IAFD 1 - Master Premiere Année, V. PAGE

TP1

Préliminaires

Vous trouverez sur le site le fichier **UCI_iris.zip**. Celui ci contient des échantillons d'éléments que vous devrez apprendre a classifier. Le fichier zip contient trois fichiers :

- les données (au format csv) (**bezdekIris.data**)
- les données (au format csv) légèrement modifiée pour vous simplifier la vie (**bezdekIris.data.v2**)
- un descriptif des données.

Ces données concernent des plantes dont on donne, pour chaque exemple, la longueur et la largeur de la tige et la longueur et largeur des pétales. Chaque plante appartient a une des trois classes possibles (Iris Setosa / Iris Versicolour / Iris Virginica).

L'objectif de ce Tp est que vous mettiez en œuvre une procédure de classification de votre choix, a l'aide de matlab, et que vous évaluiez son taux d'erreur en généralisation.

Pour lire les données d'un fichier csv, vous pouvez utiliser la fonction **csvread**. Celle ci nécessite que le fichier csv ne contienne que des valeurs numériques. C'est la raison de l'existence du fichier **bezdekIris.data.v2** dans lequel le nom de la classe a été remplacé par un chiffre (resp. 0, 1, 2).

Pour l'évaluation de ce TP, vous devrez rendre vos sources ainsi qu'un compte rendu au format pdf, incluant les courbes utiles

Contenu du TP

Question 1 : Constitution des Bases

Pour toute techniques d'apprentissage, que devez vous commencer par faire avec les exemples dont vous disposez, et pourquoi ?

Mettez donc au point une fonction matlab qui renvoie les deux tableaux correspondant respectivement à la base d'apprentissage et à la base de généralisation.

Question 2 : Visualisation des données

1. Quelle est la taille de l'espace des caractéristiques ?

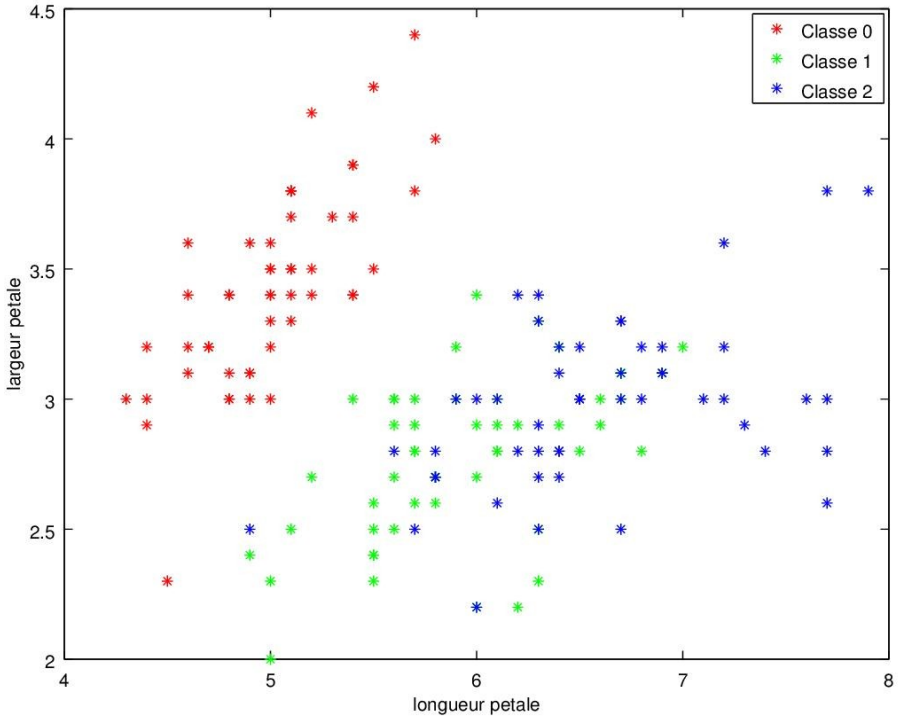
2. Visualisez vos données :

Autant que possible, regardez vos données pour comprendre comment elles sont réparties pour chaque classe. Compte tenu de la taille de votre espace de charactéristiques, est il possible de

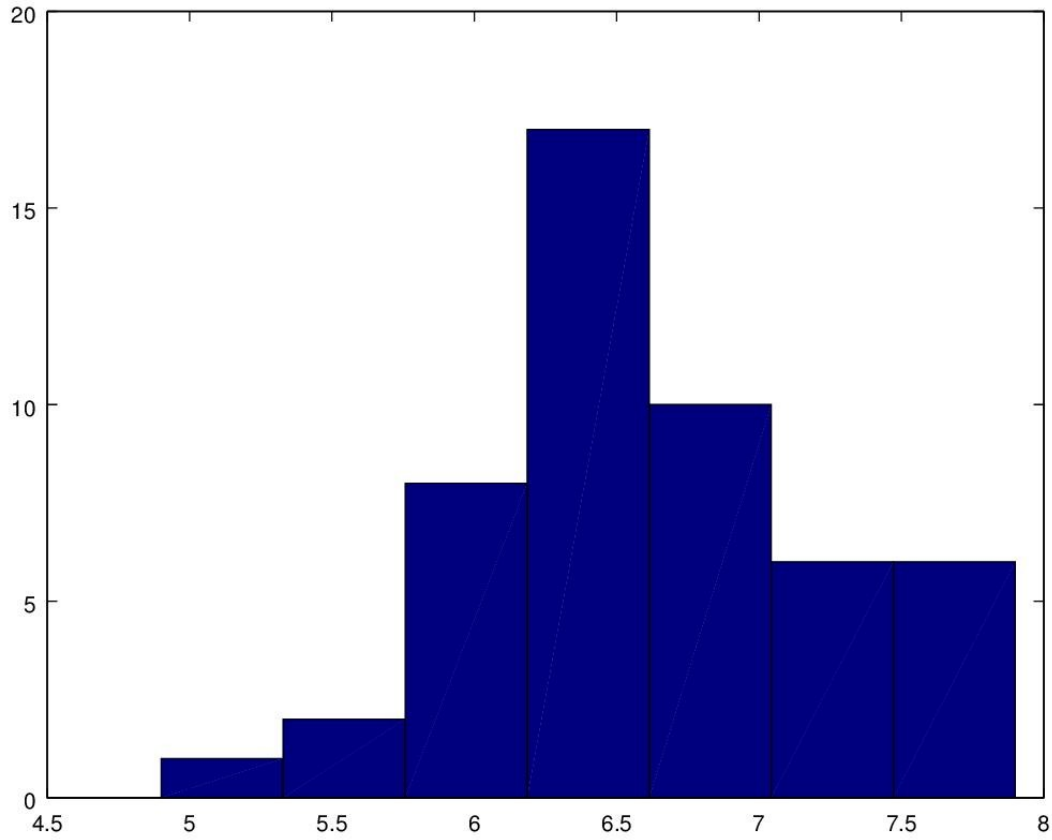
visualiser toutes les caractéristiques en meme temps ?

De ce fait, on les visualisera seulement partiellement, on regardera par exemple seulement les caractéristiques (longueur des tiges , largeur des tiges) ou (longueur des pétales, largeur des pétales) et tout autre couple pour se faire une idée aussi précise que possible de la répartition des classes dans l'espace des caractéristiques.

Pour cela, faites une fonction qui permette de visualiser 2 caractéristiques de chaque echantillon. Une couleur spécifique sera affectée a chaque classe comme dans le diagramme suivant.



Visualisez les histogrammes de chaque caractéristique, pour chaque classe, comme dans la figure suivante .



Question 3 : Choix d'une méthode de classification

Compte tenu de vos observations lors de la phase précédente :

- Les classes vous semblent-elles mono-modales ?
- Un modèle gaussien pour chaque caractéristique vous semble-t-il raisonnable ?
- Selon vous, peut on considérer que les caractéristiques sont indépendantes entre elles ?
- Si l'on s'appuie sur la théorie bayesienne de la décision, quelles quantités vous faut il calculer pour prendre une décision pour un exemple donné ?
- Disposez vous d'une formule analytique de ces quantités ?
- De quelles méthodes diposez vous pour estimer ces quantités ?

Mettez en œuvre une solution parmi celles proposées ci-dessous :

Eventuellement, vous pouvez ne baser votre classifieur que sur une seule caractéristique (c'est moins bien noté)

- Plus proches voisins.
- Classifieur Bayésien, hypothèse de densités de probabilités gaussiennes, estimation des paramètres de moyenne et de variance, hypothèse d'indépendance des charactéristiques.
- Classifieur Bayésien, hypothèse de densités de probabilités d'une loi normale multi variée,

estimation de la matrice de covariance.

- Classifieur Bayésien, estimation des densités de probabilité a l'aide de Parzen, hypothèse d'indépendance des caractéristiques
- Classifieur Bayésien, estimation des densités de probabilité a l'aide de Parzen, caractéristiques dépendantes

Estimez la probabilité d'erreur de votre classifieur.