Classification supervisée
Analyse discriminante gaussienne

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

Dans plusieurs ouvrages d'analyse de données (ou de machine learning, ou encore de statistical learning ...) l'analyse factorielle discriminante vue au TP précédent est traitée en même temps que l'analyse discriminante gaussienne. En effet, l'AFD peut être vue comme un cas particulier de l'analyse discriminante gaussienne, lorsque les groupes étudiés ont la même matrice de covariance.

Dans le premier exercice, nous allons confronter ces deux méthodes sur un jeu de données fictives, puis dans l'exercice 2 nous testerons la méthode sur un jeu de données réelles.

1 Analyse discriminante gaussienne homoscédastique ou hétéroscédastique ?

L'analyse discriminante gaussienne consiste à considérer que le nuage de points (ou l'échantillon) est constitué d'un ensemble de points issus d'une loi normale, dont les paramètres diffèrent selon le groupe. Les paramètres d'une loi normale étant sa moyenne et sa matrice de covariance, construire un modèle discriminant gaussien revient donc à **estimer les moyennes et les matrices de covariance de chaque groupe**. Deux choix sont possibles :

- les groupes ont tous des **moyennes différentes** mais **la même matrice de covariance** : c'est e modèle *homoscédastique*. Concrètement, cela signifie que les points sont dispersés de la même manière dans leurs groupes respectifs, mais que les centres de gravité des groupes sont tout simplement translatés les uns par rapport aux autres. **Ce modèle revient à une AFD, au sens où la frontière de discrimination sera la même**.
- les groupes ont tous **des moyennes ET des matrices de covariance différentes** : c'est le modèle *hétéroscédastique*. Il y a plus de liberté car on autorise les groupes à avoir des dispersions différentes selon les groupes. Ceci dit, il y a aussi plus de paramètres à estimer.

Nous allons illustrer les différences entre ces deux méthodes.

1. Importer le jeu de données `data1.txt`, et en faire une analyse descriptive (combien de variables, visualisation, ...). Identifier la variable de classe que l'on va devoir prédire et les variables explicatives.
2. Tracer le nuage de points dans le plan formé par les deux variables explicatives, et colorer les points en fonction de leur groupe.
3. Faire une analyse discriminante gaussienne homoscédastique (avec la fonction `lda`) et une analyse discriminante gaussienne hétéroscédastique (cette fois à l'aide de la fonction `qda`). Autrement dit, réaliser l'analyse et stocker les résultats dans deux objets distincts.
4. Tracer la frontière de décision obtenue avec chaque méthode. Pour cela :

— créer une grille de points qui recouvre l'étendue de vos variables. Par exemple, si la variable 1 s'étend de 0 à 1 et de même pour la variable 2, on créé une grille de points de coordonnées (i, j), avec i et j variant de 0 à 1 avec un pas de 0.01 par exemple. On peut raffiner la grille si besoin. On aura alors la grille de coordonnées suivante :

$$\begin{array}{cccc} (0,1) & (0.01,1) & \dots & (1,1) \\ (0,0.99) & (0.01,0.99) & \dots & (1,0.99) \\ \vdots & \vdots & \dots & \vdots \\ (0,0) & (0.01,0) & \dots & (1,0) \end{array}$$

N.B. : on pourra créer un data frame qui contient trois colonnes, une pour l'abscisse des points et une pour l'ordonnée. On pourra aussi utiliser la fonction `expand.grid`. **Attention à ne pas prendre trop de points (500 sur chaque direction devraient suffire) si vous faites une double boucle** ...

— prédire la classe de chacun de ces points à l'aide de la fonction `predict` (regarder l'aide de la fonction pour comprendre comment elle fonctionne)

— superposer ces points au graphique précédent, en les colorant en fonction de la classe prédite

N.B. : Pour superposer des points à un graphique existant, on peut utiliser la fonction `points`. *Pour rendre le graphe plus lisible, on pourra utiliser l'option* `cex=0.1` *ou* `pch='.'` *pour faire des points très fins.*

5. Faire la même chose avec le jeu de données `data2`.
6. `lda` signifie "linear discriminant analysis" et `qda` signifie "quadratic discriminant analysis". A votre

2 Prédiction du diabète

Dans cet exercice, on s'intéresse à la prédiction du diabète chez les femmes, en fonction de plusieurs variables cliniques. Les données se trouvent dans la table `diabetes.csv`. La variable à prédire est la variable `Outcome`.

1. Importer le jeu de données sous R. Vérifier que l'importation s'est bien passée (analyse descriptives, ...)
2. Créer un échantillon d'apprentissage (avec 80% de la base) et un échantillon test (avec les 20% restants).
3. Réaliser une analyse discriminante gaussienne homoscédastique et hétéroscédastique sur la base d'apprentissage. Comparer les performances des deux méthodes en calculant la matrice de confusion de la base d'apprentissage, et de la base de test. Les performances sont-elles similaires sur les bases d'apprentissage et de test ? Quelle méthode préférez-vous ?