

Prof. Mohamed Aly LOULY

Data Science et Machine Learning

1

Analyse Discriminante (AD)

Classification supervisée

Classement ou Scoring

LES DEUX MÉTHODES DE L'AD

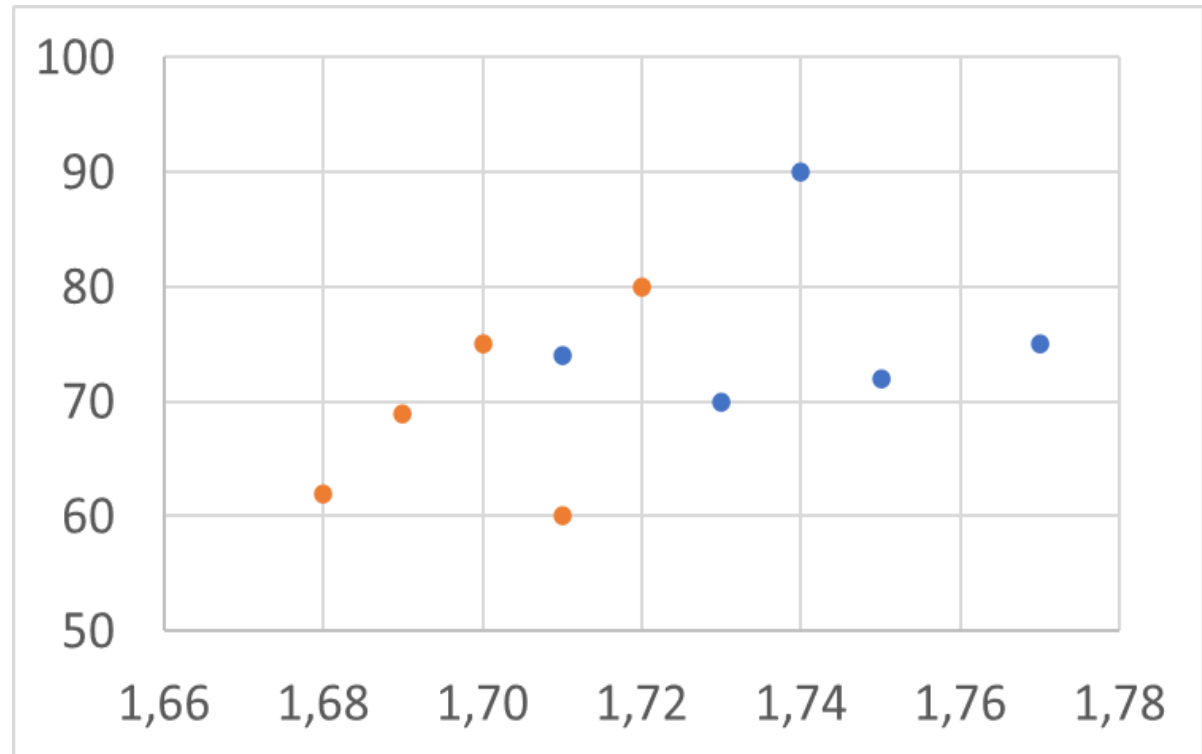
L'Analyse Discriminante (AD) contient deux méthodes :

L'Analyse Factorielle Discriminante (AFD) : une méthode géométrique descriptive qui permet de visualiser des individus sur un plan en maximisant la discrimination entre leurs classes.

L'Analyse Discriminante Décisionnelle (ADD) : une méthode géométrique et probabiliste qui explique une variable qualitative par plusieurs variables quantitatives. Elle permet de faire le classement de chaque nouvel individu à la classe qui lui correspond le mieux.

EXEMPLE DE DONNÉES

	Taille	Poids	Genre
Ind 1	1,71	60	F
Ind 2	1,68	62	F
Ind 3	1,69	69	F
Ind 4	1,73	70	H
Ind 5	1,75	72	H
Ind 6	1,71	74	H
Ind 7	1,70	75	F
Ind 8	1,77	75	H
Ind 9	1,72	80	F
Ind 10	1,74	90	H



LES DONNÉES DE L'AD

L'Analyse Discriminante (AD) fonctionne avec n individus appartenant à m classes différentes pour lesquels nous avons observé p variables.

Les observations sont données par une matrice X de taille (n, p) . Les lignes de cette matrice représentent les individus et les colonnes représentent les variables. Le coefficient x_{ij} désigne l'observation de la variable j sur l'individu i .

Les classes des individus sont données par un vecteur Y de taille n , prenant des valeurs entières entre 1 et m . La valeur y_i précise la classe de l'individu i .

DISTANCE ENTRE INDIVIDUS

Les n individus qui se trouvent sur les lignes de X peuvent être vus comme un nuage de points dans R^p . Le i -ème individu est représenté par la i -ème ligne de X : $(x_{i1}, x_{i2}, \dots, x_{ip})$. La distance entre deux individus i et k est la suivante :

$$d^2(i, k) = \|X_i - X_k\|^2 = \sum_{j=1}^p (x_{ij} - x_{kj})^2$$

$$d^2(1,2) = (1,71 - 1,68)^2 + (62 - 60)^2 = 0,03^2 + 2^2 = 0,0009 + 4 = 4,0009$$

La taille ne contribue pas réellement à la mesure de la distance.

MESURES SUR LES VARIABLES

La colonne j de X fournit les n observations de la j -ème variable X^j . Sa moyenne m_j et sa variance σ_j^2 sont les suivantes :

$$m_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

$$\sigma_j^2 = \frac{1}{n} \sum_{i=1}^n (x_{ij} - m_j)^2$$

	Taille	Poids
Moyenne	1,72	72,70
Ecartype	0,03	8,14

	Taille	Poids
Moyenne H	1,74	76
Moyenne F	1,70	69

CENTRAGE ET RÉDUCTION DES DONNÉES

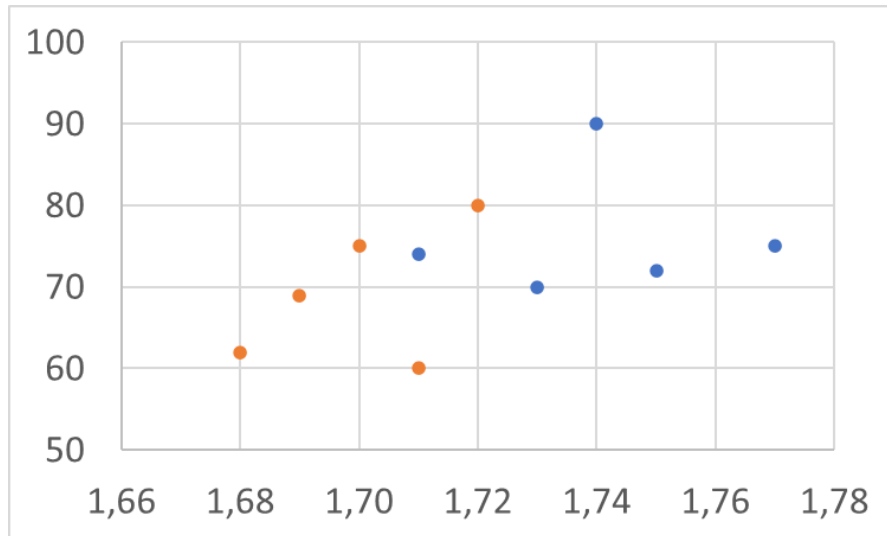
Il est parfois nécessaire de réduire toutes les variables pour uniformiser leur contribution à la distance entre les individus. Il est également plus pratique de les centrer. On note par Z le tableau des données centrées réduites :

$$z_{ij} = \frac{x_{ij} - m_j}{\sigma_j}$$

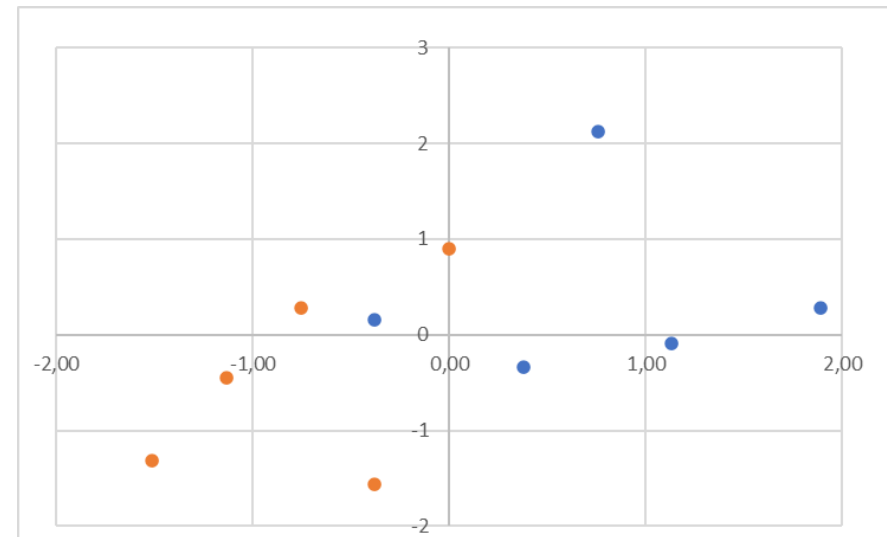
Taille	Poids	Genre
-0,38	-1,56	F
-1,51	-1,31	F
-1,13	-0,45	F
0,38	-0,33	H
1,13	-0,09	H
-0,38	0,16	H
-0,76	0,28	F
1,89	0,28	H
0,00	0,90	F
0,76	2,13	H

LES VARIABLES ONT LA MÊME IMPORTANCE DANS LES DONNÉES CENTRÉES RÉDUITES

X



Z



Chaque variable à une **variance égale à 1** dans les données centrées réduites.

VARIANCE TOTALE ET VARIANCES DES CLASSES

La matrice de variance-covariance totale est la suivante pour les données centrées réduites :

$$T = \frac{1}{n} Z^T Z = \frac{1}{n} \sum_{i=1}^n Z_i^T Z_i$$

La matrice de variance-covariance inter-classes (**Between**) est la suivante où g_k désigne la moyenne des individus de la classe C_k et n_k leur nombre :

$$B = \sum_{k=1}^m \frac{n_k}{n} g_k^T g_k \quad \text{où} \quad g_k = \frac{1}{n_k} \sum_{i \in C_k} Z_i$$

VARIANCES INTER-CLASSES ET INTRA-CLASSES

La matrice de variance-covariance intra-classes (**Within**) est la moyenne des matrices de variance des classes :

$$W = \sum_{k=1}^m \frac{n_k}{n} W_k$$

où W_k est la matrice de variance-covariance de la classe C_k , donnée par l'expression suivante :

$$W_k = \frac{1}{n_k} \sum_{i \in C_k} Z_i^T Z_i - g_k^T g_k$$

DÉCOMPOSITION DE LA MATRICE DE VARIANCE COVARIANCE

La matrice de variance covariance totale est la somme des variances inter-classes et intra-classes :

$$T = B + W$$

Variance Totale = Variance Expliquée (B) + Variances Résiduelles (W)

EXEMPLE

T

	Taille	Poids
Taille	1,00	0,47
Poids	0,47	1,00

G

	Taille	Poids
Moyenne H	0,76	0,43
Moyenne F	-0,76	-0,43

B

	Taille	Poids
Taille	0,57	0,33
Poids	0,33	0,19

W_H

	Taille	Poids
Taille	0,57	0,05
Poids	0,05	0,76

W_F

	Taille	Poids
Taille	0,29	0,25
Poids	0,25	0,87

W

	Taille	Poids
Taille	0,43	0,15
Poids	0,15	0,81

LES AXES DISCRIMINANTS

Les axes discriminants sont donnés par les vecteurs propres $u_1, u_2, \dots, u_{m-1}$ de $T^{-1}B$, associés à ses valeurs propres non nulles $\lambda_1, \lambda_2, \dots, \lambda_{m-1}$, énumérées dans l'ordre décroissant.

EXEMPLE D'AXE DISCRIMINANT

 T^{-1}

	Taille	Poids
Taille	1,29	-0,61
Poids	-0,61	1,29

 $T^{-1}B$

	Taille	Poids
Taille	0,54	0,31
Poids	0,07	0,04

 U

	U
Taille	0,992
Poids	0,129

La valeur propre associée à cet axe vaut 0,58

LA PROJECTION DES INDIVIDUS

La projection des individus sur l'axe discriminant k est alors :

$$F_k = Z u_k$$

En désignant par U la matrice dont les colonnes sont les vecteurs $u_1, u_2, \dots, u_q$, on obtient les différentes projections dans les colonnes de la matrice F suivante : $F = Z U$.

Les centres de gravité g_k peuvent être projetés de la même manière sur les axes discriminants.

PROJECTION DES INDIVIDUS

	F	Genre
Ind 1	-0,58	F
Ind 2	-1,67	F
Ind 3	-1,18	F
Ind 4	0,33	H
Ind 5	1,11	H
Ind 6	-0,35	H
Ind 7	-0,71	F
Ind 8	1,91	H
Ind 9	0,12	F
Ind 10	1,02	H

L'ANALYSE DISCRIMINANTE DÉCISIONNELLE ET PRÉDICTIVE

L'Analyse Discriminante Décisionnelle (ADD) est utilisée comme une méthode **décisionnelle** et **prédictive**. Elle permet de **classer** l'individu dans la classe qui lui correspond le mieux.

Il y a deux approches de classement : une approche **géométrique** et une approche **probabiliste**.

CLASSEMENT AVEC UNE RÈGLE GÉOMÉTRIQUE

La règle du classement géométrique consiste à affecter l'individu x à la classe C_k dont le centre de gravité g_k est le plus proche de cet individu. La distance généralement utilisée pour ce classement est celle qui est associée à la matrice T^{-1} :

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

La distance associée à la matrice W^{-1} est parfois utilisée à la place de celle de la matrice T^{-1} .

CLASSEMENT AVEC LE MODÈLE PROBABILISTE BAYÉSIEN

La règle du modèle du classement probabiliste bayésien consiste à affecter l'individu x à la classe C_k la plus probable. C'est la classe qui maximise la probabilité conditionnelle $P(C_k/x)$:

$$P(C_k/x) = \frac{P(x/C_k)P(C_k)}{P(x)}$$

Il suffira de chercher la classe C_k qui maximise le numérateur car $P(x)$ ne dépend pas de k .

CLASSEMENT AVEC LE MODÈLE BAYÉSIEN NORMAL DE MÊME VARIANCE

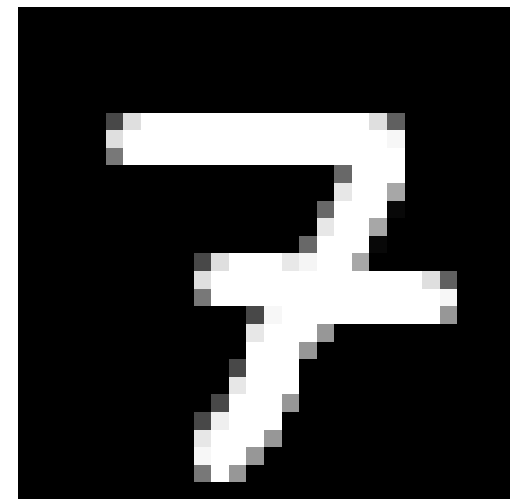
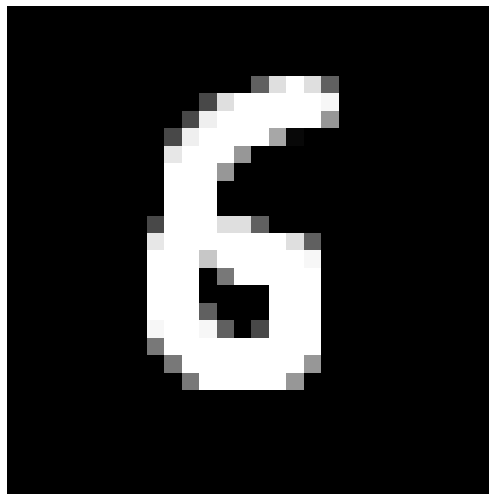
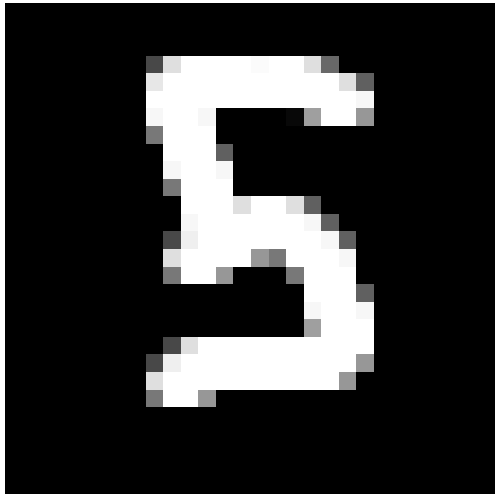
Sous l'hypothèse de matrices de variance covariance identiques W , le score à minimiser devient :

$$sc_k = (x - g_k) W^{-1} (x - g_k)^T - 2 \log(n_k)$$

On remarque que le premier terme est celui de la règle du classement géométrique avec la distance associée à la matrice W^{-1} .

Les deux classements sont donc identiques en cas de classes équiprobables, de mêmes nombres n_k .

L'AD POUR LA RECONNAISSANCE DES FORMES



L'AFD POUR LA RECONNAISSANCE DES FORMES

