Analyse discriminante géométrique

Master Modélisation Statistique et Infromatique

ibrahima SY , ()

May 14, 2022

Données & Notation

Données

On dispose d'un échantillon de n observations de Y et de $X = (X_1, \ldots, X_p)$: sur les n individus de l'échantillon, on a mesuré une variable qualitative à K modalités et p variables quantitatives. En notant y_i la valeur de la variable à expliquer mesurée sur les ième individu, on obtient le vecteur $y = (y_1, \ldots, y_n)^T \in \{1, \ldots, K\}^n$. En notant x_{ij} la valeur de la jme variable explicative mesurée sur le ime individu, on obtient ainsi la matrice de données de dimension $n \times p$

Notation

Definition

- $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \in \mathbb{R}^P$ une ligne de X décrivant le ième individu
- $x^j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T \in \mathbb{R}^n$ une colonne de X décrivant la jème variable
- E_k est le groupe des individus des l'échantillon qui possèdent la modalité k.
- $n_k = card(E_k)$ est le nombre d'individus qui possèdent la modalité k

Si les n individus sont affectés des poids $p_1, \ldots, p_n$, ((tels que $\forall i = 1, \ldots, n, p_i \leq 0 et \sum_{i=0}^{n} p_i = 1$ alors le poids de chaque groupe E_k est :

$$P_k = \sum_{i \in E_k} p_i$$

En général, on prend $p_i = \frac{1}{n}$ et donc $P_k = \frac{n_k}{n}$. On a alors les définitions suivantes :

Notation

Definition

- Le centre de gravité global est le vecteur de $\mathbb{R}^p$ défini
$$g = \sum_{i=0}^{n} p_i x_i = \frac{1}{n} \sum_{i=0}^{n} x_i$$
- Le centre de gravité du groupe E_k est le vecteur de $\mathbb{R}^p$ défini par :
$$g_k = \frac{1}{P_k} \sum_{i \in E_k} p_i x_i = \frac{1}{n_k} \sum_{i \in E_k} x_i$$
- La matrice $p \times p$ de variance-covariance globale est définie par :
$$V = \sum_{i}^{n} p_i (x_i - g)(x_i - g)^T = \frac{1}{n} \sum_{i}^{n} (x_i - g)(x_i - g)^T$$
- La matrice $p \times p$ de variance-covariance du groupe E_k est définie par :
$$V_k = \frac{1}{P_k} \sum_{i}^{n} p_i (x_i - g_k)(x_i - g_k)^T = \frac{1}{n_k} \sum_{i \in E_k} (x_i - g_k)(x_i - g_k)^T$$
- La matrice $p \times p$ de variance-covariance intra-groupe est définie par :
$$W = \sum_{k=1}^{k} P_k V_k = \sum_{k=1}^{K} \frac{n_k}{n} V_k$$

Méthode géométrique de classification

Méthode géométrique de classification

Notons x le vecteur des valeurs des p variables explicatives sur un nouvel individu dont que l'on veut classer. La règle géométrique consiste à calculer la distance de x à chacun des K centres de gravité $g_1, \ldots, g_k$ et à affecter x au groupe le plus proche. Pour cela, il faut préciser la métrique à utiliser dans le calcul des distances. La régle la plus utilisée est celle de **Mahalanobis-Fisher** qui consiste à prendre la métrique W^{-1} (ou V^{-1} ce qui est équivalent). La distance du nouvel individu au groupe k est alors

$$d^2(x, g_k) = (x - g_k)^T W^{-1}(x - g_k)$$

Fonctions linéaires discriminantes. La règle géométrique classe la nouvelle observation x dans le groupe k^* tel que :

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

ce qui se réécrit :

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

ou

$$L_k = x^T W^{-1} g_k - \frac{1}{2} g_k^T W^{-1} g_k$$

$L_k(x)$ est la fonction linéaire discriminante du groupe k (encore appelée fonction linéaire de classement).

Chaque fonction linéaire discriminante définie une fonction score qui donne une "note" à l'observation x dans chaque groupe. Cette observation est donc affectée au groupe pour lequel le score est le plus grand.

Réferences

- Cours de Marie Chavant