Analyse Discriminate Linéaire

HUYLENBROECK Florent, DELFOSSE Charly, JOSSE Thomas

7 juin 2019

Table des matières

Introduction

Qu'est-ce que l'ADL ?

Cette technique fait partie des techniques d'analyse discriminante prédictive. Le but et de pouvoir expliquer et prédire l'appartenance d'un individu à un groupe prédéfini à partir de caractéristiques qui ont été mesurées au préalable à l'aide de variables prédictives. On peut la comparer à la régression logistique.

Variables utilisées

X contient les j variables prédictives $[X = (X_1, ..., X_j)]$, n est le nombre d'observations réparties dans K groupe d'effectifs n_k, Y est notre variable à prédire qui prend des valeurs dans l'ensemble $\{y_1, ..., y_K\}$. $\pi_k = P(Y = y_k)$ et est la probabilité que Y soit dans le groupe k, $f_k(x)$ est la densité de probabilité de X dans le groupe k. De plus nous allons travailler dans le cas gaussien et donc $X \sim \mathcal{N}(\mu_k, \Sigma_k)$ dans chaque groupe k, E_k est le groupe d'individus possédant la modalité k dans l'échantillon et n_k est le cardinal de E_k

Fonctionnement

But Final

Cette règle ne prend en compte aucune hypothèse probabiliste. Elle consiste à calculer la distance de x (vecteur des variables explicatives sur un individu que l'on veut classer) à chacun des K centres de gravité $g_1, ..., g_K$ et affecter x au groupe le plus proche. Et cette distance du nouvel individu au groupe k peut être trouvée via la formule :

$$d^2(x, g_k) = (x - g_k)' \mathbf{W}^{-1} (x - g_k)$$

, Où $\mathbf{W}$ est la matrice des variance-covariance intra-groupe.

Développement

Pour pouvoir faire cela, nous allons définir notre fonction linéaire discriminante du groupe k pour savoir si x appartient au groupe k* tel que :

$$k^* = arg \max_{k=1,\dots,K} d^2(x, g_k)$$

que l'on peut réécrire :

$$k^* = arg \max_{k=1,\dots,K} L_K(x)$$

où

$$L_K(x) = x'\mathbf{W}^{-1}\mathbf{g}_k - \frac{1}{2}\mathbf{g}_k'\mathbf{W}^{-1}g_k$$

Et $L_K(x)$ est notre fonction linéaire discriminante du groupe k. Chaque $L_K(x)$ définit une fonction score qui donne une note représentant la probabilité d'appartenance au groupe de la fonction linéaire. X est donc affecté au groupe dont le score est le plus grand.

La règle bayesienne

Cette règle consiste à produire une estimation de la probabilité après notre affectation. Cela veut dire que nous devons réaliser une estimation pour une probabilité conditionnelle :

$$P(Y = y_k|X) = \frac{P(Y = y_k) \times P(X|Y = y_k)}{\sum_{i=1}^{K} P(Y = y_i) \times P(X|Y = y_i)}$$

Nous avons $P(Y = y_k)$ qui est la probabilité d'appartenance à la classe y_k et $P(X|Y = y_k)$ qui est la fonction de densité des x par rapport à l'appartenance à la classe y_k.
Et cette règle permet d'affecter une nouvelle observation x au groupe k* tel que :

$$k^* = arg \max_{k=1,\ldots,K} P(Y = y_k|X)$$

$$k^* = arg \max_{k=1,\ldots,K} \pi_k f_k(x)$$

Hypothèse d'homoscédasticité

L'homoscédasticité est le fait que les variances de chaque groupe soit équivalente (son contraire est l'hétéroscédasticité). Cela veut dire que les données sont réparties de la même manière autour de leur moyenne, ou centre de gravité.
Pour pouvoir effectuer de l'analyse discriminante linéaire c'est la principale hypothèse que l'on doit appliquer à nos données.

Développement

L'hypothèse respectée on peut réécrire la règle de Bayes telle que :

$$k^* = arg \max_{k=1,\ldots,K} x!\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \ln(\pi_k)$$

Il nous reste à estimer les paramètres sur l'échantillon d'apprentissage :

- μ_k est estimée par $g_k = \frac{1}{n_k}\sum_{i\in E_k} x_i$
- la matrice W avec Σ commune à tous les groupes devient $\mathbf{W} = \frac{1}{n}\sum_{k=1}\sum_{i\in E_k}(x_i - \mathbf{g}_k)(x_i - \mathbf{g}_k)'$ avec biais ou sans biais :

$$\mathbf{W} = \frac{1}{n-K}\sum_{k=1}\sum_{i\in E_k} x_i - \mathbf{g}_k)(x_i - \mathbf{g}_k)'$$

Développement (suite)

On obtient ainsi notre règle pour classifier nos individus par l'analyse discriminante linéaire :

$$k^* = arg \max_{k=1,\dots,K} L_k(x)$$

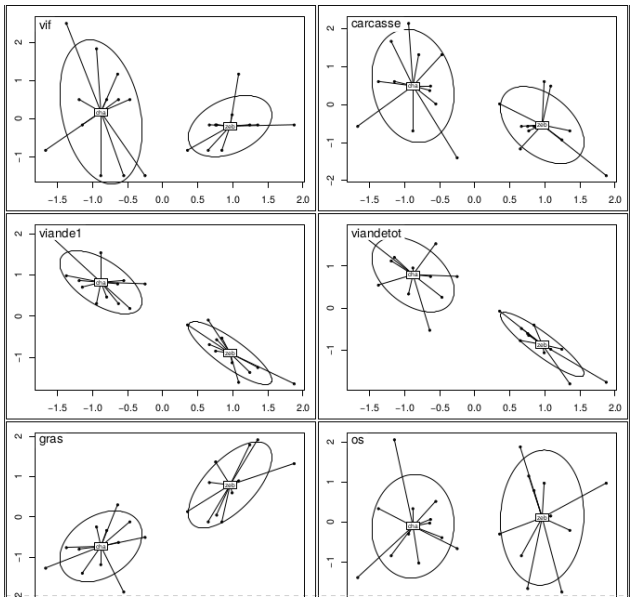
où

$$L_k(x) = x'\mathbf{W}^{-1}\mathbf{g}_k - \frac{1}{2}\mathbf{g}_k'\mathbf{W}^{-1}\mathbf{g}_k + \ln(\hat{\pi}_k)$$

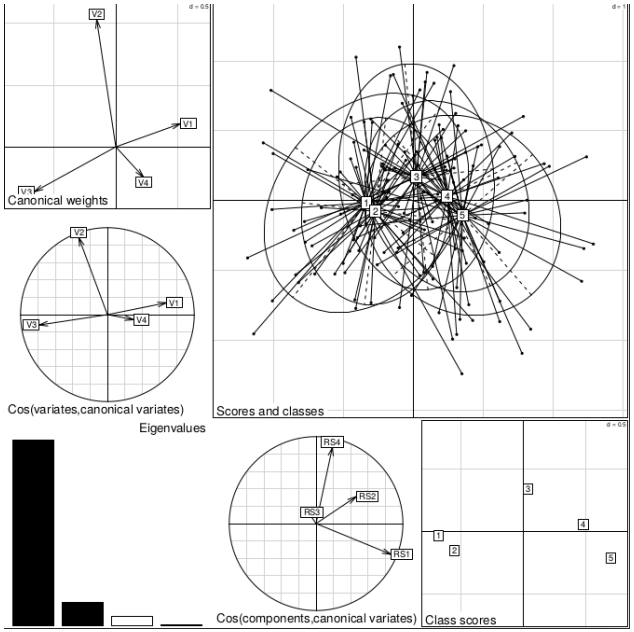
est la fonction linéaire discriminante du groupe k et où $\hat{\pi}_k = \frac{n_k}{n}$. Elle fonctionne comme dans la règle géométrique.

Exemple

Pour réaliser notre exemple, nous avons utiliser la fonction `discrimin` du package `ade4` de R qui nous a permis de réaliser l'analyse discriminante de la table de données "*chazeb*" qui est fournie avec ce package ainsi que la table "*skulls*" du même package.



(a) Chazeb



(b) Skulls