TP13 – Classification bayésienne

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

Exercice 1 : calcul de la couleur moyenne de chaque image

Dans un premier temps, vous allez classer les images selon la couleur moyenne de chaque espèce de fleurs. En chaque pixel d'une image en couleur, les trois valeurs $(R, V, B) \in [0, 255]^3$ du codage « rouge-vert-bleu » sont d'abord transformées en couleurs normalisées (r, v, b) définies de la manière suivante :

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

Le principal intérêt des couleurs normalisées est que deux valeurs parmi (r, v, b) permettent de déduire la troisième, puisque $r + v + b = 1$, sauf dans le cas exceptionnel où $(r, v, b) = (0, 0, 0)$. Par conséquent, une image est caractérisée 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*. Compte tenu des différences de couleurs moyennes entre les trois espèces de fleurs, on espère que ce vecteur suffira à les distinguer.

Écrivez la fonction `x = moyenne(nom_image)` qui retourne la couleur moyenne de l'image `nom_image`. N'oubliez pas de convertir les valeurs (R, V, B) des pixels au format `double`.

Lancez le script `exercice_1.m`, qui calcule la couleur moyenne de chaque image, la stocke dans une des trois matrices `X_pensees`, `X_oeillets` ou `X_chrysanthemes`, et affiche l'ensemble de ces couleurs moyennes 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 pour distinguer les trois espèces de fleurs ?

Exercice 2 : estimation de la vraisemblance de chaque espèce de fleurs

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

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

- 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 qui nous intéresse, la vraisemblance de la classe ω_i, qui est caractérisée par la moyenne μ_i et la matrice de variance/covariance Σ_i, peut être modélisée par une loi normale analogue à (1) :

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{d/2} (\det \Sigma_i)^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mu_i) {\Sigma_i}^{-1} (\mathbf{x} - \mu_i)^\top \right\}, \quad i \in [1, 3]$$

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

Écrivez la fonction `[mu,Sigma] = mu_Sigma(X)` qui effectue 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 `X`.

Complétez le script `exercice_2.m`, qui estime les paramètres μ_i et Σ_i des trois classes ω_i correspondant aux trois espèces de fleurs, à partir des matrices `X_pensees`, `X_oeillets` et `X_chrysanthemes`, et qui superpose la vraisemblance de chaque classe (en perspective) au nuage de points à partir de laquelle elle a été estimée.

Exercice 3 : classification par maximum de vraisemblance

Nous souhaitons maintenant 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**. Un premier type de classification consiste à affecter à $\mathbf{x}$ la classe ω_i qui maximise la vraisemblance $p(\mathbf{x}|\omega_i)$. Il s'agit alors d'un classifieur « par maximum de vraisemblance ».

Complétez le script `exercice_3.m`, qui superpose les données d'apprentissage aux classes définies selon ce principe, et qui calcule le pourcentage d'images correctement classées.

Exercice 4 : classification par maximum a posteriori

Par ailleurs, la règle de Bayes donne l'expression suivante de la **probabilité a posteriori** $p(\omega_i|\mathbf{x})$, c'est-à-dire de la probabilité pour que la classe ω_i contienne $\mathbf{x}$:

$$p(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)\, p(\omega_i)}{p(\mathbf{x})} \tag{2}$$

Il semble naturel d'affecter à $\mathbf{x}$ la classe ω_i qui maximise $p(\omega_i|\mathbf{x})$. Une telle classification est dite « par maximum a posteriori » (MAP). Sachant que, dans l'expression (2), le dénominateur $p(\mathbf{x})$ est indépendant de ω_i, il n'est pas nécessaire de le connaître pour trouver le maximum des $p(\omega_i|\mathbf{x})$. En revanche, il est nécessaire de connaître la « probabilité a priori » $p(\omega_i)$ de chaque classe ω_i, faute de quoi on fait généralement l'hypothèse que les classes sont équiprobables (l'estimateur par maximum a posteriori revient alors à un estimateur par maximum de vraisemblance).

Écrivez un script `exercice_4.m` permettant d'afficher sur une même figure les trois « régions de décision » correspondant aux trois espèces de fleurs, ainsi que les points correspondant aux couleurs moyennes de toutes les images. En jouant sur les probabilités a priori des trois classes, essayez de maximiser le pourcentage d'images correctement classées.

Exercice 5 (facultatif) : amélioration du classifieur

Même en jouant sur les probabilités a priori, le classifieur obtenu reste décevant. Or, 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 calculer la couleur moyenne des images, mais de scinder chaque image en deux parties complémentaires : le centre C (notion à préciser) et le pourtour P (complémentaire de C).

Écrivez un script `exercice_5.m` reprenant le principe du classifieur bayésien de l'exercice 4, mais utilisant trois caractéristiques pour décrire une image. Ces caractéristiques pourraient être, par exemple : le couple de valeurs $(\bar{r}, \bar{v})$ calculées sur le pourtour P, ainsi que la valeur $\bar{r}$ calculée sur le centre C.