TP3 – Classification bayésienne

On souhaite réaliser un système d'aide au diagnostic médical permettant de classer des images de lésions cutanées en deux classes : fibromes (lésions bénignes, sans gravité) et mélanomes (lésions malignes pouvant évoluer en cancer de la peau). Les données sont constituées d'images au format RVB, associées 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 :

- les données formant *l'ensemble d'apprentissage*, constituées d'un nombre n_{app} d'images et de masques de chacune des deux classes (fibromes / mélanomes) pour estimer les paramètres recherchés,
- les données formant *l'ensemble de test*, constituées d'un nombre n_{test} d'images et de masques de chacune des deux classes pour valider la généralisation du modèle trouvé avec *l'ensemble d'apprentissage*.

(a) (b) (c) (d)

FIGURE 1 – (a,b) Image et masque d'un fibrome (classe 1). (c,d) Image et masque d'un mélanome (classe 2).

À l'intérieur d'une même classe, 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.

Exercice 0 : choix des caractéristiques

Le script `exercice_0` charge les données contenues dans `donnees_carac.mat`. La matrice `X_app` contient les trois caractéristiques suivantes de chacune des n_{app} données d'apprentissage, et la matrice `X_test` contient ces mêmes caractéristiques pour les n_{test} données de test :

- 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**.

Ces données sont affichées sous la forme de deux nuages de points 3D, qui correspondent aux deux classes de lésions de la peau, et de trois figures représentant seulement deux de ces caractéristiques (cf. FIGURE 2, page suivante). La classe d'une donnée d'apprentissage est contenue dans le vecteur `Y_app`, celle d'une donnée de test dans le vecteur `Y_test` : la classe 1 correspond aux fibromes, la classe 2 aux mélanomes. 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`.

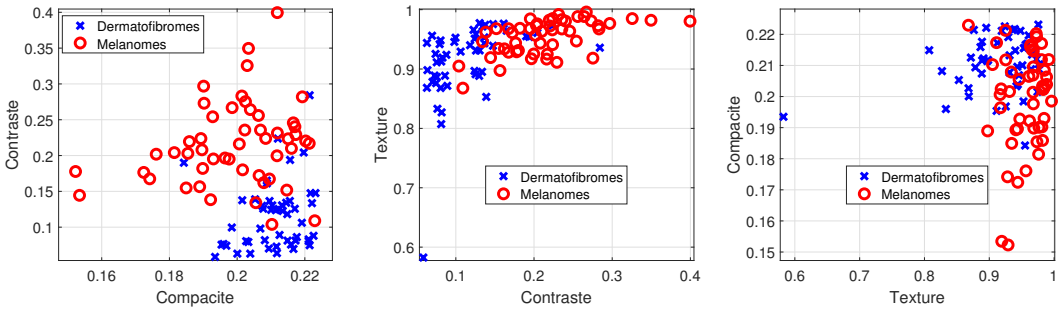

FIGURE 2 – Visualisation des données suivant toutes les possibilités de paires de caractéristiques.

Exercice 1 : apprentissage statistique pour modéliser la vraisemblance

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) :

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

Écrivez la fonction `estim_param_vraisemblance`, appelée par le script `exercice_1`, qui permet d'effectuer l'estimation empirique des paramètres μ_k et Σ_k de la loi normale bidimensionnelle (3) de la classe k, en utilisant les matrices `X_app` et `Y_app`. Complétez ensuite la fonction `modelisation_vraisemblance`, qui calcule la vraisemblance à partir d'une matrice à deux colonnes dont chaque colonne contient une des deux caractéristiques sélectionnées. Grâce à cela, le script `exercice_1` calcule la vraisemblance sur une grille régulière pour la superposer au nuage de points 2D à partir desquels elle a été estimée. La caméra de la figure de droite se déplace progressivement en vue de dessus pour faire apparaître la partition du plan en deux classes.

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 $p(\mathbf{x}|k)$. Écrivez la fonction `classification_MV`, appelée par le script `exercice_2`, qui doit retourner le vecteur de prédiction `Y_pred_MV` des données `X_app` : la valeur contenue dans `Y_pred_MV(i)` doit être égale à 1 si, pour l'individu `X_app(i,:)`, la vraisemblance de la classe 1 (fibrome) est supérieure à celle de la classe 2 (mélanome), et à 2 sinon.

Afin de valider la classification, complétez la fonction `qualite_classification` qui doit retourner les pourcentages de bonnes classifications pour l'ensemble des données, et pour chaque classe, à partir du vecteur `Y_app_pred_MV` et du vecteur `Y_app`, qui constitue la vérité terrain. Comparez les pourcentages de bonnes classifications obtenus pour les trois paires de caractéristiques possibles, en relançant à chaque fois les scripts `exercice_1` puis `exercice_2`, afin de confirmer ou non votre intuition de départ sur le choix de la paire de caractéristiques la plus discriminante.

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 une plus forte probabilité de développer un fibrome 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 la classe k, la classification par MAP revient à résoudre le problème suivant :

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

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 $p(k)$. Le cas du maximum de vraisemblance revient alors à un cas particulier du maximum a posteriori, en fixant les deux probabilités a priori $p(k)$ à 0,5.

Dans un premier temps, écrivez la fonction `classification_MAP` qui effectue la classification par le maximum a posteriori en ajoutant la probabilité a priori p_1 de la classe « fibrome » en plus de la vraisemblance (la probabilité a priori p_2 de la classe « mélanome » étant telle que $p_1 + p_2 = 1$). Écrivez ensuite la fonction `maximisation_classification_MAP`, appelée par le script `exercice_3`, dont le rôle est, pour différentes valeurs de p_1, de retourner le meilleur pourcentage de bonnes classifications des données d'apprentissage par MAP, la probabilité $p_1^{\max}$ correspondante, et le vecteur contenant les pourcentages de bonnes classifications pour l'ensemble des valeurs de p_1 testées. Sur la figure obtenue en lançant le script `exercice_3`, vous devez observer que la classification par MAP est plus performante que la classification par MV.

Une fois la probabilité a priori maximale obtenue, reportez la valeur de $p_1^{\max}$ dans le script `exercice_3bis` afin de visualiser la répartition des données d'apprentissage avec le classifieur MAP et de comparer les résultats obtenus avec le classifieur MV.

Exercice 4 : validation sur les données de test

Afin de vérifier la qualité de ces deux classifieurs, il est nécessaire de les tester sur un nouvel ensemble de données n'ayant pas servi à calculer les paramètres des vraisemblances : il s'agit de l'*ensemble de test* `X_test` contenu lui aussi dans le fichier `donnees_carac.mat`. En vous servant des scripts `exercice_2` et `exercice_3bis`, créez un script, de nom `exercice_4`, permettant de calculer et d'afficher les pourcentages de bonnes classifications, sur l'ensemble de test, des classifieurs MV et MAP, lorsque ce dernier est « optimisé », c'est-à-dire pour la probabilité a priori $p_1^{\max}$ ayant donné le meilleur pourcentage de bonnes classifications sur les données d'apprentissage.