TP3 – Classification bayésienne

On souhaite réaliser un système d'aide au diagnostic médical permettant de classer des images de lésions cutanées en deux classes : dermatofibromes (lésions bénignes, sans gravité) et mélanomes (lésions malignes pouvant évoluer en cancer de la peau). Les données sont constituées d'images au format RVB, chacune de ces images étant associée à un fichier binaire appelé « masque », qui indique l'emplacement de la lésion déterminé par un expert en dermatologie. Ces données sont scindées en deux : un même nombre `n_app` d'images de chacune des deux classes (fibromes / mélanomes), et les masques qui leur sont associés, constituent les *données d'apprentissage*, tandis que les autres images et les autres masques constituent les *données de test*.

Le script `donnees_app` lit le fichier `donnees_app.mat`, qui stocke les données d'apprentissage dans quatre matrices : `I_fibrome` contient `n_app` images de fibromes, `I_melanome` contient `n_app` images de mélanomes, tandis que `M_fibrome` et `M_melanome` contiennent les masques correspondants. À l'intérieur d'une même classe, vous constatez que les images présentent une forte variabilité, ce qui explique que seul un expert en dermatologie puisse faire un diagnostic fiable. Or, le principe de l'apprentissage statistique est le suivant : en ayant appris sur les données d'apprentissage les *caractéristiques* de chaque classe, il devient possible de classer automatiquement de nouvelles images, appelées *données de test*. Dans ce TP, deux méthodes de *classification bayésienne* sont mises en œuvre pour ce faire : le maximum de vraisemblance et le maximum a posteriori.

Choix des caractéristiques

Le script `exercice_0` calcule les trois caractéristiques suivantes de chacune des données d'apprentissage :

- Caractéristique 1 – La première caractéristique, appelée **compacité**, est égale à la racine carrée de l'aire de la tache, divisée par son périmètre (la valeur de cette caractéristique ne peut pas dépasser celle d'un disque, qui vaut $\sqrt{\pi R^2}/(2\pi\, R) \approx 0,35$).
- Caractéristique 2 – Après conversion du format RVB vers le format Ycbcr, la deuxième caractéristique, appelée **contraste**, est égale à l'écart-type de la tache dans le canal Y (canal de « luminance »).
- Caractéristique 3 – Calculée grâce à la « matrice de co-occurrence », la troisième caractéristique est appelée **texture**.

Le script `exercice_0` stocke, dans une matrice `X_app` de taille $n_{\text{app}} \times n_{\text{caractéristiques}} \times n_{\text{classes}}$, les caractéristiques de l'ensemble des données d'apprentissage, et les affiche sous la forme de deux nuages de points 3D, qui correspondent aux deux classes de lésions de la peau. Parmi ces trois caractéristiques, quelles sont les deux qui vous semblent les plus « discriminantes » ? Reportez votre choix dans les variables `ind_carac_1` et `ind_carac_2`, au début du script `exercice_1`.

Exercice 1 : apprentissage statistique

Après réduction d'une dimension dans l'espace des caractéristiques, les deux nuages de points 3D précédents deviennent des nuages de points 2D, qui peuvent être modélisés par des lois normales bidimensionnelles. Il est rappelé que la densité de probabilité d'une loi normale en dimension d s'écrit, pour $\mathbf{x} \in \mathbb{R}^d$:

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

Dans cette expression, $\mu \in \mathbb{R}^d$ désigne la moyenne et $\Sigma \in \mathbb{R}^{d\times d}$ la matrice de variance/covariance :

$$\mu = \mathbb{E}[\mathbf{x}] = \int_{\mathbf{x}\in\mathbb{R}^d} \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x} \qquad ; \qquad \Sigma = \mathbb{E}[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top] \tag{2}$$

Pour une image caractérisée par un individu $\mathbf{x} \in \mathbb{R}^2$, la vraisemblance, relativement à l'une des deux classes $k \in \{1, 2\}$, définie par les paramètres $\mu_k \in \mathbb{R}^2$ et $\Sigma_k \in \mathbb{R}^{2\times 2}$, s'obtient par une loi normale de type (1) :

$$p(\mathbf{x}|k) = \frac{1}{2\pi \left(\det \Sigma_k\right)^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \mu_k\right)^\top {\Sigma_k}^{-1} (\mathbf{x} - \mu_k)\right\} \tag{3}$$

Écrivez les fonctions `estimation_mu_Sigma` et `vraisemblance`, appelées par le script `exercice_1`, qui permettent d'effectuer l'estimation empirique des paramètres de la loi normale bidimensionnelle (3) de chaque classe, à partir des données d'apprentissage stockées dans la matrice `X_app`, et de superposer la vraisemblance de chaque classe au nuage de points 2D à partir desquels elle a été estimée.

Exercice 2 : classification par le maximum de vraisemblance

La classification par le *maximum de vraisemblance* (MV) consiste à affecter un individu $\mathbf{x}$ à la classe $k \in \{1, 2\}$ qui maximise sa vraisemblance $p(\mathbf{x}|k)$. Le script `exercice_2` affiche une partition du plan en deux couleurs correspondant aux deux classes.

Écrivez la fonction `classif_MV`, appelée par le script `exercice_2`, permettant de calculer le pourcentage d'images d'apprentissage correctement classées. Comparez les pourcentages de bonnes classifications obtenus pour les trois paires de caractéristiques possibles, en relançant à chaque fois `exercice_1` puis `exercice_2`.

Exercice 3 : classification par le maximum a posteriori

Certaines données statistiques peuvent compléter utilement les vraisemblances apprises sur des données d'apprentissage. Par exemple, il s'avère que les femmes ont plus de chances de développer un dermatofibrome que les hommes. Une telle information constitue ce que l'on appelle un *a priori*. La *règle de Bayes* donne l'expression de la *probabilité a posteriori* :

$$p(k|\mathbf{x}) = \frac{p(k)\, p(\mathbf{x}|k)}{p(\mathbf{x})} \tag{4}$$

en fonction de la *vraisemblance* $p(\mathbf{x}|k)$ et des *probabilités a priori* $p(k)$ et $p(\mathbf{x})$. La classification par le *maximum a posteriori* (MAP) consiste à chercher la classe qui maximise l'expression (4) de la probabilité a posteriori. Comme le dénominateur est indépendant de k, la classification par MAP revient à résoudre le problème suivant :

$$\widehat{k} = \underset{k=1,2}{\arg\max} \left\{p(k)\, p(\mathbf{x}|k)\right\} \tag{5}$$

En pratique, les classifieurs MV et MAP sont très similaires : la seule différence consiste à pondérer, dans le problème d'optimisation (5), la vraisemblance $p(\mathbf{x}|k)$ de la donnée de test $\mathbf{x}$ par la probabilité $p(k)$.

Écrivez la fonction `classif_MAP`, appelée par le script `exercice_3`, dont le rôle est de calculer le pourcentage de bonnes classifications des données d'apprentissage par MAP, pour une liste de valeurs de la probabilité a priori p_1 de la classe « fibrome » (la probabilité a priori p_2 de la classe « mélanome » étant telle que $p_1 + p_2 = 1$). Vous devez observer que la classification par MAP est plus performante que la classification par MV.

Exercice 4 : validation sur les données de test

Faites une copie du script `exercice_2`, de nom `exercice_4`, que vous modifierez de manière à calculer le pourcentage de bonnes classifications des classifieurs MV (exercice 2) et MAP (exercice 3), lorsque ce dernier est « optimisé », c'est-à-dire pour la probabilité a priori p_1 donnant le meilleur pourcentage de bonnes classifications sur les données d'apprentissage. Les caractéristiques des données de test sont contenues dans deux matrices `X_fibrome` et `X_melanome`, qui sont accessibles en chargeant le fichier `donnees_test.mat`.