Classification supervisée
Analyse discriminante

V. Vandewalle (vincent.vandewalle@univ-lille.fr)
C. Preda (cristian.preda@univ-lille.fr)

Polytech'Lille GIS4

Année universitaire 2019-2020

Rappel des notations et objectifs

Notations

- Y : la variable cible (**qualitative**) : $Y \in \{1, 2, \ldots, K\}$, $K \geq 2$
- $\{X_1, X_2, \ldots, X_p\}$: p prédicteurs quantitatifs des groupes

Objectifs

- mesurer le pouvoir prédictif des X_j par rapport à Y
- construire une règle de décision pour la prédiction de Y à partir des X_j

Deux méthodes d'analyse discriminante :

- **factorielle** (Fisher) : analyse factorielle discriminante (AFD)
- **probabiliste** (Bayes) : analyse discriminante linéaire (LDA) et analyse discriminante quadratique (QDA)

Sous certaines conditions l'analyse factorielle discriminante se réduit à une analyse discriminante probabiliste.

Estimation : données

n unités statistiques (individus) tels que :

- Groupe 1 : n_1 ($Y = 1$)
- Groupe 2 : n_2 ($Y = 2$)
- $\vdots$
- Groupe K : n_K ($Y = K$)

avec $n_1 + n_2 + \cdots + n_K = n$.
Soit X_{ijh} ($i = 1, \ldots, K$, $j = 1, \ldots, p$, $h = 1, \ldots, n_i$), l'observation de la variable X_j sur l'individu h dans le groupe i.

Estimation : données

	X_1	$\cdots$	X_j	$\cdots$	X_p	Y
$\vdots$						1
$\vdots$						$\vdots$
$\vdots$						1
1			$\vdots$			i
$\vdots$			$\vdots$			$\vdots$
h	$\cdots$	$\cdots$	X_{ijh}			i
$\vdots$						$\vdots$
$\vdots$						K
$\vdots$						$\vdots$
$\vdots$						K

X matrice $n \times p$, Y vecteur $n \times 1$.

Estimation : espérances

$$\bar{X}_{ij} = \frac{1}{n_i} \sum_{h=1}^{n_i} X_{ijh}$$

est un estimateur de μ_{ij} et donc

$$\bar{X}_i = (\bar{X}_{i1}, \ldots, \bar{X}_{ij}, \ldots, \bar{X}_{ip})$$

est un estimateur de μ_i.
Moyennes regroupées dans un tableau $K \times p$, noté G

	X_1	$\cdots$	X_j	$\cdots$	X_p
$Y = 1$	$\bar{X}_{11}$		$\bar{X}_{1j}$		$\bar{X}_{1p}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$Y = i$	$\bar{X}_{i1}$		$\bar{X}_{ij}$		$\bar{X}_{ip}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$Y = K$	$\bar{X}_{K1}$		$\bar{X}_{Kj}$		$\bar{X}_{Kp}$

Estimation : espérances

$$\bar{X} = \sum_{i=1}^{K} \frac{n_i}{n} \bar{X}_i$$

la moyenne globale de X, estimateur de

$$\mu = E(X_1, \ldots, X_p).$$

$$\bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p)$$

où

$$\bar{X}_j = \sum_{i=1}^{K} \frac{n_i}{n} \bar{X}_{ij} = \frac{1}{n} \sum_{i=1}^{K} \sum_{h=1}^{n_i} X_{ijh}$$

estimateur de la moyenne de la variable X_j sans la connaissance du groupe.

Estimation : variances

- W_i : matrice de variance-covariance dans le groupe $Y = i$

$$W_i[j,j'] = cov(X_j, X_{j'})_{/Y=i} = \frac{1}{n_i}\sum_{h=1}^{n_i}(X_{ijh} - \bar{X}_{ij})(X_{ij'h} - \bar{X}_{ij'})$$

- W : matrice de variance-covariance **intra-groupes**

$$W = \sum_{i=1}^{K}\frac{n_i}{n}W_i.$$

- B : la matrice de variance-covariance **inter-groupes**

$$B[j,j'] = \sum_{i=1}^{K}\frac{n_i}{n}(\bar{X}_{ij} - \bar{X}_j)(\bar{X}_{ij'} - \bar{X}_{j'})$$

- V : matrice de variance-covariance du tableau X : variance-covariance **totale**

$$V[j,j'] = cov(X_j, X_{j'}) = \frac{1}{n}\sum_{i=1}^{K}\sum_{h=1}^{n_i}(X_{ijh} - \bar{X}_j)(X_{ij'h} - \bar{X}_{j'})$$

Objectif

Chercher des facteurs (composantes) discriminants

$$d = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p$$

tels que les groupes soient les plus **séparés** les uns des autres et les données soient le plus **regroupées** possible autour du centre de gravité de leur groupe.

Critère de l'AFD

ACP (rappel) Rechercher l'axe a de plus forte variance

$$a_1 = \underset{a \in \mathbf{R}^d}{\operatorname{argmax}}\, a'Va$$

Objectif en AFD

Trouver a :

- sur lequel la variance inter-classes est maximale : maximisant $a'Ba$
- mais aussi avec des classes bien condensées : minimisant $a'Wa$

Or, ces quantités sont liées par : $a'Va = a'Wa + a'Ba$

Critère

$$\underset{a \in \mathbf{R}^d}{\operatorname{argmax}} \frac{a'Ba}{a'Va} = \underset{a \in \mathbf{R}^d}{\operatorname{argmax}} \frac{a'Ba}{a'Wa}$$

Solution de l'AFD

Solution

- Le vecteur propre a associé à la plus grande valeur propre λ_1 de $V^{-1}B$
- Les $K-1$ vecteurs propres de $V^{-1}B$ notés a_1, a_2, …, a_{K-1} permettent de calculer les composantes principales, et les valeurs propres λ_1, λ_2, …, λ_{K-1} l'inertie portée par les axes
- La i^{e} composante discriminante s'écrit alors :

$$d_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ij}X_j + a_{ip}X_p$$

Remarques

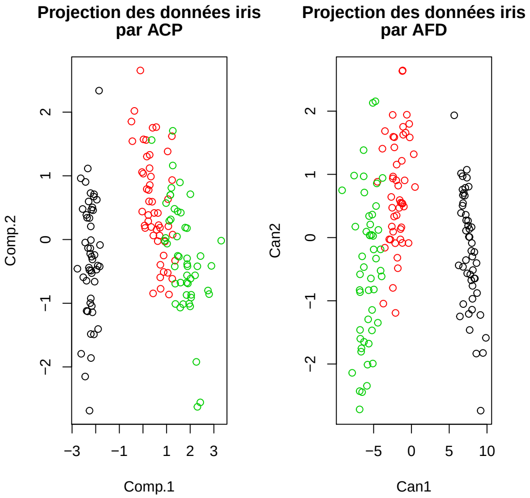
- Métrique V^{-1} et W^{-1} équivalentes
- W^{-1} la plus utilisée (Métrique de Mahalanobis) : $d^2(x,y) = (x-y)'W^{-1}(x-y)$.

AFD et ACP

Remarque :

L'ACP recherche la combinaison linéaire des variables d'origine maximisant l'inertie totale du nuage projeté (restitution de la forme globale du nuage) tandis que l'AFD recherche la combinaison linéaire des variables d'origine maximisant la seule inertie entre les classes (restitution optimale de la séparation induite par la partition).

Illustration



Régle de décision

Règle de décision donnée par

$$\hat{y} = \arg\min_i (x - \bar{X}_i)' W^{-1} (x - \bar{X}_i)$$

où $x = (x_1, \ldots, x_p)$.
Le point n'est pas affecté à la classe dont il est le plus proche en terme de distance euclidienne, mais à la classe dont il est le plus proche dans la métrique de Mahalanobis.

Dans le cas de deux groupes : individu x affecté au groupe 2 si

$$(x - \bar{X}_1)' W^{-1} (x - \bar{X}_1) > (x - \bar{X}_2)' W^{-1} (x - \bar{X}_2)$$

Règle de décision

Autrement dit, cela revient à maximiser un score $(s_i(x))$

$$\arg\min_i (x-\mu_i)'W^{-1}(x-\mu_i) \simeq \arg\min_i (x-\bar{X}_i)'W^{-1}(x-\bar{X}_i)$$

$$\arg\min_i (x-\bar{X}_i)'W^{-1}(x-\bar{X}_i) = \arg\min_i -2x'W^{-1}\bar{X}_i + \bar{X}_i'W^{-1}\bar{X}_i$$

car W est symétrique. On cherche donc

$$\arg\max_i 2x'W^{-1}\bar{X}_i - \bar{X}_i'W^{-1}\bar{X}_i = \arg\max_i s_i(x)$$

avec $s_i(x) = \alpha_{0i} + \alpha_{i1}x_1 + \cdots + \alpha_{ip}x_p$

Cadre général

Une nouvelle observation x est affectée au groupe :

$$\arg\max_i P(Y = i/X = x)$$

Remarque :

$$\sum_{i=1}^{K} P(Y = i/X = x) = 1$$

Par la formule de Bayes on a :

$$P(Y = i/X = x) = \frac{P(X = x/Y = i)P(Y = i)}{\sum_{i'=1}^{K} P(X = x/Y = i')P(Y = i')}$$

On note :

- $P(X = x/Y = i) = f_i(x)$: la densité du vecteur X dans le groupe $Y = i$, $i = 1, \ldots, K$
- $P(Y = i) = \pi_i$: probabilité *a priori* du groupe $Y = i$

Remarques

- $P(Y = i) = \pi_i$: probabilité *a priori* car calculée sans aucune information sur X.
- $P(Y = i/X = x)$: probabilité *a posteriori* car calculée à partir de l'information complémentaire $X = x$, probabilité postérieure à l'observation de $X = x$.

La règle de décision devient :

$$\underset{i}{\operatorname{argmax}} P(Y = i/X = x) = \underset{i}{\operatorname{argmax}} P(X = x/Y = i) P(Y = i)$$

$$= \underset{i}{\operatorname{argmax}} f_i(x) \pi_i$$

Cadre gaussien

X dans le groupe i suit une loi gaussienne multivariée.

$$X/Y = i \sim \mathcal{N}(\mu_i, W_i)$$

avec $\mu_i = (\mu_{i1}, \ldots, \mu_{ip})$ l'espérance et W_i la matrice de variance-covariance de $X/Y = i$.
La densité gaussienne multivariée $\mathcal{N}(\mu_i, W_i)$ a pour densité $f_i : \mathbb{R}^p \mapsto [0; +\infty[$:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}|W_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)' W_i^{-1}(x-\mu_i)}$$

Cadre gaussien

- pour $p = 1$ on retrouve la loi $\mathcal{N}(\mu_i, \sigma_i^2)$

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}$$

- pour $p = 2$ on retrouve la loi normale bivariée $\mathcal{N}((\mu_{i1}, \mu_{i2}), \sigma_{i1}^2, \sigma_{i2}^2, \rho_i)$ avec

$$W_i = \begin{pmatrix} \sigma_{i1}^2 & \rho_i \sigma_{i1} \sigma_{i2} \\ \rho_i \sigma_{i1} \sigma_{i2} & \sigma_{i2}^2 \end{pmatrix}$$

Régle de décision

Dans le cas gaussien, la règle de décision devient :

$$\underset{i}{\operatorname{argmax}} \frac{1}{(2\pi)^{p/2}|W_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)' W_i^{-1}(x-\mu_i)} \pi_i$$

équivalent à

$$\underset{i}{\operatorname{argmax}} -\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln(|W_i|) - \frac{1}{2}(x-\mu_i)' W_i^{-1}(x-\mu_i) + \ln(\pi_i)$$

Cas homoscédastique (lda)

Si on fait l'hypothèse que :

- $W_i = W, \forall i = 1, \ldots, K$ et
- $\pi_i = \pi$, probabilités *a priori* égales

alors on est ramené à la règle de décision de l'AFD

$$\arg\max_i -\frac{1}{2}(x-\mu_i)' W^{-1}(x-\mu_i) = \arg\min_i (x-\mu_i)' W^{-1}(x-\mu_i)$$

Cas homoscédastique (lda)

Dans ce cas comme en AFD la décision est basée sur une forme linéaire car

$$\arg\min_i (x-\mu_i)' W^{-1}(x-\mu_i) \simeq \arg\min_i (x-\bar{X}_i)' W^{-1}(x-\bar{X}_i) =$$

$$\arg\min_i -2x' W^{-1}\bar{X}_i + \bar{X}_i' W^{-1}\bar{X}_i =$$

$$\arg\max_i 2x' W^{-1}\bar{X}_i - \bar{X}_i' W^{-1}\bar{X}_i = \arg\max_i s_i(x)$$

avec $s_i(x) = \alpha_{0i} + \alpha_{i1}x_1 + \cdots + \alpha_{ip}x_p$

L'analyse discriminante probabiliste est dans ce cas connue sous le nom d'analyse discriminante linéaire (lda).

Remarque : Si les probabilités a priori π_i ne sont pas égales, la règle de décision reste linéaire en x mais on n'a plus équivalence avec l'AFD.

Cas hétéroscédastique (qda)

Si les matrices W_i ne sont pas égales, alors la règle devient quadratique

$$\arg\min_i \frac{1}{2}\ln(|W_i|) + \frac{1}{2}(x-\mu_i)' W_i^{-1}(x-\mu_i) - \ln(\pi_i)$$

qui est une forme quadratique en x.
On appelle alors cette analyse : quadratic discriminant analysis (qda)

Remarque : l'analyse discriminante quadratique peut produire de meilleurs résultats que l'analyse discriminante linéaire, mais est moins robuste quand le nombre de variables p est grand (plus de paramètres à estimer).