^d$: $\mu = E[\mathbf{x}] = \int \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x}$.
- Σ désigne la matrice de variance/covariance : $\Sigma = E[(\mathbf{x} - \mu)^\top (\mathbf{x} - \mu)]$.

Dans le cadre bayésien, la vraisemblance de la classe ω_k, qui est caractérisée par la moyenne μ_k et la matrice de variance/covariance Σ_k, peut être modélisée par une loi normale analogue à (1) :

$$p(\mathbf{x}|\omega_k) = \frac{1}{(2\pi)^{d/2}\,(\det \Sigma_k)^{1/2}} \exp\left\{-\frac{1}{2}\,(\mathbf{x}-\mu_k){\Sigma_k}^{-1}(\mathbf{x}-\mu_k)^\top\right\}, \qquad \forall\, k \in [1,3]$$

Il faut donc estimer les paramètres μ_k et Σ_k des trois classes correspondant aux trois espèces de fleurs.

Écrivez la fonction `estimation_mu_Sigma` permettant d'effectuer l'estimation empirique des paramètres d'une loi normale bidimensionnelle ($d = 2$) à partir des vecteurs $\mathbf{x} = [\bar{r}, \bar{v}]$ stockés dans la matrice de données `X`. Le script `exercice_3.m` est censé estimer les paramètres μ_k et Σ_k des trois classes ω_k correspondant aux trois espèces de fleurs, à partir des matrices `X_pensees`, `X_oeillets` et `X_chrysanthemes`, puis superposer la vraisemblance de chaque classe (en perspective) au nuage de points à partir de laquelle elle a été estimée.

Exercice 4 : classification par le maximum de vraisemblance

Nous souhaitons prédire à quelle espèce de fleurs une image requête $\mathbf{x}$ doit être associée. Comme nous avons utilisé des données étiquetées (chacune des images étant associée à une espèce de fleurs), il s'agit de **classification supervisée**. La *classification par maximum de vraisemblance* consiste à affecter à $\mathbf{x}$ la classe ω_k qui maximise la vraisemblance $p(\mathbf{x}|\omega_k)$.

Le script `exercice_4.m` affiche une partition du plan $\mathbb{R}^2$ en trois parties correspondant aux trois classes de fleurs. Complétez ce script de manière à afficher chaque image de la base d'apprentissage sous la forme d'une étoile dont la couleur est celle de la classe si l'image est correctement classée, ou le noir dans le cas contraire. Enfin, calculez le pourcentage d'images correctement classées.

Exercice 5 : amélioration du classifieur

L'observation attentive des images de pensées et d'œillets, dont les couleurs moyennes sont similaires, montre que ces deux espèces de fleurs ne sont pas structurées de la même façon : les pensées sont plus sombres au centre, c'est-à-dire au niveau du pistil. Cela suggère de ne pas seulement utiliser la couleur moyenne des images pour effectuer la classification.

Faites une copie du script `exercice_4.m`, de nom `exercice_5.m`, que vous modifierez de manière à utiliser, outre le couple de valeurs $(\bar{r}, \bar{v})$, une troisième caractéristique permettant d'augmenter le pourcentage de bonnes classifications.