L'Analyse Factorielle Discriminante

Département de Mathématiques
et informatique

Pr. Hasna CHAMLAL
Master-BD2C-M1

FSAC
كلية العلوم عين الشق

Introduction

- Les méthodes de discrimination sont des méthodes qui visent à séparer au mieux les classes à priori définies par une variable qualitative notée Q, et ce à partir de p variables X1,X2,…..,Xp dites variables explicatives.
- Ce sont des méthodes prévisionnelles faisant partie de l'apprentissage Statistique aussi appelées méthodes d'apprentissage supervisé (machine learning).
- Les méthodes d'apprentissage statistique passent par deux phases :
  - Phase d'apprentissage : Sur un échantillon d'apprentissage on observe X1,X2,…..,Xp et aussi Q. Ce qui conduit à la construction d'une règle prévisionnelle (affectation).
  - Phase prévisionnelle : Des individus sur lesquels on observe les Xj mais pas Q ; il s'agit d'appliquer la règle d'affectation pour prédire leur classe (leur modalité relativement à Q)
- Le passage de la phase 1 à la phase 2 est garantie par une phase intermédiaire qui est la phase de Validation(LOOCV,kfolds,AUC,CV,….).
- Domaines d'application : Crédit scoring, Finance : risque, Assurance, reconnaissance de forme et de la parole,…….
- Une variété de méthodes d'apprentissage supervisé : NB, KNN,Tree,SVM,RF, régression logistique,….
- L'AFD en fait partie c'est une méthode géométrique.

Aspect géométrique de la séparation

- Deux classes P1 et P2
  n=6, n1=3 et n2=3

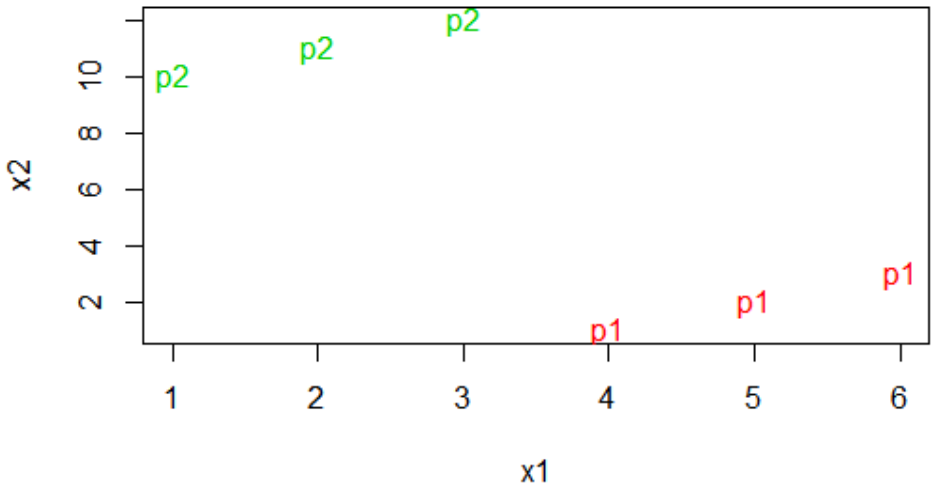
	X1	X2	Q
1	4	1	P1
2	5	2	P1
3	6	3	P1
4	1	10	P2
5	2	11	P2
6	3	12	P2

- **Caractéristiques des classes**

Classe P1
Centre de gravité :
G1=(5,2)

Classe P2
Centre de gravité :
G2=(2,11)

- Chacune des modalités de Q induit une classe : classe P1 et classe P2 décrites par les variables X1 et X2, on peut en définir un centre de gravité (noyau) G1 et G2.
- Le pouvoir discriminant de X1 et X2 se mesure par :
  - Des centres de gravité éloignées (séparés) : variance interclasse maximale (Between)
  - Pour chaque classe : les individus sont peu dispersés autour du centre de gravité : variance intraclasse com

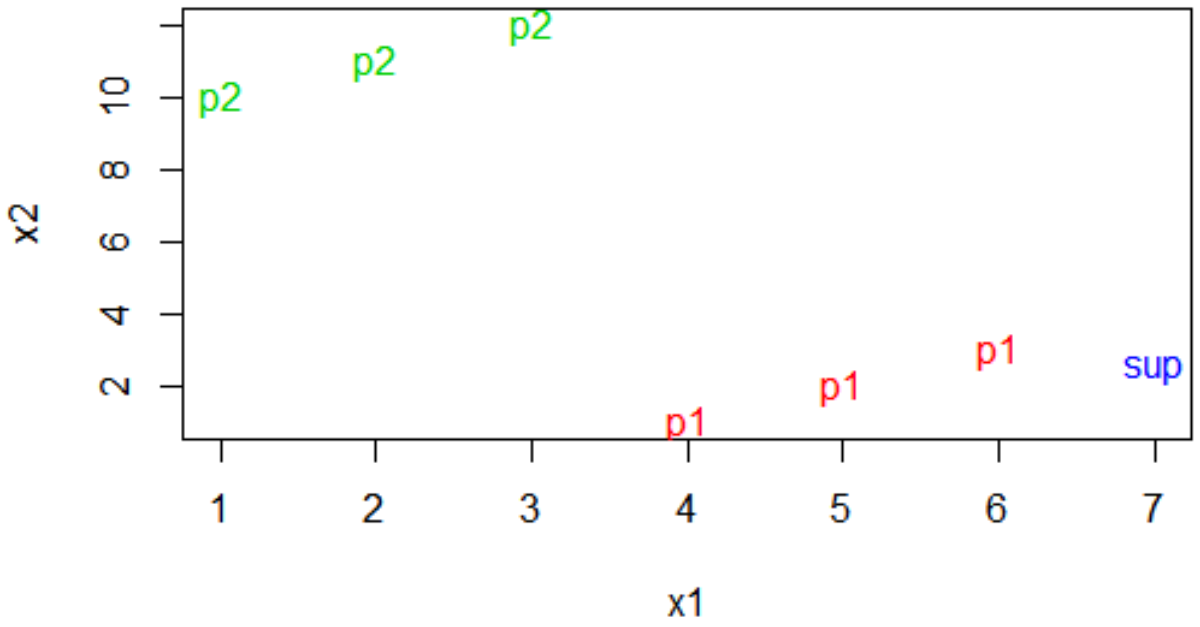


Aspect géométrique de l'affectation

Un individu sup de profil X1=7 et X2= 2.5 sa modalité de Q étant

inconnue?

La règle géométrique consiste à l'affecter à la classe dont le centre de gravité est le plus proche . Soit ici : ?



- Problème : quand p >>>2?
- La solution : méthode factorielle (de réduction)
- Critère de construction : des axes conservant au maximum la séparation des centres de gravité :
  - **Il s'agit alors d'appliquer une ACP sur les centres de gravité des classes :**

	Y1…Y2…	Yj	…Yp	Q
1				
…				
i		Yj(i)		Q(i)
n				

Tableau initial

	X1…	…Xj..	…..	Xp
1				
….				
i		Xj(i)		

X: Tableau des variables explicatives centrées

	X1	X2…	..Xj..	Xp
G1				
…				
Gk				

G: Tableau des centres de gravité

Notations et Définitions

- k=nombre de modalités de Q
- nl=nombre d'u=individus de l'échantillon d'apprentissage ayant la lième modalité de Q (l=1,…;,k)
- Matrice de variance covariance totale :$V = \frac{1}{n}\ transposé(X) * X$
- Matrice diagonale des poids Δ= $diag(\frac{n_1}{n}, \frac{n_2}{n}, \ldots\ldots, \frac{n_k}{n})$
- Matrice de variance covariance interclasse : $B = transposé(G)\Delta G$

ACP sur G

Métrique : V^{-1}
Métrique de mahalanobis

Matrice à diagonaliser : $V^{-1}B$

Systèmes de valeurs propres non nulles : λ_l

Vecteurs propres V^{-1} normés : vecteurs principaux : u_l

Composantes principales appelées composantes discriminantes : $C^l = G\,V^{-1}\,u_l$

λ_l mesure le pouvoir discriminant du lième axe discriminant

Phase d'affectation :

Un individu i de profil(Y1(i),….,Yp(i)) Q(i) étant inconnue

Centrer

X(i)=(X1(i),…..,Xp(i))

Projeter sur un plan principal(r,s)

$$C^r(i) = trans(u_r)V^{-1}X(i)$$
$$C^s(i) = trans(u_s)V^{-1}X(i)$$

Affectation

Sur le plan principal (r,s) Comparer la distance i et les différents centre de gravité

Affecter à la classe dont le centre de gravité est le plus proche