Classification supervisée
Analyse factorielle discriminante

Charlotte Baey (charlotte.baey@univ-lille.fr)

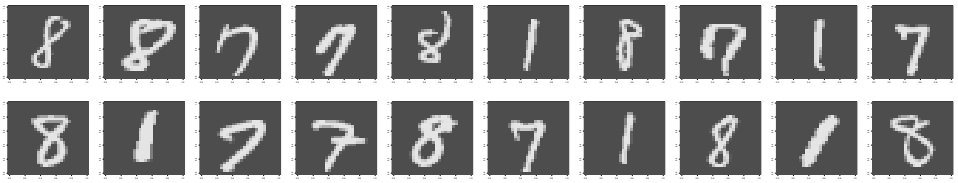
1 Les iris de Fisher

On considère le (fameux) jeu de données des iris de Fisher, disponible sous R par la commande `data(iris)`.

1. Décrivez le jeu de données.
2. Diviser le jeu de données en une partie apprentissage, contenant la moitié des individus (choisis aléatoirement), et une partie test contenant les individus restants.
3. Faites une AFD sur l'échantillon d'apprentissage, et proposer une règle de décision pour l'affectation de nouvelles données dans les différents groupes. Quel est le taux de bien classés et de mal classés sur l'échantillon test ?

2 Reconnaissance de chiffres

Dans cet exercice, on s'intéresse à une base de données contenant plusieurs chiffres manuscrits, représentés sous forme d'images et dont on présente un extrait dans la figure ci-dessous. Chaque chiffre est représenté par une matrice de taille 28×28, chaque élément de cette matrice correspondant à un niveau de gris. L'objectif de cet exercice est de décrire, puis de prédire, l'appartenance d'une image à chacune des trois classes de l'échantillon : 1, 7 ou 8.



Les données se trouvent dans la base `digits.rds`, qui contient :

— une matrice `x` contenant 3000 lignes et 784 colonnes, correspondant à l'échantillon d'apprentissage.
— une matrice `xt` contenant 1500 lignes et 784 colonnes, correspondant à l'échantillon test.
— un vecteur `y` contenant les étiquettes de l'échantillon d'apprentissage
— un vecteur `yt` contenant les étiquettes de l'échantillon test.

1. Importer le jeu de données sous R.
2. Tracer quelques images pour vous familiariser avec la base de données et son format.
3. Réaliser une AFD sur la base d'apprentissage. On utilisera la fonction `discrimin` du package `ade4`. Combien d'axes discriminants obtient-on ?

4. Tracer le nuage de points dans le premier plan discriminant. Commenter.
5. Proposer une règle de décision pour affecter un nouveau point à l'un des trois groupes, et appliquer cette règle à l'échantillon test. Quel est le taux de mal classés ? Que remarque t-on ? Pourquoi ?

3 Classification d'espèces d'insectes

On considère le jeu de données publié par Lubischew en 1962, et comportant des relevés de 6 variables morphologiques sur un ensemble de 74 insectes appartenant à 3 groupes différents. Ce jeu de données est disponible dans le package `amap`, sous le nom `lubisch`, ou fourni dans le fichier texte `insectes.txt`.

1. Importer le jeu de données, et réaliser une analyse descriptive de la base de données. Quelle est la variable la plus dispersée ? la moins dispersée ? Combien y a t-il d'insectes dans chaque groupe ?
2. Calculer les centres de gravité de chaque groupe, ainsi que le centre de gravité de l'ensemble des points.
3. Réaliser une AFD. Combien d'axes gardez-vous ? Pourquoi ?
4. Représenter les insectes sur le premier plan de l'AFD, et les variables discriminantes sur le cercle des corrélations correspondant. Commentez.